THE FIRST INTERNATIONAL CONFERENCE

on

LANGUAGE TECHNOLOGIES FOR LOW-RESOURCE LANGUAGES

LaTeLL'2026

P R O C E E D I N G S

Edited by:

Salima Lamsiyah, Tharindu Ranasinghe, Saad Ezzini,
Ernesto Luis Estevanell-Valladares, Ruslan Mitkov

Fes, Morocco
September 28–30, 2026
www.latell.org/2026/

THE FIRST INTERNATIONAL CONFERENCE
ON LANGUAGE TECHNOLOGIES FOR
LOW-RESOURCE LANGUAGES
(LaTeLL'2026)

PROCEEDINGS

28–30 September, 2026
Fes, Morocco
https://latell.org/2026/

© 2026 The authors of the individual papers, for their respective contributions.

The individual papers in this volume are licensed by their authors under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

LaTeLL'2026 is organised by the University of Alicante (Spain), Lancaster University (UK), King Fahd University of Petroleum and Minerals (Saudi Arabia) and the University of Luxembourg, with the local support of Mohammed VI Polytechnic University (Morocco).

The organisers gratefully acknowledge the support of AutoTrust AI, Diamond Sponsor of LaTeLL'2026.

Online ISBN 978-954-452-112-7
e-book site: www.acl-bg.org, www.latell.org/2026/

Designed by INCOMA Ltd.
Shoumen, BULGARIA

Message from the Conference Chair

Welcome to Fes and to LaTeLL'2026, the First International Conference on Language Technologies for Low-resource Languages.

Most of the world's languages are low-resource languages. They account for at least 85% of the world's linguistic diversity and are often spoken by smaller or marginalised communities, yet the language technology that now shapes how people read, write, search and communicate has been built almost entirely for a few dozen languages with abundant data. For the rest, the basic ingredients are missing: large, well-annotated datasets, standardised tools and reliable linguistic resources. Large language models have narrowed some of these gaps and widened others. They perform impressively in the languages they were trained on and much less well in the languages they were not. Closing this gap is a scientific problem, but it is also a question of fairness: the speakers of low-resource languages should not be the last to benefit from language technology.

Several workshops have addressed natural language processing for low-resource languages over the years and have done valuable work. LaTeLL is the first international conference dedicated specifically to the automatic processing of these languages. Our aim in launching it is to give this research community a venue of its own, where work on any low-resource language, from any part of the world, can be presented, compared and built upon. Fes, home to one of the oldest universities in the world and set in a region where Arabic dialects, Amazigh and other languages are spoken side by side, seemed to us the right place to begin.

The papers in this volume cover the whole range of the field: language resources and corpora; core language technologies such as morphology, syntax and machine translation; applications; multimodal and grounded approaches; multilinguality and cross-lingual methods; human-centred approaches; and the ethical and social questions that arise when technology reaches communities it has previously ignored. The programme also features two special tracks, one on building applications based on large language models for low-resource languages and one on Modern Standard Arabic and Arabic dialects.

Many people made this conference possible. We thank the authors who submitted their work and presented it, in Fes or by video, and the members of the Programme Committee, who reviewed every submission carefully and gave the authors the feedback that improved the final versions of their papers. We are grateful to Professor Nizar Habash of New York University Abu Dhabi for accepting our invitation to give the keynote address. The Programme Committee Chairs, Saad Ezzini, Salima Lamsiyah and Tharindu Ranasinghe, put the programme together, and the quality of the papers in this volume owes much to them. We are also grateful to AutoTrust AI, the Diamond Sponsor of LaTeLL'2026, for its support of the conference.

Finally, we thank the members of the Organising Committee, whose work over many months made LaTeLL'2026 happen: Ismail Berrada, our Local Chair, and Radouane Achamlal, who took care of the local arrangements in Fes; Salmane Chafik, Houdaifa Atou, Milica Ikonić Nešić, Ernesto Luis Estevanell-Valladares and Maram Alharbi; and Nikolai Nikolov, our conference administrator.

Welcome to LaTeLL'2026. We hope you enjoy the conference and your stay in Fes.

Ruslan Mitkov
LaTeLL'2026 Conference Chair

Message from the Programme Committee Chairs

This volume contains the papers accepted for presentation at LaTeLL'2026, the First International Conference on Language Technologies for Low-resource Languages, held in Fes, Morocco, on 28–30 September 2026.

This volume brings together the papers accepted for LaTeLL'2026, the first edition of the conference. Every submission was reviewed by three members of the Programme Committee, who gave their time generously and often under tight deadlines, and the authors responded to the reviews with care, improving their papers as a result. Some papers whose authors could not travel to Fes are presented as recorded videos on the conference website.

The submissions showed how wide the field has become. They deal with languages from Africa, Asia, the Middle East, Europe, the Americas and Oceania, with work on building corpora and annotation schemes, machine translation and cross-lingual transfer, morphology and syntax, speech, and applications that range from the detection of hate speech to the processing of clinical text. The two special tracks, on building applications based on large language models for low-resource languages and on Modern Standard Arabic and Arabic dialects, are both represented in the programme; papers on Arabic and its dialects make up a substantial part of this volume, as befits a conference held in Morocco.

We thank the members of the Programme Committee for their reviews, written to a tight schedule, and in particular those who took on extra reviews at short notice for the late submissions. We are grateful to Nizar Habash for the keynote address, to Ruslan Mitkov for his guidance throughout, to the Organising Committee, and to Ernesto Luis Estevanell-Valladares, who prepared these proceedings.

The proceedings will be included in the ACL Anthology, where the papers will remain freely available. We hope they prove useful to everyone working to bring language technology to the languages that have so far had the least of it.

Saad Ezzini, Salima Lamsiyah and Tharindu Ranasinghe
LaTeLL'2026 Programme Committee Chairs

Organising Committee

Conference Chair
Ruslan Mitkov, Lancaster University, UK and University of Alicante, Spain

Programme Committee Chairs
Saad Ezzini, King Fahd University of Petroleum and Minerals, Saudi Arabia
Salima Lamsiyah, University of Luxembourg, Luxembourg
Tharindu Ranasinghe, Lancaster University, UK

Local Chair
Ismail Berrada, Mohammed VI Polytechnic University, Morocco

Chair of Local Arrangements
Radouane Achamlal

Organising Committee Members
Salmane Chafik, Mohammed VI Polytechnic University, Morocco
Houdaifa Atou, Mohammed VI Polytechnic University, Morocco
Milica Ikonić Nešić, University of Belgrade, Serbia
Ernesto Luis Estevanell-Valladares, University of Alicante, Spain and University of Havana, Cuba
Maram Alharbi, Lancaster University, UK

Programme Committee

Gavin Abercrombie, Heriot-Watt University, UK
Abdelghafour Aboukacem, Mohammed VI Polytechnic University, Morocco
Shadi Abudalfa, King Fahd University of Petroleum and Minerals, Saudi Arabia
Cengiz Acartürk, Jagiellonian University, Poland
Amneh Al Abdi, King Fahd University of Petroleum and Minerals, Saudi Arabia
Hamza Alami, Sidi Mohamed Ben Abdellah University, Morocco
Itziar Aldabe, HiTZ Center, University of the Basque Country UPV/EHU, Spain
Maram Alharbi, Lancaster University, UK
Hassina Aliane
Salim Al Mandhari, Lancaster University, UK
Reem Ali Alzahrani, Imam Abdulrahman Bin Faisal University, Saudi Arabia
Mullosharaf Kurbonovich Arabov, Kazan Federal University, Russia
Ibrahim Aref, Lancaster University, UK
Houdaifa Atou, Mohammed VI Polytechnic University, Morocco
Soran Badawi, Charmo University, Iraq
Timo Baumann, Ostbayerische Technische Hochschule Regensburg, Germany
Abdessamad Benlahbib, Sidi Mohamed Ben Abdellah University, Morocco
Laura Bernardy, University of Luxembourg, Luxembourg
Mariam Biltawi, Al Hussein Technical University, Jordan
Happy Buzaaba, Princeton University, USA
Salmane Chafik, Mohammed VI Polytechnic University, Morocco
Abdelhalim Hafedh Dahou, GESIS – Leibniz Institute for the Social Sciences, Germany
Daniel Dakota, Indiana University, USA
Sudhansu Bala Das, Insight Research Ireland Centre for Data Analytics, University of Galway, Ireland

Tabea De Wille, University of Limerick, Ireland
Liviu Dinu, University of Bucharest, Romania
Damith Dola Mullage, Lancaster University, UK
Omar El Bachyr, University of Luxembourg, Luxembourg
Abdessalam Elhabbash, Lancaster University, UK
Noura El Moussa, Franklin University, USA
Ernesto Luis Estevanell-Valladares, University of Alicante, Spain and University of Havana, Cuba
Saad Ezzini, King Fahd University of Petroleum and Minerals, Saudi Arabia
Christiane Fellbaum, Princeton University, USA
Aadil Gani Ganie, King Fahd University of Petroleum and Minerals, Saudi Arabia
Nadia Ghezaiel Hammouda, University of Ha'il, Saudi Arabia
Yoan Gutiérrez, University of Alicante, Spain
Amal Haddad Haddad, University of Granada, Spain
Hicham Hammouchi, University of Luxembourg, Luxembourg
Ogtay Hasanov, King Fahd University of Petroleum and Minerals, Saudi Arabia
Milica Ikonić Nešić, University of Belgrade, Serbia
Anton Karl Ingason, University of Iceland, Iceland
Danka Jokić, University of Belgrade, Serbia
Nouran Khallaf, University of Leeds, UK
Gražina Korvel, Vilnius University, Lithuania
Sobha Lalitha Devi, AU-KBC Research Centre, Anna University, India
Salima Lamsiyah, University of Luxembourg, Luxembourg
Cedric Lothritz, Luxembourg Institute of Science and Technology, Luxembourg
Aleksandra Marković, Institute for the Serbian Language of SASA, Serbia
Javier Martín Arista, University of La Rioja, Spain
Brigitte Mathiak, GESIS – Leibniz Institute for the Social Sciences, Germany
So Miyagawa, University of Tsukuba, Japan
Simeon Monov, Plovdiv University, Bulgaria
Nikola S. Nikolov, University of Limerick, Ireland
Ivelina Nikolova, Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Bulgaria
Temilola Oloyede
Partha Pakray, National Institute of Technology Silchar, India
Nikolay Pavlov, Plovdiv University, Bulgaria
Scott Piao, Lancaster University, UK
Alistair Plum, University of Luxembourg, Luxembourg
Jihad R'baiti, Mohammed VI Polytechnic University, Morocco
Sara Renji
Branislava Šandrih Todorović, NLB DigIT, Belgrade, Serbia
Nasredine Semmar, CEA, France
Ranka Stanković, University of Belgrade, Serbia
Alina Trubnikova, University of Granada, Spain
Menno van Zaanen, South African Centre for Digital Language Resources, South Africa
Andreas Vlachidis, University College London, UK
Sumit Yadav, IOE Pulchowk Campus, Nepal
Jihad Zahir, Cadi Ayyad University, Morocco
Katerina Zdravkova, Ss. Cyril and Methodius University in Skopje, North Macedonia

Table of Contents

A Kazakh–Russian Corpus of Child-Directed Language for Low-Resource Languages
Albina Mukusheva, Achille Fusco and Cristiano Chesi 1

Data Curation, Annotation Quality, and Error Patterns in Old English Automatic Lemmatisation
Javier Martín Arista 8

Creating the RozMuz Corpus: Applying Ethics and Technology in Language Research
Alicja Helena Derych, Bartłomiej Alberski, Hubert Jankowski and Paweł Dembowski 18

Towards a Universal Dependencies Treebank for Amazigh: A Tarifit Pilot Study
Azzeddine Afrouni, Fadoua Ataa Allah and Jamal Abarnous 24

Arabic Dialect-to-MSA Translation: A Comparative Evaluation of Large Language Models and Neural Machine Translation
Maram I. Alharbi, Jihad R'baiti and Ruslan Mitkov 30

Cross-Lingual Transfer from Portuguese to Nheengatu: Evidence for Contact-Induced Convergence as a Computational Bridge
Rafael Macario Fernandes 40

Retrieval-Augmented Machine Translation for Bohairic Coptic: A Pilot Case Study
So Miyagawa 50

Automating Mwotlap morphology using formal grammar
Fabio Meroni, Alexandre François and Max Silberztein 59

Dependency Parsing Across the Resource Spectrum: Evaluating Architectures on High and Low-Resource Languages
Kevin Guan, Happy Buzaaba and Christiane Fellbaum 68

A finite-state model of Innu verbal inflection
Loïc Daignault-Pichette and François Lareau 80

Morphological Operations in the Persian Verbal System
Marzieh Rabiei and Max Silberztein 89

A Massive Open-Source Corpus for Valencian: Over 4.7 Billion Tokens for Low-Resource Language Modelling
Yoan Gutiérrez, Juan Pablo Consuegra-Ayala, Robiert Sepúlveda-Torres and Rafael Muñoz Guillena 96

LuxIT: A Luxembourgish Instruction Tuning Dataset from Monolingual Seed Data
Julian Valline, Cedric Lothritz, Siwen Guo and Jordi Cabot Sagrera 102

Topic Modeling for Moroccan Darija: A Comparative Study of Classical Machine Learning Approaches
Salma Mekaoui, Ilham Chaker, Arsalane Zarghili and Nikola S. Nikolov 121

Distil and Evolve: Compact Adaptive Document Classification with Continual Learning for Low-Resource Settings
Mehdi Soufiane, Mourad Hammale, Kawtar Zerhouni, Adil Ahidar-Coutrix and Rachid Benouini 131

Do Large Language Models Style-Shift? Register-Conditioned Stylistic Fronting in AI-Generated Icelandic
Anton Karl Ingason, Johanna Mechler and Lilja Björk Stefánsdóttir 140

Benchmarking Large Language Models on Mandarin Proverb Explanation and Contextual Matching
Xiaojing Zhao, Salima Lamsiyah, Emmanuele Chersoni and Han Xu 150

LUXDIAG-RAG: Diagnostic Evaluation of Retrieval-Augmented Generation for Luxembourgish Reading Comprehension
Keerthana Murugaraj, Hedi Tebourbi, Christophe Friezas Gonçalves and Salima Lamsiyah 161

Assessment of Human-in-the-Loop Multi-LLMs for Low-Resource Educational Data Expansion
Christophe Friezas Gonçalves, Hedi Tebourbi, Christoph Schommer and Salima Lamsiyah 172

Human-Centred Approaches in Low-Resource Languages for Educational Applications and Language Learning: A Systematic Literature Review
Noura El Moussa 181

SimuLe-Lux: A Neuro-Symbolic Teaching System for Diagnosing L2 Reading Comprehension
Christophe Friezas Gonçalves, Christoph Schommer and Salima Lamsiyah 191

Towards Readability Assessment for Under-Resourced Arabic Dialects: A Study of Moroccan Darija
Houdaifa Atou, Nouran Khallaf, Salima Lamsiyah and Ruslan Mitkov 199

Why Is Current XAI Not Enough for Arabic NLP? A Critical Survey of the Explainability Gap
Salima Lamsiyah and Ruslan Mitkov 226

Benchmarking Speech Foundation Models and LLMs for Sentiment Analysis in Low-Resource Najdi Arabic of Saudi Arabia
Nadia Ghezaiel Hammouda, Maram I. Alharbi and Ruslan Mitkov 237

Spatial Entity Extraction Methods from Arabic Texts in the Context of Epidemiological Surveillance: A Comparative Study
Fatima Ezzahra El Houbri, Najlae Idrissi, Mathieu Roche and Sarah Valentin 246

Building an Aligned Speech Corpus for Northern Russian
Kseniia Protonina and Daniil Ignatev 257

Probing Geographic Performance Gaps in Moroccan ASR with ALAMA: Annotated Local Audio of Moroccan Arabic
Avery Cole Kanel, Christian Schuler, Bouazza Laracha, Imrane Lbouhli, Yassine Chaouri, Yusser Al Ghussin and Timo Baumann 264

Generator-Guided Amount Recovery for Voice-Based Financial Record-Keeping in Mooré-French Code-Switched Speech
Maimouna Ouattara, El-Hacen Diallo, Fred Philippy, Abdoul Kader Kaboré, Jacques Klein and Tegawendé F. Bissyandé 272

LëtzCross: A Cross-Lingual Page-Level Benchmark for Multimodal Retrieval over Luxembourgish Documents
Omar El Bachyr, Fred Philippy, Laura Maria Bernardy, Saad Ezzini, Jacques Klein and Tegawendé Bissyandé 284

Improving OCR for a Latvian Pronunciation Dictionary
Viesturs Jūlijs Lasmanis 295

LLM-based conversational systems for Sinhala: Investigating model limitations and exploring a framework for agentic retrieval
Malithi P. Alahapperuma, Andreas Vlachidis and Antonis Bikakis 304

Clinical Entity Recognition from Electronic Health Records and Linking to Biomedical Knowledge Bases in Low-resource Language Settings
Eno-Martin Lotman, David Hübner, Kris Collins, Svetla Boytcheva and Ivelina Nikolova-Koleva 309

Beyond RAG: A Multi-Graph, Multi-Agent, Recursive Retrieval Architecture for Traceable and Source-Grounded Legal Answers
Assia Bouamir, Marie Bonnin, Youssef Al Mouatamid and Jihad Zahir 318

FLICK: Few-Label Incremental Learning for Low-Resource Dialects
Ali Almutairi, Abdullah Alsuhaibani, Shoaib Jameel, Aditya Joshi, Gelareh Mohammadi and Imran Razzak 326

Linguistic Proximity Enables Pivot-Based Machine Translation for an Under-Resourced Tribal Language
Pooja Singh, M Kaab Bin Shahid, Atai Waris Khan, Aryan Kumar Jha and Sandeep Kumar . . . 339

Cross-Lingual Hate Speech Detection in Low-Resource Languages: The Case of Persian and Kurdish
Shahin Yousefi, Ernesto Luis Estevanell-Valladares and Ruslan Mitkov 350

Is Overconfidence Language-Specific? Cross-Lingual Calibration and Recalibration Transfer in Base and Instruction-Tuned LLMs
Divya Gupta and Ruslan Mitkov 361

Knowledge Tracing for Early Childhood Learners with Minimal Telemetry from Low-Connectivity Environments: Evidence from India's Anganwadi Ecosystem
Badmavasan Kirouchenassamy, Rahul Singh, Vishnu Dev, Meenakshi Garg, Sukhna Sawhney and Sudeep Gowrishankar 370

Kuwain 1.5B: An Arabic SLM via Language Injection
Khalil Hennara, Sara Chrouf, Mohamed Motasim Hamed, Zeina Aldallal and Safwan AlModhayan 378

A Neuro-Symbolic RAG System for Marine Environmental Law: from Domain Ontology to Knowledge Graphs and Context-Enriched Legal Question Answering
Assoumana Souley Hadiza, Youssef Al Mouatamid, Marie Bonnin and Jihad Zahir 388

Evaluating Model-Task Fit in Arabic Word Sense Disambiguation
Yousef Younes, Abdelhalim Hafedh Dahou and Brigitte Mathiak 398

Enhancing Urdu ASR with Whisper v3: Fine-Tuning on Latest Datasets and Realistic Multi-Speaker Evaluation with SLM Post-Processing
Zehra Ahmed, Farah Inayat, Zuha Aqib and Sajjad Haider 408

Baseer: A Vision-Language Model for Arabic Document-to-Markdown OCR
Khalil Hennara, Muhammad Hreden, Mohamed Motasim Hamed, Ahmad Bastati, Zeina Aldallal, Sara Chrouf and Safwan AlModhayan 418

Worldview Annotation for Low-Resource Languages
Tetiana Ilman 429

A Multilingual Sentence-Transformer Baseline for Serbian News Topic Classification
Saša Petalinkar, Milica Ikonić Nešić, Ranka Stanković and Jelena Graovac 438

Wasm: A Pipeline for Constructing Structured Arabic Interleaved Multimodal Corpora
Khalil Hennara, Ahmad Bastati, Muhammad Hreden, Mohamed Motasim Hamed, Zeina Aldallal, Sara Chrouf and Safwan AlModhayan 449

Linguistic Specialization of Arabic in Text Embedding Models via Multi-Teacher Knowledge Distillation
Ahmad Abdelfattah, Zeina Aldallal, Sara Chrouf and Safwan AlModhayan 459

Conference Program

Monday September 28, 2026

10:50–11:50 **Session 1: Corpora, Annotation and Resource Development**
Chair: Yoan Gutiérrez

10:50–11:10 *A Kazakh–Russian Corpus of Child-Directed Language for Low-Resource Languages*
Albina Mukusheva, Achille Fusco and Cristiano Chesi

11:10–11:30 *Data Curation, Annotation Quality, and Error Patterns in Old English Automatic Lemmatisation*
Javier Martín Arista

11:30–11:40 *Creating the RozMuz Corpus: Applying Ethics and Technology in Language Research*
Alicja Helena Derych, Bartłomiej Alberski, Hubert Jankowski and Paweł Dembowski

11:40–11:50 *Towards a Universal Dependencies Treebank for Amazigh: A Tarifit Pilot Study*
Azzeddine Afrouni, Fadoua Ataa Allah and Jamal Abarnous

11:50–12:30 **Session 2: Machine Translation and Cross-Lingual Transfer**
Chair: Salima Lamsiyah

11:50–12:10 *Arabic Dialect-to-MSA Translation: A Comparative Evaluation of Large Language Models and Neural Machine Translation*
Maram I. Alharbi, Jihad R'baiti and Ruslan Mitkov

12:10–12:20 *Cross-Lingual Transfer from Portuguese to Nheengatu: Evidence for Contact-Induced Convergence as a Computational Bridge*
Rafael Macario Fernandes

12:20–12:30 *Retrieval-Augmented Machine Translation for Bohairic Coptic: A Pilot Case Study*
So Miyagawa

Monday September 28, 2026 (continued)

14:20–15:30 **Session 3: Morphology, Syntax and Formal Modelling**
Chair: Javier Martín Arista

14:20–14:40 *Automating Mwotlap morphology using formal grammar*
Fabio Meroni, Alexandre François and Max Silberztein

14:40–15:00 *Dependency Parsing Across the Resource Spectrum: Evaluating Architectures on High and Low-Resource Languages*
Kevin Guan, Happy Buzaaba and Christiane Fellbaum

15:00–15:20 *A finite-state model of Innu verbal inflection*
Loïc Daignault-Pichette and François Lareau

15:20–15:30 *Morphological Operations in the Persian Verbal System*
Marzieh Rabiei and Max Silberztein

15:50–17:10 **Session 4: Training Data and Model Adaptation**
Chair: Saad Ezzini

15:50–16:10 *A Massive Open-Source Corpus for Valencian: Over 4.7 Billion Tokens for Low-Resource Language Modelling*
Yoan Gutiérrez, Juan Pablo Consuegra-Ayala, Robiert Sepúlveda-Torres and Rafael Muñoz Guillena

16:10–16:30 *LuxIT: A Luxembourgish Instruction Tuning Dataset from Monolingual Seed Data*
Julian Valline, Cedric Lothritz, Siwen Guo and Jordi Cabot Sagrera

16:30–16:50 *Topic Modeling for Moroccan Darija: A Comparative Study of Classical Machine Learning Approaches*
Salma Mekaoui, Ilham Chaker, Arsalane Zarghili and Nikola S. Nikolov

16:50–17:10 *Distil and Evolve: Compact Adaptive Document Classification with Continual Learning for Low-Resource Settings*
Mehdi Soufiane, Mourad Hammale, Kawtar Zerhouni, Adil Ahidar-Coutrix and Rachid Benouini

Tuesday September 29, 2026

09:00–10:00 **Session 5: Language Model Behaviour and Evaluation**
Chair: Ernesto Luis Estevanell-Valladares

09:00–09:20 *Do Large Language Models Style-Shift? Register-Conditioned Stylistic Fronting in AI-Generated Icelandic*
Anton Karl Ingason, Johanna Mechler and Lilja Björk Stefánsdóttir

09:20–09:40 *Benchmarking Large Language Models on Mandarin Proverb Explanation and Contextual Matching*
Xiaojing Zhao, Salima Lamsiyah, Emmanuele Chersoni and Han Xu

09:40–10:00 *LUXDIAG-RAG: Diagnostic Evaluation of Retrieval-Augmented Generation for Luxembourgish Reading Comprehension*
Keerthana Murugaraj, Hedi Tebourbi, Christophe Friezas Gonçalves and Salima Lamsiyah

10:50–11:50 **Session 6: Human-Centred NLP and Education**
Chair: Salima Lamsiyah

10:50–11:10 *Assessment of Human-in-the-Loop Multi-LLMs for Low-Resource Educational Data Expansion*
Christophe Friezas Gonçalves, Hedi Tebourbi, Christoph Schommer and Salima Lamsiyah

11:10–11:30 *Human-Centred Approaches in Low-Resource Languages for Educational Applications and Language Learning: A Systematic Literature Review*
Noura El Moussa

11:30–11:50 *SimuLe-Lux: A Neuro-Symbolic Teaching System for Diagnosing L2 Reading Comprehension*
Christophe Friezas Gonçalves, Christoph Schommer and Salima Lamsiyah

Tuesday September 29, 2026 (continued)

11:50–12:50 **Session 7: Arabic and Dialect NLP**
Chair: Salmane Chafik

11:50–12:10 *Towards Readability Assessment for Under-Resourced Arabic Dialects: A Study of Moroccan Darija*
Houdaifa Atou, Nouran Khallaf, Salima Lamsiyah and Ruslan Mitkov

12:10–12:30 *Why Is Current XAI Not Enough for Arabic NLP? A Critical Survey of the Explainability Gap*
Salima Lamsiyah and Ruslan Mitkov

12:30–12:40 *Benchmarking Speech Foundation Models and LLMs for Sentiment Analysis in Low-Resource Najdi Arabic of Saudi Arabia*
Nadia Ghezaiel Hammouda, Maram I. Alharbi and Ruslan Mitkov

12:40–12:50 *Spatial Entity Extraction Methods from Arabic Texts in the Context of Epidemiological Surveillance: A Comparative Study*
Fatima Ezzahra El Houbri, Najlae Idrissi, Mathieu Roche and Sarah Valentin

14:20–15:30 **Session 8: Speech and Spoken-Language Technologies**
Chair: Nadia Ghezaiel

14:20–14:40 *Building an Aligned Speech Corpus for Northern Russian*
Kseniia Protonina and Daniil Ignatev

14:40–15:00 *Probing Geographic Performance Gaps in Moroccan ASR with ALAMA: Annotated Local Audio of Moroccan Arabic*
Avery Cole Kanel, Christian Schuler, Bouazza Laracha, Imrane Lbouhli, Yassine Chaouri, Yusser Al Ghussin and Timo Baumann

15:00–15:20 *Generator-Guided Amount Recovery for Voice-Based Financial Record-Keeping in Mooré-French Code-Switched Speech*
Maimouna Ouattara, El-Hacen Diallo, Fred Philippy, Abdoul Kader Kaboré, Jacques Klein and Tegawendé F. Bissyandé

Tuesday September 29, 2026 (continued)

15:50–17:00 **Session 9: Retrieval, RAG and Document Processing**
Chair: Houdaifa Atou

15:50–16:10 *LëtzCross: A Cross-Lingual Page-Level Benchmark for Multimodal Retrieval over Luxembourgish Documents*
Omar El Bachyr, Fred Philippy, Laura Maria Bernardy, Saad Ezzini, Jacques Klein and Tegawendé Bissyandé

16:10–16:30 *Improving OCR for a Latvian Pronunciation Dictionary*
Viesturs Jūlijs Lasmanis

16:30–16:40 *LLM-based conversational systems for Sinhala: Investigating model limitations and exploring a framework for agentic retrieval*
Malithi P. Alahapperuma, Andreas Vlachidis and Antonis Bikakis

16:40–16:50 *Clinical Entity Recognition from Electronic Health Records and Linking to Biomedical Knowledge Bases in Low-resource Language Settings*
Eno-Martin Lotman, David Hübner, Kris Collins, Svetla Boytcheva and Ivelina Nikolova-Koleva

16:50–17:00 *Beyond RAG: A Multi-Graph, Multi-Agent, Recursive Retrieval Architecture for Traceable and Source-Grounded Legal Answers*
Assia Bouamir, Marie Bonnin, Youssef Al Mouatamid and Jihad Zahir

Session VID: Video Presentations (on the conference website)

FLICK: Few-Label Incremental Learning for Low-Resource Dialects
Ali Almutairi, Abdullah Alsuhaibani, Shoaib Jameel, Aditya Joshi, Gelareh Mohammadi and Imran Razzak

Linguistic Proximity Enables Pivot-Based Machine Translation for an Under-Resourced Tribal Language
Pooja Singh, M Kaab Bin Shahid, Atai Waris Khan, Aryan Kumar Jha and Sandeep Kumar

Cross-Lingual Hate Speech Detection in Low-Resource Languages: The Case of Persian and Kurdish
Shahin Yousefi, Ernesto Luis Estevanell-Valladares and Ruslan Mitkov

Is Overconfidence Language-Specific? Cross-Lingual Calibration and Recalibration Transfer in Base and Instruction-Tuned LLMs
Divya Gupta and Ruslan Mitkov

Knowledge Tracing for Early Childhood Learners with Minimal Telemetry from Low-Connectivity Environments: Evidence from India's Anganwadi Ecosystem
Badmavasan Kirouchenassamy, Rahul Singh, Vishnu Dev, Meenakshi Garg, Sukhna Sawhney and Sudeep Gowrishankar

Tuesday September 29, 2026 (continued)

Kuwain 1.5B: An Arabic SLM via Language Injection
Khalil Hennara, Sara Chrouf, Mohamed Motasim Hamed, Zeina Aldallal and Safwan AlModhayan

A Neuro-Symbolic RAG System for Marine Environmental Law: from Domain Ontology to Knowledge Graphs and Context-Enriched Legal Question Answering
Assoumana Souley Hadiza, Youssef Al Mouatamid, Marie Bonnin and Jihad Zahir

Evaluating Model-Task Fit in Arabic Word Sense Disambiguation
Yousef Younes, Abdelhalim Hafedh Dahou and Brigitte Mathiak

Enhancing Urdu ASR with Whisper v3: Fine-Tuning on Latest Datasets and Realistic Multi-Speaker Evaluation with SLM Post-Processing
Zehra Ahmed, Farah Inayat, Zuha Aqib and Sajjad Haider

Baseer: A Vision-Language Model for Arabic Document-to-Markdown OCR
Khalil Hennara, Muhammad Hreden, Mohamed Motasim Hamed, Ahmad Bastati, Zeina Aldallal, Sara Chrouf and Safwan AlModhayan

Worldview Annotation for Low-Resource Languages
Tetiana Ilman

A Multilingual Sentence-Transformer Baseline for Serbian News Topic Classification
Saša Petalinkar, Milica Ikonić Nešić, Ranka Stanković and Jelena Graovac

Wasm: A Pipeline for Constructing Structured Arabic Interleaved Multimodal Corpora
Khalil Hennara, Ahmad Bastati, Muhammad Hreden, Mohamed Motasim Hamed, Zeina Aldallal, Sara Chrouf and Safwan AlModhayan

Linguistic Specialization of Arabic in Text Embedding Models via Multi-Teacher Knowledge Distillation
Ahmad Abdelfattah, Zeina Aldallal, Sara Chrouf and Safwan AlModhayan

A Kazakh–Russian Corpus of Child-Directed Language for Low-Resource Languages

Albina Mukusheva
IUSS Pavia
Piazza della Vittoria 15
Pavia, Italy
albina.mukusheva@iusspavia.it

Achille Fusco
IUSS Pavia
Piazza della Vittoria 15
Pavia, Italy
achille.fusco@iusspavia.it

Cristiano Chesi
IUSS Pavia
Piazza della Vittoria 15
Pavia, Italy
cristiano.chesi@iusspavia.it

Abstract

Morphologically rich and low-resource languages present challenges for tokenization and other natural language processing (NLP) tasks. Standard tokenization algorithms often fail to segment words into meaningful morphological units. At the same time, datasets for many low-resource languages remain limited. We introduce a new [1]dataset of child-directed language in Kazakh and Russian compiled from multiple sources including child–adult dialogue transcripts, fairy tales, cartoons and children's literature texts. The dataset is manually collected and contains 1M tokens in Kazakh and 2.3M tokens in Russian. Dialogue data includes metadata such as child age and speaker identity, enabling research on language acquisition and linguistic development. We describe the process of data collection, corpus organization, and basic statistical properties of the dataset. The corpus provides a new resource for research on low-resource NLP, language acquisition, and morphologically-aware tokenization methods.

1 Introduction

Both agglutinative and morphologically rich languages present persistent challenges for natural language processing because extensive word formation increases surface-form variation and contributes to data sparsity. Kazakh and Russian provide an informative contrast in this respect. Kazakh, an agglutinative Turkic language, forms words by attaching sequences of affixes to stable stems, whereas Russian exhibits complex inflectional morphology across nouns, verbs, and adjectives. These linguistic characteristics make both languages valuable for computational and comparative linguistic research.

At the same time, progress in NLP for low-resource languages is often constrained by the lack of suitable datasets. Although a number of corpora are available for Russian, resources for Kazakh remain limited, particularly in domains that capture natural child-directed language.

Child-directed texts, such as children's literature, provide linguistically accessible material while still preserving productive morphological patterns. In addition, transcripts of child–adult interaction offer valuable insight into language acquisition and developmental linguistic patterns that are not typically represented in standard written corpora. In Kazakhstan, many children are exposed to both Kazakh and Russian from an early age, requiring them to acquire and process two typologically distinct morphological systems simultaneously. A bilingual corpus of child-directed input therefore provides not only a resource for NLP but also a foundation for investigating how bilingual children encounter and learn different patterns of word formation across the two languages.

The dataset presented in this work combines narrative texts and dialogue transcripts to support research in low-resource NLP, child language modeling, and tokenization methods. The contributions of this work are:

1. A new Kazakh-Russian corpus of child-directed language.

2. A dataset combining narrative texts and child-adult dialogue transcripts.

3. A statistical and linguistic analysis of the corpus.

2 Related Work

Several corpora have been developed to support research on child language acquisition. One of the most widely used resources is CHILDES (MacWhinney, 2000), which provides transcripts of child–adult interactions across multiple languages. Russian child language data is available

[1]https://huggingface.co/datasets/esimijoq/Kazakh-Russian-Child-Directed-Speech-Corpus

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 1–7
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_001

within this resource, though its size remains relatively limited for large-scale NLP experiments.

For low-resource languages, the development of linguistic resources remains a major research challenge. Many NLP studies rely on large datasets such as news corpora or web crawls, which are rarely available for languages with smaller digital presence. Kazakh, despite being widely spoken, still lacks large publicly available corpora suitable for training and evaluating NLP models.

Children's literature texts have often been used as sources for linguistic datasets because they provide grammatically rich but accessible language. Such texts typically contain simpler syntactic structures while preserving complex morphological patterns. The corpus introduced in this work complements existing resources by combining child speech transcripts and children's literature texts in Kazakh and Russian.

3 Linguistic Characteristics

Kazakh and Russian differ substantially in how grammatical information is encoded.

Kazakh morphology is characterized by its agglutinative structure, in which affixes are added sequentially to a stable root while preserving clear morpheme boundaries, and each affix typically expresses a single grammatical meaning (Zhanpeyisov, 2002). The language distinguishes between root and affixal morphemes, with affixes functioning either derivationally to form new words or inflectionally to mark grammatical categories (Kaliakbarova, 2013). Important morphological features include the extensive use of possessive, case, and personal endings to indicate grammatical relationships and syntactic roles within the sentence (Momynova et al., 2024). Kazakh morphology is also notable for the absence of grammatical gender and for expressing meanings that many other languages convey through prepositions by means of case endings and postpositions. Kazakh additionally employs multiple writing systems across regions and historical contexts. In Kazakhstan, both Cyrillic and Latin scripts are used, while Kazakh communities in China, primarily use an Arabic-based script (Abish, 2016). The Kazakh corpus used in this study is written in Cyrillic script.

Russian exhibits a fusional synthetic morphological system, in which grammatical meanings are encoded mainly within the word itself through inflection, with a single ending often simultaneously expressing categories such as number, gender, case, person, and tense (Bogdanov et al., 2009). An important feature of this system is the use of zero inflections, where the absence of an overt ending still carries grammatical meaning. Alongside these synthetic patterns, Russian also employs analytic constructions, where grammatical meaning is conveyed by function words, as well as combined forms, where meaning is expressed through both a preposition and an inflected word. In some cases, Russian uses suppletion, where different grammatical forms of the same lexeme derive from different roots (Kukatova, 2020). Word forms are created not only through affixation but also by sound alternation, stem modification, and stress shift, and the complete set of these forms constitutes a word's paradigm, which may be either complete or defective (Kamynina, 1999).

4 Dataset Construction

To reflect the bilingual linguistic environment of many children in Kazakhstan, the corpus includes child-directed materials in both Kazakh and Russian. By representing child-directed input in both languages, the corpus provides a comparative resource for examining how bilingual children process two typologically distinct morphological systems simultaneously. Because the corpus combines dialogue and narrative material, it can support broader natural language processing tasks, including language modeling of early linguistic input, bilingual lexical prediction, morphological analysis, and cross-lingual representation learning.

4.1 Data Sources

The corpus was compiled from multiple sources of child-directed language in Kazakh and Russian.

The Kazakh portion of the dataset contains approximately 1 million tokens collected from:

- Kazakh fairy tales,
- Cartoon subtitles (from OpenSubtitles [2] and manually collected),
- Texts from children's journals,
- Stories from children's books,
- Educational materials.

These texts are primarily written narratives and instructional materials directed toward young children. The table 1 below provides a quantitative

[2] https://www.opensubtitles.org/

summary of the Kazakh corpus, including the number of documents, total tokens, and unique types for each source category.

source	docs	tokens	types
Fairy_tales	176	81609	16375
Cartoon_subtitles	97	143780	28466
Children_journals	1021	219704	41655
Children_books	3	166490	35711
Educational_materials	358	460334	66485
total	1655	1071917	127773

Table 1: Quantitative summary of the Kazakh corpus.

The Russian portion of the dataset consists of approximately 2.3 million tokens collected from three main sources:

- Child–adult dialogue transcripts (from CHILDES (Lelik et al., 2025; Bar-Shalom and Snyder, 1997, 1998) and from *From the First Words to the First Class* (Gvozdev, 1981)),
- Russian folk and fairy tales,
- Fairy tales from other cultures translated into Russian.

The data cover five children recorded across different developmental periods: Varvara (1;6–2;10), Yasha (0;10–3;10), Tosya (0;10–3;10), Tanja (2;5–2;11), and Zhenya (1;0–7;9).

The Russian texts were organized into three groups. The first group contains age-specific fairy tales divided into the following age categories: 0–2, 3, 4, 5, 6, 7, and 8–11 years old. The second group consists of traditional Russian folk tales. The third group includes fairy tales translated into Russian from a wide range of source languages, from Abkhaz to Yakut.

The table 2 below shows a quantitative summary of the Russian corpus, including the number of documents, total tokens, and unique types for each source category.

4.2 Data Format

The dialogue transcripts are stored in spreadsheet format containing three columns: child age, speaker identity, and the utterance text. An example is shown in Table 3; the transliteration and English translation presented in the table were added for illustrative purposes and are not included in the original dataset.

source	docs	tokens	types
Dialogues	1	294608	29623
Tales_by_ages	182	213869	30939
Folk_tales	200	171642	22920
Translated_tales	1605	1589193	84693
total	1988	2269312	116611

Table 2: Quantitative summary of the Russian corpus.

Age	Speakers	Sentences
02;05;14	INTERVIEWER	Что? Chto? “What?”
02;05;14	CHILD	Не может открыть. Ne mozhet otkryt’. “He can’t open it.”
02;05;14	MOTHER	А рот не может открыть. A rot ne mozhet otkryt’. “And he can’t open his mouth.”
02;05;14	MOTHER	А почему? A pochemu? “And why?”
02;05;14	CHILD	Не может. Ne mozhet. “He can’t.”

Table 3: Example of Russian dialogue transcript

The narrative texts are stored in txt format, with each document containing the title followed by the narrative body which separated by a blank line.

5 Corpus Statistics

5.1 Overall corpus statistics

For statistical analysis, each narrative text file in both corpora was treated as an individual document, whereas the Russian dialogue spreadsheet was treated as a single document containing all dialogue utterances. Analyses were conducted on surface word forms in order to capture the full range of lexical and morphological variation present in child-directed input.

Table 4 presents lexical diversity measures for each corpus. To obtain a comprehensive assessment of lexical diversity, several complementary metrics were employed. Type-token ratio (TTR) (Templin, 1957) was included as a traditional and intuitive measure of vocabulary variation, calculated as the proportion of unique word types relative to the total number of tokens. However, because TTR decreases as text length increases, it is

Corpus	TTR	MATTR	MTLD	HLR	Guiraud's Index	D Index (Vocd-D)
Kazakh	0.1192	0.9134	1.0000	0.5374	123.4123	0.9833
Russian	0.0514	0.8411	1.0000	0.4336	77.4092	0.9299

Table 4: Lexical diversity measures for Kazakh and Russian corpora

primarily suitable for preliminary comparisons between similarly sized texts.

To address this limitation, moving-average type-token ratio (MATTR) (Covington and McFall, 2010) was also calculated. MATTR computes the average TTR across fixed-length moving windows and is substantially less sensitive to document length, making it more reliable for comparing corpora of unequal size. Measure of textual lexical diversity (MTLD) (McCarthy, 2005) was additionally included because it is considered as one of the most robust length-independent lexical diversity measures. The hapax legomena ratio (HLR) (Popescu and Altmann, 2008), representing the proportion of words occurring only once, was used to assess vocabulary novelty and lexical richness. Guiraud's Index (Guiraud, 1954) was included as a normalized alternative to TTR, reducing text-length effects by scaling the number of word types by the square root of the total number of tokens. Finally, the D Index (Vocd-D) (Malvern et al., 2004) was calculated as a probabilistic estimate of lexical diversity based on Monte Carlo sampling procedures.

The Kazakh corpus demonstrates consistently higher lexical diversity than the Russian corpus across nearly all measures. In particular, Kazakh shows higher values for TTR, MATTR, HLR, Guiraud's Index, and Vocd-D, indicating a broader and more varied vocabulary. The MTLD values are identical across the two corpora, suggesting comparable levels of length-independent lexical diversity according to this metric.

5.2 Source-level statistics

To provide a more detailed characterization of the dataset, the same statistical analysis was performed on lemmatized texts to reduce inflation of diversity measures caused by morphological variation and to improve comparability across genres.

Lemmatization was performed using the Stanza Python NLP library (Qi et al., 2020) for both the Kazakh and Russian corpora. The choice of Stanza was motivated primarily by the limited availability of lemmatization tools for Kazakh. We do not expect these models to perform perfectly, given the challenges associated with processing under-resourced languages. It was therefore considered important to report performance statistics for Stanza in these languages and to acknowledge the possibility of lemmatization errors.

Lang	Acc (%)	aLd	Prec	Rec	F1
Kazakh	81.40	0.346	0.835	0.814	0.818
Russian	83.20	0.332	0.857	0.832	0.834

Table 5: Lemmatization accuracy on manually annotated samples.

To assess the reliability of automatic lemmatization, a random sample of 500 tokens from each corpus was manually annotated with gold-standard lemmas and compared with the lemmas predicted by Stanza. Performance was evaluated using exact-match accuracy, average Levenshtein distance, and weighted precision, recall, and F1 score.

Morph. Seg.	Analysis	POS
қожа+мыз+дан	Gold: қожа	NOUN
master+POSS.1PL+ABL	Pred.: қожамыз	
'from our master'		
қаш+ай+ық	Gold: қаш	VERB
flee+HORT+1PL	Pred.: қаша	
'let us flee'		
ой+ы+мыз+ға	Gold: ой	NOUN
thought+POSS.1PL+DAT	Pred.: ойымыз	
'to our thought'		

Table 6: Kazakh lemmatization error examples with glosses and English translations.

Many existing Kazakh morphological tools rely on stemming-based methodologies (Rakhimova and Turganbayeva, 2019; Sultanova et al., 2019; Bogdanchikov et al., 2021), which often fail to preserve lexical consistency and may introduce additional noise into lexical diversity measurements. Although manual review indicated that Stanza's Kazakh lemmatizer is not entirely error-free, it of-

Corpus	Source category	TTR	MATTR	MTLD	HLR	Guiraud's Index	D Index (Vocd-D)
KZ	Fairy_tales	0.0882	0.7870	1.0010	0.4720	24.7368	0.9121
KZ	Cartoons_subtitles	0.0980	0.8259	1.0007	0.5347	36.9920	0.9168
KZ	Children_journals	0.0804	0.8280	1.0010	0.4822	37.7399	0.9590
KZ	Children_books	0.0945	0.8864	1.0023	0.4891	38.6663	0.9541
KZ	Educational_materials	0.0630	0.8330	1.0002	0.4964	41.9438	0.9402
RU	Dialogues	0.0688	0.6290	1.0001	0.5412	37.3533	0.8849
RU	Tales_by_ages	0.0730	0.8179	1.0002	0.4357	33.7357	0.9077
RU	Folk_tales	0.0712	0.8035	1.0000	0.4265	29.4929	0.9153
RU	Translated_tales	0.0226	0.8186	1.0001	0.3518	28.4351	0.9005

Table 7: Lexical diversity measures by source category in Kazakh and Russian corpora

fers a unified framework capable of performing full lemmatization for both languages within a single, integrated pipeline. Examples of lemmatization errors in the Kazakh corpus are shown in Table 6.

Table 7 presents the lexical diversity measures calculated for each source category in the Kazakh and Russian corpora. In the Kazakh corpus, *Children_books* exhibit the highest lexical diversity overall, with the highest MATTR, MTLD, and relatively high Guiraud's Index values, suggesting that this category contains the broadest and most varied vocabulary among the Kazakh sources. *Educational_materials* show the highest Guiraud's Index, indicating substantial vocabulary growth relative to text length, although their TTR values are comparatively lower. *Cartoons_subtitles* and *Children_journals* display similarly high MATTR scores, reflecting relatively rich local lexical variation. By contrast, *Fairy_tales* show lower MATTR and Guiraud's Index values, suggesting comparatively less lexical diversity across texts.

In the Russian corpus, lexical diversity measures are more homogeneous across source categories. *Translated_tales* show the highest MATTR value, while *Tales_by_ages* display slightly higher Guiraud's Index scores than the other Russian written sources. *Folk_tales* and *Tales_by_ages* exhibit broadly similar lexical diversity profiles across most measures. Although *Translated_tales* have a notably lower TTR value, their MATTR and MTLD scores remain comparable to the other Russian categories, indicating that lexical diversity remains relatively stable when text-length effects are controlled. Overall, the Russian written materials demonstrate less variation between source categories than the Kazakh corpus.

6 Future Work

Future work will focus on expanding the corpus with additional child-directed materials in both Kazakh and Russian. In particular, we plan to collect more spoken and written data from age-specific sources such as children's conversations, educational materials, storybooks, and audiovisual transcripts. Increasing the amount and diversity of child-directed language will improve the linguistic coverage of the corpus and provide a broader representation of developmental language patterns across both languages.

Another direction of future work concerns the development of evaluation benchmarks for morphological generalization. Inspired by the SIGMORPHON 2022 shared tasks (Batsuren et al., 2022), we plan to investigate the construction of benchmark datasets for Kazakh based on child-directed language resources. Such benchmarks would provide a standardized framework for evaluating language models and tokenization methods on morphologically rich and low-resource languages.

In addition, the corpus offers an opportunity to investigate tokenization strategies tailored to distinct morphological systems. Because the dataset contains parallel challenges from agglutinative and fusional languages, it provides a suitable foundation for comparing and developing tokenization approaches that can better capture morphological structure. Recent work in the BabyLM shared task has highlighted the importance of tokenization choices for both intrinsic and extrinsic evaluation (Fusco et al., 2025). Building on these findings, we plan to further investigate how different tokenization methods affect text representations and language model performance in low-resource set-

tings.

Finally, the dataset is intended to support the training of language models for low-resource settings. We plan to investigate how child-directed corpora can improve small-scale multilingual and monolingual language models in Kazakh and Russian, particularly in tasks that require better morphological generalization. This resource may therefore contribute not only to tokenization research but also to broader NLP development for underrepresented languages.

References

Aynur Abish. 2016. *Modality in Kazakh as Spoken in China*, 1 edition. Harrassowitz Verlag.

Eva Bar-Shalom and William Snyder. 1997. Optional infinitives in russian and their implications for the pro-drop debate. In Marina Lindseth and Steven Franks, editors, *Formal Approaches to Slavic Linguistics: The Indiana Meeting 1996*, pages 38–47. Michigan Slavic Publications, Ann Arbor.

Eva Bar-Shalom and William Snyder. 1998. Root infinitives in child russian: A comparison with italian and polish. In Richard Shillcock, Antonella Sorace, and Caroline Heycock, editors, *Language Acquisition: Knowledge Representation and Processing. Proceedings of GALA '97*. The University of Edinburgh, Edinburgh, UK.

Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinovic, Kyle Gorman, Zdeněk Žabokrtský, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, Fausto Giunchiglia, Ryan Cotterell, and Ekaterina Vylomova. 2022. The SIGMORPHON 2022 shared task on morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 103–116, Seattle, Washington. Association for Computational Linguistics.

A. V. Bogdanchikov, O. A. Baimuratov, and D. A. Ayazbayev. 2021. Stemming of kazakh language. *News of the National Academy of Sciences of the Republic of Kazakhstan*.

S. I. Bogdanov, V. B. Evtyukhin, Yu. P. Knyazev, et al. 2009. *Морфология современного русского языка [Morphology of Modern Russian]*. Русский мир: учебники для высшей школы [Russian World: Textbooks for Higher Education]. Faculty of Philology and Arts, Saint Petersburg State University, Saint Petersburg.

Michael A. Covington and Joe D. McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of Quantitative Linguistics*, 17(2):94–100.

Achille Fusco, Maria Letizia Piccini Bianchessi, Tommaso Sgrizzi, Asya Zanollo, and Cristiano Chesi. 2025. Linguistic units as tokens: Intrinsic and extrinsic evaluation with babylm. In *Proceedings of the First BabyLM Workshop*, pages 496–507. Association for Computational Linguistics.

P. Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Pierre Guiraud. 1954. *Les caractères statistiques du vocabulaire: essai de méthodologie*. Presses universitaires de France, Paris.

A. N. Gvozdev. 1981. *От первых слов до первого класса: дневник научных наблюдений [From the First Words to the First Class: A Journal of Scientific Observations]*. Saratov State University, Saratov. (In Russian).

Gulnur Kaliakbarova. 2013. *Қазақ тілі (қысқаша грамматикалық анықтағыш) [Kazakh Language (A Brief Grammatical Reference Guide)]*. Institute for the Development of the State Language, Almaty.

A. A. Kamynina. 1999. *Современный русский язык. Морфология [Modern Russian Language. Morphology]: Учебное пособие для студентов филологических факультетов государственных университетов [Study Guide for Students of Philological Faculties of State Universities]*. Moscow State University Press, Moscow.

O. A. Kukatova. 2020. *Современный русский язык. Морфология [Modern Russian Language. Morphology]*. Innovatsion Rivozhlanish Nashriyot-Matbaa Uyi, Tashkent.

Viktoria Lelik, Maria Diachkova, Svetlana V. Dorofeeva, Anastasia Lopukhina, Olga Dragoy, and Irina A. Sekerina. 2025. Ruslan-m corpus. Retrieved from `https://childes.talkbank.org`.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*, 3 edition. Lawrence Erlbaum Associates, Mahwah, NJ.

D. D. Malvern, B. J. Richards, N. Chipere, and P. Durán. 2004. *Lexical diversity and language development: quantification and assessment*. Palgrave Macmillan, Basingstoke.

Philip M McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, The University of Memphis.

B. Momynova, A. Soltanbekova, U. Anesova, T. Ramazanov, M. Imangazina, and Q. Slyambekov. 2024. *Қазақ тілінің кестелі грамматикасы. Морфология [Tabular Grammar of the Kazakh Language. Morphology]*. Asyl ZhK, Almaty.

Ion-Iovitz Popescu and Gabriel Altmann. 2008. Hapax legomena and language typology. *Journal of Quantitative Linguistics*, 15(4):370–378.

Prime Minister of the Republic of Kazakhstan. 2021. Askar mamin chairs meeting of national commission on translation of kazakh alphabet into latin script.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Diana Rakhimova and Aliya Turganbayeva. 2019. Lemmatization of big data in the kazakh language. In *Proceedings of the 5th International Conference on Engineering and MIS*, ICEMIS '19, New York, NY, USA. Association for Computing Machinery.

Nazerke Sultanova, Kanat Kozhakhmet, Ruslan Jantayev, and Azhar Botbayeva. 2019. Stemming algorithm for kazakh language using rule-based approach. In *2019 15th International Conference on Electronics, Computer and Computation (ICECCO)*, pages 1–4.

Mildred C Templin. 1957. Certain language skills in children; their development and interrelationships.

Erbol Zhanpeyisov. 2002. *Қазақ грамматикасы. Фонетика, сөзжасам, морфология, синтаксис [Kazakh Grammar: Phonetics, Word Formation, Morphology, Syntax]*. Astana.

Data Curation, Annotation Quality, and Error Patterns in Old English Automatic Lemmatisation

Javier Martín Arista
Universidad de La Rioja
javier.martin@unirioja.es

Abstract

Old English (c. 450–1150 CE), preserved in approximately three million words, is a historical language that poses distinctive problems for natural language processing: pervasive spelling variation, rich inflectional morphology, and heterogeneous annotated sources. This paper provides an analysis of the data curation pipeline, annotation quality, and linguistically motivated error patterns for Old English lemmatisation. We describe the extraction, harmonisation, and validation of training data from three historical sources, including a parsed corpus, a scholarly dictionary, and a historical dictionary, each with different formats, POS conventions, and coverage. Quantitative analysis shows that 11.3% of word types receive inconsistent lemma assignments, identity transformations account for 28.3% of training data, and the *ge-* prefix affects 11.7% of annotated instances. Error analysis organised by morphological category shows that strong-verb ablaut accounts for approximately 25% of lemmatisation errors, spelling variation for 20%, and prefix handling for 15%. We contextualise these findings within the broader landscape of low-resource NLP and formulate practical guidelines for building lemmatisation systems for historical low-resource languages, including recommendations on annotation consistency, POS completeness, and source harmonisation.

1 Introduction

Applying modern Natural Language Processing (NLP) to historical languages is hampered by limited annotated data, spelling variation, and the absence of standardised orthography (Piotrowski, 2012). While much attention has gone to neural architectures for lemmatisation, POS tagging, and parsing, far less has addressed data preparation, annotation quality, and systematic error analysis—factors that often determine the ceiling of model performance more directly than architectural choices.

This paper focuses on these under-explored aspects using Old English (c. 450–1150 CE) as a case study. Old English is the earliest documented form of English, preserved in approximately three million words digitised as the Dictionary of Old English Corpus (DOEC; Healey, 2012). It is a paradigmatic low-resource language: generalised spelling variation across scribal traditions, rich morphology, a small annotated corpus relative to the complexity of the language, and multiple digital resources with barely compatible annotation conventions.

Character-level Transformer models achieve the highest lemmatisation accuracy reported to date for Old English: the best model of Martín Arista (2026) reaches approximately 68% lemma accuracy on development data and 67.58% on the held-out test set (the error analysis in Section 5 uses the test figure). This is well below accuracies for better-resourced historical languages such as Latin (∼85–90%; Passarotti, 2004) and Middle Dutch (∼80–94%; Kestemont et al., 2017). These figures come from different datasets, annotation schemes, and evaluation settings, and are indicative rather than directly comparable; nevertheless, the size of the gap motivates looking beyond model architecture to the training data themselves.

Lemmatisation—the mapping of inflected forms to their dictionary headwords—is fundamental to corpus linguistics and the digital humanities: without it, scholars must manually normalise thousands of variant spellings and inflections before any study. For Old English it enables concordance generation, frequency analysis, and dictionary linking, among other tasks.

2 Old English as a Low-Resource Language

Old English is a West Germanic language attested across roughly seven centuries of scribal activity. Four properties make it both relevant and challeng-

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 8–17
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_002

ing for NLP: inflectional morphology, spelling variation, character variants, and limited annotated resources.

Old English marks case, gender, number, and tense through complex, largely explicit paradigms. A single lemma may surface in dozens of forms: the verb *bēon* 'to be' has over 15, including *eom*, *eart*, *is*, *sind*, *bēo*, *wæs*, and *wǣron*. Nouns inflect for five cases and two numbers, and strong verbs exhibit ablaut (systematic vowel alternation) that changes the stem rather than attaching a suffix.

Orthographically, Old English lacked established spelling conventions. The same word may appear in different forms across or even within a manuscript: 'king' may be spelt *cyning*, *kyning*, or *cyng*. It also employs characters absent from Latin script—<þ> (thorn), <ð> (eth), and <æ> (ash): thorn and eth are used interchangeably for dental fricatives, and ash alternates with <a> or <e> across traditions.

Regarding the availability of annotated resources, several digital resources support the computational analysis of Old English, although they differ widely in scope, format, and annotation depth. The York-Toronto-Helsinki Parsed Corpus of Old English Prose (Taylor et al., 2003) and the York-Helsinki Parsed Corpus of Old English Poetry (Pintzuk and Plug, 2001) together provide a POS-tagged and parsed corpus of approximately 1.5 million words, known collectively as the YCOE. The Dictionary of Old English (DOE; Healey, 2018) offers comprehensive digitised dictionary entries for the letters A–I. The Bosworth-Toller dictionary (Crist and Tichý, 2014) is a fully digitised historical Anglo-Saxon dictionary covering the entire alphabet. ParCorOEv3 (Martín Arista et al., 2023) provides a parallel corpus of Old English with Universal Dependencies annotation. The verbal lemmas from the YCOE prose have been semi-automatically gathered by means of morphological generation (Metola Rodríguez, 2015; García Fernández, 2018; Tío Sáenz, 2019), and the lemmas from poetical texts have been compiled by Hamdoun Bghiyel (2025). Finally, the VARIOE morphological dictionary (Cichosz et al., 2021) provides manually compiled lemmas for the YCOE. These resources, while individually valuable, are not mutually compatible: they adopt different lemma inventories, spelling conventions, POS tagsets, and treatments of derivational morphology, which give rise to dire incompatibilities when attempting to combine them for NLP training.

Lemmatisation of historical languages ranges from rule-based morphological analysers (Lindén and Pirinen, 2009) to dictionary lookup (Passarotti, 2004). Neural approaches have advanced quickly: Manjavacas et al. (2019) showed that sequence-to-sequence models outperform rule-based systems for non-standard varieties, and Kestemont et al. (2017) applied character-level convolutions with word embeddings for Middle Dutch, treating lemmatisation as classification over a fixed lemma set. Bergmanis and Goldwater (2018) introduced context-sensitive lemmatisation with bidirectional LSTMs and attention, handling unseen words effectively, and Kondratyuk and Straka (2019) presented UDPipe 2.0 using contextualised embeddings for multi-task morphological analysis. For Old English, Martín Arista and Metola Rodríguez (2026) showed that character-level processing with custom embeddings outperforms multilingual models for POS tagging, and Martín Arista (2026) found the Transformer to be the best architecture for lemmatisation and established a direct relationship between POS-annotation completeness and lemmatisation accuracy, a finding we return to in Section 4.3.

Low-resource NLP has been widely discussed: Joshi et al. (2020) documented the global inequality in resource availability, Magueresse et al. (2020) reviewed mitigation strategies, Hedderich et al. (2021) surveyed data augmentation, and Anastasopoulos and Neubig (2019) showed that character-level noise injection benefits morphological inflection—particularly relevant for historical languages with pervasive spelling variation. For historical texts specifically, Piotrowski (2012) surveys NLP challenges (tokenisation, normalisation, orthographic variation), while Universal Dependencies (Nivre et al., 2020) has standardised morphosyntactic annotation cross-linguistically.

Annotation quality has been studied mainly for contemporary languages. Plank et al. (2014) found inter-annotator disagreement in POS annotation to be largely systematic and concentrated in debatable cases, so part of what is scored as model error may reflect legitimate alternative analyses; Reiss et al. (2020), by contrast, found substantial proportions of genuinely incorrect labels in benchmarks such as CoNLL-2003. For historical languages such analysis is sparse: Cichosz et al. (2021) built a morphological dictionary for the YCOE, one of the few efforts addressing Old English annotation consistency at scale. We contribute the first quantitative

Language	Ann. tokens	Spell. var.	Lem. acc.
Modern English	>100 M	Low	>95%
Latin	∼11 M	Moderate	85–90%
Middle Dutch	130–575 K	High	80–94%
Old English	∼84 K	Very high	∼68%
Old High German	<50 K	Very high	—

Table 1: Old English in the landscape of historical-language lemmatisation. Accuracy figures are taken from the cited studies and were obtained on different datasets, annotation schemes, and evaluation settings; they provide orientation rather than a controlled comparison. The Old English figure is the development-set accuracy of Martín Arista (2026).

analysis of annotation inconsistency in Old English and its impact on neural lemmatisation.

Table 1 situates Old English within the wider context of historical language NLP. Annotated tokens refers to the size of the training split available for lemmatisation experiments. The combination of generalised spelling variation, rich morphology, and a small annotated corpus places Old English among the most challenging settings for automatic lemmatisation. Studies of annotation consistency suggest that human agreement on Old English lemmatisation is itself in the range of roughly 90–95% (Cichosz et al., 2021), which places an empirical ceiling on achievable accuracy.

3 Data Curation Pipeline

Building training data for Old English lemmatisation requires integrating material from sources that were compiled for different purposes, at different times, and following different conventions. We describe the three sources and the harmonisation steps applied. The overall pipeline proceeds in four stages: (i) extraction of raw word–lemma–POS triples from each source; (ii) format validation and column-order correction; (iii) POS tag harmonisation to UPOS; and (iv) quality checks including duplicate detection, encoding normalisation, and consistency analysis.

3.1 Sources

The YCOE contains approximately 1.5 million words of Old English prose and poetry, annotated with phrase-structure trees and POS tags. We extract word–lemma–POS triples from these trees. The corpus is distributed as two subcorpora with different internal formats: Prose (∼85,500 entries) with columns ordered as LEMMA–WORD–POS, using uppercase UPOS-style tags (e.g., NOUN, VERB); and Poetry (∼16,200 entries) with columns ordered as WORD–LEMMA–POS, using title-case category names (e.g., Noun, Verb). Failing to detect the reversed column order in the prose subcorpus would silently swap words and lemmas, producing a training set in which the model learns the inverse transformation. This kind of silent format mismatch is a recurrent hazard in historical NLP, where datasets are often manually curated and lack machine-readable metadata.

Source	Original	UPOS
YCOE Prose	NOUN	NOUN
YCOE Poetry	Noun	NOUN
DOE	noun	NOUN
BT	noun	NOUN
YCOE Prose	CCONJ	CCONJ
DOE	conjunction	CCONJ
BT	foreign word	X

Table 2: POS tag mapping examples across sources.

The DOE (Healey, 2018) provides comprehensive lemmatised entries for the letters A–I, each listing inflectional forms alongside the headword and its lexical category. We extracted approximately 65,700 inflectional forms, expanding each headword entry into individual word–lemma pairs. For instance, *cyning* 'king' yields training pairs from forms *cyning*, *cyninges*, *cyninge*, *cyningas*, *cyninga*, and *cyningum*, all mapped to the lemma *cyning*. Lexical categories follow dictionary conventions (e.g., noun, verb, adjective) rather than the UPOS tagset.

The Bosworth-Toller dictionary (Crist and Tichý, 2014) is a historical Anglo-Saxon dictionary covering the full alphabet. We extracted approximately 16,000 inflectional forms with lemmas and POS information, providing coverage for lexical items in the letters L–Y not yet published by the DOE.

3.2 Harmonisation

Integrating these three sources requires addressing five interrelated problems. Regarding POS tag mapping, Table 2 illustrates the diversity of POS conventions across sources. All tags were mapped to the Universal Dependencies tagset (UPOS; Nivre et al., 2020), with empty or unresolvable values assigned the tag X (unknown).

We adopt UPOS rather than a finer-grained, Old English-specific tagset for three reasons. First, UPOS is the most detailed level at which the three sources—the YCOE's extended morphosyntactic tags and the dictionaries' traditional lexical-

category labels—can be aligned reliably; any finer inventory would have to be imputed for much of the dictionary data, introducing noise. Second, with only ∼84K instances a fine-grained tagset would fragment already sparse evidence per tag. Third, UPOS ensures interoperability with the Universal Dependencies ecosystem (Nivre et al., 2020) and with resources such as ParCorOEv3 (Martín Arista et al., 2023). The cost is that morphological features (case, number, tense, mood) become unavailable to the model; encoding them as UD FEATS in CoNLL-U is a natural extension we leave for future work.

A unified character vocabulary of 98 symbols was constructed, including Old English-specific characters (<þ>, <ð>, <æ>), macronised vowels (<ā>, <ē>, <ī>, <ō>, <ū>, <ȳ>), and standard Latin script; out-of-vocabulary characters map to an unknown symbol at inference time.

The *ge-* prefix, one of the most productive in Old English (past participles, perfective verbs, and some nouns and adjectives), is treated inconsistently across sources: some lemmas include it, some exclude it, and some use a parenthetical notation *(ge)lyfan* for optionality. In the corpus data, 12,275 instances (11.7% of the total) involve *ge-*, and 221 lemmas use the *(ge)* notation. We normalised these 221 parenthetical cases to canonical forms (selecting the prefixed or non-prefixed variant by predominant corpus usage), but residual inconsistencies remain: unmarked lemmas that are prefixed in one occurrence and unprefixed in another (e.g., *gelȳfde* as *gelȳfan* vs. *lȳfan*) cannot be conflated automatically without risking the merger of genuinely distinct *ge-*/non-*ge-* pairs, and near-even splits make the canonical choice partly arbitrary. These cases surface as the prefix-inconsistency category of Section 4.1.

Derivational morphology poses similar problems: whether *unriht* 'injustice' is a single entry or *un-* + *riht* varies across dictionaries, producing incompatible inventories. Dictionary headwords also follow editorial conventions (e.g., macrons for length: *cȳning* vs. corpus *cyning*). We normalised case but preserved macrons, as they encode phonological information relevant to morphological class.

Quality checks included duplicate detection: exact duplicates of complete word–lemma–POS triples were removed before splitting, so no identical triple spans partitions. Entries where the same word form receives different lemma and/or POS values were deliberately retained, as they represent either genuine ambiguity or the inconsistencies analysed in Section 4. Deduplication scripts and logs are in the project repository (Section 5).

Statistic	**Value**
Total annotated instances	105,063
Training set	84,050
Development set	10,506
Test set	10,507
Unique word types	90,752
Unique lemmas	37,627
Character vocabulary	98 symbols
Identity transf. (% of train)	23,748 (28.3%)
Unknown POS (X) (% of total)	4,605 (4.4%)
ge- prefix (% of total)	12,275 (11.7%)
Inconsistent types (% of types)	10,256 (11.3%)

Table 3: Dataset statistics after harmonisation (corpus-only condition). The reference base for each percentage is indicated per row.

After harmonisation, the corpus-only training set comprises 84,050 instances, 10,506 development instances, and 10,507 test instances (80/10/10 random split, seed 42), covering 90,752 unique word types and 37,627 unique lemmas. Table 3 provides the detailed statistics.

4 Annotation Quality Analysis

Annotation inconsistency is a recurrent problem in historical corpora, where editorial decisions about lemma assignment are often subjective and where different annotators may follow different conventions. In this section, we quantify the types and prevalence of inconsistencies in the Old English data.

4.1 Types of inconsistency

Inconsistency was operationalised at the level of the word type: after Unicode normalisation and case folding, with macrons preserved and no further spelling normalisation, a type was counted as inconsistent if two or more distinct lemma strings are assigned to its tokens anywhere in the harmonised corpus data. POS tags were not used to partition types at this stage, so forms whose different lemmas correspond to different POS (homographs) are included in the count; the manual categorisation below separates such genuine ambiguity from true inconsistency.

By this criterion, we identified 10,256 word types (11.3% of the vocabulary) assigned different lemmas across occurrences. A random sample of 200 inconsistent types was manually categorised by a single expert annotator following written guide-

lines with prototypical examples (in the project repository; see Section 5); as a single annotator was involved, the proportions are informed estimates rather than adjudicated figures (see Limitations). Four main categories emerge.

Genuine ambiguity (homographs). Some word forms correspond to multiple lemmas depending on context. For example, *þā* functions as a demonstrative pronoun (lemma: *sē*) or as an adverb (lemma: *þā*); *mæg* can be a verb 'can' (lemma: *magan*) or a noun 'kinsman' (lemma: *mæg*). These account for approximately 35% of inconsistent types and represent true linguistic ambiguity rather than annotation error.

Variant lemma forms. Different annotators assign different but related lemma forms: *hālgan* may be mapped to *hālig* or *hālga*, which are spelling variants of the same lexeme. This category accounts for approximately 25% of cases.

Prefix inconsistency (*ge-*). The same verb may be lemmatised with or without the *ge-* prefix in different occurrences: *gelȳfde* mapped to *gelȳfan* in one instance and to *lȳfan* in another. This accounts for approximately 20% of inconsistencies.

Annotation errors. A residual category of clear errors, e.g. a wrong POS tag or a phonologically unrelated lemma. This accounts for approximately 20% of cases.

With $n = 200$, the 95% confidence interval for a 35% proportion is about ± 6.6 points, enough for the coarse ranking but not fine-grained comparisons.

4.2 Identity transformations

A notable feature of the data is the high proportion of identity transformations, where the inflected form equals the lemma (e.g., *cyning* → *cyning*): 23,748 instances (28.3% of training data), including uninflected forms, nominative singular nouns, and verb infinitives. Though trivially easy, they teach the model the default 'copy' behaviour for invariable words and anchor the distinction between inflected and uninflected forms.

4.3 Impact of POS completeness

The corpus contains 4,605 instances (4.4% of the total) with unknown POS tags (X). Prior experiments on the same dataset (Martín Arista, 2026) demonstrated that the proportion of unknown POS tags in the training data has a direct and measurable impact on joint lemma–POS model performance: each percentage point of unknown POS correlates with approximately 0.2% lemma accuracy degradation. When dictionary sources lacking POS annotations are added (raising the unknown POS proportion to 52.9%), lemma accuracy degrades by 13–23 absolute percentage points (Martín Arista, 2026). This finding advises never to train historical-language lemmatisers on data lacking POS annotations.

Property	**YCOE**	**DOE**	**BT**
Spelling conv.	Corpus	Editorial	Historical
POS format	UPOS	Lex. cat.	Lex. cat.
Macrons	Rare	Systematic	Variable
ge-	Variable	Systematic	Variable
Coverage	A–Z	A–I	A–Z

Table 4: Key discrepancies across the three data sources.

4.4 Inter-source consistency

The three data sources differ not only in format but in the lemma inventories they presuppose. Table 4 summarises key discrepancies.

The discrepancies span five dimensions. The YCOE records lemmas close to manuscript forms, the DOE uses normalised editorial headwords, and Bosworth-Toller follows nineteenth-century practice, so one lexeme may enter under three spellings. Only the YCOE maps transparently to UPOS; both dictionaries use lexical-category labels, and BT often lacks usable categories. Macrons are systematic in the DOE, rare in the YCOE, and variable in BT, so a naive merge treats *cyning* and *cȳning* as distinct. *ge-* handling is systematic only in the DOE. Coverage also differs (the DOE is restricted to A–I), so harmonisation problems are unevenly distributed across the lexicon.

Dictionary augmentation therefore introduces not just more data but a distributional shift: dictionaries document paradigmatic variation (all inflected forms), whereas the corpus reflects syntagmatic variation (forms attested in running text). We hypothesise that this mismatch, together with POS incompleteness, underlies the finding that adding dictionary data does not improve—and may degrade—accuracy (Martín Arista, 2026, see Section 6.1 and Table 8): the added entries shift the input distribution while raising the unknown-POS proportion from 4.4% to 52.9%. Disentangling the two factors would require ablations we leave for future work.

Category	%	Typical example
Strong verb ablaut	25	*sang* → **singan*
Spelling variation	20	*forðon* → **forðon*
Prefix handling	15	*gelȳfde* → **gelȳfan*
Rare words	15	Hapax legomena
Annotation incon.	10	Gold standard conflicts
Homographs	10	Context-dependent
Other	5	Miscellaneous

Table 5: Distribution of lemmatisation errors by morphological category, based on a single-annotator categorisation of 500 sampled errors. * marks incorrect model predictions. Gold lemmas: *singan*, *forþam*, *lȳfan*, respectively.

Cl.	Lemma	Inf.	Past	Past pl.	Ptcp.
I	*rīdan*	ī	ā	i	i
II	*bēodan*	ēo	ēa	u	o
III	*drincan*	i	a	u	u
IV	*beran*	e	æ	ǣ	o
V	*sprecan*	e	æ	ǣ	e
VI	*faran*	a	ō	ō	a
VII	*hātan*	ā	ē	ē	ā

Table 6: Ablaut patterns of the seven Old English strong-verb classes. Each column shows the stressed vowel in the relevant principal part.

5 Morphological Error Analysis

We analysed the errors of the best character-level Transformer lemmatiser of Martín Arista (2026), which achieves 67.58% lemma accuracy on the corpus-only test set of Section 4 (the test-set counterpart of the ∼68% development figure), examining only the lemma prediction component. The lemmatiser is a character-level encoder–decoder Transformer that maps the inflected form's character sequence, plus its UPOS tag as a special input token, to the lemma's character sequence; it was trained on the 84,050-instance corpus-only set of Section 4. Full specifications, hyperparameters, ablations, and test-set predictions are in the project repository at `https://huggingface.co/Nerthus-Project`, allowing independent assessment.

5.1 Error distribution by category

Table 5 presents the error distribution across morphological categories, from a random sample of 500 errors categorised by a single expert annotator following written guidelines (see the repository); as in Section 4.1, the proportions are informed estimates. With $n = 500$, the 95% confidence interval for a 25% proportion is about ± 3.8 points, enough for the coarse ranking discussed here.

5.2 Strong-verb ablaut

Strong verbs are the single largest source of error. Old English has seven strong-verb classes, each with a distinct ablaut pattern (vowel alternation across tense forms); Class III, for instance, alternates i–a–u–u (*drincan* 'to drink': *drincþ*, *dranc*, *druncon*, *gedruncen*). The model must learn seven such patterns while also handling spelling variation within each. Table 6 shows the four principal parts per class: the vowel of a past-tense form maps back to a different infinitive vowel, and this transformation is class-dependent and not recoverable from a single form without knowing the class.

Irregular verbs compound the difficulty. Suppletive verbs (*bēon*/*wesan* 'to be', *gān* 'to go') have stems so different that character-level matching fails—no cue links *eom* to *bēon*—while preterite-present verbs (*cunnan* 'to know', *magan* 'to be able') add further irregularity, as their present forms follow strong-verb past patterns.

5.3 Spelling variation

Spelling variation accounts for 20% of errors and affects all POS categories. Three subtypes stand out:

- Thorn/eth interchange: <þ> and <ð> are used interchangeably across (and within) manuscripts. A model trained on *forþam* may fail on *forðam*.
- Vowel variants: *æ* alternates with *a* or *e*; *y* alternates with *i* (e.g., *cyning*/*kyning*).
- Consonant doubling: inconsistent gemination (*mann*/*man*; *wolcenn*/*wolcen*).

Character-level data augmentation—simulating these alternations during training—has been shown to improve accuracy by approximately 2 percentage points (Martín Arista, 2026), but cannot fully resolve variation patterns not represented in the augmentation rules.

5.4 Prefix handling

The *ge-* prefix is the most frequent source of prefix-related errors. Because *ge-* was highly productive in Old English and its inclusion in the lemma varies across lexicographic traditions, the model faces systematic uncertainty. In 15% of errors, the model correctly identifies the stem but incorrectly includes or excludes the prefix.

POS	Approx. acc.	Main difficulty
ADP	∼85%	Few errors
CCONJ	∼83%	Few errors
DET	∼80%	Paradigm size
ADV	∼75%	Spelling variation
NOUN	∼68%	Declension complexity
ADJ	∼65%	Weak/strong inflection
VERB	∼55%	Ablaut, *ge-* prefix
PROPN	∼45%	Idiosyncratic forms

Table 7: Approximate accuracy by POS category. Estimates based on error analysis of a stratified sample.

Other productive prefixes causing errors include *un-* (negation), *for-* (intensification/destruction), *be-* (transitivisation), and *ā-* (completion). Unlike *ge-*, these prefixes typically change the meaning and should be retained in the lemma, but the model sometimes strips them by overgeneralising the *ge-* removal pattern.

5.5 Rare words and hapax legomena

Old English has many low-frequency words; many test items appear only once or twice, giving minimal training signal. Proper nouns (personal and place names) are especially affected, with idiosyncratic spellings and inconsistent lemmatisation. The model falls back on character-level heuristics that succeed only when a form closely resembles a more frequent training example.

5.6 Error analysis by POS category

Table 7 shows accuracy broken down by POS category. Verbs have the lowest accuracy, driven by ablaut patterns and the *ge-* prefix. Proper nouns are the next most difficult category. Function words (adpositions, conjunctions, determiners) achieve the highest accuracy because they are frequent and have few inflected forms.

6 Discussion: Lessons for Low-Resource Languages

The challenges documented above are not unique to Old English. Many low-resource languages—historical and contemporary—share limited annotated data, rich morphology, spelling variation or the absence of a standard orthography, and heterogeneous sources with incompatible conventions. We now draw lessons from the case study.

6.1 Data quality and data quantity

The experiments of Martín Arista (2026) show that augmenting training data with dictionary entries—though it roughly doubles the dataset—does not improve and may substantially degrade performance (Table 8). The critical factor is not data volume but quality, consistency, and relevance to the target distribution: corpus data (attested forms in running text) provides a richer signal than decontextualised dictionary paradigms.

Condition	Train	Unk. POS	Test acc.
Corpus only	84,050	4.4%	67.58%
+ DOE + BT	∼165,750	52.9%	44.6–54.6%

Table 8: Corpus-only vs. dictionary-augmented training. Results from Martín Arista (2026); the augmented condition is 13–23 pp below corpus-only across configurations, with per-configuration figures documented in the project repository.

6.2 Annotation consistency

If 11.3% of word types are inconsistently lemmatised, no model can exceed ∼89% accuracy even with perfect learning, and the effective ceiling is lower because inconsistencies concentrate in high-frequency ambiguous forms. Low-resource projects should run inter-annotator agreement studies and document explicit annotation guidelines before building NLP tools.

6.3 POS annotations

The degradation from missing POS tags (4.4% to 52.9% unknown POS, costing 13–23 points; Martín Arista, 2026) shows POS information is not optional: joint lemma–POS models learn conditioned patterns (e.g., *-e* maps differently for nouns vs. adjectives), and incomplete POS disrupts this. Projects should prioritise POS annotation alongside lemmatisation.

6.4 Format standardisation

The sources' heterogeneity—differing column orders, POS conventions, character inventories, and lemma spellings—required preprocessing that consumed a disproportionate share of effort. Digitisation projects should adopt standard formats from the outset (e.g., CoNLL-U) and document conventions explicitly.

6.5 Morphology-specific augmentation

Character-level augmentation mimicking attested variation (thorn/eth interchange, vowel alternation, consonant gemination) yields meaningful gains

(∼2 points) at zero annotation cost. For other languages with known orthographic variation, philologically grounded augmentation rules are a cost-effective strategy.

6.6 Character-level processing

Our experiments indicate character-level processing is both necessary and sufficient here: pretrained word embeddings and subword (BPE) tokenisation both reduced accuracy (Martín Arista and Metola Rodríguez, 2026). A plausible reading is that Old English lemmatisation is primarily a formal task in which distributional-semantic information adds noise; whether this extends to other morphologically rich or orthographically variable languages is an empirical question.

6.7 Cross-linguistic applicability

The conclusions may generalise. Historical Germanic languages (Old High German, Old Saxon, Old Norse) share rich inflection, orthographic instability, and scarce resources; contemporary languages without codified writing face analogous variation; and agglutinative languages (Finnish, Turkish, Quechua) present comparable morphological complexity, so our finding that character-level models beat word- and subword-level approaches may transfer, though we have not tested this. The key question is whether lemmatisation in a given language is primarily morphological (character-level pattern recognition) or semantic (distributional disambiguation); where morphology dominates, we expect—subject to verification—character-level Transformers with morphology-specific augmentation to be a strong baseline.

7 Conclusion

This paper has analysed the data-side challenges of building an NLP lemmatisation pipeline for Old English, a paradigmatic low-resource historical language. Three conclusions stand out. First, annotation inconsistencies are pervasive: the 11.3% of word types with inconsistent lemma assignments set a hard accuracy ceiling and explain many apparent model errors, and the four categories we identify (genuine ambiguity, variant lemma forms, prefix inconsistency, annotation errors) each need different fixes. Second, data provenance matters as much as volume: combining the three sources without harmonisation introduces distributional shift and POS incompleteness that degrade performance, with the jump in unknown POS from 4.4% to 52.9% costing 13–23 points. Third, errors concentrate in well-defined categories—strong-verb ablaut, spelling variation, and prefix handling account for 60%—and are unevenly distributed (function words ∼85% vs. verbs ∼55%), so targeted strategies could yield real gains.

These implications extend beyond Old English: for any low-resource language, the quality, consistency, and completeness of annotated data set a ceiling no architecture can exceed. Future work should pursue formal inter-annotator agreement studies for Old English, measure the impact of correcting inconsistencies on downstream accuracy, develop verb-class-aware lemmatisation, and test whether these guidelines transfer to languages such as Old High German, Old Norse, and Old Irish.

Limitations

This study has several limitations. First, it is a single-language case study; the cross-linguistic claims of Section 6 remain hypotheses to be tested. Second, the categorisations of inconsistencies (Section 4.1) and errors (Section 5) rest on manual samples of 200 and 500 items, with 95% confidence intervals of roughly ±7 and ±4 points for mid-range proportions, and were carried out by a single expert annotator, so the proportions are informed estimates rather than adjudicated figures; a formal multi-annotator study is future work. Third, the error analysis presupposes the model of the companion paper (Martín Arista, 2026), whose full specifications are in the project repository at `https://huggingface.co/Nerthus-Project`. Fourth, the cross-language accuracies in Table 1 come from studies with different corpora and protocols and are orientation only. Fifth, harmonisation is lossy: mapping to UPOS discards morphological features, and *(ge)*-normalisation resolves only explicitly marked cases. Finally, the random 80/10/10 split does not control for lexeme overlap between training and test, so accuracy on genuinely unseen lexemes may be lower than reported.

Acknowledgements

This work was supported by grant PID2023149762NB-100 (Agencia Estatal de Investigación) and grant PRXX24/00108 (Ministerio de Ciencia, Innovación y Universidades), which are gratefully acknowledged.

References

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 984–996, Hong Kong, China. Association for Computational Linguistics.

Toms Bergmanis and Sharon Goldwater. 2018. Context sensitive neural lemmatization with Lematus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1391–1400, New Orleans, Louisiana. Association for Computational Linguistics.

Anna Cichosz, Piotr Pęzik, Marcin Grabski, Magdalena Adamczyk, Paulina Rybińska, and Alicja Ostrowska. 2021. The VARIOE online morphological dictionary for YCOE.

Sean Crist and Ondřej Tichý. 2014. Bosworth-toller's anglo-saxon dictionary online.

Laura García Fernández. 2018. *The Lemmatisation of Old English Preterite-Present, Contracted, Anomalous and Class VII Strong Verbs*. Ph.D. thesis, University of La Rioja.

Yosra Hamdoun Bghiyel. 2025. *The Lemmatisation of The York-Toronto-Helsinki Parsed Corpus of Old English Poetry with a Relational Database of Old English Dictionaries*. Ph.D. thesis, University of La Rioja.

Antonette diPaolo Healey. 2012. The dictionary of old english web corpus. With John Price Wilkin and Xin Xiang.

Antonette diPaolo Healey. 2018. The dictionary of old english: A to i.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293. Association for Computational Linguistics.

Mike Kestemont, Guy de Pauw, Renske van Nie, and Walter Daelemans. 2017. Lemmatization for variation-rich languages using deep learning. *Digital Scholarship in the Humanities*, 32(4):797–815.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795. Association for Computational Linguistics.

Krister Lindén and Tommi Pirinen. 2009. Weighted finite-state morphological analysis of finnish compounding with HFST-LEXC. In *Proceedings of the 17th Nordic Conference on Computational Linguistics (NODALIDA 2009)*, pages 89–95, Odense, Denmark. Northern European Association for Language Technology (NEALT).

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Enrique Manjavacas, Ákos Kádár, and Mike Kestemont. 2019. Improving lemmatization of non-standard languages with joint learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1493–1503. Association for Computational Linguistics.

Javier Martín Arista. 2026. Neural architectures for old english lemmatisation: A comparative study of corpus-based and dictionary-augmented approaches. *Procesamiento del Lenguaje Natural*. Forthcoming.

Javier Martín Arista, Sara Domínguez Barragán, Luisa Fidalgo Allo, Laura García Fernández, Yosra Hamdoun Bghiyel, Miguel Lacalle Palacios, Raúl Mateo Mendaza, Carmen Novo Urraca, Ana Elvira Ojanguren López, Esaúl Ruíz Narbona, Roberto Torre Alonso, and Raquel Vea Escarza. 2023. ParCorOEv3. an open access annotated parallel corpus old english–english.

Javier Martín Arista and Darío Metola Rodríguez. 2026. Outperforming multilingual models: Character-level processing and custom word embeddings for old english. *International Journal of Humanities and Arts Computing*, 20(1):18–35.

Darío Metola Rodríguez. 2015. *Lemmatisation of Old English Strong Verbs on a Relational Database*. Ph.D. thesis, University of La Rioja.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043. European Language Resources Association.

Marco Passarotti. 2004. Development and perspectives of the latin morphological analyser LEMLAT. *Linguistica Computazionale*, 20–21:397–414.

Susan Pintzuk and Leendert Plug. 2001. The york-helsinki parsed corpus of old english poetry.

Michael Piotrowski. 2012. *Natural Language Processing for Historical Texts*, volume 5 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511. Association for Computational Linguistics.

Frederick Reiss, Hong Xu, Bryan Cutler, Karthik Muthuraman, and Zachary Eichenberger. 2020. Identifying incorrect labels in the CoNLL-2003 corpus. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 215–226. Association for Computational Linguistics.

Ann Taylor, Anthony Warner, Susan Pintzuk, and Frank Beths. 2003. The york-toronto-helsinki parsed corpus of old english prose.

Marta Tío Sáenz. 2019. *The Lemmatisation of Old English Weak Verbs on a Relational Database*. Ph.D. thesis, University of La Rioja.

Creating the *RozMuz* Corpus: Applying Ethics and Technology in Language Research

Alicja Derych
Wrocław University of Science and Technology / Poland
alicja.derych@pwr.edu.pl

Bartłomiej Alberski
Wrocław University of Science and Technology / Poland
bartlomiej.alberski@pwr.edu.pl

Hubert Jankowski
University of Warsaw / Poland
hd.jankowski@uw.edu.pl

Paweł Dembowski
Wrocław University of Science and Technology / Poland
pawel.dembowski@pwr.edu.pl

Abstract

The paper aims to present work in progress: the corpus *Rozmowy przy muzyce – korpus języka mówionego RozMuz* '*Conversations over music – RozMuz speech corpus*.' It contains authentic spoken material from instrument lessons (group and individual), recording sessions, and rehearsals of musical groups. *RozMuz* serves as a subcorpus of the specialized musical corpus being constructed for the purpose of creating the Musical Subwordnet (an experimental resource dependent on plWordNet). The preparation methodology of the spoken subcorpus consists of three stages: 1) recordings followed by file preparation (including anonymization of audio files with pre-recorded labels), 2) their semi-automatic transcription with manual verification, and 3) generation of output documents: a situation description card and a metadata card compliant with Dublin Core.

1 Introduction

The aim of this paper is to present work in progress: the corpus *Rozmowy przy muzyce – korpus języka mówionego RozMuz* '*Conversations over music – RozMuz speech corpus*' containing authentic spoken data from rehearsals of musical groups, recording sessions, and group and individual instrument lessons. Although several large spoken corpora, such as Spokes (Pęzik, 2015), have been compiled to date, currently there is no Polish-language corpus of spoken texts representing musical professional language that we know of. Corpora, particularly those annotated with meaning, are used, for example, in Word Sense Disambiguation (Scarlini et al., 2020; Vial et al., 2018; Raileanu et al., 2002). It is important, however, to keep in mind the possibility of addressing other research questions, so that the corpus being designed can be reused in the future. The aim is to ensure that it can also serve to answer questions that we are not yet able to formulate today (de Marneffe and Potts, 2017; Zaśko-Zielińska, 2025; Gajek, 2007; Kamasa, 2014), and make it possible to reuse existing resources (Karkowska and Bielińska, 2019). From this perspective, this will be a unique resource, given that it consists of spoken and specialized data at the same time (Zaśko-Zielińska, 2025)[1].

2 Musical corpus and specialized jargony vocabulary

RozMuz serves as a spoken subcorpus of the specialized musical corpus being constructed for the purpose of creating the Musical Subwordnet. This is an experimental resource dependent on plWordNet (a lexical-semantic database, an electronic relational dictionary for the Polish language, containing mainly, but not solely, vocabulary of the general variety of Polish) (Rudnicka et al., 2025). The recordings contained in *RozMuz* as well as their transcriptions will be used for the extraction of marked lexical units (LUs), that is, the ones that do not belong to the general variety of Polish. We are particularly interested in non-terminological units of language, as terminology is present in written sources. For example, LU *prymka 'a lead sheet'* is a graphically notated form of a vocal piece that includes the melody, lyrics, and chords. It is a professional-language equivalent of the phrase *simplified notation of a vocal piece*. Other good examples of such vocabulary are: LU *drewienka 'claves'*

[1] From the perspective of computational linguistics, Polish, in its standard variety, is not a low-resource language. However, specialized, jargony professiolect of a relatively niche topic could be treated as one.

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 18–23
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_003

(only in plural form) which is a percussion instrument in which sound is produced by striking two pieces of wood against each other, and *rosół 'Polish soup made of chicken broth'* which is a humorous name for a commission to perform live music during a wedding ceremony.

As we mentioned above, *RozMuz* is intended to be a part of a larger corpus being developed for the purpose of the extraction of specialized and jargony LUs of the music and musicology domains. Thanks to the construction of the spoken resource, it has been possible to identify LUs that are not present in the text data. An example of such an LU might be *paczka 'speaker'* which is a humorous name for the speaker cabinet. The corpus contains a dialogue in which this word appears: *"Piece mi nie grają." "Przynajmniej paczki masz dobre, nie?" "Co z tego, piece nie grają, to nic nie pomoże." 'My amps aren't working.' 'At least your speaker cabinets are good, aren't they?' 'So what? The amps aren't working, so that won't help.'* Another example is the LU *kwit 'music sheet'* which is a graphical representation of a piece of music, a functional notation, chord symbols or musical notation. In the corporus it is used in a sentence *"Czekaj sobie wezmę tylko kwit, ostatnio robiłem sobie skany." 'Just wait a moment, I'll just grab the music sheet, I've been scanning them myself recently.'*

Ultimately, the musical corpus will consist of several types of sources. At this stage of construction, it has been divided into three subcorpora based on the criteria of medium (oral/written) and author-recipient group (specialists/practitioners), namely: spoken subcorpus (*RozMuz*), theoretical subcorpus, and practical subcorpus (see Table 4). Each text introduced is described with a set of metadata including basic information about its title, author, and for the purposes of further use, copyright. The resource is intended to be compiled with the use of an offline application enabling to dynamically search for specific word forms and lemmas (Piasecki et al., 2025). Decision on the choice of the tool was based on the primary, operational aim of the first stage of corpus development, which is to facilitate the construction of the Musical Subwordnet.

3 Preparation of *RozMuz*: processing and transcription of recordings

The recordings are obtained ethically. The speakers (or their legal guardians) consciously sign a consent to be recorded, and the consent specifies the scope of the use of the obtained data. Those selected for the recording had to be involved in music either as professionals or as students. Including participants with varying levels of proficiency made it possible to collect data from speakers with different levels of familiarity with the jargon. As this was the initial stage of compiling the resource, the data were obtained from available participants (recruited using the snowball sampling method) of various ages (from 7 to 37, with the majority being school-age children) and with varying levels of musical proficiency (from complete beginners to advanced learners and professionals). Above all, they had to be willing to take part in the collection of data for the corpus. Some of the participants were minors, so they provided only verbal consent, while the relevant documentation was duly completed by their legal guardians. The conditions of being recorded may possibly trigger an observation effect that could influence the data obtained (content). Limiting the recording to audio only (without video footage) can significantly reduce this effect. Furthermore, music schools frequently use recording equipment to effectively document learning progress, so the situation was neither unusual nor stressful for young participants involved in the project.

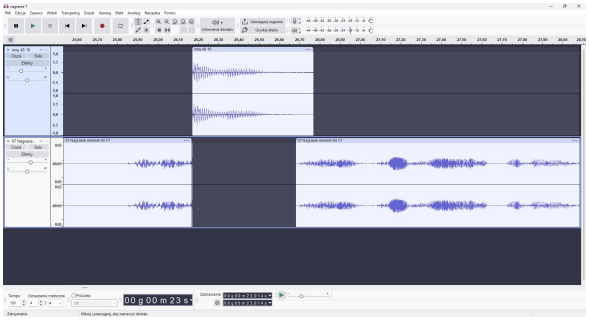


Figure 1: Anonymization process in Audacity (using anonymization tags)

The anonymization process involved removing sensitive information and replacing that sections with pre-recorded audio anonymization labels (using the Audacity application, ver. 3.7.7[2], see Figure 1) at the audio file processing level. For this purpose, we used a prepared set of labels that replace not only personal information such as names or addresses but also any other details about private life that may be considered sensitive. We employed the

[2]https://www.audacityteam.org/

following labels: *address*, *email*, *address*, *website*, *date*, *first name*, *last name*, *phone number*, *social security number*, *username*, *location*, and *personal information*.

Moreover, we erased larger, non-verbal passages both to reduce the size of the resource for further processing and sharing, and to facilitate the perception of audio material for future corpus users. Longer segments of silence, however, are removed and marked with corresponding labels: *music*, *silence*. As Table 1 shows, this procedure allowed us to significantly decrease the size of the corpus which was mostly the result of cutting out musical fragments. Consequently, the natural course of conversations was preserved.

	before anonymization and cuts	**after anonymization and cuts**
file size	5.18 GB	3.09 GB
number of files	19	19
duration	08:03:04	04:48:07

Table 1: Audio size of *RozMuz*

To ensure better sound quality and volume balance, the recordings are also prepared from the sound engineering perspective, compressed, and enhanced in Digital Audio Workstation (Cakewalk Sonar 2025.09). All files were processed in 48kHz, 16-bit WAV format with maximum loudness at approximately -1 dB, approximately -23LUFS. The recordings comply with the EBU 128R loudness standard[3].

The first stage of data preparation revealed some issues with the accuracy of anonymization as some of the personal data were missed in the process and remained audible in the recordings. Therefore, we have decided to introduce an additional stage of cross-check to ensure complete anonymization which led to establishing the procedure of triple check. Personal data is thus reviewed and/or removed in the phase of:

1. anonymization and music/silence cut made by one editor;
2. cross-check – another editor's listening (if some remaining personal data is identified, corrections are then introduced);
3. transcription – during which eventual omissions in the anonymization process are further noted (if some remaining personal data is identified, corrections are then introduced).

The transcription process is carried out in a semi-automatic manner. We use the Transcribe tool in Microsoft Word[4], the output of which, in the form of a transcription, is then verified and manually corrected by an annotator.

tool	**pros**	**cons**
Korpusomat.eu	relatively good processing quality, especially with good audio input; fast; wide range of supported formats	requires manual revision; audio size limit 200 MB; common errors during processing
CLARIN ASR	output files in JSON convenient for further processing; fast; recordings available in the cloud	very low transcription quality on this type of audio sample; requires manual revision; does not enable parallel work with audio and transcription; does not support WAV format, conversion to MP3 required
Microsoft Word	highly automated; supports WAV format; divides and labels utterances by speakers and time (timestamps)	omits parts of utterances; not fully accurate transcription requires manual revision; does not always adequately recognize speakers; uses capital letters and punctuation by default; audio size limit 300 MB
Whisper	highly automated; supports WAV format; divides and labels utterances by speakers and time	omits parts of utterances; inaccurate transcription requires thorough manual revision; requires programming skills

Table 2: Automatic transcription tools tested

Apart from this tool, at the initial stage of methodology preparation we tested three other applications dedicated to manual transcription: ELAN (Sloetjes and Wittenburg, 2008), EXMARaLDA (Schmidt, 2012; Schmidt and Wörner, 2009), and Annotation Pro (Klessa et al., 2013), and three automatic tools: Whisper (Radford et al., 2022), an automatic module built into the tool used to compile corpora: Korpusomat.eu[5] (Saputa et al., 2023), and CLARIN ASR application[6] (Piasecki

[3]https://tech.ebu.ch/docs/r/r128.pdf

[4]https://www.office.com/

[5]https://korpusomat.eu/

[6]https://services.clarin-pl.eu/app/services/asr

et al., 2025).

tool	pros	cons
Annotation Pro	accuracy and detail of transcription; high quality of transcription; supports XML formats; supports WAV format; enables precise parallel work with audio and transcription; convenient work	requires extensive manual work; subjective output
EXMARaLDA	accuracy and detail of transcription; supports WAV format; enables precise parallel work with audio and transcription; EXAKT corpus tool; wide range of possible export file types; enables uploading automatic Whisper transcription; possibility of transcribing overlapping utterances; enables annotation in a separate line	time-consuming manual work; uncut music sections make manual transcription challenging; EXAKT corpus tool without export options; annotation line not exported into TXT output file; subjective output
ELAN	ease of use; supports WAV format; possibility of transcribing overlapping utterances; accuracy and detail of transcription; enables precise parallel work with audio and transcription	uncut music sections make manual transcription challenging; does not enable annotation; time-consuming manual work; subjective output

Table 3: Manual transcription tools tested

The test was conducted on a sample of a WAV recording (a part of a rehearsal, duration: 00:02:32) from a practical perspective, in a qualitative manner[7]. Its results are considered mainly regarding: supported input and output file types; transcription quality (sound-text adequacy), objectivity and need of manual revisions; division into speaker turns; general ease of use (including programming skills requirement); and time consumption.

In summary, all of the manual transcription tools seem to produce the most accurate results, but the process is also highly time consuming (see Table 3). Automatic transcription using Whisper requires a number of adjustments and did not produce satisfactory transcription ready to be included in the corpus. Transcriptions obtained through Korpusomat.eu seemed promising, although the size of the input audio file in the application is limited. CLARIN ASR tool provides transcription that requires editing and conversion of the input recordings into MP3 format (conversion from WAV to MP3 results in a loss of quality, see Table 2). All key limitations and advantages of the tools under review are presented in detail in Tables 2 and 3.

Consequently, after the evaluation of pros and cons of each solution, taking into account also the primary purpose of the subcorpus construction (mainly a source of specialized lexis), the Microsoft tool was selected due to mostly satisfactory accuracy of the source transcription and convenience of use. It requires a number of manual corrections and adjustments but, overall, the whole process appears to be relatively less time consuming and the results are adequate enough (detailed description of pros and cons is presented in Table 3). In addition to the text of the transcription, we also retain a DOC document containing timestamps and automatic speaker division that facilitates searching the audio material.

Manual verification consists of checking for completeness, adequacy, division into speaker turns, editing speaker IDs, and adjusting orthographic layer of the text. The following transcription convention results from the primary goal of the corpus construction which is intended to serve as a source of specialized vocabulary (e.g., speech turns are of lesser importance):

- each speaker's utterance begins with their ID, a colon and a tabulator;
- segmentation into sentences (punctuation, capitalization) is omitted;
- labels <*MUSIC*> and <*SILENCE*> are written in uppercase, in angle brackets, in separate lines, without speaker's IDs (as they do not concern specific utterances);
- anonymization labels (e.g., <*LOCATION*>, <*NAME*>) are written in uppercase, in angle

[7]The choice of this more practical method of tool comparison followed from a goal of the whole project: to produce a resource that would be of use to editors of the subwordnet. A quantitative comparison is possible, though it would require a certain degree of standardization. As the test was performed by a team, we would have to consider individual differences, measure parameters such as, e.g., exact time of work, percentage of tokens/segments/sentences transcribed accurately and mistakes made; and account for technical differences between the tools (manual and automatic).

brackets, within utterances (as parts of these utterances);

- <*?*> label is used to mark inaudible/unclear fragments;
- filled pauses (e.g., *eee*, *yyy*) and laughter are not transcribed, and pauses in individual utterances are not marked;
- conventionalized non-articulated speech signals (e.g., *hmm*) and onomatopoeia are transcribed (e.g., *la la la*);
- numbers are written out in full;
- overlapping utterances and interruptions are transcribed in chronological order (if multiple begin simultaneously, the order is arbitrary);
- in general, non-standard word forms (e.g., pronunciation variants) are standardized.

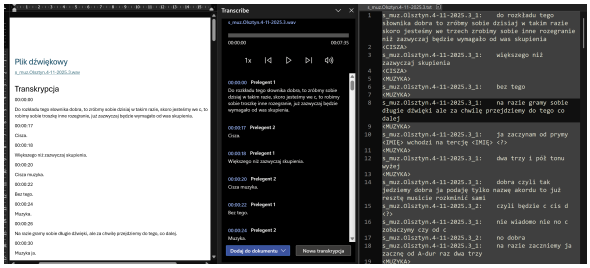


Figure 2: Semi-automatic transcription process

Final transcriptions are stored as TXT files (see Figure 2). Another accompanying document for the recording is a situation description card[8] containing basic data about the interlocutors and the circumstances of the recording, e.g., *age*, *relationships between speakers*, their *education*, *location of the recording*, or *type of situation*. For each recording, we prepare a corresponding metadata card compliant with Dublin Core[9], a widely used metadata scheme. The card indicates the identifiers of the recording and transcription as well as basic information about the file. The resource is prepared in accordance with the FAIR standard[10] and will be made available, among others, in the DSpace repository[11], CoRD[12], and Disciplinary Open Research Data Repositories[13].

4 Limitations and further work

We are aware that the scope and methodology of our project have certain limitations. Firstly, the manual construction of a spoken resource of this kind is time-consuming (although necessary to ensure complete anonymization and satisfactory quality of audio material) and burdened with a certain degree of subjectivity. To address these issues, we employed the initial stage of automated transcription and cross-check of the audio material. Secondly, the question of future usability of the resource emerges (also following from the tool selected for the transcription process). To ensure interoperability, file types employed aim to comply with the FAIR standard[14]. TXT is a broadly compatible format, and the originally generated DOC files enable to supplement transcriptions with timestamps for possible future projects, and align the transcripts with the recordings. Finally, one may question the representativeness of the subcorpus as a source of professional vocabulary. In order to overcome this problem, in two other subcorpora we aim to include also texts of other types (see Table 4). What is more, it is crucial to note that representativeness is extremely difficult to achieve for this type of geographically, idiolectally, and situationally varied material.

theoretical publications	**practical texts**
handbooks	thematic forums
scientific releases	tips for musicians
music critique	equipment reviews

Table 4: Types of sources

As we mentioned above, the musical corpus, along with its subparts, including *RozMuz* is still under construction. We intend to collect, edit, and transcribe (and, later on, publish) more audio materials that will be used as sources of professional vocabulary in the Musical Subwordnet and as usage examples in the resource (which will include at least 1000 LUs representing the musical domain). We presented primary advantages of employing specialized, spoken corpus in subwordnet construction. Its final evaluation can be conducted upon completion of the lexical resource.

[8] Based on (Zaśko-Zielińska et al., 2020).

[9] Based on: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/ and (Zaśko-Zielińska et al., 2020).

[10] https://howtofair.dk/

[11] https://clarin-pl.eu/dspace/

[12] https://varieng.helsinki.fi/CoRD/

[13] https://drodb.icm.edu.pl/materialy-2/

[14] https://howtofair.dk/

Acknowledgments

The work is being conducted as part of the project the European Funds for a Modern Economy 2021–2027 (FENG), Priority 2: "Innovation-Friendly Environment", call FENG.02.04-IP.04-001/24 Research Infrastructure for a Modern Economy – pathway for projects from the Polish Roadmap for Research Infrastructures (PMIB) and as part of the investment: "CLARIN ERIC – European Research Infrastructure Consortium: Common Language Resources and Technology Infrastructure" (period: 2024-2026) funded by the Polish Minister of Science under the programme: "Support for the participation of Polish scientific teams in international research infrastructure projects", agreement number 2024/WK/01.

References

Elżbieta Gajek. 2007. Korpusy językowe w kształceniu nauczycieli języków obcych. *Przegląd Glottodydaktyczny*, 24:19–42.

Victoria Kamasa. 2014. Techniki językoznawstwa korpusowego wykorzystywane w krytycznej analizie dyskursu. przegląd. *Przegląd Socjologii Jakościowej*, 10(2):100–117.

Marta Karkowska and Magdalena Bielińska. 2019. Po co nam „stare wywiady"? O ponownym wykorzystaniu danych z badań jakościowych (rewizytach, reanalizach i nie tylko). *Przegląd Socjologii Jakościowej*, 15(2):12–39.

Katarzyna Klessa, Maciej Karpiński, and Agnieszka Wagner. 2013. Annotation Pro – a new software tool for annotation of linguistic and paralinguistic features. In *Proceedings of the Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop*.

Marie-Catherine de Marneffe and Christopher Potts. 2017. Developing linguistic theories using annotated corpora. In *Handbook of linguistic annotation*, pages 411–438. Springer.

Maciej Piasecki, Agnieszka Dziob, Arkadiusz Janz, Jan Kocoń, Tomasz Naskręt, Marcin Oleksy, Ewa Rudnicka, Tomasz Walkowiak, Jan Wieczorek, and Krzysztof Hwaszcz. 2025. CLARIN-PL: a user centred language technology infrastructure. *Language Resources and Evaluation*, 59(4):4493–4528.

Piotr Pęzik. 2015. Spokes – a search and exploration service for conversational corpus data. In *Selected papers from the CLARIN 2014 Conference*, pages 99–109.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Diana Raileanu, Paul Buitelaar, Spela Vintar, and Jörg Bay. 2002. Evaluation corpora for sense disambiguation in the medical domain. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. European Language Resources Association (ELRA).

Ewa Rudnicka, Bartłomiej Alberski, and Maciej Piasecki. 2025. plWordNet 5.0 – challenges of a lifelong wordnet development process. In *Proceedings of the 13th Global Wordnet Conference*, pages 162–170. Global Wordnet Association.

Karol Saputa, Aleksandra Tomaszewska, Natalia Zawadzka-Paluektau, Witold Kieraś, and Łukasz Kobyliński. 2023. Korpusomat. eu: A multilingual platform for building and analysing linguistic corpora. In *International Conference on Computational Science*, pages 230–237. Springer.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. Sense-annotated corpora for word sense disambiguation in multiple languages and domains. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5905–5911. European Language Resources Association.

Thomas Schmidt. 2012. EXMARaLDA and the FOLK tools-two toolsets for transcribing and annotating spoken language. In *LREC 2012 Proceedings*, pages 236–240. European Language Resources Association (ELRA).

Thomas Schmidt and Kai Wörner. 2009. EXMARaLDA – creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics*, 19:565–582.

Han Sloetjes and Peter Wittenburg. 2008. Annotation by category: ELAN and ISO DCR. In *6th international Conference on Language Resources and Evaluation (LREC 2008)*. European Language Resources Association (ELRA).

Loïc Vial, Benjamin Lecouteux, and Didier Schwab. 2018. UFSAC: Unification of sense annotated corpora and tools. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Monika Zaśko-Zielińska. 2025. *Mały korpus – wielkie wyzwania: Small corpora – big challenges*. Universitas.

Monika Zaśko-Zielińska, Anna Majewska-Tworek, Marta Śleziak, and Artur Tworek. 2020. *Od rozmowy do korpusu, czyli jak zbierać i archiwizować dane mówione*. Quaestio.

Towards a Universal Dependencies Treebank for Amazigh: A Tarifit Pilot Study

Azzeddine Afrouni[1], **Fadoua Ataa Allah**[2], **Jamal Abarnous**[1]

[1]Faculty of Letters and Human Sciences, Mohammed First University, Oujda, Morocco
[2]Center of Computer Science Studies, Information Systems and Communication, Royal Institute of Amazigh Culture, Rabat, Morocco

{azzeddine.afrouni24,abarnous.jamal.pr}@ump.ac.ma
ataaallah@ircam.ma

Abstract

This paper presents a pilot study on the construction of a Universal Dependencies (UD) treebank for Tarifit, a low-resource variety of Moroccan Amazigh. To bridge the syntax-resource gap, we develop a manually annotated corpus comprising 123 sentences collected from textbooks and digital sources, including IRCAM and MAP portals. The selected texts were chosen for their orthographic consistency to ensure annotation quality and reliability. Morphosyntactic annotation has been completed for the full corpus, while dependency annotation has been finalized for 50 sentences. Annotated under UD guidelines v2, the study focuses on language-specific phenomena including pronominal cliticization and word order flexibility. We detail the annotation challenges in adapting the UD framework to Tarifit and propose cross-linguistically consistent solutions within the UD ecosystem.

1 Introduction

Treebanks are linguistically annotated corpora that represent the syntactic and semantic relations of a language. They serve as essential resources for training and evaluating parsers and language models, thereby enabling advanced NLP applications such as machine translation and information extraction (Indurkhya and Damerau, 2010). Amazigh is an Afro-Asiatic language family spoken across North Africa, officially recognized in Morocco, and integrated into its educational system. It encompasses three major regional varieties: Tarifit in the north, this study's focus, Tamazight in central regions, and Tachelhit in the south (Ameur et al., 2004), with ongoing efforts aiming at a unified written form. Research on Moroccan Amazigh NLP has produced vital computational resources (Ataa Allah and Bouhjar, 2019). Early work introduced POS tagsets and corpora via the IRCAM portal[1] (Ataa Allah et al., 2014), while morphological analysis evolved from ANMorph (Raiss and Cavalli-Sforza, 2012) to finite-state approaches (Nejme et al., 2016), followed by machine learning and neural POS tagging (Amri and Zenkouar, 2018; Maarouf and El Ayachi, 2021). Structurally, Taghbalout et al. (2018) addressed syntactic and semantic aspects via a UNL-based machine translation approach. To the best of our knowledge, no complete dependency treebank currently exists for Amazigh. However, recent UD trends show that small, manually annotated corpora for morphologically rich, low-resource languages, such as Breton (Tyers and Ravishankar, 2018), Sinhala (Liyanage et al., 2023), and Uzbek (Akhundjanova and Talamo, 2025) constitute invaluable foundational resources. Following these models, this work constructs a pilot UD treebank for Tarifit. In this context, the present paper introduces a pilot study aimed at constructing a UD-compliant dependency treebank for Amazigh, focusing on the Tarifit variety spoken in northern Morocco. The study presents the corpus selection and annotation methodology, together with an analysis of major annotation challenges. The main contributions of this work are: i the creation of a manually annotated UD-style dataset for Amazigh, and ii an examination of the challenges involved in adapting the UD framework to Tarifit, along with proposed solutions. The remainder of this paper is organized as follows. Section 2 provides a linguistic overview of Amazigh and the Tarifit variety. Section 3 describes the corpus construction and annotation methodology. Section 4 presents the main annotation challenges. Section 5 discusses possible solutions and perspectives related to these annotation challenges. Finally, Section 6 concludes the paper and outlines directions for future work.

[1]https://tal.ircam.ma/talam/corpus.php

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 24–29
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_004

2 Linguistic overview of Amazigh

From a linguistic perspective, Amazigh is characterized by a rich morphology (Ennaji and Sadiqi, 2004). It exhibits pronominal cliticization, particularly with prepositions and kinship terms, producing single orthographic units that correspond to multiple syntactic elements. This property is shared with other Afro-Asiatic, Chamito-Semitic languages in terms of nominal and pronominal morphology (Chaker, 1990). In addition, certain functional particles display significant categorial ambiguity; for instance, the particle ⵏ d can function as a coordinating conjunction, a copular particle, or a directional marker depending on the syntactic context (Boukhris et al., 2008). Amazigh also relies on aspectual particles that co-occur with verbs to express temporal and aspectual values, such as the particle ⴰⵏ ad, which is used with the aorist verb form to express futurity (Boukhris et al., 2008). By contrast, specifically in Tarifit, the language employs the auxiliary verb ⵜⵓⵖⴰ tuɣa in combination with the perfective aspect (Cadi, 2005). As noted by (Greenberg, 1963), Afro-Asiatic languages often allow variation in constituent order, a property also observed in Amazigh. It is generally characterized as a head-initial language associated with a dominant Verb-Subject-Object (VSO) order, where heads precede their complements (El Moujahid, 2022; Ouhalla, 1988). However, this property varies across Amazigh varieties. For Tarifit, previous studies report that VSO is the most frequent order, followed by Subject-Verb-Object (SVO), while Verb-Object-Subject (VOS) is less common (Cadi, 2005). This word order flexibility must therefore be considered in syntactic annotation and dependency analysis.

3 Research methodology

The methodological framework of the present study covers the data sources and selection criteria for the pilot corpus, as well as the annotation process followed to produce the Tarifit dependency treebank.

3.1 Data Sources

The pilot corpus comprises sentences extracted from Amazigh school textbooks for primary education levels one through four, written in accordance with the standardized orthographic and linguistic norms established by IRCAM (Ameur et al., 2006). These textbooks cover the three regional varieties of Amazigh used in Morocco. Given that the present study focuses on Tarifit, the selection was restricted to texts representative of this variety (Boudris et al., 2022; Elbaghdadi et al., 2022). Additionally, texts from digital resources, namely the IRCAM and MAP portals, were selected based on a targeted analysis to ensure their linguistic proximity and structural alignment with Tarifit variety. Table 1 summarizes the main characteristics of the pilot corpus, including the distribution of tokens and sentences across sources, as well as the average sentence length in characters.

Source	Tokens	Sentences	Avg. Length
Textbooks	1084	63	17.20
IRCAM portal	1059	34	31.14
MAP portal	1033	26	39.73

Table 1: Corpus Source and Average Length Statistics

3.2 Selection Criteria

Three main criteria guided the selection of texts for the pilot corpus. First, linguistic suitability was ensured by selecting texts from pedagogical materials and edited resources provided by the IRCAM[2] and MAP[3] portals. These sources offer a high degree of linguistic clarity, providing a controlled environment ideal for validating the adaptation of the UD framework to Tarifit. Second, orthographic and normative consistency was ensured by restricting selection to texts adhering strictly to the standardized orthography and linguistic rules established by IRCAM, thereby mitigating the effort typically spent on correcting non-standardized texts (Agnaou et al., 2017). Third, the selection prioritized structural and functional completeness, focusing on complete communicative acts rather than fragmented or isolated sentences, so as to enable the analysis of syntactic patterns within cohesive textual environments.

3.3 Annotation Process

The annotation of the pilot corpus was carried out entirely manually, with no automatic pre-annotation or parser-assisted tagging at any stage. Before starting the annotation process, text preprocessing was performed entirely by hand, involving sentence splitting followed by word tokenization. All morphosyntactic labels, dependency heads, and dependency relations were then

[2] https://www.ircam.ma/amz
[3] https://www.map.ma/amz/

assigned by the first author through direct linguistic analysis of each sentence. Specifically, the annotation is conducted in two phases: the first phase completed the morphosyntactic tagging for 123 sentences comprising 3,176 tokens, while the ongoing second phase has finalized the dependency relations for 50 sentences to date. It was conducted using the Arborator-Grew platform[4], a web-based tool designed specifically for building and visualizing dependency treebanks (Guibon et al., 2020). Throughout the annotation process, the UD guidelines v2[5], were systematically consulted to ensure cross-linguistic consistency and compatibility with the UD ecosystem. While the annotation was carried out solely by an author, the annotated corpus was subsequently reviewed by a specialist in Amazigh language. The review focused on verifying the consistency of the morphosyntactic and dependency annotations, and annotation decisions were discussed until a consensus was reached. As this is an ongoing pilot project, a formal inter-annotator agreement evaluation has not yet been conducted. We consider this a limitation of the current study and plan to incorporate a more systematic validation procedure in future stages of the project.

4 Annotation Challenges

The annotation of the Tarifit pilot corpus raised several morphological and syntactic challenges. Although many of these challenges are not unique to Tarifit, they are examined here under four main headings: word-based tokenization, lemmatization, annotation decisions, and word order flexibility.

4.1 Word-based tokenization

A key challenge in tokenization concerns the behavior of pronominal clitics, which in accordance with IRCAM writing conventions are orthographically attached to their host. For instance, the preposition ⵏⵖ dg "in" combines with the pronoun ⵙ s "him/her/it", triggering a vowel insertion that yields the single orthographic unit dags "in him/it". In a context like ⵖⵖⵉⵏ ⵏⴰⵖⵙ ggin dags "they put in it", this creates a syntactic challenge in dependency annotation: following UD guidelines, the pronoun ⵙ s must be directly linked to the verbal head ⵖⵖⵉⵏ ggin via the oblique relation obl, while the preposition ⵏⵖ dg attaches to the pronoun via a prepositional relation case. This segmentation prevents the orthographic unit as a whole from being directly linked to the verbal head, an analysis that would not be sanctioned by the UD guidelines. The corresponding CoNLL-U representation of this analysis is shown in Figure 1. The same treatment applies to kinship terms: ⴱⴰⴱⴰⵙ babas "his father" consists of the kinship noun ⴱⴰⴱⴰ baba "father" and the pronominal clitic ⵙ s "his". Segmenting this form into two components allows ⴱⴰⴱⴰ baba to be linked to its syntactic head via the appropriate dependency relation, while the clitic ⵙ s attaches to it via the nominal modifier relation nmod:poss, reflecting its possessive function. This analysis is more consistent with UD guidelines than treating the orthographic unit as a whole as a single syntactic node.

4.2 Lemmatization

The tokenization of clitic-host constructions discussed above also raises lemmatization challenges. These challenges do not concern the host forms themselves, since ⵏⵖ dg "in" and ⴱⴰⴱⴰ baba "father" correspond to their citation forms. The difficulty instead concerns the clitic s after tokenization. When separated from the prepositional construction ⵏⴰⵖⵙ dags, ⵙ s functions as a pronominal complement attached to the preposition ⵏⵖ dg "in". By contrast, when tokenized from the kinship construction ⴱⴰⴱⴰⵙ babas, it expresses possessive reference associated with the noun ⴱⴰⴱⴰ baba "father". Although both forms surface identically after tokenization, they correspond to different syntactic and semantic functions, creating a challenge for lemma assignment within the annotation framework.

4.3 Annotation Decisions

The particle d presents a notable annotation challenge within the UD framework, particularly regarding the mapping between the language-specific tagset XPOS established by IRCAM and the Universal Part Of Speech tags UPOS. Although d displays multiple syntactic functions in Tarifit, not all of its uses raise equivalent annotation challenges within the UD framework. The primary difficulty concerns its predicative use, where the annotator must determine how the language-specific IRCAM category should be mapped onto universal syntactic categories. For instance, when d functions as a copular particle, as in ⵜⵍⴰ ⵏ ⴰⵙⵎⵍⴰⵏ uma d aslmad

[4] https://arborator.grew.fr/#

[5] https://universaldependencies.org/guidelines.html

```
1       ⵣⵣⵉⵍ   ⵣⵣ      VERB    V.ACC_POS.3MP   Aspect=Perf|Polarity=Pos|Voice=Act      0       root
        _       _
2-3     ⴰⵙⵣⵙ   _       _       _       _       _       _       _       _
2       ⴰⵙ     ⴰⵣ      ADP     AP.LIEU _       3       case    _       _
3       ⵙ       ⵏⵜⵜⵓ    PRON    PR.AP.M.S.3     _       1       obl     _       _
```

Figure 1: CoNLL-U format illustrating the segmentation of a preposition–pronominal clitic construction.

"my brother is a teacher", the IRCAM tagset provides the fine-grained label P.PRED predicative particle. Within the UD framework, this function is mapped to the UPOS tag AUX and linked to the predicate through the cop dependency relation (Nivre et al., 2020). This challenge does not reflect a contradiction between the two systems; rather, it requires careful annotation decisions to ensure that the structural distinctions encoded in the IRCAM tagset are appropriately represented within UD. By contrast, the coordinative and directional uses of ⴰ d do not pose any particular annotation difficulty, as they map more directly onto the corresponding UPOS categories. A similar case involves aspectual particles and auxiliaries. In Tarifit, ⵜⵓⵖⵓ tuɣa works as an auxiliary verb. It is used in the perfective aspect, while ⵓⴰ ad serves as an aspectual particle with the aorist form to express futurity. In the language-specific tagset published by IRCAM, a dedicated tag for auxiliary verbs is currently unavailable; they are, generally, annotated as a verb in the perfective aspect. While the particle ⵓⴰ ad is appropriately covered by the label P.ASP aspectual particles. Within the UD framework, this distinction is already accommodated by assigning both elements the UPOS tag AUX and linking them to the main verb through the aux dependency relation (Nivre et al., 2020). Rather than a contradiction, this alignment demonstrates how UD complements the current IRCAM tagset, allowing these distinct functional elements to be systematically represented within universal syntactic standards.

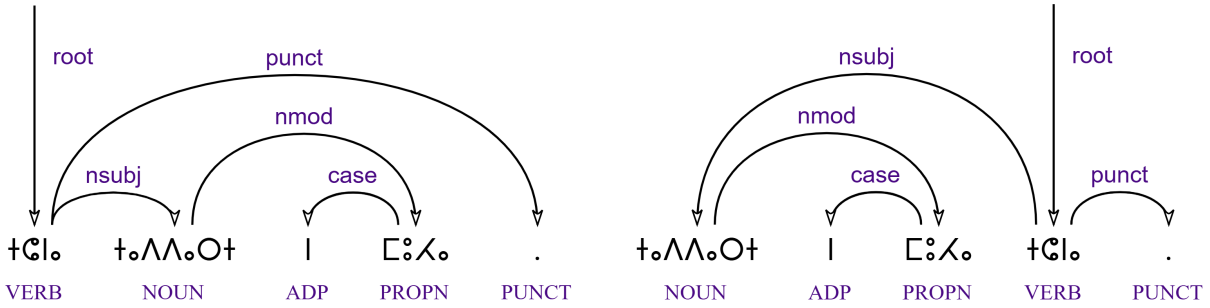
4.4 Word Order Flexibility

A further challenge concerns the flexible word order of Tarifit. While the language is predominantly head-initial with a canonical VSO order, the attested variation in constituent order, particularly SVO and VOS structures, introduces cases where the verb may appear in sentence-final position, potentially leading to misidentification of its syntactic category. For instance, while ⵜⵛⵍⵓ ⵜⵓⴰⴰⵓⵔ ⵏ ⵎⵓⵃⴰ "the house of Muha is beautiful" unambiguously identifies the main verb ⵜⵛⵍⵓ tcna sentence-initially, the variant ⵜⵓⴰⴰⵓⵔ ⵏ ⵎⵓⵃⴰ ⵜⵛⵍⵓ tadart n muḥa tcna places it finally, where it may superficially resemble an adjectival predicate, despite retaining its verbal status. The corresponding dependency annotations are shown in Figure 2, illustrating how the UD framework consistently analyzes both word-order variants while preserving the verbal status of tcna.

5 Discussion

With regard to the segmentation of compound forms involving attached pronouns, such as prepositional constructions and kinship terms, this process is necessary to ensure consistency with the UD guidelines, which allow the segmentation of clitics and contracted forms, treating such constructions as multi-word tokens, as we identified 22 multi-word token instances across the 123 morphosyntactically annotated sentences. This is notably the case in French, where the fused form au "to" is analyzed as a multi-word token segmented into à + le "to + the" (Nivre et al., 2020). A similar approach has been adopted in Arabic, a language belonging to the same Afro-Asiatic family as Amazigh, where pronouns attached to prepositions are likewise segmented; for instance, the form fīhi "in him/it" is treated as a multi-word token[6]. This type of cross-linguistic segmentation is highly consistent with the principle of parallelism endorsed by UD, which aims to promote greater cross-linguistic consistency in treebank construction (De Marneffe et al., 2021). As for lemmatization, and still within the framework of the same principle of parallelism, the UD guidelines specify that the lemma should correspond to the canonical dictionary form of the item (Nivre et al., 2020).This principle is reflected in the previously mentioned in Arabic example, where the attached pronoun -hi is restored to its independent lexical lemma huwa "he". On this basis, the pronoun ⵙ s attached to the preposition in the Amazigh example ⴰⵙⵣⵙ dags is lemmatized as either ⵏⵜⵜⵓ ntta "he" or ⵏⵜⵜⵓⵜ nttat "she" de-

[6] https://universaldependencies.org/treebanks/ar_padt/index.html



(a) Sentence-initial verb.

(b) Sentence-final verb.

Figure 2: UD dependency annotations of two word-order variants in Tarifit.

pending on the context, whereas the pronoun ⵙ s attached to the kinship noun in ⵜⴰⴱⴰⵙ babas is lemmatized by restoring it to its canonical form ⵏⵏⵙ nns "his/her/its". To address the annotation challenges associated with the particle ⴷ d, the auxiliary verb ⵜⵓⵖⴰ tuγa, and the aspectual particle ⴰⴷ ad, this study adopts a dual-layer annotation approach that combines UPOS tags with the IRCAM XPOS tagset. This strategy effectively aligns the descriptive strengths of both frameworks, resolving any systematic divergences. Specifically, the copular, auxiliary, and aspectual functions are uniformly assigned the UPOS tag AUX to ensure cross-linguistic consistency—utilizing the cop relation for the copula and the aux relation for both auxiliary and aspectual particles —while their distinct structural classifications are faithfully preserved within the XPOS layer. The displacement of the verb from sentence-initial position in Tarifit is not arbitrary but reflects a topic-prominence strategy whereby the most salient constituent is fronted at the expense of the canonical VSO order (Cadi, 2005). We must therefore resist interpreting sentence-final verbal forms as adjectival or nominal predicates on the basis of linear position alone, and rely instead on morphosyntactic evidence to determine the correct status of each token.

6 Conclusion

This paper presents a pilot study toward constructing a Universal Dependencies treebank for Tarifit Amazigh. We delineate the corpus selection and annotation methodology, analyzing core challenges in tokenization, lemmatization, annotation decisions, and word order flexibility. By reconciling the language-specific IRCAM tagset with the cross-linguistic UPOS framework, this work establishes a foundational baseline for Amazigh syntactic annotation. Future work will focus on progressively expanding the treebank toward a sufficiently large and representative resource for Tarifit Amazigh, refining annotation guidelines for structural variations, training automated dependency parsers to support downstream NLP applications, and making the annotated corpus together with its annotation guidelines publicly available in CoNLL-U format upon completion of the project.

References

Fatima Agnaou, Khalid Ansar, Fadoua Ataa Allah, Aicha Bouhjar, and Siham Boulaknadel. 2017. L'amazighe dans les sciences du numérique: Expérience de l'IRCAM. In *24th Conférence sur le Traitement Automatique des Langues Naturelles*, pages 1–4.

Arofat Akhundjanova and Luigi Talamo. 2025. Universal dependencies treebank for Uzbek. In *Proceedings of the 3rd Workshop on Resources and Representations for Under-Resourced Languages and Domains*, pages 1–6.

Meftaha Ameur, Aicha Bouhjar, Fatima Boukhris, Ahmed Boukous, Abdallah Boumalk, Mohamed Elmedlaoui, and El Mehdi Iazzi. 2006. *Graphie et Orthographe de l'Amazighe*. Institut Royal de la Culture Amazighe.

Meftaha Ameur, Aicha Bouhjar, Fatima Boukhris, Ahmed Boukous, Abdallah Boumalk, Mohamed Elmedlaoui, and Hamid Souifi. 2004. *Initiation à la Langue Amazighe*. El Maârif al Jadida.

Said Amri and Lahbib Zenkouar. 2018. Amazigh POS tagging using TreeTagger: A language independent model. In *Proceedings of the International Conference on Advanced Intelligent Systems for Sustainable Development*, pages 622–632, Cham. Springer International Publishing.

Fadoua Ataa Allah and Aicha Bouhjar. 2019. The IRCAM realizations for the amazigh preservation and revitalization in morocco. In *Proceedings of the Language Technologies for All*, pages 215–218, Paris, UNESCO Headquarters. ELRA.

Fadoua Ataa Allah, Siham Boulaknadel, and Hamid Souifi. 2014. Jeu d'Étiquettes morphosyntaxiques de la langue amazighe. *Asinag*, 9(9):171–184.

Belaid Boudris, Radia Aitmenait, Abdallah Farkouch, Jamal Belhaj, Faysal Mouttaqi, Abdessaleh Lamrini, Mohamed Elbaghdadi, Hamid Souifi, and Fouad Elhabib. 2022. *Abrid n Umurs 3: Student Textbook for the Third Year of Primary Education*. Ministry of National Education and Okad.

Fatima Boukhris, Abdallah Boumalk, El Houssain El Moujahid, and Hadda Souifi. 2008. *La Nouvelle Grammaire de l'Amazighe*. Institut Royal de la Culture Amazighe.

Kaddour Cadi. 2005. *Transitivité et Diathèse en Tarifit: Analyse de Quelques Relations de Dépendances Lexicale et Syntaxique*. Institut Royal de la Culture Amazighe.

Salem Chaker. 1990. Les bases de l'apparentement chamito-sémitique du berbère. *Etudes et Documents Berbères*, 7(1):28–57.

Marie-Catherine De Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.

El Houssain El Moujahid. 2022. *Grammaire Générative de l'Amazighe: Morphologie et Syntaxe du Nom*, 2 edition. Institut Royal de la Culture Amazighe.

Mohamed Elbaghdadi, Mustapha Oumouch, and Hassan Bouzit. 2022. *Tifawin a TamaziŏŒ263t 2: Student Textbook for the Second Year of Primary Education*. Ministry of National Education and Dar Al-Ouma.

Moha Ennaji and Fatima Sadiqi. 2004. *A Grammar of Amazigh*. Faculty of Dhar El Mehraz, Fès.

Joseph H. Greenberg. 1963. *Universals of Language*. MIT Press.

Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5291–5300.

Nitin Indurkhya and Fred J. Damerau. 2010. *Handbook of Natural Language Processing*. Chapman and Hall/CRC.

Chirantha Liyanage, Kamal Sarveswaran, Tharindu Nadungodage, and Ruvan Pushpananda. 2023. Sinhala dependency treebank (STB). In *Proceedings of the 6th Workshop on Universal Dependencies (UDW, GURT/SyntaxFest)*, pages 17–26.

Otman Maarouf and Rachid El Ayachi. 2021. Part-of-speech tagging using long short term memory (LSTM): Amazigh text written in Tifinaghe characters. In *Proceedings of the 6th International Conference on Business Intelligence*, pages 3–17, Cham. Springer International Publishing.

Fatima Zahra Nejme, Siham Boulaknadel, and Driss Aboutajdine. 2016. AmAMorph: Finite state morphological analyzer for amazighe. *Journal of Computing and Information Technology*, 24(1):91–110.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Jan Hajic, Christopher D. Manning, Sampo Pyysalo, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043.

Jamal Ouhalla. 1988. *The Syntax of Head Movement: A Study of Berber*. Ph.D. thesis, University of London.

Hanai Raiss and Violetta Cavalli-Sforza. 2012. ANMorph: Amazigh nouns morphological analyzer. In *Proceedings of the 5th International Conference on Information and Communication Technologies for Amazigh*.

Imane Taghbalout, Fadoua Ataa Allah, and Mohamed El Marraki. 2018. Towards UNL-based machine translation for Moroccan amazigh language. *International Journal of Computational Science and Engineering*, 17(1):43–54.

Francis M. Tyers and Vinit Ravishankar. 2018. A prototype dependency treebank for Breton. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles*, pages 197–204.

Arabic Dialect-to-MSA Translation: A Comparative Evaluation of Large Language Models and Neural Machine Translation

Maram I. Alharbi[1,2], **Jihad R'baiti**[3], **Ruslan Mitkov**[1,4],
[1]School of Computing and Communications, Lancaster University, UK
[2]Jazan University, Saudi Arabia
[3]International Artificial Intelligence Center of Morocco,
Mohammed VI Polytechnic University, Morocco
[4]University of Alicante, Spain
m.i.alharbi@lancaster.ac.uk, jihad.rbaiti@um6p.ma

Abstract

Arabic dialect to Modern Standard Arabic (MSA) translation remains challenging due to the rich linguistic variation across dialects and the limited availability of parallel corpora. Recent advances in Large Language Models (LLMs) have created new potential to address these challenges. However, their effectiveness in Arabic dialect translation remains underexplored. In this paper, we present a comparative evaluation of Arabic language models, multilingual language models, and dedicated neural machine translation under zero-shot, few-shot, and fine-tuning settings on the ADOR dataset, a parallel corpus covering Saudi, Moroccan Darija, Egyptian, and Jordanian Arabic. Results show clear differences between model families, with Fanar achieving the strongest overall performance. Fine-tuning consistently improved translation quality, whereas prompting produced more variable results across dialects.

1 Introduction

Machine Translation (MT) has advanced considerably with the development of Neural Machine Translation (NMT) and large-scale language models, enabling high-quality translation across many languages and domains. Despite these advances, Arabic remains a challenging language for MT due to its diglossic nature (Almahairi et al., 2016; Harrat et al., 2019).

The translation of Arabic dialects into Modern Standard Arabic (MSA) is important for a wide range of applications, including information retrieval, sentiment analysis, and the processing of user-generated content. This need is particularly evident in social media and online reviews, where most content is written in dialectal Arabic rather than MSA (Alabdullah et al., 2026). However, dialectal Arabic presents several challenges for automatic processing. The absence of standardised orthography, substantial variation across dialects, and the influence of other languages often result in highly variable written forms that increase ambiguity and data sparsity (Alharbi, 2026; Alhafni et al., 2024).

These challenges are further compounded by the limited availability of parallel dialect-MSA corpora. Although recent advances in NMT and Large Language Models (LLMs) have improved translation quality for low-resource languages, the effectiveness of different modelling approaches for dialect-to-MSA translation remains insufficiently explored, particularly across multiple Arabic dialects setting.

In this study, we investigate dialect-to-MSA translation across multiple Arabic dialects and analyse the effectiveness of different modelling approaches under a unified evaluation framework. The main contributions of this work are as follows:

- We compare the performance of Arabic LLMs, multilingual language models, and a dedicated NMT model for Arabic dialect-to-MSA translation using the ADOR dataset.
- We investigate the effect of few-shot prompting and LoRA-based fine-tuning on translation quality across different Arabic dialects.
- We conduct a comprehensive evaluation using automatic metrics, human evaluation and qualitative error analysis.
- We analyse the strengths and limitations of the evaluated models and identify challenges in Arabic dialect translation.

The remainder of this paper is organised as follows. Section 2 reviews related work on Arabic dialect translation and large language models. Section 3 describes the dataset used in this study. Section 4 presents the experimental setup, models, and

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 30–39
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_005

evaluation procedure. Section 5 provides the implementation details of the training models. Section 6 reports the experimental results. Section 7 discusses the main findings of the study. Section 8 analyses the translation errors observed across the evaluated models. Finally, Section 9 concludes the paper and outlines directions for future work.

2 Related Work

The scarcity of parallel corpora remains a major challenge for Arabic dialect machine translation. Khered et al. (2025) introduced a large-scale dialect-to-MSA corpus collected from social media and demonstrated its usefulness for dialect normalisation and translation tasks. El Faida and Yousfi (2025) reviewed rule-based, statistical, and neural approaches to dialect translation and discussed challenges related to dialectal variation and non-standard orthography. The transformation of dialectal Arabic into MSA has also been studied as a formality transfer task, where preserving semantic meaning while adapting linguistic form is a central requirement (Abudalfa et al., 2024). Neural approaches have become the dominant paradigm for Arabic dialect translation. R'baiti et al. (2026) proposed BERT-TransDial, which integrates BERT embeddings into a Transformer architecture for translation between Maghrebi dialects and MSA. Hamed et al. (2025) extended the task beyond dialect-to-MSA translation by investigating dialect-to-dialect translation across multiple Arabic varieties and exploring a question-answering formulation of the task.

Large language models (LLMs) have also been evaluated for Arabic dialect translation. Alabdullah et al. (2025) evaluated prompting and fine-tuning strategies for dialect-to-MSA translation and reported competitive results using few-shot prompting and multi-dialect training. Alharbi (2026) evaluated LLMs on Saudi dialect translation, while Alabdullah et al. (2026) examined translation quality through human-centred error analysis. Prompting has also been compared with fine-tuning; Obeidat et al. (2025) reporting improvements from few-shot prompting over fine-tuning in a low-resource setting, whereas Beidas et al. (2025) observed that prompting effectiveness varied across dialects and evaluation metrics. Abdul-Nabi et al. (2024) surveyed neural machine translation methods for low-resource Arabic dialects and highlighted the continued impact of data scarcity on translation performance. Similarly, Alimi et al. (2024) showed that dialect grouping strategies and carefully selected training data can improve dialect-to-MSA translation quality.

Despite these developments, most previous work has focused on dialect-to-MSA translation. Comparisons between Arabic LLMs, multilingual LLMs, and specialised machine translation systems under zero-shot, few-shot, and parameter-efficient fine-tuning settings remain limited.

3 Data

In this study, we use the ADOR dataset (Alharbi et al., 2025), a parallel Arabic dataset consisting of 538 hotel review sentences in MSA aligned with four dialectal variants: Saudi, Moroccan Darija, Egyptian, and Jordanian Arabic. We reformulate the dataset by treating each dialectal sentence as the source and its corresponding MSA sentence as the target translation. Additionally, we retain the source dialect label and the corresponding sentiment annotation for each instance. This transformation results in a total of 2,152 dialect-to-MSA parallel sentence pairs. The dataset was then randomly split into 80% training and 20% test sets.

4 Experimental Methods

We formulate the task as a parallel machine translation problem from four Arabic dialects: Saudi, Moroccan Darija, Egyptian, and Jordanian to MSA. The objective is to evaluate the ability of different models to generate accurate and fluent MSA translations while preserving the meaning of the original dialectal input.

We evaluated a diverse set of Arabic and multilingual models. The evaluated Arabic LLMs included Jais-Family-2.7B (Sengupta et al., 2023), AceGPT-7B and AceGPT-7B-chat (Huang et al., 2023), Fanar-1-9B (Team et al., 2025), Atlas-Chat-9B (Shang et al., 2024), and ALLaM-7B-Instruct (Alzahrani et al., 2024). We also evaluated Qwen2.5-3B-Instruct (Team, 2024), a multilingual instruction-tuned model, and NLLB-200-3.3B (Costa-Jussà et al., 2022), a multilingual NMT system supporting MSA and several Arabic dialects, including those represented in the ADOR dataset.

The following three experimental settings were investigated: zero-shot prompting, few-shot prompting, and parameter-efficient fine-tuning.

Zero-shot prompting: The models were evaluated in a zero-shot setting. Each model was provided with a dialectal Arabic sentence and an instruction to translate it into MSA. No task-specific examples were included in the prompt.

Few-shot prompting: The same set of models were evaluated using few-shot prompting. Three dialect-to-MSA translation examples were included in the prompt before the test sentence. In addition, the prompt contained explicit instructions to preserve the original meaning, sentiment, and contextual information while producing a fluent MSA translation.

Fine-tuning: The models were adapted using parameter-efficient fine-tuning with Low-Rank Adaptation (LoRA) (Hu et al., 2021). The task was formulated as an instruction-following translation problem, where the input consisted of a translation instruction, the dialect label, and the dialectal sentence, while the target output corresponded to the reference MSA translation. The performance of the fine-tuned models was then compared with the zero-shot and few-shot prompting settings.

In addition to the LLMs, we evaluated NLLB-200-3.3B model in both its pretrained and fine-tuned configurations, allowing comparison between a dedicated NMT and general-purpose language models under the same dialect-to-MSA translation task.

5 Implementation details

Throughout all the finetuning experiments, we used a fixed set of hyperparameters: 8 training epochs, a learning rate of (2×10^{-5}), a per-device batch size of 1, and 16 gradient accumulation steps. For LoRA-based adaptation, the rank was set to (r = 16), the scaling factor to ($\alpha = 32$), and the dropout rate to 0.05. During inference, translations were generated using a greedy decoding strategy with a maximum of 128 new tokens. All experiments were conducted on NVIDIA A100-SXM4 GPUs with 80 GB of memory.

6 Experimental Results

This section presents the BLEU (Papineni et al., 2002), BERTScore precision (Zhang et al., 2020), ChrF (Popović, 2015), and METEOR (Banerjee and Lavie, 2005) results obtained on the test set under three experimental settings: zero-shot prompting, few-shot prompting, and fine-tuning. In Tables 1-3, bold values denote the best overall score for each evaluation metric across all models and dialects, while underlined values indicate the highest-scoring dialect for each individual model.

6.1 Zero-Shot Prompting

Table 1 summarises the zero-shot translation performance across the four Arabic dialects. Clear differences were observed across both models and dialects. Fanar-1-9B, ALLaM-7B, and NLLB-3.3B consistently achieved the strongest results across the four evaluation metrics, whereas Qwen2.5-3B, AceGPT-7B-chat, and Jais-Family-2.7B generally produced lower scores.

Across almost all models, Jordanian Arabic proved to be the easiest dialect to translate, achieving the highest scores for Fanar-1-9B, ALLaM-7B, Atlas-Chat-9B, Qwen2.5-3B, and AceGPT-7B-chat. Fanar-1-9B achieved the highest overall performance, reaching a BLEU score of 16.04, a BERTScore of 82.60, a ChrF score of 51.38, and a METEOR score of 32.71 on Jordanian Arabic. NLLB-3.3B showed a slightly different trend, obtaining its highest BLEU, BERTScore, and METEOR scores on Egyptian Arabic while achieving its highest ChrF score on Jordanian Arabic. In contrast, Jais-Family-2.7B consistently performed best on Saudi Arabic.

Moroccan Darija was the most challenging dialect for all evaluated models, producing the lowest BLEU, ChrF, and METEOR scores in every case. Although Egyptian and Saudi Arabic generally yielded stronger performance than Darija, their relative ranking varied across models. Overall, the results indicate that Arabic-specific models consistently outperformed the multilingual models under zero-shot prompting, with Fanar-1-9B achieving the strongest performance across most dialects and evaluation metrics.

6.2 Few-Shot Prompting

Table 2 presents the few-shot translation results. Unlike expectations, incorporating three in-context translation examples did not improve translation quality. Instead, performance declined across most models and evaluation metrics compared with the zero-shot setting.

Despite this overall reduction, the relative behaviour of the dialects remained largely unchanged. Jordanian Arabic continued to achieve the highest scores for most models, whereas Moroccan Darija remained the most difficult dialect. Atlas-Chat-9B

Model	Dialect	BLEU	BERTScore	ChrF	METEOR
NLLB-3.3B	Darija	8.24	78.88	39.92	23.47
	Egyptian	14.88	82.13	46.00	31.48
	Jordanian	12.36	80.77	46.82	30.30
	Saudi	11.68	81.70	45.18	30.56
Atlas-Chat-9B	Darija	6.91	78.10	37.90	21.03
	Egyptian	8.12	77.49	37.46	21.04
	Jordanian	9.38	79.46	43.48	25.98
	Saudi	7.09	78.83	39.17	22.35
ALLaM-7B	Darija	9.02	79.16	41.85	22.77
	Egyptian	11.15	80.08	46.07	27.40
	Jordanian	14.94	81.85	50.63	31.12
	Saudi	11.22	81.32	47.66	28.58
Fanar-1-9B	Darija	8.04	78.99	41.33	22.59
	Egyptian	12.78	80.84	45.32	28.79
	Jordanian	**16.04**	**82.60**	**51.38**	**32.71**
	Saudi	12.34	81.95	48.22	31.45
Qwen2.5-3B	Darija	4.48	75.60	35.24	16.63
	Egyptian	6.97	75.68	37.63	19.65
	Jordanian	9.88	78.67	43.04	24.56
	Saudi	5.81	76.49	39.19	21.06
AceGPT-7B	Darija	4.23	77.33	34.49	19.76
	Egyptian	8.63	80.55	36.22	26.28
	Jordanian	11.26	82.22	41.53	29.60
	Saudi	10.20	82.24	41.16	31.13
AceGPT-7B-chat	Darija	5.35	77.17	35.27	18.02
	Egyptian	7.26	78.10	34.85	20.31
	Jordanian	10.29	80.57	41.35	26.18
	Saudi	8.14	80.04	38.45	24.31
jais-family-2.7b	Darija	5.46	75.86	33.21	16.61
	Egyptian	9.27	78.13	35.82	21.45
	Jordanian	10.46	79.00	39.79	24.15
	Saudi	10.66	80.49	42.55	27.74

Table 1: Zero-shot dialect-to-MSA translation performance across four Arabic dialects using BLEU, BERTScore, ChrF, and METEOR.

achieved the strongest overall performance in the few-shot setting, obtaining the highest BERTScore 78.93 on Saudi Arabic together with the highest ChrF 46.16 and METEOR 27.70 scores on Jordanian Arabic. Fanar-1-9B produced the highest BLEU score 11.74, although its performance was noticeably lower than in the zero-shot setting. ALLaM-7B maintained relatively consistent performance across the four dialects, with its strongest results again observed on Jordanian Arabic.

The remaining models followed similar trends. Qwen2.5-3B, AceGPT-7B, and AceGPT-7B-chat all achieved their best results on Jordanian Arabic despite reductions across the four evaluation metrics. Jais-Family-2.7B remained the weakest-performing model, with particularly low scores on Moroccan Darija. Overall, the results suggest that few-shot prompting was insufficient to improve dialect-to-MSA translation and often reduced translation quality compared with zero-shot prompting.

6.3 Fine-Tuning

Table 3 presents the results obtained after parameter-efficient fine-tuning. In contrast to prompting, fine-tuning substantially improved translation quality for every evaluated model across BLEU, BERTScore, ChrF, and METEOR, demonstrating the effectiveness of supervised adaptation for dialect-to-MSA translation.

Although all dialects benefited from fine-tuning, their relative difficulty remained largely unchanged. Jordanian Arabic consistently achieved the highest scores, while Moroccan Darija continued to be the most challenging dialect. Fanar-1-9B and ALLaM-7B achieved the strongest overall performance. Fanar-1-9B obtained the highest BLEU 23.24 and METEOR 46.06 scores on Jordanian Arabic, whereas ALLaM-7B achieved the highest

Model	Dialect	BLEU	BERTScore	ChrF	METEOR
Atlas-Chat-9B	Darija	5.24	78.00	37.49	19.27
	Egyptian	9.52	78.33	39.34	22.45
	Jordanian	**12.81**	**80.27**	**46.16**	**27.70**
	Saudi	8.02	78.93	40.07	22.51
ALLaM-7B	Darija	7.61	77.30	39.43	22.13
	Egyptian	8.16	77.03	42.25	22.76
	Jordanian	10.16	78.22	46.03	26.29
	Saudi	7.79	78.34	42.71	23.39
Fanar-1-9B	Darija	6.12	73.46	32.10	16.01
	Egyptian	9.42	74.78	34.97	20.85
	Jordanian	11.74	76.45	40.29	24.74
	Saudi	9.36	77.05	38.01	24.20
Qwen2.5-3B	Darija	4.10	75.92	35.70	16.59
	Egyptian	7.15	76.97	38.46	20.79
	Jordanian	9.62	78.32	44.01	25.41
	Saudi	5.17	76.92	39.56	21.83
AceGPT-7B	Darija	3.57	77.49	32.18	17.36
	Egyptian	7.59	78.96	33.05	22.72
	Jordanian	7.88	79.70	35.29	24.01
	Saudi	5.72	78.33	30.44	21.37
AceGPT-7B-chat	Darija	4.58	76.68	34.96	17.19
	Egyptian	7.03	77.67	34.35	21.12
	Jordanian	9.70	78.16	38.22	23.19
	Saudi	7.54	77.33	34.79	22.42
jais-family-2.7b	Darija	3.76	74.05	26.64	10.64
	Egyptian	6.60	76.24	28.63	15.78
	Jordanian	8.15	78.08	34.89	18.63
	Saudi	4.31	76.67	28.35	14.63

Table 2: Few-shot translation performance across the four evaluated Arabic dialects.

BERTScore 86.04 and ChrF 55.84 scores on the same dialect. NLLB-3.3B also benefited substantially from fine-tuning and remained competitive across all four dialects.

The remaining models also showed consistent improvements. Jais-Family-2.7B exhibited one of the largest gains relative to its prompting performance, particularly on Egyptian and Jordanian Arabic. Qwen2.5-3B improved across all evaluation metrics but continued to lag behind Fanar-1-9B, ALLaM-7B, and NLLB-3.3B. AceGPT-7B and AceGPT-7B-chat likewise achieved higher scores than in both prompting settings; however, they remained the lowest-performing fine-tuned models, especially on Moroccan Darija and Egyptian Arabic.

6.4 Human Evaluation

To complement the automatic evaluation, we conducted a human assessment of the translations generated by the fine-tuned models. Three native Arabic speakers independently evaluated the translation quality. All annotators were fluent in MSA and familiar with the four Arabic dialects included in this study.

For each model, five test sentences were randomly selected from each dialect, resulting in 20 translations per model and a total of 160 translations across the eight evaluated models.

Following the evaluation framework of Arnold et al. (1994), each translation was assigned a single quality score on a four-point ordinal scale, where 4 indicates a perfectly clear, grammatical, and fully intelligible translation; 3 indicates a generally clear translation with only minor inaccuracies; 2 indicates that the general meaning can be understood only after careful reading due to grammatical or lexical errors; and 1 indicates an unintelligible translation that fails to reliably convey the intended meaning.

The annotators evaluated each translation independently. During the evaluation, they were presented with the dialectal source sentence, the reference MSA translation, and the model-generated translation.

Table 4 presents the average human evaluation scores. Fanar-1-9B achieved the highest average score 3.72, followed by NLLB-3.3B 3.68 and Jais-

Model	Dialect	BLEU	BERTScore	ChrF	METEOR
NLLB-3.3B	Darija	11.28	81.61	43.67	30.31
	Egyptian	15.52	83.56	48.50	35.65
	Jordanian	18.59	85.16	53.21	41.71
	Saudi	14.55	84.37	49.32	37.91
Atlas-Chat-9B	Darija	14.30	81.80	46.08	32.67
	Egyptian	15.02	83.40	47.03	37.71
	Jordanian	19.32	85.16	53.19	43.08
	Saudi	13.98	84.13	49.63	38.48
ALLaM-7B	Darija	14.05	83.01	47.11	33.87
	Egyptian	18.11	84.70	50.60	41.96
	Jordanian	22.68	**86.04**	**55.84**	44.52
	Saudi	19.06	85.25	52.46	41.72
Fanar-1-9B	Darija	13.75	82.57	46.02	33.65
	Egyptian	18.00	84.15	49.54	40.18
	Jordanian	**23.24**	85.87	54.89	**46.06**
	Saudi	18.22	85.28	52.45	42.35
Qwen2.5-3B	Darija	8.25	79.94	40.14	25.26
	Egyptian	11.96	81.46	43.64	30.24
	Jordanian	15.29	83.97	49.67	37.05
	Saudi	9.68	81.91	43.55	29.08
AceGPT-7B	Darija	6.98	78.47	36.25	26.19
	Egyptian	8.69	81.31	38.20	31.24
	Jordanian	12.49	81.96	42.48	34.92
	Saudi	8.16	81.16	38.74	31.54
AceGPT-7B-chat	Darija	6.28	78.02	33.31	22.72
	Egyptian	6.12	80.45	33.65	27.33
	Jordanian	10.77	82.47	40.39	31.57
	Saudi	7.61	81.39	35.97	28.99
jais-family-2.7b	Darija	13.14	81.69	44.40	32.59
	Egyptian	17.26	84.10	48.84	39.67
	Jordanian	18.44	85.10	52.85	41.68
	Saudi	15.60	84.95	51.28	39.72

Table 3: Translation performance after LoRA fine-tuning across the four evaluated Arabic dialects.

Table 4: Average human evaluation scores on the four-point quality scale.

Model	Human Score
Fanar-1-9B	**3.72**
NLLB-3.3B	3.68
Jais-Family-2.7B	3.67
AceGPT-7B	3.53
ALLaM-7B	3.52
Atlas-Chat-9B	3.45
Qwen2.5-3B	3.45
AceGPT-7B-chat	3.15

Family-2.7B 3.67. AceGPT-7B and ALLaM-7B obtained comparable scores of 3.53 and 3.52, respectively, while Atlas-Chat-9B and Qwen2.5-3B each achieved an average score of 3.45. AceGPT-7B-chat received the lowest average score 3.15 among the evaluated models.

Table 5: Inter-annotator agreement for the human evaluation.

Metric	Score
Krippendorff's α (ordinal)	0.465
Weighted Cohen's κ (mean)	0.560

6.5 Inter-Annotator Agreement

Table 5 reports the agreement among the three annotators. Krippendorff's α yielded a value of 0.465, indicating moderate agreement. The mean weighted Cohen's κ was 0.560, further demonstrating a moderate level of consistency when accounting for the ordinal nature of the evaluation scale.

7 Discussion

The results demonstrate that model adaptation plays a more important role than prompting alone for dialect-to-MSA translation. While zero-shot

prompting provided a reasonable baseline, few-shot prompting did not consistently improve performance across the evaluated models. In contrast, fine-tuning substantially improved translation quality across all automatic evaluation metrics, highlighting the effectiveness of parameter-efficient adaptation even with a relatively small parallel dataset.

The limited effectiveness of few-shot prompting may be attributed to the small number of in-context examples provided during inference. In contrast, fine-tuning enables the models to adapt their internal representations to dialect-specific patterns, leading to consistent improvements across all evaluation metrics.

Differences were also observed between model families. Arabic-specific models, particularly ALLaM-7B and Fanar-1-9B, consistently achieved the strongest automatic evaluation results after fine-tuning, while NLLB-3.3B remained highly competitive despite being designed as a dedicated multilingual NMT system rather than an instruction-following language model. In contrast, the smaller general-purpose language models showed more limited gains after adaptation, suggesting that model architecture and pre-training data continue to influence translation quality even after fine-tuning.

Translation quality also varied across dialects. Jordanian Arabic consistently produced the highest scores for most models, whereas Moroccan Darija remained the most challenging dialect. This pattern is consistent with the greater lexical and linguistic variation found in Darija compared with MSA, highlighting the difficulty of handling dialect-specific vocabulary and non-standard lexical forms.

The human evaluation broadly supported the conclusions drawn from the automatic evaluation. Fanar-1-9B achieved the highest average human score, followed closely by NLLB-3.3B and Jais-Family-2.7B, while AceGPT-7B-chat consistently received the lowest ratings. Although the overall ranking of the strongest and weakest models remained largely consistent, some differences were observed among the middle-performing systems. In particular, Jais-Family-2.7B received higher human ratings than suggested by the automatic metrics, indicating that lexical overlap-based evaluation metrics may not fully capture perceived translation quality. These findings highlight the complementary role of human evaluation when assessing Arabic dialect-to-MSA translation systems.

8 Error Analysis

Manual inspection of the generated translations revealed several recurring error patterns across the evaluated models. The most common errors involved incorrect interpretation of dialect-specific vocabulary, semantic shifts, incomplete normalisation into MSA, and information omission.

Lexical interpretation errors were particularly common in Moroccan Darija. Several models failed to correctly interpret dialect-specific words and instead generated approximate or unrelated MSA equivalents. This behaviour was most noticeable in the zero-shot setting and frequently resulted in changes to the original meaning of the sentence. For example, in a review containing the phrase مكيف الهواء ما هو إلا عبارة عن مروحة لا يمكنها تبريد الغرفة, Atlas-Chat-9B generated المكيّاج هو كل ما تبقى من المروحة لا يمكنها تبريد الغرفة, incorrectly interpreting مكيف الهواء and producing a semantically unrelated translation. Similarly, Qwen2.5-3B-Instruct translated مقارنة بالهند بالنسبة لهنغاريا as, replacing the reference to Hungary with India. In a Saudi review, the colloquial expression ما انصحك تجرب هذا الفندق كانت تجربة سيئة اني اخترت هذا الفندق was translated by AceGPT-7B-chat as كنت أتوقع أكثر من ذلك، لذا اخترت فندقًا آخر, introducing information that was not present in the source text and altering the intended meaning.

Semantic shifts were observed across all prompting-based models. In these cases, the generated sentence remained fluent but did not preserve the intended meaning of the source text. Hotel review sentences containing subjective evaluations were especially prone to this problem, with models occasionally altering the strength of positive or negative opinions. For example, in a Jordanian review, the reference MSA states اتمنى ان يتأكدوا اكثر من الفنادق والصور الموجودة, expressing a request for the booking company to verify the accuracy of hotel descriptions and images. Atlas-Chat-9B generated أتمنى أن تضمنوا المزيد من الفنادق والصور, changing the meaning to a request for more hotels and images, while ALLaM-7B-Instruct produced أتمنى أن يتم تقديم المزيد من الخدمات, introducing a different request altogether. A similar shift appears in a Saudi review containing the strongly negative statement أسوأ فندق من

الفنادق ولا يستحق نجمة واحدة. Jais-Family-2.7B generated الفندق عادي ولا يستحق السعر, weakening the original criticism and changing the focus from overall quality to value for money. In another example, the reference sentence الإفطار في الفندق عادي was translated by Atlas-Chat-9B as الإفطار في الفندق جيد, strengthening a neutral evaluation into a positive one.

Another recurring issue involved partial normalisation. Rather than producing a fully standardised MSA sentence, some models retained dialectal words or colloquial expressions in the output. These errors were particularly noticeable in Moroccan Darija and Egyptian, where dialect-specific vocabulary was sometimes copied unchanged into the generated translation. The problem was less frequent for Saudi and Jordanian Arabic due to their closer lexical similarity to MSA. For example, Moroccan Darija expressions such as بزاف and ديال were sometimes preserved instead of being converted to their MSA equivalents. Similar behaviour was observed in Egyptian Arabic, where colloquial forms such as كويس and أوي occasionally remained unchanged in the output. Saudi and Jordanian translations also exhibited partial normalisation, although less frequently, with expressions such as مره and مو sometimes appearing in otherwise standardised sentences.

Information omission was also common, particularly in sentences describing multiple aspects of a hotel review. Models occasionally translated the main opinion correctly while omitting secondary details related to room facilities, staff behaviour, location, or pricing.

The distribution of errors varied across dialects. Moroccan Darija produced the largest number of lexical interpretation errors, whereas Egyptian Arabic more frequently exhibited semantic shifts involving evaluative expressions. Saudi and Jordanian Arabic generally produced fewer severe lexical errors, although omission and partial normalisation errors remained present.

Fine-tuning substantially reduced lexical interpretation and normalisation errors across all dialects. In contrast, the few-shot setting did not consistently improve translation quality over zero-shot prompting and continued to exhibit many of the same error patterns observed in the baseline setting.

9 Conclusion and Future work

This paper presented a comparative evaluation of Arabic language models, multilingual language models, and an NMT for dialect-to-MSA translation. Experiments were conducted on the ADOR dataset under three settings: zero-shot prompting, few-shot prompting, and parameter-efficient fine-tuning.

The results showed that fine-tuning consistently produced the best performance across all evaluation metrics, substantially outperforming both zero-shot and few-shot prompting. Among the evaluated models, ALLaM-7B-Instruct-preview and Fanar-1-9B achieved the strongest overall results after fine-tuning, while NLLB-200-3.3B remained a competitive NMT. In contrast, few-shot prompting did not consistently improve translation quality and often performed similarly to or worse than the zero-shot setting.

Analysis of the generated translations showed that dialect-specific vocabulary, semantic shifts, partial normalisation, and information omission remain common challenges, particularly for dialects that are lexically more distant from MSA. Although fine-tuning reduced many of these issues, accurately handling highly dialectal and idiomatic expressions remains difficult.

Future work will investigate additional Arabic dialects and alternative adaptation strategies. We also plan to explore dialect-aware prompting and translation approaches.

References

Razan Abdul-Nabi, Rasha Obeidat, and Anas Bsoul. 2024. A survey on machine translation of low-resource arabic dialects. In *2024 15th International Conference on Information and Communication Systems (ICICS)*, pages 1–6. IEEE.

Shadi I Abudalfa, Fahad J Abdu, and Maad M Alowaifeer. 2024. Arabic text formality modification: A review and future research directions. *IEEE Access*, 12:185117–185148.

Abdullah Alabdullah, Lifeng Han, and Chenghua Lin. 2025. Advancing dialectal arabic to modern standard arabic machine translation.

Abdullah Alabdullah, Lifeng Han, and Chenghua Lin. 2026. Ara-HOPE: Human-centric post-editing evaluation for dialectal Arabic to Modern Standard Arabic translation. In *Proceedings*

of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects, pages 157–171, Rabat, Morocco. Association for Computational Linguistics.

Bashar Alhafni, Sarah Al-Towaity, Ziyad Fawzy, Fatema Nassar, Fadhl Eryani, Houda Bouamor, and Nizar Habash. 2024. Exploiting dialect identification in automatic dialectal text normalization. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 42–54, Bangkok, Thailand. Association for Computational Linguistics.

Ghada Alharbi. 2026. Saudi dialects to msa machine translation: A systematic evaluation of llms. *Procedia Computer Science*, 275:464–473. 7th International Conference on AI in Computational Linguistics.

Maram I Alharbi, Saad Ezzini, Hansi Hettiarachchi, Tharindu Ranasinghe, and Ruslan Mitkov. 2025. Ador: Dataset for arabic dialects in hotel reviews: A human benchmark for sentiment analysis. In *Proceedings of the First Workshop on Advancing NLP for Low-Resource Languages*, pages 187–191.

Tahar Alimi, Rahma Boujebane, Wiem Derouich, and L Belguith. 2024. Fine-tuned transformers for translating multi-dialect texts to modern standard arabic. *Int. J. Cogn. Lang. Sci*, 18:679–684.

Amjad Almahairi, Kyunghyun Cho, Nizar Habash, and Aaron Courville. 2016. First result on arabic neural machine translation. *arXiv preprint arXiv:1606.02680*.

Norah A. Alzahrani, Nouf M. Alotaibi, Hisham A. Alyahya, Sultan AlRashed, Faisal A. Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, et al. 2024. Allam: Large language models for arabic and english. *arXiv preprint arXiv:2407.15390*.

Douglas J. Arnold, Lorna Balkan, Siety Meijer, R. Lee Humphreys, and Louisa Sadler. 1994. *Machine Translation: An Introductory Guide*. NCC Blackwell, London.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Ayah Beidas, Kousar Mohi, Fatme Ghaddar, Imtiaz Ahmad, and Sa'Ed Abed. 2025. Cross-dialectal arabic translation: comparative analysis on large language models. *Frontiers in Artificial Intelligence*, 8:1661789.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Hafsa El Faida and Abdellah Yousfi. 2025. Machine translation of dialects into standard arabic: A systematic review and comparative analysis. In *2025 International Conference on Intelligent Systems: Theories and Applications (SITA)*, pages 1–7. IEEE.

Mohamed Motasim Hamed, Muhammad Hreden, Khalil Hennara, Zeina Aldallal, Sara Chrouf, and Safwan AlModhayan. 2025. Lahjawi: Arabic cross-dialect translator. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 12–24.

Salima Harrat, Karima Meftouh, and Kamel Smaili. 2019. Machine translation for arabic dialects (survey). *Information Processing Management*, 56(2):262–273. Advance Arabic Natural Language Processing (ANLP) and its Applications.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Zhen Huang, Shichao Sun, Weizhe Yuan, Ruofan Wang, Jiahui Gao, Yuxin Jiang, Xiuying Chen, Zhiyang Teng, Qin Jin, Kam-Fai Wong, and Xiangliang Zhang. 2023. Acegpt, localizing large language models in arabic.

Abdullah Khered, Youcef Benkhedda, and Riza Batista-Navarro. 2025. Dial2MSA-verified: A multi-dialect Arabic social media dataset for neural machine translation to Modern Standard

Arabic. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 50–62, Abu Dhabi, UAE. Association for Computational Linguistics.

Rasha Obeidat, Luay Alawneh, and Yara Al-Harahsheh. 2025. Exploring prompting for dialectical machine translation: a focus on north jordanian arabic. *PeerJ Computer Science*, 11:e3209.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Jihad R'baiti, Youssef Hmamouche, and Amal El Fallah Seghrouchni. 2026. Maghrebi dialects – arabic bidirectional translation: an improved transformer with transfer learning. *Natural Language Processing*, 32(2):204–223.

Spandan Sengupta, Amith Sankar, Fady El-Kazaz, and et al. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Guokan Shang, Hadi Abdine, Yousef Khoubrane, Amr Mohamed, Yassine Abbahaddou, Sofiane Ennadir, Imane Momayiz, Xuguang Ren, Eric Moulines, Preslav Nakov, Michalis Vazirgiannis, and Eric Xing. 2024. Atlas-chat: Adapting large language models for low-resource moroccan arabic dialect.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, Majd Hawasly, Mus'ab Husaini, Soon-Gyo Jung, Ji Kim Lucas, Walid Magdy, Safa Messaoud, Abubakr Mohamed, Tasnim Mohiuddin, Basel Mousi, Hamdy Mubarak, Ahmad Musleh, Zan Naeem, Mourad Ouzzani, Dorde Popovic, Amin Sadeghi, Husrev Taha Sencar, Mohammed Shinoy, Omar Sinan, Yifan Zhang, Ahmed Ali, Yassine El Kheir, Xiaosong Ma, and Chaoyi Ruan. 2025. Fanar: An arabic-centric multimodal generative ai platform.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Cross-Lingual Transfer from Portuguese to Nheengatu: Evidence for Contact-Induced Convergence as a Computational Bridge

Rafael Macário Fernandes
University of São Paulo
Department of Linguistics, São Paulo, Brazil
rafael.macario@usp.br

Abstract

Typological distance predicts that Portuguese-to-Nheengatu transfer should fail: the languages are genetically unrelated and typologically divergent (URIEL+ genetic 1.0, syntactic 0.61). We argue that 400 years of contact-induced structural convergence may provide a computational bridge invisible to these measures, so that transfer should succeed at the levels where contact has operated most deeply. Across five experiments on a parallel corpus of about 5,000 sentence pairs, we find a progressive pattern: word-level and lexical-semantic methods fail, while fine-tuned XLM-RoBERTa reaches Precision@1 of 24.7% and MRR of 0.371 on cross-lingual sentence retrieval—more than double a BM25 character-*n*-gram baseline and 1.7× that of zero-shot LaBSE, the strongest off-the-shelf retriever we test. Crucially, Portuguese beats Spanish—typologically near-identical and no less represented in pretraining, but without comparable contact history—on every metric and all five seeds, and the two sources share substantial lexical overlap, so lexical overlap alone is unlikely to explain the gap. These results suggest that contact history can distinguish transfer sources that typological distance treats as equivalent.

1 Introduction

Nheengatu (ISO 639-3: yrl)[1] is a Tupian language and the only living descendant of Tupinambá, the former lingua franca of colonial Brazil (Moore, 2014; g1 Ceará, 2025). According to the 2022 Brazilian census, it is spoken by 13,070 people in Brazil, making it the country's eighth most spoken Indigenous language (IBGE, 2024). It is co-official with Portuguese in São Gabriel da Cachoeira (Amazonas) and is also spoken in neighboring Colombia (Moore, 2014; g1 Ceará, 2025). In the Upper Rio Negro, where it serves as a supra-ethnic lingua franca, it has been adopted by several Indigenous groups, while community-led revitalization initiatives are expanding its use in Pará (Cintra, 2018; Silva Meirelles, 2020). Despite its demographic importance and ongoing revitalization, Nheengatu remains severely under-resourced: existing computational research relies on essentially the same small set of textual resources as ours (Adelani et al., 2024; de Alencar, 2024; Pinhanez et al., 2024).

Portuguese is the natural candidate for cross-lingual transfer: every parallel resource in our corpus pairs Nheengatu with it, and Adelani et al. (2024) find that zero-shot POS transfer into Nheengatu is stronger from Brazilian Portuguese than from English (31.5 vs. 28.9 accuracy). Typologically, however, the pair is unfavorable. URIEL+ (Khan et al., 2025) places the two at maximal genetic distance (1.0) and at syntactic distance 0.61, against 0.38 for Spanish–Portuguese.[2] Since transfer is "substantially less effective . . . for distant languages" (Lauscher et al., 2020) and typological distance generally guides transfer-source selection (Lin et al., 2019; Rice et al., 2025; Ng et al., 2026), transfer for this language pair should be weak.

What these measures miss is 400 years of intense language contact. Portuguese conjunctions have grammaticalized in Nheengatu (adversative *ma*, explanatory *porke*), each more fixed than in the source (Cândido et al., 2021), and Finbow and O'Neill (2022) document pronominal restructuring toward Portuguese, borrowed preverbal modals (*pudéri* <

[1] Also written *Yẽgatu*, or *Yẽgatú* with stress marked (Finbow and O'Neill, 2022), following the orthography the Rio Negro communities unified with FOIRN/UFAM in 2016; the language is also known as *Língua Geral Amazônica*. We use the spelling that is standard in the computational literature (Alencar, 2023).

[2] Computed with the `urielplus` package (database integration plus imputation). We report URIEL+ rather than URIEL (Littell et al., 2017), which has no observed syntactic features for Nheengatu and returns the same distance for every language with an empty syntax vector. Sparsity is systemic: 87% of URIEL+'s typological matrix is imputed (Ng et al., 2025).

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 40–49
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_006

poder 'can'), and a rapidly expanding GO-future under pressure from Brazilian Portuguese *ir* + infinitive. Alves (2026) finds the same auxiliary constructions in annotated corpus data. In our corpus, hundreds of loanword types—roughly 6% of the vocabulary—include legal borrowings such as *estado* → *istadu* 'state' (Cruz, 2011; Avila, 2021). These convergences motivate our central claim:

> Contact-induced structural convergence acts as a *de facto* typological bridge that may enable cross-lingual transfer—but not uniformly. For Portuguese and Nheengatu, transfer should be strongest at the linguistic levels where such convergence has been independently documented. Computationally, this predicts a *progressive* pattern: methods operating at shallower levels should fail, whereas methods exploiting supralexical representations—those spanning more than the word—should succeed.

We make three contributions: (i) a *contact-informed hierarchy* linking independently documented levels of contact to the computational levels that probe them; (ii) the first evaluation of cross-lingual sentence representations for Nheengatu; and (iii) evidence from a controlled Portuguese–Spanish comparison for a contact-related signal beyond typological similarity. Five experiments follow: four probing successive contact levels, one a source-language control.

2 Related Work

Cross-lingual transfer. Multilingual transformers such as mBERT (Pires et al., 2019) and XLM-RoBERTa (Conneau et al., 2020) enable cross-lingual transfer in low-resource settings (Lauscher et al., 2020). Transfer is strongest when the low-resource target has a structurally similar high-resource partner: Senel et al. (2024) show that using a high-resource cousin language substantially boosts NLU across five low-resource Turkic languages. For Indigenous languages of the Americas, Ebrahimi et al. (2022) show that zero-shot transfer is weak but fine-tuning on small amounts of target text substantially improves results. Adelani et al. (2024) demonstrate that Brazilian Portuguese outperforms English as a transfer source for low-resource Brazilian languages, motivating Portuguese as a plausible candidate. When direct parallel data is scarce, synthetic training data produced by machine translation ("translate-train") is a standard control (Conneau et al., 2018b), though it can introduce translationese artifacts (Artetxe et al., 2020).

Contact linguistics and transfer. Contact linguistics provides the theoretical basis for our hypothesis. Thomason and Kaufman (1988, 35) hold that contact outcomes are primarily determined by the sociolinguistic history of speakers rather than by the structure of their languages—so contact can reshape grammar in ways typological distance metrics cannot anticipate. Law (2013) documents contact-induced borrowing of bound morphemes in Lowland Mayan, producing structural convergence unlikely between unrelated languages. For Portuguese–Nheengatu specifically, such convergence at the clause-combining and constructional levels is well documented (Cândido et al., 2021; Finbow and O'Neill, 2022). Alves (2026) measures that convergence: by word order entropy over 24 syntactic features, Nheengatu patterns with Portuguese rather than Tupinambá in subject, object, and root position; in auxiliaries and copulas, which Tupinambá lacks; and in subordinator placement. He quantifies where it is detectable; we test whether it aids transfer. More broadly, cross-lingual transfer has been shown to succeed where source and target languages share deep structural, phonological, and morphological properties (Bankula and Bankula, 2025; Haddad-Najjar and Abu-Rabia, 2024; Barking et al., 2025), and to operate differentially across linguistic levels (Cosma, 2026; Blevins et al., 2022).

Nheengatu NLP. Prior computational work on Nheengatu is limited and largely rule-based or resource-building. de Alencar (2021) implements a Nheengatu grammar fragment in Grammatical Framework with Portuguese and English modules—structural overlap made computationally exploitable, but through hand-crafted rules rather than learned representations. The same line of work yields the Yauti analyzer (Alencar, 2023) and the first Universal Dependencies (UD) treebank for the language (de Alencar, 2024; Alexandre and Alencar, 2026). Alexandre et al. (2021) compile a tagged Nheengatu corpus. Neural work is scarcer: Pinhanez et al. (2024) fine-tune translators from Nheengatu into English, and prior work on Indigenous-language generation has used Spanish

as a pivot with NLLB-200 (Fernandes, 2026). All of this takes the transfer source as given; we ask *which* source transfers, and why. Ours is the first evaluation of cross-lingual sentence representations for Nheengatu, and the first source comparison that holds typological distance constant—Adelani et al. (2024) contrast English with Portuguese, which differ in typology as well as contact history.

3 Corpus

The corpus comprises 4,997 Portuguese–Nheengatu sentence pairs spanning legal, narrative, historical, and pedagogical registers, drawn from four sources: the Brazilian Constitution translated into Nheengatu (2,020 pairs); historical fragments from the Tycho Brahe corpus (Veronesi, 2026) (597 pairs); and Navarro's pedagogical grammar (Navarro, 2016) (2,139 pairs) and Casasnovas's grammatical notes (Casasnovas, 2006) (241 pairs), both obtained via the CompLin/nheengatu repository (Alencar, 2023). Quality filtering removed 31 misaligned pairs via length-ratio outlier detection ($\pm 3\sigma$) followed by manual verification. All resources—code, datasets, trained models, and the machine-translated Spanish control set—are publicly available.[3] The *hapax rate* (the proportion of word types occurring exactly once) is reported in Table 1. Nheengatu's hapax rate of 60.6% greatly exceeds Portuguese's 45.5%, reflecting its productive agglutinative morphology and the resulting lexical sparsity. All four sources are written and edited; no spoken or informal Nheengatu is represented, so the contact features we measure are those that have entered the written standard.

Metric	Portuguese	Nheengatu
Total word tokens	66,990	59,334
Unique word types	6,653	8,149
Hapax rate (freq. = 1)	45.5%	60.6%
Type/token ratio	9.9%	13.7%

Table 1: Vocabulary statistics for the cleaned corpus (4,997 pairs).

Beyond aggregate statistics, the corpus directly attests the contact phenomena our hypothesis concerns (Table 2): phonologically adapted loanwords, the borrowed adversative *ma* (61 tokens), preverbal modal auxiliaries of the *pudéri* family (408 tokens), the *sú* + Verb construction (37 pairs), and the relativizer *waá* (476 tokens).

Portuguese	Nheengatu
LEXICAL	
polícia	***pulisia***
'police'	
CLAUSE-COMBINING	
... ***mas*** *ele não foi*	*...* ***ma*** *aé nti usú*
'...but he did not go'	
CONSTRUCTIONAL: MODAL	
...não ***poderá*** *ser superior...*	*...ti* ***upuderi*** *usasá...*
'...may not exceed...'	
CONSTRUCTIONAL: RELATIVIZER	
...cidade grande ***que*** *está...*	*...tauauasu uiku* ***uaá****...*
'...a big city that is...'	

Table 2: Contact phenomena attested in the corpus, with English glosses. Corresponding items in **bold**. The borrowed modal *upuderi* (< *poder* 'can') precedes the verb, where native Nheengatu auxiliaries follow it. The relativizer *uaá*/*waá* (< Tupi *-ba'e*) occurs with any predicate without nominalizing it, like Portuguese *que*, whereas Tupi–Guaraní subordination nominalizes the verb (Finbow and O'Neill, 2022).

4 Experiments

Task and metrics. Our reference evaluation is *cross-lingual sentence retrieval*: given a Portuguese query sentence, rank all 1,000 Nheengatu sentences in the test set and retrieve the correct translation. *Precision@k* (P@k) measures whether the correct translation appears in the top k results; *Mean Reciprocal Rank* (MRR) is the mean reciprocal rank of the first correct retrieval. Retrieval measures whether representations from two languages share a semantic space; it is standard in XTREME via BUCC and Tatoeba (Hu et al., 2020; Artetxe and Schwenk, 2019), and needs no target-language annotation beyond parallel sentences—relevant here, since no NLI or QA benchmark exists for Nheengatu, which AmericasNLI does not cover (Ebrahimi et al., 2022). Experiments 1–4 use the same 80/20 train/test split (3,997 training pairs; 1,000 test pairs). Experiments are grouped by the linguistic level each method probes (Figure 1), mirroring the documented contact-informed hierarchy. *Constructional* names the documented phenomenon—schematic patterns with open slots, such as *sú* + [Verb]; *supra-lexical* names what we measure, since retrieval sees whole sentences and cannot isolate constructions from other signal above the word. (In Construction Grammar words

[3] Code: https://github.com/rmaacario/nheengatu-crosslingual Data and checkpoints: https://huggingface.co/rmacario/nhengatu-xlmr

and morphemes are constructions too, so the morphological level covers bound affixation: *-itá*, *ta-*.)

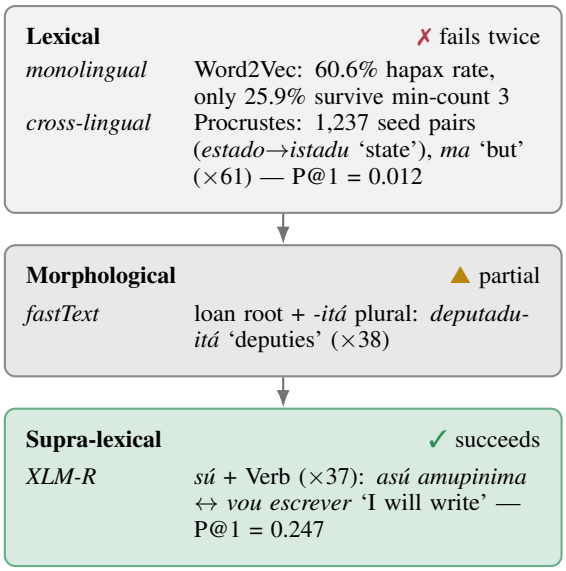


Figure 1: Contact-informed hierarchy: each linguistic level, the methods probing it, and examples attested in our corpus ($\times n$ = corpus token/pair counts). The lexical level is probed twice—within Nheengatu and across the language pair—and produces only a weak signal (P@1 = 0.012, some 110× the 1/9,048 random baseline), insufficient for retrieval; morphological patterns are captured by fastText but do not scale to sentence retrieval. Strong transfer sufficient for retrieval emerges only at the supra-lexical level.

Experiment 1: Word2Vec. *Goal*: Establish whether word-level semantic spaces are learnable from this corpus. Skip-gram models (100d, window sizes 3 and 10, min_count=3, 10 epochs) were trained separately on each language. At min_count=3, Word2Vec retains only 2,108 Nheengatu types (25.9% of 8,149 unique types); the discarded 74.1% include morphologically productive forms occurring only once or twice. Retained types do not organize semantically: IsoScore 0.023 (Portuguese 0.012), against 0.955 for an isotropic Gaussian of the same dimensions (Rudman et al., 2022). The corpus is therefore too sparse for reliable word-level semantic learning.

Experiment 2: fastText. *Goal*: Test whether character n-grams capture the morphological level of contact. Skip-gram models with character n-grams (minn=3, maxn=6, min_count=1, 20 epochs) (Bojanowski et al., 2017) were trained on each language. The model encodes Nheengatu affixes as consistent directions: for five of eight morphemes documented by Cruz (2011) and Avila (2021), the derived-minus-root offset is far more coherent within a morpheme than across morphemes (plural *itá*, $d = 2.10$; 3pl *ta-*, $d = 2.37$), and borrowed roots participate: *istadu* → *istaduita* 'states', *prupusta* → *prupustaita* 'proposals'. fastText captures morphological patterns consistent with documented contact phenomena, but mean-pooled fastText sentence vectors achieve P@1 = 0.000, MRR = 0.009: morphological signal does not scale to sentence retrieval.

Experiment 3: Orthogonal Procrustes alignment. *Goal*: Test whether lexical-semantic alignment is achievable given contact-induced lexical overlap. We learn a linear mapping from Portuguese fastText vectors (Common Crawl, 300d) to Nheengatu vectors (trained from scratch, 300d), seeding it with 1,237 automatically extracted Portuguese–Nheengatu candidate correspondence pairs (including many non-loan look-alikes)[4] (Artetxe et al., 2018). The alignment fails: on 248 held-out seed pairs P@1 = 0.012 and P@5 = 0.028, and even fitted and scored on all 1,237 pairs it reaches only P@1 = 0.039—it cannot recover its own training correspondences.[5] These rates are some 110× the 1/9,048 random baseline (fastText's case-sensitive vocabulary), so a weak lexical-semantic signal is present, as the documented borrowing predicts, but far too sparse for retrieval. The from-scratch Nheengatu space clusters by surface form rather than meaning (neighbors dominated by morphological variants), and linear alignment needs semantically organized spaces on both sides.

Experiment 4: XLM-R fine-tuning. *Goal*: Test whether sentence-level representations succeed where word-level methods fail. Sentence encoders can exploit supra-lexical information—sentence embeddings have been shown to encode linguistic properties such as tense and clause coordination (Conneau et al., 2018a), as well as argument frames (Kann et al., 2019)—of the kind restructured through contact (Table 2) (Cândido et al., 2021; Finbow and O'Neill, 2022), without re-

[4]Generated by applying Nheengatu phonological adaptation rules (Cruz, 2011) to Portuguese vocabulary and keeping pairs whose adapted form occurs in our Nheengatu corpus; this high-recall seed is deliberately broader than the validated loanword inventory used in Experiment 5.

[5]We fit on 989 pairs and score 248 held out, so the figures measure generalization, not memorization. Matched capacity does not help: retraining Portuguese from scratch on the same corpus gives P@1 = 0.02 against 0.01 for the pretrained source, under leave-one-out on 100 shared seed pairs.

quiring semantically organized word spaces. We fine-tune XLM-RoBERTa **base** (270M parameters) (Conneau et al., 2020) using Multiple Negatives Ranking Loss (batch size 16, lr 2×10^{-5}, 3 epochs, max length 128, in-batch negatives yielding 15 negatives per anchor, no early stopping), which maps parallel sentence pairs to nearby points in a shared embedding space while separating non-parallel pairs (Figure 2). As baselines we evaluate BM25 over character 3-grams and two multilingual sentence encoders zero-shot—LaBSE (Feng et al., 2022) and paraphrase-multilingual-MiniLM-L12-v2 (Reimers and Gurevych, 2020)—neither of which covers Nheengatu. Table 3 reports the seed-42 run; single runs of this recipe vary, with five seeds under the Experiment 5 harness giving P@1 $= 0.281 \pm 0.012$ on the same test split. We report seed 42 throughout, since the domain and qualitative analyses below derive from that model.

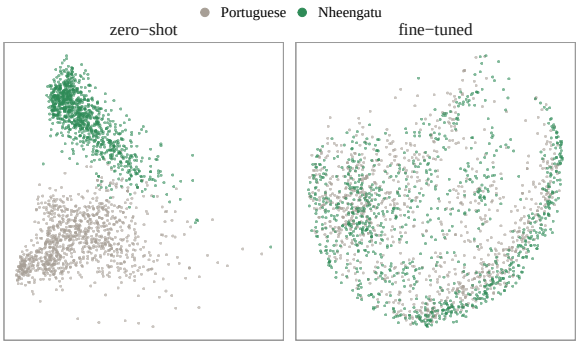


Figure 2: PCA of test-set sentence embeddings (n = 1,000 pairs). Before fine-tuning, Portuguese and Nheengatu occupy disjoint regions of the space; after fine-tuning on the training set, they interleave.

System	P@1	P@5	P@10	MRR
Random	0.001	0.005	0.010	0.007
fastText	0.000	0.006	0.013	0.009
XLM-R (z.s.)	0.011	0.042	0.062	0.034
MiniLM (z.s.)	0.036	0.085	0.116	0.066
BM25 (3-gram)	0.104	0.218	0.278	0.165
LaBSE (z.s.)	0.145	0.277	0.346	0.215
XLM-R (ft.)	**0.247**	**0.504**	**0.621**	**0.371**

Table 3: Sentence retrieval results (test set n = 1,000; reference run, seed 42).

Fine-tuned XLM-R outperforms every baseline (Figure 3): more than double BM25 over character 3-grams (0.104) and 1.7× LaBSE (0.145), despite LaBSE's larger size (471M vs. 270M parameters) and broad multilingual pretraining. Zero-shot XLM-R (0.011) is weaker still, reflecting the

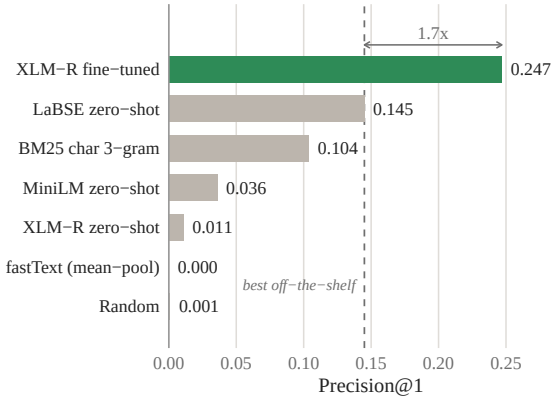


Figure 3: Precision@1 for all retrieval systems (seed 42). Off-the-shelf systems—lexical and neural alike—top out at LaBSE's 0.145 (dashed); fine-tuning XLM-R on the training set clears that ceiling by 1.7×, consistent with the hypothesis that contact history may provide a transfer advantage.

known degeneracy of MLM-only sentence embeddings rather than anything specific to this pair. The advantage is not lexical: on the 799 test queries whose Nheengatu target contains no loanword from our inventory, fine-tuned XLM-R still reaches P@1 = 0.233—close to its 0.247 over the full test set—while BM25 falls to 0.080. And since LaBSE sits only 4.1 points above the character-level baseline where fine-tuned XLM-R sits 14.3, what in-domain fine-tuning adds is not available either from surface matching or from general multilingual pretraining—which is what we would expect of a supra-lexical signal.

Experiment 5: Source-language control. *Goal*: Help isolate contact from generic multilingual transfer. We compare Portuguese against Spanish, which is essentially as distant from Nheengatu as Portuguese is (URIEL+ genetic 1.0 for both; syntactic 0.66 for Spanish against 0.61 for Portuguese, the latter on 183 Grambank features observed for Nheengatu where URIEL has none) and no less represented in XLM-R's pretraining data (9,374M vs. 8,405M tokens in CC-100; Conneau et al. 2020, Table 6), so pretraining coverage cannot explain a gap, and the two sources are themselves far closer to each other (0.38) than either is to Nheengatu. Following the translate-train paradigm (Conneau et al., 2018b), we machine-translate the Portuguese side of every pair to Spanish with NLLB-200-distilled-600M (Team et al., 2022) and re-run the pipeline,

varying only the source language.[6] Portuguese wins on every metric and in all five seeds (Table 4); the P@1 gap of $+0.050 \pm 0.007$ is roughly seven times its standard deviation, and at the query level (reference run, seed 42) is significant under approximate randomization for P@1 ($p = 0.00063$) and paired bootstrap for MRR ($p < 10^{-5}$).[7] Lexical forms are substantially shared: for 56% of the 133 adapted loanwords the Spanish form is string-identical to the Portuguese (*dekretu* < *decreto*, unchanged in Spanish), and mean character similarity to the Nheengatu form is 0.746 for Portuguese against 0.705 for Spanish. Lexical overlap alone therefore does not account for the gap; and because Spanish is derived from the Portuguese source, any translationese (Artetxe et al., 2020) biases it toward Portuguese, making the gap conservative.

Metric	PT (contact)	ES (control)	Gap
P@1	**0.281** $\pm$ 0.012	0.232 $\pm$ 0.005	+0.050
P@5	**0.545** $\pm$ 0.034	0.453 $\pm$ 0.009	+0.091
P@10	**0.673** $\pm$ 0.030	0.568 $\pm$ 0.006	+0.104
MRR	**0.406** $\pm$ 0.018	0.340 $\pm$ 0.006	+0.066

Table 4: Source-language control (mean $\pm$ std over 5 seeds; XLM-R fine-tuning, shared fixed test split, $n = 1{,}000$). Spanish is machine-translated from the Portuguese source, so the gap is a conservative estimate of any contact-related advantage.

Domain analysis. Performance varies by register (Table 5). The Tycho historical corpus achieves the highest P@1 (0.387). Sentence length does not explain this—the Constitution's sentences are longer (19.0 vs. 12.3 words) yet score lower—and the two are comparably distinctive against the 1,000 candidates. What separates Tycho is an unusually tight correspondence between the two sides' token counts ($r = 0.96$ vs. 0.86 for the Constitution): a property of its translation regime, which favors retrieval and bounds how far this figure generalizes. Casasnovas performs lowest (P@1 = 0.057); we report it for completeness ($n = 35$; not interpretable).

[6]Experiments 4 and 5 share the fixed 80/20 split. Table 3 is a single run from our initial implementation, Table 4 five-seed means from the control harness; absolute scores differ between the two, but the paired PT–ES comparison is unaffected, since both conditions share each run's randomness.

[7]A paired t-test over the five seeds also favors Portuguese ($t = 15.7$, $p < 0.001$), but at $n = 5$ Wilcoxon and the sign test both bottom out at $p = 0.0625$, their minimum achievable value, so we test at the query level instead.

Domain	P@1	Test pairs
Tycho (historical)	**0.387**	124
Constitution (legal)	0.252	401
Navarro (pedagogical)	0.218	440
Casasnovas[†]	0.057	35
Weighted average	0.247	1,000

Table 5: XLM-R fine-tuned performance by domain. [†]$n = 35$; not statistically reliable.

Qualitative analysis: the existential construction. Table 6 groups the ten test sentences containing Portuguese *há* by their Nheengatu rendering. The nine *aikué* existentials reach a median rank of 4, and confusions can stay within the construction: for one query, the top-ranked candidate is another *aikué* existential. The lexical exception *Akuera*, by contrast, falls to rank 137. This contrast is consistent with the model representing the existential construction rather than simply matching lexical items, although the sample is too small for firm conclusions.

'há' rendered as	Ranks of correct translation
Existential *aikué/Aikwé* ($n = 9$)	1, 1, 2, 2, 4, 11, 21, 52, 70 (median 4)
e.g.	*Aikué igara igarapé upé* 'There is a canoe in the creek'
Lexical exception *Akuera* 'long ago' ($n = 1$)	137
e.g.	*Akuera iaueuera aé* 'He has long been used to it'

Table 6: Retrieval ranks (among 1,000 candidates) for the ten test sentences containing Portuguese *há*, grouped by how Nheengatu renders it. The model seems to track the *aikué* construction (median rank 4); the sentence where *há* is not existential falls to rank 137.

5 Discussion

The progressive pattern matches the hypothesis. Word2Vec and Procrustes fail because a 60.6% hapax rate leaves the from-scratch Nheengatu space too anisotropic to organize semantically—the fitted map cannot retrieve even its own training pairs—and fastText's morphological signal does not scale to sentence retrieval. Sparsity is part of that explanation, but not of the source-language gap: the Nheengatu side is the same text in both conditions, and on the source side Spanish is if anything slightly less sparse than Portuguese (43.4%

vs. 45.5% hapax over the same 4,997 pairs), so sparsity would favor the control—yet Portuguese wins. Fine-tuning succeeds at the supra-lexical level, aligning sentence representations without needing organized word spaces. The Portuguese advantage (Table 4) points to long-term contact rather than generic Romance typology or model capacity, and the domain results agree—register-varied sentences (Tycho, Constitution) outperform the short, formulaic ones (Navarro), while BM25 and LaBSE bound the lexical-overlap contribution. P@1 = 0.247 is modest, but retrieval among 1,000 candidates is far harder than bilingual classification. For practical use P@10 = 0.621 matters more: the correct translation is in the top ten for nearly two-thirds of queries, enough for translation-memory lookup or parallel-text alignment, though not for unattended retrieval.

6 Conclusion

Our results highlight a factor standard typological resources do not encode: *language contact*—a gap URIEL+'s own maintainers note (Shipton et al., 2026). Two findings support this. First, transfer from a Romance source to Nheengatu succeeds well above lexical and zero-shot baselines despite maximal genetic distance, so typological distance alone under-predicts transfer. Second, of two comparably distant sources, the one with a documented contact history—Portuguese—transfers significantly better, across all metrics and five seeds. The progressive pattern across linguistic levels further suggests that what multilingual sentence encoders exploit is supra-lexical, consistent with contact-induced convergence though not localized to it by our design. We do not claim contact is necessary for transfer, nor that this generalizes beyond this pair and task, but it is a measurable factor in transfer-source quality, worth encoding alongside genealogy and typology.

Future work. Future work will (i) expand the source-language comparison with non-contact and genealogically related (Tupi–Guaraní) controls, complementing Alves (2026); (ii) evaluate additional tasks, including morphological tagging and dependency parsing on the UD_Nheengatu-CompLin treebank (de Alencar, 2024) against the baselines of Alexandre and Alencar (2026), and a Nheengatu natural language inference set following AmericasNLI (Ebrahimi et al., 2022); and (iii) collect and compare spoken Nheengatu against the written registers studied here.

Ethics Statement

This research uses publicly available written texts and is motivated by supporting Nheengatu language technology for revitalization. Deployment of any models should be conducted in consultation with Nheengatu-speaking communities.

References

David Ifeoluwa Adelani, A. Seza Doğruöz, André Coneglian, and Atul Kr. Ojha. 2024. Comparing LLM prompting with cross-lingual transfer performance on indigenous and low-resource Brazilian languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 34–41, Mexico City, Mexico. Association for Computational Linguistics.

Leonel Figueiredo de Alencar. 2021. Uma gramática computacional de um fragmento do nheengatu / a computational grammar for a fragment of nheengatu. *Revista de Estudos da Linguagem*, 29(3):1717–1777.

Leonel Figueiredo de Alencar. 2024. A Universal Dependencies treebank for nheengatu. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 2*, pages 37–54, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Leonel Figueiredo de Alencar. 2023. Yauti: A tool for morphosyntactic analysis of nheengatu within the Universal Dependencies framework. In *Proceedings of the 14th Brazilian Symposium in Information and Human Language Technology*, pages 144–154, Belo Horizonte, Brazil. Association for Computational Linguistics.

Dominick Alexandre, Juliana Gurgel, and Leonel Araripe. 2021. Compilação de um corpus etiquetado da língua geral amazônica. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 427–431, Porto Alegre, RS, Brasil. SBC.

Dominick Maia Alexandre and Leonel Figueiredo de Alencar. 2026. Parsing nheengatu: Performance gains for a Brazilian indigenous Universal Dependencies treebank. In *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 2*, pages 210–219, Salvador, Brazil. Association for Computational Linguistics.

Diego Alves. 2026. Measuring the effects of language contact using word order entropy: the case of nheengatu. *Linguistics Vanguard*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Marcel Twardowsky Avila. 2021. *Proposta de dicionário nheengatu-português*. Tese (doutorado), Universidade de São Paulo, São Paulo. Departamento de Letras Modernas, Área de concentração: Estudos da Tradução. Orientador: Eduardo de Almeida Navarro. https://teses.usp.br/teses/disponiveis/8/8160/tde-10012022-201925/pt-br.html.

Ajitesh Bankula and Praney Bankula. 2025. Cross-linguistic transfer in multilingual nlp: The role of language families and morphology.

M. Barking, Maria Mos, and Ad Backus. 2025. Investigating language transfer from a usage-based perspective. *International Journal of Bilingualism*, 29(2):415–438.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2022. Analyzing the mono- and cross-lingual pretraining dynamics of multilingual language models. *arXiv preprint*. ArXiv:2205.11758 [cs.CL].

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

A. Casasnovas. 2006. *Noções de língua geral ou Nheengatú: gramática, lendas e vocabulário*. Faculdade Salesiana Dom Bosco.

Maria Sílvia Cintra. 2018. Nheengatu language and its role in the tactics of construction of indigenous identity. *Advances in Anthropology*, 8(4):289–301.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018a. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018b. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Adrian Cosma. 2026. An in-vitro study on cross-lingual generalization in language models. *arXiv preprint*. ArXiv:2605.26683 [cs.CL].

Aline da Cruz. 2011. *Fonologia e Gramática do Nheengatú: A língua geral falada pelos povos Baré, Warekena e Baniwa*. LOT Dissertation Series. LOT, Utrecht. https://www.lotpublications.nl/Documents/280_fulltext.pdf.

Gláucia Vieira Cândido, Aline da Cruz, and Giovana Alves de Oliveira. 2021. Empréstimos de conjunções e subordinadores do português em nheengatu. *Estudos Linguísticos (São Paulo. 1978)*, 50(1):91–107.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Rafael Fernandes. 2026. USP at AmericasNLP 2026 shared task: Culturally-aware image captioning for indigenous languages via vision-language models and fine-tuned neural machine translation. In *Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 264–271, San Diego, California, USA. Association for Computational Linguistics.

Tom Finbow and Paul O'Neill. 2022. Koinéization and language contact: The social causes of morphological

change in and with portuguese. In Adam Ledgeway, John Charles Smith, and Nigel Vincent, editors, *Periphrasis and Inflexion in Diachrony: A View from Romance*. Oxford University Press.

g1 Ceará. 2025. Vídeo: Conheça o nheengatu, uma das línguas indígenas mais faladas no brasil. https://g1.globo.com/ce/ceara/noticia/2025/10/24/video-conheca-o-nheengatu-uma-das-linguas-indigenas-mais-faladas-no-brasil.ghtml. Accessed: 2026-08-01.

Laila Haddad-Najjar and Salim Abu-Rabia. 2024. Exploring cross-linguistic transfer of phonology, morphology, syntax, and orthography from Arabic (L1) to English (FL): An intervention study. *International Journal of Bilingualism*, 29(6):1657–1680.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization.

IBGE. 2024. *Censo Demográfico 2022: Indígenas: principais características das pessoas e dos domicílios, por situação urbana ou rural do domicílio: resultados do universo*. Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro.

Katharina Kann, Alex Warstadt, Adina Williams, and Samuel R. Bowman. 2019. Verb argument structure alternations in word and sentence embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL)*, pages 287–297, New York City, New York. ArXiv:1811.10773 [cs.CL].

Aditya Khan, Mason Shipton, David Anugraha, Kaiyao Duan, Phuong H. Hoang, Eric Khiu, A. Seza Doğruöz, and En-Shiun Annie Lee. 2025. URIEL+: Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6937–6952, Abu Dhabi, UAE. Association for Computational Linguistics.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Danny Law. 2013. Inherited similarity and contact-induced change in mayan languages. *Journal of Language Contact*, 6(2):271 – 299.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Denny Moore. 2014. Historical development of nheengatu (língua geral amazônica). In Salikoko S. Mufwene, editor, *Iberian Imperialism and Language Evolution in Latin America*. University of Chicago Press.

Eduardo de Almeida Navarro. 2016. *Curso de Língua Geral (Nheengatu ou Tupi Moderno): A língua das origens da civilização amazônica*, 2 edition. Centro Angel Rama, Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo. 2.ª edição corrigida e aperfeiçoada.

York Hay Ng, Phuong Hanh Hoang, and En-Shiun Annie Lee. 2025. Less is more: The effectiveness of compact typological language representations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25805–25816, Suzhou, China. Association for Computational Linguistics.

York Hay Ng, Aditya Khan, Xiang Lu, Matteo Salloum, Michael Zhou, Phuong Hanh Hoang, A. Seza Doğruöz, and En-Shiun Annie Lee. 2026. Modality matching matters: Calibrating language distances for cross-lingual transfer in URIEL+. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 110–130, Rabat, Morocco. Association for Computational Linguistics.

Claudio Pinhanez, Paulo Cavalin, and Julio Nogima. 2024. Human evaluation of the usefulness of fine-tuned English translators for the Guarani mbya and nheengatu indigenous languages. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 2*, pages 32–36, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Enora Rice, Ali Marashian, Hannah Haynie, Katharina von der Wense, and Alexis Palmer. 2025. Untangling the influence of typology, data, and model architecture on ranking transfer languages for cross-lingual POS tagging. In *Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025)*, pages 22–31, Albuquerque, New Mexico. Association for Computational Linguistics.

William Rudman, Nate Gillman, Taylor Rayne, and Carsten Eickhoff. 2022. IsoScore: Measuring the uniformity of embedding space utilization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3325–3339, Dublin, Ireland. Association for Computational Linguistics.

Lütfi Kerem Senel, Benedikt Ebing, Konul Baghirova, Hinrich Schuetze, and Goran Glavaš. 2024. Kardeş-NLU: Transfer to low-resource languages with the help of a high-resource cousin – a benchmark and evaluation for Turkic languages. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1672–1688, St. Julian's, Malta. Association for Computational Linguistics.

Mason Shipton, York Hay Ng, Aditya Khan, Phuong H. Hoang, Xiang Lu, A. Seza Doğruöz, and Annie En-Shiun Lee. 2026. Simple additions, substantial gains: Expanding scripts, languages, and lineage coverage in URIEL+. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pages 11045–11059, Palma de Mallorca, Spain. ELRA Language Resource Association.

Sâmela Ramos da Silva Meirelles. 2020. *A reinscrição de uma língua destituída: o Nheengatu no Baixo Tapajós*. Tese (doutorado em linguística), Universidade Estadual de Campinas, Campinas, SP.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Sarah Grey Thomason and Terrence Kaufman. 1988. *Language Contact, Creolization, and Genetic Linguistics*. Anthropology: Linguistics. University of California Press, Berkeley.

Luiz Henrique Lima Veronesi. 2026. *A Plataforma Tycho Brahe: um sistema para corpora sintaticamente anotados*. Tese (doutorado em linguística), Universidade Estadual de Campinas, Instituto de Estudos da Linguagem, Campinas. 211 f.

Retrieval-Augmented Machine Translation for Bohairic Coptic: A Pilot Case Study

So Miyagawa
Institute of Humanities and Social Sciences, University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8571, Japan
miyagawa.so.kb@u.tsukuba.ac.jp

Abstract

Coptic is extremely low-resource in natural language processing despite its importance for philology, historical linguistics, and heritage-language learning. This paper reports a pilot case study of Bohairic Coptic–English translation using one passage from the *Vita Sinuthii*, one published English reference, and one retained output from each of seven deployed systems. THOTH AI, a retrieval-augmented system built with Coptological lexical, corpus, and grammatical resources, obtained 40.63 BLEU, compared with 31.40 for a separately accessed Gemini 3.0 Pro baseline. Results from chrF, ROUGE, TER, and METEOR are also reported, together with an illustrative expert reading of lexical errors and unsupported generation. These figures are descriptive, not a general benchmark or a causal retrieval ablation: the systems were accessed through different deployed interfaces, exact retrieval traces were not retained, and the evaluation contains only one passage. Three outputs contained clear passage-scale fabricated material. The study documents this bounded case and the evidence needed for a larger, reproducible evaluation.

1 Introduction

Egyptian has a documented history extending over more than four millennia, and Coptic is conventionally treated as its latest historical phase (Loprieno, 1995; Grossman and Richter, 2014). Coptic is written in an alphabet derived principally from Greek and supplemented by Demotic-derived characters. It is attested in several dialects rather than as a single uniform variety. Bohairic is the principal liturgical dialect of the Coptic Orthodox Church today, whereas Sahidic predominates in many modern grammars and digitally annotated corpora (Layton, 2000; Till, 1961). This imbalance has direct computational consequences: a model trained or retrieved primarily on Sahidic data cannot be assumed to have adequate Bohairic lexical, orthographic, or grammatical coverage.

After the Arab conquest of Egypt, the domains of written and spoken Coptic gradually contracted as Arabic became dominant, although the chronology and regional course of this shift were not uniform (Rubenson, 1996). Coptic nevertheless remains indispensable for Egyptian historical linguistics, biblical and patristic studies, and the textual traditions of Egyptian Christianity. It also continues to be taught and used liturgically. Computational tools for Coptic thus serve several overlapping constituencies: philologists working with manuscripts and editions, linguists examining the Egyptian–Coptic continuum, and learners seeking access to Bohairic religious texts.

For natural language processing (NLP), however, Coptic is extremely low-resource. Available texts are modest in comparison with major modern languages, dialects and orthographies vary, and reliable parallel data are limited. Hagiographic translation adds culturally specific vocabulary, embedded discourse, and formulaic narrative patterns that can encourage a generative model to produce fluent but source-unsupported material. Recent work has begun to apply neural machine translation and large language models (LLMs) to Coptic, but these constraints remain substantial (Enis and Megalaa, 2024; Wannaz and Miyagawa, 2024).

Retrieval-augmented generation (RAG) offers a practical way to place specialist lexical, corpus, and grammatical evidence in a general-purpose model's context at inference time. Yet an apparent gain from a deployed RAG application cannot automatically be ascribed to retrieval: differences in system prompts, provider interfaces, decoding defaults, and model revisions can all affect an output. The present study therefore asks a deliberately narrow observational question: what behavior is found when one deployed Coptic RAG system and

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 50–58
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_007

six general-purpose interfaces translate the same Bohairic passage?

The paper makes three bounded contributions. First, it documents the passage, prompt, retained-output procedure, deployed retrieval settings, interface labels, and automatic scores for a pilot case. Second, it compares THOTH AI descriptively with six separately accessed systems while making the interface confound and missing retrieval trace explicit. Third, it supplies an operational error taxonomy and an illustrative expert analysis of every retained translation, including three passage-scale cases of unsupported narrative. The study does not claim a representative benchmark, a same-code RAG ablation, a model-wide hallucination rate, or a validated annotation dataset.

2 Related Work

2.1 Digital Resources for Coptic

Coptic digital infrastructure includes Coptic SCRIPTORIUM, which provides linguistically annotated corpora and an NLP pipeline (Schroeder and Zeldes, 2016; Zeldes and Schroeder, 2016); the Coptic Dictionary Online (Feder et al., 2018); and the Comprehensive Coptic Lexicon (CCL), which integrates Egyptian-Coptic lexical material with Greek loanword data (Burns et al., 2020). The Corpus dei Manoscritti Copti Letterari represents an earlier major application of digital methods to Coptic manuscript research (Orlandi, 1990). Community resources such as Tasbeha.org additionally make liturgical texts searchable for church users and learners (Tasbeha.org, 2026).

These resources are complementary rather than interchangeable. A manuscript catalogue, an annotated corpus, a lexicon, a reference grammar, and a liturgical website differ in structure, evidential status, and intended task. Their dialect balance also matters: substantial Sahidic resources do not automatically resolve Bohairic forms. This motivates a retrieval design that can present dialect-sensitive lexical evidence together with grammatical and textual context. It does not, by itself, demonstrate that retrieval improves translation; that question requires a controlled comparison beyond the observational case reported here.

2.2 Coptic and Low-Resource Machine Translation

For Coptic machine translation, Enis and Megalaa (2024) investigate a specialized low-resource neural pipeline using normalization, romanization, backtranslation, and domain restrictions. Wannaz and Miyagawa (2024) evaluate LLMs on Coptic and Ancient Greek ostraca and identify source-grounding problems in Coptic outputs. More broadly, massively multilingual translation work shows both the promise and persistent limitations of transfer to low-resource languages (Costa-jussà et al., 2022). The present paper instead examines a Bohairic hagiographic passage and a retrieval-based deployed system.

The setting differs from ordinary modern-language machine translation in several respects. The source language is historical and dialectally diverse; spelling and editorial practice vary; parallel data are small; and a published literary translation may be more interpretive than a sentence-aligned benchmark reference. Consequently, surface-overlap metrics can penalize legitimate philological choices while still rewarding outputs that reproduce the wording of one translator. These properties motivate combining automatic scores with source-based expert reading.

2.3 Retrieval-Augmented Generation

RAG conditions generation on documents retrieved from an external collection (Lewis et al., 2020). THOTH AI previously applied this general approach to Middle Egyptian translation (Miyagawa, 2025). The system evaluated here is not a new retrieval algorithm: it is a domain composition of standard vector retrieval and a foundation model, implemented in Dify (Dify, n.d.), with Coptological resources selected and configured for translation. The contribution is the documented Bohairic case and failure analysis, not an architectural innovation over standard RAG.

3 System and Comparison Design

3.1 THOTH AI and Author Role

THOTH AI is developed and maintained by the author, who selected its sources, configured the deployed system, collected the comparison outputs, and performed the qualitative analysis. This dual role creates a potential conflict of interest and increases the importance of independently reproducible evaluation. No THOTH AI output was manually post-edited before scoring.

The evaluated deployment used the interface label *Gemini 3.0 Pro Preview* as its generator. This is the label preserved in the study record, not a claim

Setting	Recorded value
Chunk structure	General
Index method	High quality
Embedding model	`text-embedding-3-large`
Retrieval	Vector search
Reranker	None configured
Top-k	3
Score threshold	0

Table 1: Recorded Dify settings for the CCL knowledge base.

that the provider still exposes an identical archival snapshot. THOTH AI adds retrieved material to the model context before generation. Its principal Coptic sources were:

- a system-internal representation based on CCL v1.2, containing Egyptian-Coptic lexical material and Greek loanwords (Burns et al., 2020);
- Coptic SCRIPTORIUM corpus data and linguistic annotation (Schroeder and Zeldes, 2016; Zeldes and Schroeder, 2016); and
- digitized reference grammars, including Layton (2000), Reintges (2018), and the dialectal description of Till (1961).

This evidence mix is intended to address Bohairic lexical ambiguity and the uneven dialectal coverage of individual resources. It should not be understood as a homogeneous training corpus: the sources differ in format, dialect, and scholarly purpose, and are used as retrieval-time context rather than as newly fine-tuned model parameters.

3.2 Recorded Retrieval Configuration

The deployment was built in Dify. Table 1 reports the settings preserved for the CCL knowledge base. Dify's general chunk structure was used; parent–child and structured question–answer indexing were not used for this resource.

The exact retrieved chunks, source identifiers, ranks, similarity scores, and complete context passed for the evaluated query were not archived. The test passage and Bell's translation were not intentionally loaded as evaluation-specific records. Nevertheless, without the retrieval trace, incidental overlap with indexed materials cannot be categorically excluded. The present paper therefore makes neither a passage-level no-leakage claim nor a causal attribution to particular retrieved evidence. The settings in Table 1 describe the deployment, but they are not a reconstruction of the missing query trace.

3.3 Nature of the Comparison

THOTH AI and the nominally corresponding Gemini baseline were accessed through different deployed interfaces. System prompts, provider-side preprocessing, hidden defaults, and model revisions may therefore differ. This is a comparison of deployed systems, not a same-code RAG-on/RAG-off experiment. A causal ablation would require a shared API model identifier, identical decoding and system prompts, fixed seeds where available, saved retrieval traces, and repeated generations.

4 Experimental Design

4.1 Evaluation Passage

The source is a Bohairic passage from the *Vita Sinuthii*, following *Sinuthii archimandritae vita et opera omnia*, volume I, §88 (Leipoldt, 1906). The English reference is Bell's translation of §88 (pp. 67–68; Bell, 1983). Because the exact page or line range in the 1906 edition has not been rechecked against a physical or page-image copy, the paragraph identifier rather than an unverified page number is reported for the Coptic source.

The scored unit is one passage-level segment. By whitespace-delimited surface counting, the preserved source contains 53 Coptic surface tokens in four sentence-like units; the reference contains 102 English tokens in five sentence-like units. These are descriptive counts for the exact strings supplied to the metric implementation, not claims about morphologically segmented Coptic words. No sentence-level score distribution can therefore be inferred from the table.

The passage concerns the curse and punishment of Kesios/Gesios and contains eschatological vocabulary, embedded speech, and polysemous body-part terminology. It was selected as a philologically difficult case, not sampled to represent Coptic genres or Bohairic generally.

4.2 Prompt and Collection

Each interface received the same user-facing request:

> Please translate and analyze this passage: [Bohairic Coptic source passage].

No in-context examples or model-specific hints were added. For scoring, the English translation portion was manually separated from any surrounding analysis. One retained generation per model was scored; outputs were collected manually through the deployed interfaces in January 2026. The preserved record does not contain stable API model IDs, exact per-output timestamps, provider system prompts, temperature, top-p, maximum-token settings, random seeds, or a complete retry history. Provider defaults applied where controls were not exposed. These omissions prevent exact regeneration and are treated as limitations rather than reconstructed retrospectively.

4.3 Evaluated Systems

The labels below reproduce those exposed by the interfaces at collection time; commercial preview labels and aliases are not archival identifiers. The evaluated systems were THOTH AI (Gemini 3.0 Pro Preview), Gemini 3.0 Pro, Claude 4.5 Opus, Grok 4.1, DeepSeek V3.2, Kimi K2, and Qwen3 Max. Provider documentation is cited for the corresponding model families and interfaces (Google, n.d.; Anthropic, n.d.; xAI, n.d.; DeepSeek, n.d.; Moonshot AI, n.d.; Alibaba Cloud, n.d.). The manuscript does not imply that the providers expose reproducible snapshots under these labels. An eighth condition labelled "GPT 5.2 Pro" in an earlier draft was excluded from the camera-ready analysis because its original output string could not be recovered and its preserved scores could therefore not be independently audited. No output from a later GPT model was substituted.

4.4 Metrics

The retained translation strings were compared with the single Bell reference using BLEU (Papineni et al., 2002), chrF (Popović, 2015), ROUGE-1/2/L (Lin, 2004), TER (Snover et al., 2006), and METEOR (Banerjee and Lavie, 2005), through Hugging Face Evaluate.[1] In Evaluate, the `chrf` and `ter` modules are backed by sacreBLEU, whereas `bleu` and `sacrebleu` are separate modules. The scoring script, Evaluate version, and module revisions were not archived, so the exact BLEU module and configuration used for the retained values cannot now be verified. The score is therefore reported simply as BLEU, following the caution about implementation-sensitive BLEU reporting in Post (2018). BLEU emphasizes exact word n-gram precision; chrF is based on character n-grams; ROUGE-1/2 measure unigram and bigram overlap; ROUGE-L uses longest-common-subsequence overlap; TER estimates the edits needed to match the reference; and METEOR allows limited English stemming and synonym matching. Scores are presented on a 0–100 scale; lower TER is better, and TER can exceed 100 when the required edits outnumber reference tokens.

The metrics have different formal emphases, but they are all computed from the same retained hypothesis and the same single reference. They are therefore correlated descriptive views rather than independent confirmations of translation quality. No aggregate average is calculated, and the analysis does not treat a small numerical difference as a reliable system ranking.

5 Results

Table 2 gives the preserved scores. In this single case, THOTH AI obtained 40.63 BLEU and the separately accessed Gemini baseline obtained 31.40, a difference of 9.23 points. THOTH AI also has the highest chrF, ROUGE, and METEOR values and the lowest TER in the table. These observations describe the one retained set of strings; Section 3 explains why they cannot be interpreted as an isolated retrieval effect. Qwen3 Max's TER of 102.94 means that the edit count exceeded the reference length, not that the metric calculation failed.

The seven metrics move broadly together in this case, which is unsurprising because they are computed from the same strings and reference. For example, the differences between Gemini and Claude are much smaller than those between either of them and the three outputs with passage-scale unsupported narrative. Even this contrast is passage-specific: the table should not be used to infer population-level rankings or provider-wide quality. Stable comparison would require more texts, multiple retained generations, controlled inference conditions, and human assessment.

6 Illustrative Expert Error Analysis

6.1 Definition and Annotation Scope

An *unsupported addition* is defined here as a target-language proposition for which the annotator could identify no plausible corresponding source span and which was not a reasonable explicitation required for intelligible English. A *genre-driven fab-*

[1] https://huggingface.co/docs/evaluate

System	BLEU	chrF	R-1	R-2	R-L	TER↓	METEOR
THOTH AI (Gemini 3.0 Pro Preview)	40.63	63.84	76.85	48.76	69.95	44.12	70.35
Gemini 3.0 Pro	31.40	60.20	70.87	42.16	66.02	50.00	66.78
Claude 4.5 Opus	29.97	56.91	70.14	40.19	63.51	50.98	64.70
Grok 4.1	16.32	49.36	59.33	29.95	53.59	64.71	48.33
Kimi K2	6.64	40.88	48.87	15.53	32.58	90.20	36.66
DeepSeek V3.2	4.76	33.69	39.79	10.58	27.23	83.33	24.08
Qwen3 Max	3.38	30.55	37.44	3.69	19.18	102.94	22.20

Table 2: Descriptive scores for one passage-level segment and one retained output per system. R-1/2/L are ROUGE F1 scores.

rication is a sequence of unsupported propositions that together form a coherent hagiographic episode. These phenomena differ from (i) mistranslation, where an identifiable source item is rendered incorrectly; (ii) omission, where source content has no target counterpart; and (iii) interpretive variation, where the target remains plausibly licensed by the source but selects one philological or cultural interpretation.

The sole author, a researcher in Coptology and Egyptian linguistics, inspected all seven retained translation strings against the Coptic source and reference. There was no second annotator, adjudication, or inter-annotator-agreement calculation. Accordingly, this section is an illustrative expert reading, not a validated annotation study. At output level, three of seven retained translations—DeepSeek V3.2, Kimi K2, and Qwen3 Max—contained at least one clear passage-scale genre-driven fabrication. Category frequencies below the output level are not reported because no pre-registered span annotation or independent labels were available.

Table 3 records the codebook used for this reading. The categories are analytical distinctions, not mutually exclusive labels for an entire output: a single clause may contain, for example, both a lexical error and a structural divergence. The examples are drawn from this passage and are not proposed as an exhaustive taxonomy for Coptic machine translation.

6.2 Lexical and Interpretive Cases

***amenti* (the West/netherworld).** Crum glosses the Bohairic form *amenti* as "hades, literally western place" (Crum, 1939, 8b). The expression therefore denotes the western place or netherworld, understood as the realm of the dead. Bell renders it as "hell," placing the expression in a Christian English interpretive frame. THOTH AI, Gemini 3.0 Pro, and Claude 4.5 Opus retained "Amenti" or "Amente," sometimes with an explanatory gloss, whereas Grok 4.1 used "Hades." The three strategies—transliteration, Christianizing translation, and a classicizing substitution—are not equivalent, but neither can their relative philological adequacy be decided by single-reference overlap alone. An automatic metric will favor Bell's wording even when transliteration is a defensible editorial choice.

***ini ... nte tef caloč* (great toe of his foot).** The relevant phrase contains two body-part lexemes. Crum glosses Bohairic *ini* as "thumb; great toe" and cites this exact *Vita* phrase in that entry (Crum, 1939, 81a); *caloč* is separately glossed "foot, knee" (Crum, 1939, 814a). The construction *e-ti-ini nte tef caloč* therefore means "to the great toe of his foot": *ini* supplies "great toe," while *nte tef caloč* supplies "of his foot." Bell's rendering specifies the "big toe." THOTH AI likewise produced "big toe," while Gemini 3.0 Pro and Claude 4.5 Opus used "heel." Grok 4.1 instead produced "root of his tongue," which does not identify the source body part correctly. This contrast is compatible with the possibility that domain retrieval supplied useful lexical evidence. It is not proof of that mechanism: because the retrieved chunks were not archived, the study cannot show that a particular lexicon entry was retrieved or that it caused THOTH AI's selection.

***helēnos* (normalized *hellēnos*; Greek/pagan).** The source form is a Greek-derived ethnonym, within the Coptic–Greek loanword domain represented in CCL (Burns et al., 2020). More directly, the Coptic Dictionary Online glosses normalized Coptic *hellēnos* as "pagan, gentile" and relates it to *Hellēn*, for which both the ethnic sense "Greek" and the Christian religious senses "gentile, pagan" are registered (KELLIA, n.d.). Patristic Greek lexicography corroborates this religious extension (Lampe, 1961, s.v. *Hellēn*). Bell's "pagan" is there-

Category	Operational question	Illustration in this case
Lexical mistranslation	Is an identifiable Coptic lexical item mapped to an English item not licensed in context?	Rendering *ini ... nte tef caloč* as "heel" although *ini* denotes the great toe and *caloč* the foot.
Cultural or theological interpretation	Does the output select a culturally situated interpretation while remaining plausibly source-licensed?	Transliteration "Amenti," Bell's "hell," and the classicizing "Hades" represent different treatments of the netherworld term.
Morphosyntactic or structural divergence	Does the output alter predicate–argument relations, negation, tense/aspect, reported speech, or clause linkage?	Reassigning an action or embedded utterance to a participant not licensed by the source syntax.
Omission	Does a source proposition or participant lack a target counterpart?	Leaving out part of the curse or punishment sequence without replacing it elsewhere.
Unsupported addition	Does the target assert a proposition with no plausible source span and no need for English explicitation?	DeepSeek's added imperative "Let them do my will."
Genre-driven fabrication	Do multiple unsupported propositions form a fluent new hagiographic episode?	Qwen's statement that "Christ appeared to him in the dream" and the narrative developed around it.

Table 3: Operational taxonomy used for the illustrative expert reading. Quoted fragments are evidence from retained outputs, not reference translations.

fore lexicographically and contextually well supported. The general interfaces' "Greek" remains etymologically and semantically related, so it is not an unsupported addition, but it is less context-sensitive in this passage.

Taken together, these items show why the expert reading is not reducible to marking every deviation from Bell as an error. The relevant question is whether the English output is licensed by the Coptic source, and, if so, what interpretive commitment it makes. Multiple references or a human adequacy protocol would represent this distinction more faithfully than a single surface form.

6.3 Unsupported Narrative

Three retained outputs contain passage-scale material that meets the operational definition of genre-driven fabrication. DeepSeek V3.2 adds "Let them do my will," an imperative for which no corresponding source proposition was identified. Qwen3 Max introduces an episode in which "Christ appeared to him in the dream." Kimi K2 mixes fragments related to the passage with extended hagiographic prose that cannot be aligned to the Coptic sequence. These are not simply free translations of difficult words: each output adds propositions or an episode while maintaining the tone expected of a saint's life.

This fluency makes the failure important for educational and scholarly use. A visibly broken translation invites checking, whereas a genre-compatible addition may be accepted by a reader who cannot consult the source. At the same time, the evidence remains deliberately limited. The study establishes three observed output-level cases among seven retained translations of this passage; it supplies neither a per-proposition error rate nor evidence about the providers' models across other texts. Without repeated sampling, independent annotators, and a registered span-level scheme, the cause and prevalence of the behavior remain unknown.

7 Discussion

7.1 What This Case Can Support

The observed score pattern and the *ini ... nte tef caloč* example are consistent with specialist retrieval being useful for Bohairic translation. They do not establish that retrieval alone produced the difference. THOTH AI and the general Gemini interface did not share a fully controlled inference stack, and the missing trace prevents connecting an output token to a specific retrieved record. The defensible result is therefore narrower: in one preserved deployed-system comparison, the THOTH AI output was closer to Bell's translation across the reported metrics and avoided the passage-scale unsupported narrative observed in three other outputs.

This narrower claim remains informative as a system case study. It identifies a concrete evaluation item, exposes the configuration and provenance information that a deployed RAG system should retain, and demonstrates that an apparently straightforward comparison becomes difficult to interpret when commercial interfaces hide model and decoding details. For future ancient-language NLP experiments, logging retrieval evidence is not merely an engineering convenience. It is part of

the scholarly apparatus needed to evaluate whether a proposed translation is grounded in a lexicon, grammar, parallel passage, or potentially leaked reference.

7.2 Metrics and Philological Interpretation

The lexical cases illustrate a basic mismatch between single-reference metrics and philological translation. Bell's "hell" and "pagan" are interpretive choices. A model that returns "Amenti" or "Greek" may diverge from the reference while remaining connected to the source. Conversely, a fluent output may share religious vocabulary with the reference while adding an unsupported event. BLEU, chrF, ROUGE, TER, and METEOR can summarize formal similarity, but none can determine whether a theological or cultural interpretation is appropriate.

For this reason, the metrics are best read at coarse resolution in the present paper. They make the large surface differences among these seven strings visible; they do not establish a definitive quality order or replace source-based assessment. A stronger protocol would combine multiple translations where feasible with expert ratings of semantic adequacy, fluency, named entities, grammatical relations, and source support. Span-level links between an English proposition, the Coptic source, and any retrieved evidence would make the analysis auditable.

7.3 Educational and Revitalization Context

Modern efforts to extend the learning and use of Coptic have a history well preceding AI. Nineteenth- and twentieth-century initiatives associated with Coptic church education and figures such as Claudius Labib produced pedagogical materials and new vocabulary, while modern digital corpora and dictionaries have broadened access to scholarly resources (Meinardus, 2002; Miyokawa, 2017). Retrieval-based tools belong to this longer history of mediation; they do not by themselves constitute language revitalization.

Their plausible educational value is narrower and practical. A system may help a learner locate lexical alternatives, compare a draft translation with grammatical evidence, or formulate questions for further study. That value depends on showing provenance and uncertainty. Ordinary mistranslation may be noticed by an advanced learner, whereas fluent religious narrative can appear authoritative. A responsible interface should therefore separate source text, retrieved evidence, generated translation, and commentary, and should encourage consultation of editions, dictionaries, and qualified readers rather than presenting generated prose as an authoritative ecclesiastical or scholarly translation.

8 Limitations and Reproducibility

The evaluation has one passage-level segment, one dialect, one hagiographic work, one English reference, and one retained generation per system. It provides no variance or uncertainty estimate and cannot establish population-level rankings. The passage was chosen for philological difficulty rather than randomly or stratifiably sampled, so its vocabulary and narrative structure may disproportionately affect the observed contrast.

The deployed-system comparison confounds retrieval with interface, hidden system instructions, provider-side preprocessing, decoding defaults, and model revision. Exact API snapshot identifiers and exact per-output timestamps were not preserved. Temperature, top-p, maximum-token settings, seeds, and a complete retry history are also unavailable. The interface labels are therefore historical labels from the collection record, not reproducible model identifiers.

The retrieval evidence needed for a causal or leakage analysis was not archived. Complete chunks, source identifiers, ranks, scores, and the context actually passed to the generator are missing. Although the passage and Bell's translation were not intentionally ingested as evaluation-specific items, this does not prove that related or overlapping text was absent from the indexed resources or the foundation model's training data. The paper consequently cannot distinguish assistance from lexical evidence, incidental textual overlap, hidden interface behavior, or general model competence.

The qualitative analysis was conducted by the author, who also develops THOTH AI. Domain expertise permits close source reading, but involvement creates a risk of expectation bias. There was no independent annotator, adjudication, or inter-annotator-agreement estimate, and no preregistered span-level error inventory. The 3/7 observation is therefore an output-level description, not a general hallucination frequency. Automatic metrics add a separate limitation: all use one published English translation and cannot determine philological adequacy for culturally loaded renderings.

A reproducible follow-up should use a fixed API model snapshot with RAG toggled within the same code path; archive all prompts, parameters, output attempts, and retrieval traces; run multiple generations; and evaluate a planned sample of Bohairic and Sahidic texts across genres. At least two independent Coptologists should annotate source support and translation adequacy using a fixed codebook, with disagreements adjudicated and agreement reported. Multiple references or acceptable-variant annotations should be added where copyright and scholarly practice permit. Specialized Coptic MT systems should also be included alongside general LLM interfaces.

9 Conclusion

This paper reports a pilot case study of seven deployed systems translating one Bohairic passage. In the preserved comparison, THOTH AI produced the strongest single-reference automatic scores, while three other retained outputs contained clear passage-scale unsupported narrative. Detailed lexical cases also show that disagreement with one reference may reflect a defensible philological choice rather than simple error.

Because the sample is minimal, the interfaces are confounded, and the retrieval trace is missing, the findings are not a general benchmark or a causal demonstration of RAG effectiveness. The contribution is instead a transparent account of what was observed, an operational taxonomy for distinguishing mistranslation from unsupported narrative, and a specification of the controls and provenance needed for a defensible larger study of Coptic machine translation.

Conflict of Interest and Author Role

The author develops and maintains THOTH AI and configured and evaluated the system reported here. This constitutes a potential conflict of interest. Independent replication with archived retrieval traces and separately annotated outputs is needed.

Ethical Considerations

Machine translation of historical religious texts can mislead scholarly, educational, or faith communities when fluent output is mistaken for a source-grounded translation. The systems studied here should be used as aids, not substitutes for philological review. Reporting unsupported generation and the limits of the evidence is therefore central to responsible use.

References

Alibaba Cloud. n.d. Model Pricing: Qwen3 Max. Official Model Studio documentation. Accessed January 2026.

Anthropic. n.d. Claude Opus 4.5 System Card. System card. Accessed January 2026.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, MI. Association for Computational Linguistics.

David N. Bell. 1983. *The Life of Shenoute*. Number 73 in Cistercian Studies Series. Cistercian Publications, Kalamazoo, MI.

Dylan Michael Burns, Frank Feder, Katrin John, and Maxim Kupreyev. 2020. Comprehensive Coptic lexicon: Including loanwords from ancient Greek v 1.2. Dataset, Freie Universität Berlin.

Marta R. Costa-jussà et al. 2022. No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672.

Walter E. Crum. 1939. *A Coptic Dictionary*. Clarendon Press, Oxford.

DeepSeek. n.d. DeepSeek API Updates: DeepSeek-V3.2 Release. Official API documentation. Accessed January 2026.

Dify. n.d. Knowledge Retrieval. Online documentation. Accessed January 2026.

Maxim Enis and Andrew Megalaa. 2024. Ancient voices, modern technology: Low-resource neural machine translation for Coptic texts. Technical report, Coptic Translator. 15 pages; accessed August 2026.

Frank Feder, Maxim Kupreyev, Emma Manning, Caroline T. Schroeder, and Amir Zeldes. 2018. A linked Coptic dictionary online. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2018)*, pages 12–21, Santa Fe, NM. Association for Computational Linguistics.

Google. n.d. Gemini 3 Pro Preview. Online documentation. Accessed January 2026.

Eitan Grossman and Tonio Sebastian Richter. 2014. The egyptian-coptic language: Its setting in space, time and culture. In Eitan Grossman, Martin Haspelmath, and Tonio Sebastian Richter, editors, *Egyptian-Coptic Linguistics in Typological Perspective*, volume 55 of *Empirical Approaches to Language Typology*, pages 69–102. De Gruyter Mouton.

KELLIA. n.d. Coptic Dictionary Online: TLA Lemma C11084 (hellēnos). Online lexicon entry. Accessed August 11, 2026.

G. W. H. Lampe, editor. 1961. *A Patristic Greek Lexicon*. Clarendon Press, Oxford.

Bentley Layton. 2000. *A Coptic Grammar: With Chrestomathy and Glossary. Sahidic Dialect*. Harrassowitz, Wiesbaden.

Johannes Leipoldt, editor. 1906. *Sinuthii archimandritae vita et opera omnia, Vol. I: Sinuthii vita Bohairice*. Number 41 in CSCO. E Typographeo Reipublicae, Paris. Adiuvante W. E. Crum.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona. Association for Computational Linguistics.

Antonio Loprieno. 1995. *Ancient Egyptian: A Linguistic Introduction*. Cambridge University Press, Cambridge.

Otto F. A. Meinardus. 2002. *Two Thousand Years of Coptic Christianity*. American University in Cairo Press, Cairo.

So Miyagawa. 2025. RAG-enhanced neural machine translation of Ancient Egyptian text: A case study of THOTH AI. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 33–40, Albuquerque, USA. Association for Computational Linguistics.

Hiroko Miyokawa. 2017. The revival of the Coptic language and the formation of Coptic ethnoreligious identity in modern Egypt. In Nelly van Doorn-Harder, editor, *Copts in Context: Negotiating Identity, Tradition, and Modernity*, pages 151–156. University of South Carolina Press, Columbia, SC.

Moonshot AI. n.d. Kimi K2. Official GitHub repository. Accessed January 2026.

Tito Orlandi. 1990. The Corpus dei Manoscritti Copti Letterari. *Computers and the Humanities*, 24:397–405.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 311–318, Philadelphia, PA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT 2015)*, pages 392–395, Lisbon. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191. Association for Computational Linguistics.

Chris H. Reintges. 2018. *Coptic Egyptian (Sahidic Dialect): A Learner's Grammar*, 2nd edition. Rüdiger Köppe, Cologne.

Samuel Rubenson. 1996. The transition from Coptic to Arabic. *Égypte/Monde arabe*, 27–28:77–92.

Caroline T. Schroeder and Amir Zeldes. 2016. Raiders of the lost corpus. *Digital Humanities Quarterly*, 10(2).

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231. Association for Machine Translation in the Americas.

Tasbeha.org. 2026. Coptic church liturgical resources. Website. Accessed January 2026.

Walter C. Till. 1961. *Koptische Dialektgrammatik: Mit Lesestücken und Wörterbuch*, 2nd edition. C. H. Beck, Munich.

Audric-Charles Wannaz and So Miyagawa. 2024. Assessing large language models in translating Coptic and Ancient Greek ostraca. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 463–471, Miami, USA. Association for Computational Linguistics.

xAI. n.d. Grok 4.1. Official announcement. Accessed January 2026.

Amir Zeldes and Caroline T. Schroeder. 2016. An NLP pipeline for Coptic. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2016)*, pages 146–155, Berlin. Association for Computational Linguistics.

Automating Mwotlap morphology using formal grammar

Fabio Meroni
University of the Basque Country
fmeroni001@ikasle.ehu.eus

Alexandre François
Lattice, CNRS-ENS-PSL-USN
alexandre.francois@ens-psl.eu

Max Silberztein
Marie and Louis Pasteur University
max.silberztein@univ-frcomte.fr

Abstract

This paper presents a rule-based computational formalization of vowel alternations in Mwotlap morphology using NooJ. Mwotlap displays several morphophonological processes at the prefix-root boundary, including vowel copy, epenthesis, transfer, elision, and blockage. To model these first three, which cannot be derived by simple string concatenation within NooJ's finite-state framework, we introduce a new string-edit operator which swaps adjacent segments. This operator unifies vowel copy, insertion, and transfer under a single formal mechanism while preserving linguistic generalization.

1 Introduction to non-concatenative morphology

Morphology is not always strictly linear (as with suffixes and prefixes), and its interaction with phonology can give rise to complex surface patterns, a phenomenon known as non-linear or non-concatenative morphology (Kroeger, 2005, pp. 304ff). In many languages, phonotactic constraints and prosodic requirements intersect with morphological boundaries in ways that trigger processes such as vowel epenthesis and vowel copying. These strategies may serve to repair illicit consonant clusters, satisfy minimal word conditions, or preserve preferred syllable structures, often yielding outputs that reflect both the local phonological environment at morpheme boundaries and the morphological category of the affixes involved.

Olson (2001) described that in Mono (Ubangian, Democratic Republic of the Congo), a well-formed word needs to contain at least two morae. To appear in isolation, a monomoraic root thus requires the addition of an extra syllable, which takes the form of a copy of the root vowel (1a). Likewise, obstruent-liquid consonant clusters are broken by inserting an epenthetic vowel, whose quality and tone are borrowed from the following vowel (1b). And if such a cluster is found in a monomoraic root in isolation, like *gré* in (1c), then the two rules operate in sequence, resulting in the insertion of two vowels (Hall, 2006, p. 6). Similarly, Kekchi (Mayan, Guatemala) avoids CC clusters at morpheme boundaries with the insertion of a copied vowel (Campbell, 1974, p. 271).

(1)
a.	*ʒī* → ***ī****ʒī*	‘tooth’
b.	*gàfrū* → *gàf****ū****rū*	‘mortar’
c.	*gré* → ***é****g****é****ré*	‘big’

Another phonotactic constraint, this time of Lucazi (Bantu, Angola and Zambia), requires every word to end with a vowel (Fleisch, 2000). If such a vowel is not brought by the root itself, it must be added via epenthesis. Its default form is /a/ (2a) but, when the preceding morpheme is a lexical root, the suffix borrows its quality from the previous vowel (Marten & Chongo Kula, 2002, p. 4) (2b).

(2) a. *tu-na-fúm-**i**n-**a***
SUBJ.1PL-ANT-come-APPL-FV
‘we came for (it)’

b. *tu-na-f**ú**m-**u***
SUBJ.1PL-ANT-come-FV
‘we came’

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 59–67
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_008

A single language may employ multiple repair strategies depending on the phonological environment. This study examines such vowel rules in Mwotlap (Oceanic, Vanuatu), a language spoken by 2100 people on Mota Lava, one of the Banks Islands in northern Vanuatu. Its vowel rules are described in François (1999, 2000, 2001), and will be summarized here.

Mwotlap has a strict (C)V(C) syllable structure, which disallows tautosyllabic consonant clusters. If a lexical root has an underlying shape CCVC, phonotactic rules entail that its first two Cs must be distributed across two distinct syllables. In case the root combines with a CV- prefix, the resulting pattern #CV-C|CVC allows the CCVC root to surface unchanged, as in (3a). But in the absence of a supporting prefix, the #CC onset must be broken by an epenthetic vowel, whose quality is borrowed from the following syllable (3b).

(3) a. *na-syip* < *nA-* + *syip*
ART-waterfall
'a/the waterfall'

b. *s**i**yip* < *syip*
waterfall
'waterfall'

Another rule, *vowel copy*, concerns several "copiable" prefixes, including the article *na-* that we saw in (3a). When the following root starts with a single C, the prefix alters its underlying vowel to borrow that of the following syllable. Thus in (4), the combination of *na-* with the noun *qoyqoy* 'dream' surfaces as a form *no-qoyqoy*, showing vowel copy:

(4) *n**o**-q**o**yqoy* < *nA-* + *q**o**yqoy*
ART-dream
'a/the dream'

Finally, a small number of irregular C-initial roots do not trigger vowel copy, but *vowel transfer*. That is, instead of just lending its quality to the vowel of the prefix, the first vowel of the root migrates to the previous syllable, replacing the (copiable) prefix's underlying vowel – as in (5). Vowel transfer often involves high vowels /i/ or /u/, yet it is not entirely predictable; it is a property of a handful of lexemes.

(5) *n**u**-wmet* < *nA-* + *w**u**met*
ART-bowl
'a/the bowl'

When vowel copy, epenthesis, and transfer co-occur within a single morphological system, any computational model must be able to distinguish the three processes, apply them conditional on the radical's phonological shape, and remain general enough to avoid enumerating each vowel quality as a separate rule. The challenge is therefore to find a level of formalization that is both expressive enough to capture the phonological generalization and constrained enough to remain computationally tractable, that is, to stay within a class of formal grammars that guarantees decidability and linear-time processing.

This paper presents a possible solution to this problem, achieved with a single bounded stack operator conceived in the domain of finite-state Natural Language Processing (NLP). The project presented here aimed specifically at formalizing Mwotlap morphology for subsequent text and corpus analysis, and succeeded with the use of such a tool. In the following pages we present the software environment chosen to develop this idea, within a theoretical framework and a methodological choice that favored linguistic soundness above the aesthetics of the results.

2 Theoretical and methodological background

Today, most NLP projects rely on empirical methods (statistical, deep learning, LLMs) and thus do not require linguists to develop hand-crafted linguistic resources to perform the various tasks they deploy. However, studying and building formalized descriptions of natural languages remains a legitimate scientific goal in and of itself. Ultimately, accurate descriptions of natural languages enable NLP applications to perform more reliably and in a more controlled manner. Most importantly, NLP applications should attempt to emulate the theory that underlies the language's structures, as defined by the linguist independently of any computational implementation (Silberztein, 2024). Rule-based approaches remain a principled choice for this goal, precisely because they make the linguist's theoretical commitments explicit and verifiable. The significant results this approach has produced in recent work (e.g. Barclay, 2026)

Prefix	Property	Prefixed to
nA-	Article	Nouns
bE-	Adverbializer	
lE-	Locativizer	
mE-	Perfect	Verbs, nouns, adjectives
nE-	Stative	
tE-	Future	
tE-	Adjectivizer	Locatives
vĒ-	Enumerator	Numerals

Table 1: The copiable prefixes of Mwotlap. Capital letters mark the vowel as variable.

is itself evidence that linguistic explicitness and computational rigor are not in conflict.

The NooJ linguistic software (Silberztein, 2016) allows linguists to formalize a variety of phenomena (orthography, morphology, syntax and semantics), using the four types of Chomsky-Schützenberger's (1963) hierarchy of grammars (regular, context-free, context-sensitive and unrestricted grammars). In the present study, we are focusing on a particular aspect of Mwotlap morphology concerning vowel rules, following the linguistic description by François (1999, 2000, 2001).

Non-concatenative morphological processes have been handled in other finite-state frameworks. In XFST (Beesley & Karttunen, 2003) and Foma (Hulden, 2009), such operations can be modelled via compile-replace or flag diacritics (see Amsalu & Demeke, 2006; Cohen-Sygal & Wintner, 2006; Wintner, 2014 inter alia). Before that, two-level morphology, originally proposed by Koskenniemi (1983) and later extended, addressed similar phenomena through the interaction of surface and lexical levels constrained by correspondence rules (see Kay, 1987; Trost, 1990; Pisceldo et al., 2008 inter alia).

Both traditions have produced substantial descriptive coverage across typologically diverse languages, but in a different methodological orientation. An XFST or two-level implementation would require the linguist to write a separate transducer for each non-concatenative process, which is then recomposed with the lexicon externally. Our NooJ-based solution, instead, expresses the operation inline, within the paradigm rule that is already stored in the lexicon as an inflectional or derivational property, at the cost of introducing a single new primitive. Thanks to this, NooJ can deploy the same resource both to generate inflected forms and to analyse running text and corpora, without any additional recomposition step.

To support this approach, we created a new swap operator (`<W>`) for the description of Mwotlap, dedicated to local reordering and allowing adjacent segments to be swapped when required by phonotactic or morphophonological constraints.

3 The Mwotlap language

Mwotlap (endonym *na-vap to-M̄otlap*, ISO 639-3: mlv, Glottocode: motl1237) is spoken by an estimated 2100 people, primarily on Mota Lava and Vanua Lava, in northern Vanuatu. It is the most vibrant language within the Torres-Banks linkage of Oceanic languages, a subgroup of the Austronesian family (François, 2012). Because Mwotlap is spoken by all generations and intergenerational transmission has – so far – run uninterrupted, UNESCO (2025) lists it as a "Safe / Not Endangered" entry in its atlas of endangered languages; however, the *Endangered Languages Project* (ELP, n.d.) recognizes it as "threatened".

François produced a vast body of research to document the Mwotlap language, beginning in 1997. First and foremost, his doctoral dissertation (François, 2001) constitutes the first complete description of Mwotlap ever written. Before that date, the only publications about this language were a grammar sketch (Codrington, 1885, pp. 310-322), an article about its phonology (Kasarhérou, 1962), and a wordlist (Tryon, 1976).

3.1 Prefixation in Mwotlap

The language distinguishes two types of productive prefixes: invariable prefixes, whose form is constant regardless of the phonological shape of the radical, and copiable prefixes, whose initial vowel slot is sensitive to the radical's vowel. Table 1 lists the eight copiable prefixes of Mwotlap and the Parts of Speech (PoS) they can be associated to (after François, 2001, p. 100).

Mwotlap has seven phonological vowels:/i, ɪ, ɛ, a, ɔ, ʊ, u/, spelled respectively as ⟨i, ē, e, a, o, ō, u⟩. By convention, the 8 copiable prefixes (first column of Table 1) are noted with a capitalized vowel: this does not denote an archiphoneme, but rather flags the *default* realization of the prefix in the absence of copy, or of elision (François 2000). Five distinct processes govern the surface form of

Process	Radical	Unprefixed	Prefixed
Copy	QOYQOY	*qoyqoy*	*no-qoyqoy*
Insertion	SYIP	*siyip*	*na-syip*
Transfer	WUMET	*wumet*	*nu-wmet*
Elision	EP	*ep*	*n-ep*
Blockage	N̄EY	*n̄ey*	*na-n̄ey*

Table 2: The five vowel rules at the prefix-radical boundary of words exemplified with the addition of the article *nA-*

this vowel at the prefix–radical boundary (François, 2005, p. 118), as shown in Table 2.

The first three vowel rules were presented and illustrated in Section 1. The last rule, namely *vowel blockage*, concerns several #CV- roots that do not undergo vowel copy nor vowel transfer, and instead take the default form of copiable prefixes shown in Table 1. For example, if *n̄ey* 'almond' were a regular C-initial root, we would expect a prefixed form **ne-n̄ey*, but we observe *na-n̄ey* instead. While this behavior is expected for #CC- roots like SYIP (see 3a), its occurrence with #C- radicals is exceptional, and historically vestigial (François, 2000). To represent these cases on NooJ, it is sufficient to move to the left boundary of a word and attach the default realization of a prefix. As for #V- roots, they do not need any special term in NooJ: the corresponding automation moves the cursor to the left boundary of the word form and directly adds the thematic consonant C_0 of the prefix in that position, with no vowel.

As mentioned before, #CC- roots simply trigger the default realization of the prefix. For these forms, the rule to be applied does not concern the prefixed form, but the unprefixed one (see Table 1), as this is when they require vowel insertion: the epenthetic vowel V breaking #CC- into #CVC- inherits the quality of the vowel in the following syllable, and this is where a new operator for non-concatenative morphology is needed.

4 Formalizing morphology with NooJ

To describe the morphology of a language, NooJ proposes two formal tools: inflectional and derivational grammars (described in .nof files), and morphological grammars (described in .nom files) used to describe other phenomena such as agglutinations (e.g., to analyze word forms such as *antidisestablishmentarianism* and productive neologisms like in "The *retrumpization* of the US policy"). Both systems are represented by

Operator	Function
<L>*	Move cursor left
<R>*	Move cursor right
<E>	Add empty string (ε)
<C>	Change case
<D>	Duplicate
<B>*	Delete preceding character
<S>	Delete next character
<N>*	Go to end of subsequent word
<P>*	Go to end of previous word

Table 3: Generic operators in NooJ .nof ECFGs (Silberztein, 2026). Operators marked with "*" accept arguments.

Enhanced Context-Free Grammars (ECFGs). As inflectional and derivational phenomena are not recursive, they are transparently converted into efficient Finite-State Transducers (FSTs) by NooJ. Linguists therefore can write very expressive and natural grammars while not losing any efficiency during parsing, which allows them to parse large corpora of texts in real time.

Inflectional and derivational paradigms are described by grammar rules that represent sets of affixes that must be added to each lexical entry to produce their corresponding inflected and derived forms, associated with linguistic properties described in the property definition file (.def file). For example, (6) represents the `PEN` inflectional paradigm in English.

(6) `PEN = <E>/+Singular | s/+Plural ;`

This rule states that by adding the empty string (ε, encoded as `<E>` in NooJ) to a lexical entry, one produces its singular form, while adding *-s* forms the plural. Affixes may contain orthographic operators that can perform special operations at the cursor position. For instance, the operator `<B>` deletes the current character. To inflect the verb *to eat*, one can use the paradigm in (7).

(7)
```
EAT = <E>/+Base | s/+3+Singular |
en/+PastParticiple | ing/+Gerund
| <B3>ate/+Preterite ;
```

In this paradigm, `<B3>ate` deletes the last 3 characters of the lexical entry, and then adds *ate*: *eat* + `<B3>ate` → *ate*. The resulting form is associated with property `+Preterite`.

Although the `<B>` operator allows the system to produce any wordform from any lexical entry, it is not satisfying in practice as, in certain cases, it

would force linguists to use multiple grammar rules to formalize a single morphological paradigm.

NooJ offers several other operators, listed in Table 3. <B>, <L>, <N>, <P>, <R>, <S> accept arguments to perform their assigned operation more than once, or even until reaching the edge of the current string (e.g. <L2> 'move left twice', <LW> 'move to the left boundary of this word'). On top of these generic operators, operators to describe phenomena specific to many languages were developed, such as <Q> for reduplication in Asian and Amerindian languages, <F> for consonant finalization in Semitic languages, <I> to add a iota to the current Greek letter, etc. To describe Mwotlap morphology, the new operator <W> (swap letters) was added.

The approach presented here relies on a fundamental assumption worth making explicit: that the orthography of the language being formalized is phonetically transparent, meaning that there is a close (ideally one-to-one) correspondence between graphemes and phonological segments, as is the case for Mwotlap (François, 2001, p. 51). A digraph representing a single segment, a diacritic encoded as a separate code point, or a character sequence whose phonological value is context-dependent would all silently shift the cursor to the wrong position, producing incorrect forms without raising an error.

This limitation was evident in the development of the Mwotlap module: the letters ⟨m̄⟩ and ⟨n̄⟩, which represent /n͡mʷ/ and /ŋ/, have no single precomposed Unicode code point and are encoded as a base character followed by a combining macron (U+0304). A specific NooJ update (v7.3, update b20250410) was required to ensure that each base letter and its diacritic were treated as a single unit by the morphological engine. This is analogous to the handling of Hungarian digraphs such as ⟨sz⟩ and underscores the need for linguists to verify that every grapheme in their language is correctly processed by NooJ. For languages with rich diacritic systems, non-Latin scripts, or orthographic conventions that routinely encode single phonemes as multi-character sequences, additional preprocessing steps of this kind should be anticipated in the module development pipeline.

5 Introducing the Mwotlap operator <W>

In Mwotlap, vowels may be swapped with prefixes under certain conditions – as in the rule of vowel transfer – e.g., (*na-*) + *wumet* → *nuwmet*. To describe all potential swapping operations with NooJ generic operators, we would need to implement 8 prefixes × 7 vowels, i.e., 56 grammar rules. While technically functional, that is not acceptable from a linguistic point of view, as it would obscure the unitary nature of the process by treating what is, in effect, a single productive rule as 56 idiosyncratic rules. In particular, this would ignore François' (1999, 2000, 2001) description of vowel behavior in Mwotlap and regress to the inaccurate generalizations of Kasarhérou (1962) and Crowley (2002).

As anticipated, the prefixation of #CCV- radicals and blocking #CV- radicals was easily described as the juxtaposition of two strings of text (namely, the radical and the prefix). What was left to do was computationally describing all the rest, *i.e.,* the prefixation of #CV- radicals, vowel insertion in unprefixed #CCV- radicals, and vowel transfer in prefixed #CV- radicals that undergo it. To do so, we added to NooJ the operator <W> to swap characters in the right position when performing vowel copy, vowel insertion, and vowel transfer.

5.1 Vowel Copy

The rule (8) for vowel copy works such that, given a #C_1V_1 radical, V_1 is taken as the vowel of the CV- prefix to be put before it (François, 2001, p. 100):

(8) C_0V_0- + #$C_1V_1C_2$- → #C_0V_1-$C_1V_1C_2$-

The corresponding implementation in NooJ moves the cursor to the left boundary of the word form (<LW>), then moves it right twice (<R2>). At this point, the cursor is positioned after the vowel subject to copy (V1) and, therefore, such a vowel must be duplicated (<D>). To move the copied vowel to the beginning of the word form, we move the cursor left once (<L>) and then swap the two preceding characters (<W>). We then return to the left boundary of the word form (<LW>) and insert the thematic consonant C_0. The resulting word form is the correct prefixed form of the starting word form, now morphologically connotated with

the trait that C_0 brought. The final result is the fully inflected form. For example, *qoyqoy* 'dream' is prefixable as *no-qoyqoy* ('the/a dream') since the vowel of a radical that undergoes vowel copy is copied into the prefix, and the process is described as (9) in NooJ.

(9) `qoyqoy<LW><R2><D><L><W><LW>n`

5.2 Vowel Insertion

The strict (C)V(C)|(C)V(C) syllabic structure of Mwotlap, which does not allow syllables of CCV type to exist in the language, was anticipated in Section 3.1. This is why #CCV- radicals, when appearing alone, show an epenthetic vowel that breaks the two consonants with a necessary syllable boundary. For example, *siyip* 'waterfall' is SYIP as a radical, prefixable as *na-syip* 'the/a waterfall', and usable as an unprefixed word as *siyip*. The inserted vowel of a radical that undergoes vowel insertion is in fact not part of the underlying radical, but added via such a process.

The formal rule (10) for vowel insertion makes sure that, when appearing without any prefix, a wordform with radical $\#C_1C_2V_1$-, has V_1 inserted between the two initial consonants (François, 2001, p. 79):

(10) $\#C_1C_2V_1\text{-} \rightarrow \#C_1V_1C_2V_1\text{-}$

Among all the formal rules shown in this section, this is the only one that points at unprefixed ("bare") forms as its objective, starting from the radical contained in a compiled NooJ dictionary (.nod file). The corresponding grammar rule moves the cursor to the left boundary of the word form (`<LW>`), then moves it to the right three times (`<R3>`). At this point, the cursor is positioned after V_1, that must be inserted between C_1 and C_2. To do so, the vowel is duplicated (`<D>`) and, as the cursor moves left once (`<L>`), it is then swapped with the preceding character (`<W>`). The resulting form is the correct unprefixed form of the starting word form, obtained in NooJ as in (11).

(11) `syip<LW><R3><D><L><W>`

5.3 Vowel Transfer

The vowel of a prefix is truly found to be "intermittent" (François, 2000; 2001) in those lexemes that are not easily categorized as either "copiable" or "blocking" in the sense of vocalic copy, and undergo vowel transfer instead. For example, *wumet* 'bowl' is prefixable as *nu-wmet* 'the/a bowl', with *nA-* taking on the radical's /u/, and usable as an unprefixed word as *wumet*. The transferred vowel of a radical that undergoes vowel transfer, is in fact part of the radical and transferred to the prefix via vowel transfer, which follows the rule in (12) (François, 2001, p. 115):

(12) $C_0V_0\text{-} + \#C_1V_1C_2V_2\text{-} \rightarrow \#C_0V_1\text{-}C_1C_2V_2\text{-}$

The corresponding grammar rule moves the cursor to the left boundary of the word form (`<LW>`), then moves it to the right twice (`<R2>`). The cursor is then positioned after the vowel (V1) that must be moved to the prefix; it is duplicated (`<D>`), then moves left once (`<L>`), and is swapped with the preceding character (`<W>`). The transferred vowel is now in its correct position, and the cursor can now delete it (`<S>`). At this point, one can move the cursor to the left boundary of the wordform (`<LW>`) and then insert the thematic consonant C_0. The resulting form is the correct prefixed form corresponding to the starting lexical entry, associated with the linguistic property for C_0. In NooJ, this is obtained by what is exemplified in (13):

(13) `wumet<LW><R2><D><L><W><S><LW>n`

6 Evaluation

The NooJ Mwotlap module comes with separate dictionaries for grammatical and lexical morphemes (Meroni, 2026, p. 5). The latter is LWOTLAP.nod, a 2241-entry electronic dictionary in the form of a lexicon-grammar table, containing all monomorphemic lexical roots. It was automatically extracted from François (2026) using a Python pipeline, with particular attention to the conventions it uses: e.g. in a headword, [V] stands for vowel insertion, ‹V› stands for vowel transfer, and a preposed ° stands for vowel blockage; all other consonant-initial headwords are treated as undergoing vowel copy, while all other vowel-initial headwords are treated as undergoing vowel elision. In LWOTLAP.nod sit 1199 nouns, 602 verbs, 155 adjectives, 19 prefixable locatives and the 12 basic numerals (the terms for designating the numbers 1-10, 100 and

```
bem,bem,V+Trans+V2+EN="carry"+FR="porter"+FLX=VC_PRED+BARE
bem,bem,V+Trans+V2+EN="carry"+FR="porter"+FLX=VC_PRED+AO
nibem,bem,V+Trans+V2+EN="carry"+FR="porter"+FLX=VC_PRED+AO+3+s
malbem,bem,V+Trans+V2+EN="carry"+FR="porter"+FLX=VC_PRED+IAM
etbem,bem,V+Trans+V2+EN="carry"+FR="porter"+FLX=VC_PRED+NEG+REALIS
titbem,bem,V+Trans+V2+EN="carry"+FR="porter"+FLX=VC_PRED+NEG+IRREALIS
mebem,bem,V+Trans+V2+EN="carry"+FR="porter"+FLX=VC_PRED+PFT
tebem,bem,V+Trans+V2+EN="carry"+FR="porter"+FLX=VC_PRED+FUT
ben,ben,V+Trans+N0VN1Inani+EN="light"+FR="allumer"+FLX=VC_PRED+BARE
ben,ben,V+Trans+N0VN1Inani+EN="light"+FR="allumer"+FLX=VC_PRED+AO
niben,ben,V+Trans+N0VN1Inani+EN="light"+FR="allumer"+FLX=VC_PRED+AO+3+s
malben,ben,V+Trans+N0VN1Inani+EN="light"+FR="allumer"+FLX=VC_PRED+IAM
etben,ben,V+Trans+N0VN1Inani+EN="light"+FR="allumer"+FLX=VC_PRED+NEG+REALIS
titben,ben,V+Trans+N0VN1Inani+EN="light"+FR="allumer"+FLX=VC_PRED+NEG+IRREALIS
meben,ben,V+Trans+N0VN1Inani+EN="light"+FR="allumer"+FLX=VC_PRED+PFT
teben,ben,V+Trans+N0VN1Inani+EN="light"+FR="allumer"+FLX=VC_PRED+FUT
bēnwar,bēnwar,N+Weak+NHum+Inani+Conc+Flora+EN="benoa"+FR="boisnoir"+FLX=VC_N+DRV=VC:<E>+BARE
malbēnwar,bēnwar,N+Weak+NHum+Inani+Conc+Flora+EN="benoa"+FR="boisnoir"+FLX=VC_N+DRV=VC:<E>+IAM
etbēnwar,bēnwar,N+Weak+NHum+Inani+Conc+Flora+EN="benoa"+FR="boisnoir"+FLX=VC_N+DRV=VC:<E>+NEG+REALIS
titbēnwar,bēnwar,N+Weak+NHum+Inani+Conc+Flora+EN="benoa"+FR="boisnoir"+FLX=VC_N+DRV=VC:<E>+NEG+IRREALIS
nēbēnwar,bēnwar,N+Weak+NHum+Inani+Conc+Flora+EN="benoa"+FR="boisnoir"+FLX=VC_N+DRV=VC:<E>+ART
nēbēnwar,bēnwar,N+Weak+NHum+Inani+Conc+Flora+EN="benoa"+FR="boisnoir"+FLX=VC_N+DRV=VC:<E>+STA
mēbēnwar,bēnwar,N+Weak+NHum+Inani+Conc+Flora+EN="benoa"+FR="boisnoir"+FLX=VC_N+DRV=VC:<E>+PFT
tēbēnwar,bēnwar,N+Weak+NHum+Inani+Conc+Flora+EN="benoa"+FR="boisnoir"+FLX=VC_N+DRV=VC:<E>+FUT
bēnwar,bēnwar,+INITIALCATEGORY=N+Weak+NHum+Inani+Conc+Flora+EN="benoa"+FR="boisnoir"+FLX=VC_N+DRV=VC:<E>
lēbēnwar,bēnwar,LOC+INITIALCATEGORY=N+Weak+NHum+Inani+Conc+Flora+EN="benoa"+FR="boisnoir"+FLX=VC_N+DRV=VC:<E>+NounDerived
tēbēnwar,bēnwar,ADJ+INITIALCATEGORY=N+Weak+NHum+Inani+Conc+Flora+EN="benoa"+FR="boisnoir"+FLX=VC_N+DRV=VC:<E>+NounDerived
tēlēbēnwar,bēnwar,ADJ+INITIALCATEGORY=N+Weak+NHum+Inani+Conc+Flora+EN="benoa"+FR="boisnoir"+FLX=VC_N+DRV=VC:<E>+LOC
bēbēnwar,bēnwar,ADV+INITIALCATEGORY=N+Weak+NHum+Inani+Conc+Flora+EN="benoa"+FR="boisnoir"+FLX=VC_N+DRV=VC:<E>+NounDerived
bep,bep,N+Weak+NHum+Inani+Conc+Flora+EN="chilli"+FR="piment"+FLX=VC_N+DRV=VC:<E>+BARE
malbep,bep,N+Weak+NHum+Inani+Conc+Flora+EN="chilli"+FR="piment"+FLX=VC_N+DRV=VC:<E>+IAM
etbep,bep,N+Weak+NHum+Inani+Conc+Flora+EN="chilli"+FR="piment"+FLX=VC_N+DRV=VC:<E>+NEG+REALIS
```

Figure 1: Snippet of inflected forms of LWOTLAP.nod, the main lexicon-grammar of lexical morphemes in the NooJ Mwotlap module.

1000, then composable in all others), all susceptible of being prefixed (see Table 1).

By using the "Inflect" option of NooJ's "Lab" menu on LWOTLAP.nod, 44380 inflected forms are obtained. Among these, all of the PoS susceptible of being prefixed are correctly produced by the engine, which applies the inflectional and derivational grammars associated with each entry (see Figure 1). This confirms that the implementation successfully captures the formal behavior of the prefix-root boundary while remaining fully productive across the lexicon. Importantly, this approach avoids having to resort to separate lexicalized rules for individual vowel qualities, as was anticipated in Section 2: across the lexicon, the operator correctly derives forms involving local vowel displacement or duplication; and it does so while preserving the phonotactic constraints of Mwotlap, especially the language's strict (C)V(C) syllable structure. Besides, the implementation of `<W>` makes it possible to generate forms that would otherwise require extensive rule duplication, thereby reducing coding redundancy while preserving descriptive accuracy. From a computational perspective, this confirms that local segmental reordering can be modeled efficiently within NooJ's finite-state architecture, without moving outside a tractable formal framework.

A demo text, *Mayanag t-Aliwan.not* (346 types, 1711 tokens; TTR = 0.202), also comes with the NooJ Mwotlap module. Performing a linguistic analysis on this text, thus applying all the linguistic resources present in the module, will yield all correct analyses of every prefixable PoS, confirming 100% recall and 100% precision – every token receives at least one analysis, and every analysis set contains the correct reading.

The NooJ ambiguity log returns 285 form types with two or more analyses. Of these, 131 are false positives since they carry exactly one lexical or morphological analysis accompanied by one or more syntactic annotations, which do not constitute competing morphological readings. The remaining 154 types (44.5% of the type inventory) are genuinely ambiguous. Their ambiguity arises from well-defined sources unrelated to the vowel rules formalized in this paper. One of them is accidental homophony – e.g. *et* can be a noun 'person', a verb 'to see', or a clitic of negation. Another cause for ambiguity is lexical flexibility, i.e. the ability, for a single lexeme, to fill different grammatical functions (see François, 2017): e.g. the lexeme *bah* 'finish+' may adopt the grammatical behavior of an intransitive verb, of a transitive verb, of an adverb, and of a TAM marker. While these sources of ambiguity reflect the linguistic profile of Oceanic languages more generally, they also combine with some idiosyncrasies of the Mwotlap language: e.g., the often identical surface realization – due to Vowel Copy [§5.1] – of the noun article *nA-*, the stative *nE-* and/or of the 3rd person singular aorist prefix *ni-*. Table 4 addresses this evaluation quantitatively.

Recognition performance	
Unique wordforms	**346**
Recall	346 (100%)
Precision	346 (100%)

Ambiguity profile	
Unambiguous	192 (55.5%)
Ambiguous (true positives)	154 (44.5%)
Total	**346**

Table 4: Evaluation of the NooJ Mwotlap module on *Mayanag t-Aliwan.not* (1711 tokens, 346 types): recall, precision, and ambiguity by source.

Disambiguation of these readings requires syntactic context that the current module does not yet have access to. For the scope of this study, this natural language-based evaluation confirms not only the correctness of the generated forms, but also the ability of the analyzer to recognize and parse naturally occurring prefixed word forms in running text. At the current stage of development, the module prioritizes maximal recall at the morphological level, returning all linguistically valid analyses for a given form.

Developing syntactic disambiguation grammars (.nog files), which are needed to filter competing analyses given their surrounding context, is a primary direction for future work.

7 Concluding remarks

Thanks to the creation of the new `<W>` operator, three superficially distinct linguistic processes can now be formalized as a single unified paradigm. In the case proposed here, vocalism in Mwotlap has a unitary sensitivity to the radical's phonological shape, not an arbitrary list of vowel-specific correspondences.

The three morphophonological processes described in this study are not specific to Mwotlap, as analogous phenomena are attested across typologically and genealogically unrelated languages (see Section 1). In all cases covered, the shared structural property is the same, and the new `<W>` operator (perform a swap of two adjacent letters), allows linguists to unify their formalization. It is therefore reasonable to expect that any language where morphophonological generativity involves local segment reordering, vowel spreading across a consonant boundary, or epenthesis conditioned by a following nucleus, could benefit from this operator, or a variant of it. Formalizing such phenomena would no longer require linguists to enumerate separate rules for each vowel quality, as the operator abstracts over segmental identity entirely. Future work should test this expectation against other languages.

Glossing conventions

1PL	1st person plural	ART	article
ANT	anterior	FV	final vowel
APPL	applicative	SUBJ	subject

References

Saba Amsalu and Girma A. Demeke. 2006. Non-concatenative finite-state morphotactics of Amharic simple verbs. *ELRC Working Papers*, *2*(3): 304-325.

Peter Barclay. 2026. A Rule-Based Computational Model for Gàidhlig Morphology. In *Proceedings of the 18th International Conference on Agents and Artificial Intelligence*, Volume 1. Scitepress, pages 683-693. https://doi.org/10.5220/0014654900004052

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications, Stanford, CA (distributed by the University of Chicago Press).

Lyle Campbell. 1974. Theoretical implications of Kekchi phonology. *International Journal of American Linguistics*, *40*(1): 269-278. https://www.jstor.org/stable/1265001

Noam Chomsky and Marcel P. Schützenberger. 1963. The algebraic theory of context free languages. In *Computer Programming and Formal Systems*. North Holland, pages 118-161.

Robert H. Codrington. 1885. *The Melanesian Languages*. Clarendon Press, Oxford, United Kingdom.

Yael Cohen-Sygal and Shuly Wintner. 2006. Finite-State Registered Automata for Non-Concatenative Morphology. *Computational Linguistics*, *32*(1):49–82. https://doi.org/10.1162/coli.2006.32.1.49

Terry Crowley. 2002. Mwotlap. In *The Oceanic Languages*. Curzon Press, pages 587-598.

ELP. n.d. *Motlav*. https://endangeredlanguages.com/elp-language/motlav

Axel Fleisch. 2000. *Lucazi: A Morphosemantic Analysis*. Köppe, Köln, Germany.

Alexandre François. 1999. Mouvements et clonages de voyelles en motlav: entre phonologie et morphologie. *Bulletin de la Société de Linguistique*, *94*(1): 437-486.

Alexandre François. 2000. Vowel shifting and cloning in Motlav: historical explanation vs. formal description. In *Proceedings of AFLA 7 (The Seventh Meeting of Austronesian Formal Linguistics Association)*. Vrije Universiteit Amsterdam, pages 49-68.

Alexandre François. 2001. *Contraintes de structures et liberté dans l'organisation du discours. Une description du mwotlap, langue océanienne du Vanuatu*. Doctoral Dissertation. Université Paris-IV Sorbonne, Paris, France.

Alexandre François. 2005. A typological overview of Mwotlap, an Oceanic language of Vanuatu. *Linguistic Typology*, *9*(1): 115-146. https://doi.org/10.1515/lity.2005.9.1.115

Alexandre François. 2012. The dynamics of linguistic diversity: Egalitarian Multilingualism and Power Imbalance among Northern Vanuatu Languages. *International Journal of the Sociology of Language*, *38*(214): 85-110. https://doi.org/10.1515/ijsl-2012-0022

Alexandre François. 2017. The Economy of Word Classes in Hiw, Vanuatu: Grammatically Flexible, Lexically Rigid. *Studies in Language*, *41*(2): 294-357. https://doi.org/10.1075/sl.41.2.03fra

Alexandre François. 2026. *A Mwotlap–English–French cultural dictionary*. CNRS, Paris, France. https://tiny.cc/Mwotlap-dict

Nancy Hall. 2006. Cross-linguistic patterns of vowel intrusion. *Phonology*, *23*(3): 387-429. https://doi.org/10.1017/S0952675706000996

Mans Hulden. 2009. Foma: a Finite-State Compiler and Library. In *Proceedings of the Demonstrations Session at EACL 2009*. Association for Computational Linguistics, pages 29-32.

Jacqueline Kasarhérou. 1962. Les changements vocaliques de trois préfixes en motlav. *Te reo*, 5(1): 32-34.

Martin Kay. 1987. Nonconcatenative Finite-State Morphology. In *Third Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 2-10.

Kimmo Koskenniemi. 1983. Two-level Model for Morphological Analysis. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, Volume 2. BRD, pages 683-685.

Paul R. Kroeger. 2005. *Analyzing Grammar: An Introduction*. Cambridge University Press, Cambridge, United Kingdom.

Lutz Marten and Nancy Chongo Kula. 2002. Semantic transparency in phonology: telicity and vowel copying in Dciriku. In *Phonologica, Proceedings of the 20th International Phonology Meeting*. Cambridge University Press, pages 1-17.

Fabio Meroni. 2026. Introducing Mwotlap Language Resources for NooJ. In *Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities. 19th International Conference, NooJ 2025, Paris, France, June 11–13, 2025, Revised Selected Papers*. Springer, pages 3-15. https://doi.org/10.1007/978-3-032-17103-0_1

Kenneth Olson. 2001. *The phonology and morphology of Mono*. Doctoral Dissertation. University of Chicago, Chicago, IL.

Femphy Pisceldo, Rahmad Mahendra, Ruli Manurung, and I Wayan Arka. 2008. A Two-Level Morphological Analyser for the Indonesian Language. In *Proceedings of the Australasian Language Technology Association Workshop 2008*. Association for Computational Linguistics, pages 142-150.

Max Silberztein. 2016. *Formalizing Natural Languages: the NooJ Approach*. Wiley, Hoboken, NJ.

Max Silberztein. 2024. Preface. In *Linguistic Resources for Natural Language Processing*. Springer Nature, pages xviii-xvi. https://doi.org/10.1007/978-3-031-43811-0

Max Silberztein. 2026. *NooJ v7 Manual*. Université Marie et Louis Pasteur, Besançon, France. https://nooj.univ-fcomte.fr/tutorials.html

Harald Trost. 1990. The application of two-level morphology to non-concatenative German morphology. In *COLING 1990*, Volume 2. Association for Computational Linguistics, pages 371-376.

Darrell T. Tryon. 1976. *New Hebrides Languages: An internal classification*. Pacific Linguistics, Canberra, Australia.

UNESCO. 2025. *World Atlas of Languages*. https://en.wal.unesco.org/

Shouly Wintner. 2014. Morphological Processing of Semitic Languages. In *Natural Language Processing of Semitic Languages. Theory and Applications of Natural Language Processing*. Springer, pages 43-66. https://doi.org/10.1007/978-3-642-45358-8_2

Dependency Parsing Across the Resource Spectrum: Evaluating Architectures on High and Low-Resource Languages

Kevin Guan Happy Buzaaba Christiane Fellbaum
Princeton University
{kevinguan, happy.buzaaba, fellbaum}@princeton.edu

Abstract

Transformer-based models achieve state-of-the-art dependency parsing for high-resource languages, yet their advantage over simpler architectures in low-resource settings remains poorly understood. We evaluate four parsers—the Biaffine LSTM, Stack-Pointer Network, AfroXLMR-large, and RemBERT—across twelve typologically diverse languages, with a focus on low-resource African languages. We find that the Biaffine LSTM consistently outperforms transformer models in low-resource regimes, with transformers recovering their advantage as training data increases. The crossover falls within a resource range typical of treebanks for under-resourced languages. Morphological complexity (measured via MATTR) emerges as a significant secondary predictor of transformers' relative disadvantage after controlling for corpus size. These results indicate that the Biaffine LSTM may be better suited for syntactic tool development in low-resource regimes until sufficient annotated data is available to leverage the representational capacity of pre-trained transformers.

1 Introduction

Dependency parsing—the task of identifying grammatical relations between words in a sentence—is a foundational task in NLP with broad downstream applications, including machine translation, information extraction, and semantic role labeling. Transformer-based models such as BERT and its multilingual variants now dominate dependency parsing benchmarks for high-resource languages, where abundant pre-training corpora and large treebanks allow these parameter-rich architectures to perform extremely well (Kondratyuk and Straka, 2019; Li et al., 2019; Grünewald et al., 2021).

This advantage, however, does not transfer uniformly to low-resource languages. Many languages—particularly African languages such as Xhosa and Wolof, spoken by millions but severely underrepresented in large text corpora—have dependency treebanks containing only a few hundred to a few thousand sentences, compared to tens of thousands available for English (Jha et al., 2020; Zhou et al., 2025). Pre-trained transformers require substantial in-domain data to fine-tune effectively, and the computational demands of fine-tuning present additional barriers in low-resource settings (Wang et al., 2020; Wu and Dredze, 2020).

We hypothesize that traditional architectures—particularly LSTM-based parsers—remain competitive with or outperform transformers when training data is scarce, for three reasons. First, **generalization**: transformer models contain orders of magnitude more learnable parameters than LSTM-based parsers. With limited training data, smaller models may generalize more reliably and overfit less (Gessler and Zeldes, 2022; Lindenmaier et al., 2025). Second, **inductive bias**: LSTM networks encode stronger built-in assumptions about sequential and hierarchical structure, which may serve as useful guidance when data is scarce (Tran et al., 2018; Dehghani et al., 2018). Third, **stability**: LSTM-based models tend to exhibit more consistent behavior across training runs, a meaningful advantage when working with small, noisy datasets (McCoy et al., 2020; Dodge et al., 2020).

To test this hypothesis, we systematically evaluate two LSTM-based parsers (Biaffine and Stack-Pointer) and two transformer-based parsers (AfroXLMR-large and RemBERT) across a diverse set of languages spanning a wide range of treebank sizes. Our central questions are: (1) Do LSTM-based parsers outperform transformer-based parsers on low-resource languages, while transformers regain their advantage at higher resource levels? (2) Does the transformers' relative error disadvantage decrease as training data grows? We measure performance using labeled and unla-

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 68–79
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_009

beled attachment scores (LAS/UAS), and we operationalize the transformer's relative advantage as its change in error rate normalized against the Biaffine LSTM baseline—a measure we call relative error rate (RER). This measure accounts for the fact that LAS gains are not comparable across languages and settings with very different baseline difficulties.

2 Related Work

Most African languages remain severely under-resourced in NLP due to their near-absence from web-crawled corpora such as Wikipedia and CommonCrawl, producing a digital language divide in which standard NLP pipelines fail to generalize to African linguistic contexts (Adelani, 2025; Bella et al., 2023; Nekoto et al., 2020).

Recent efforts to close this gap have proceeded along two complementary tracks: dataset construction and model adaptation. In the data domain, various initiatives have produced word embeddings (Adelani, 2022; Bojanowski et al., 2017), human-curated parallel corpora, and labeled benchmarks for machine translation (Nsumba et al., 2026; Kalejaiye et al., 2025), sentiment analysis (Muhammad et al., 2023), and news classification (Adelani et al., 2023) across a wide range of languages. On the architectural front, AfriBERTa (Ogueji, 2022) and AfroXLMR (Alabi et al., 2022) adapt multilingual transformer encoders to African morphology and orthography. AfriBERTa is trained from scratch, while AfroXLMR leverages continued pre-training on African-language corpora; both outperform general baselines such as mBERT and XLM-R on African NLP tasks. Masakhane, a grassroots research community, has been central to both efforts, providing named-entity and part-of-speech benchmarks across 20 languages through MasakhaNER (Adelani et al., 2022) and MasakhaPOS (Dione et al., 2023), among other contributions.

3 Data

Syntactic resources for African languages have developed more slowly than the broader benchmarks described above. Treebanks such as Wolof-WTB (Dione, 2019) and Naija-NSC (Caron et al., 2019) establish gold-standard dependency annotations for individual languages, but broad cross-lingual coverage has remained limited. AfriSUD (Buzaaba et al., 2026) represents a significant step forward, providing human-annotated treebanks in the Surface-Syntactic Universal Dependencies (SUD) framework across nine African languages. This collection forms the basis of the present study, with seven of the nine languages—Swahili, Kinyarwanda, Nigerian Pidgin, Wolof, Yoruba, Xhosa, and Hausa—included in our evaluation.

Languages and resource groupings. While the AfriSUD collection provides data for nine languages, we exclude Efik and Runyankore from our evaluation because pre-trained static word embeddings were unavailable. To span a broader resource spectrum and establish high-resource baselines, we supplement the remaining seven AfriSUD languages with five additional SUD treebanks: French, Dutch, Romanian, Turkish, and Afrikaans. This yields a final evaluation set of twelve languages: Swahili (swa), Kinyarwanda (kin), Nigerian Pidgin (pcm), Wolof (wol), Yoruba (yor), Hausa (hau), Xhosa (xho), French (fra), Dutch (nld), Romanian (ron), Turkish (tur), and Afrikaans (afr).

Treebank sources and splits. Our treebanks are SUD-converted versions of standard UD releases: French GSD (Abeillé et al., 2019), Alpino for Dutch (Bouma et al., 2001), Romanian RRT (Barbu Mititelu and Irimia, 2016), Turkish IMST (Sulubacak et al., 2016), AfriBooms for Afrikaans (Augustinus et al., 2016), and the AfriSUD collection (Buzaaba et al., 2026) for remaining languages. Splits follow an 80–10–10 ratio. Because the French and Dutch training sets are larger than needed to characterize high-resource behavior, we randomly subsample them to 10,000 sentences; dev and test sets remain unchanged. Sentence counts after subsampling are reported in Table 1.

Language	Code	Train	Dev	Test
French	fra	10,000*	1,634	1,634
Dutch	nld	10,000*	1,360	1,361
Romanian	ron	7,619	953	952
Turkish	tur	4,508	563	564
Afrikaans	afr	1,547	194	193
Swahili	swa	1,187	149	148
Kinyarwanda	kin	1,098	137	137
Nigerian Pidgin	pcm	1,085	135	136
Wolof	wol	736	92	92
Yoruba	yor	694	87	87
Hausa	hau	690	86	86
Xhosa	xho	369	46	46

Table 1: Treebank splits (sentence counts). *French and Dutch training sets subsampled from 13,074 and 10,882, respectively, to 10,000; dev/test sets unchanged.

Morphological complexity. To examine whether typological properties of the target language modulate architecture-level differences in parsing performance, we include morphological complexity as a covariate in our statistical models. We operationalize complexity via the Moving-Average Type–Token Ratio (MATTR) (Covington and McFall, 2010), which corrects the well-known sensitivity of raw type–token ratio to text length by averaging type–token ratios over a sliding window. Languages with richer morphology produce more distinct surface forms per window, yielding higher MATTR values. We use the gold tokenization provided by our treebanks and a window size of 500 tokens. MATTR scores are computed over each language's training split and reported in Table 2.

The resulting rankings align with established characterizations of morphological complexity. Xhosa, Turkish, and Kinyarwanda—languages with productive agglutinative morphology (Probert, 2019; Briscoe and Carroll, 1993; Nzeyimana, 2024)—rank highest, reflecting the large inventory of distinct surface forms generated by concatenative affixation. Romanian ranks fourth, consistent with its retention of nominal case morphology distinguishing it from its Romance relatives (Maiden et al., 2021). At the other extreme, Nigerian Pidgin and Afrikaans score lowest, in line with the morphological reduction characteristic of pidginization and Afrikaans's well-documented morphological simplification from Dutch (McWhorter, 2001; Donaldson, 1993).

Language	Tokens	MATTR	z-score
Xhosa	6,282	0.800	+2.241
Turkish	46,239	0.732	+1.389
Kinyarwanda	26,700	0.673	+0.658
Romanian	173,946	0.634	+0.175
Swahili	29,781	0.626	+0.076
Dutch	154,187	0.609	−0.130
Yoruba	10,422	0.586	−0.409
French	244,944	0.570	−0.617
Wolof	15,415	0.567	−0.651
Hausa	13,874	0.566	−0.662
Nigerian Pidgin	26,489	0.541	−0.974
Afrikaans	39,199	0.531	−1.095

Table 2: MATTR scores and standardized values, computed over training splits with window size $w = 500$.

4 Models

We evaluate three primary dependency parsing architectures: the Biaffine LSTM parser (Dozat and Manning, 2017), transformer-based biaffine parsers, and the Stack-Pointer Network (Ma et al., 2018). The biaffine parsers are implemented using the SuPar Python package (Zhang, 2024), and the Stack-Pointer parsers are implemented using the NeuroNLP2 repository (Ma, 2024).

Biaffine LSTM. The deep biaffine attention parser (Dozat and Manning, 2017) is our graph-based baseline and the de facto standard for neural dependency parsing. It encodes sentences with a BiLSTM and scores arcs via a biaffine function, decoding the optimal tree with the Chu-Liu/Edmonds algorithm (Chu, 1965; Edmonds, 1967). We use this model with static FastText embeddings.

Stack-Pointer Network. The Stack-Pointer Network (Ma et al., 2018) is a transition-based parser with a BiLSTM-CNN encoder that decodes top-down and depth-first via a pointer network, incorporating ancestor and sibling features for higher-order context. This approach avoids the directional bias of traditional transition parsers while remaining more efficient than MST-based decoding.

AfroXLMR-large. AfroXLMR-large (Alabi et al., 2022) is a domain-adapted variant of XLM-R (Conneau et al., 2020) that continues pre-training on 17 African languages, yielding richer representations for the morphologically complex languages in our evaluation set. It is paired with SuPar's biaffine parsing head and fine-tuned end-to-end.

RemBERT. RemBERT (Chung et al., 2020) is a multilingual transformer encoder that decouples input and output embeddings, allocating a larger input layer to better handle morphologically rich languages. Like AfroXLMR-large, it is paired with a biaffine head in SuPar and fine-tuned jointly during training.

5 Experiments

Hyperparameter optimization. We tune each model via grid search with a fixed seed. To keep the search tractable, high-resource languages are downsampled to 5,000 training sentences, and models are trained at 25% of the full epoch budget with halved early-stopping patience. For the Biaffine LSTM we search over learning rate, decay rate, and decay steps (12 runs per language); for the transformer-based models (AfroXLMR-large and RemBERT) we search over learning rate, warmup steps, and the ratio between the encoder and task-head learning rates (18 runs); and for Stack-Pointer

we additionally tune unknown-word replacement probability and gradient clipping (24 runs). All other hyperparameters are kept at default values as defined by the respective implementation libraries.

Word embeddings. The Biaffine LSTM and Stack-Pointer models use 300-dimensional pretrained static word vectors. For languages in the official FastText distribution (Grave et al., 2018) we use vectors trained on CommonCrawl and Wikipedia. For African languages not covered there, we fall back to embeddings provided by Adelani's AfricaNLP resources repository (Adelani, 2022). As noted in Section 3, Efik and Runyankore are excluded from our analysis because no pretrained vectors could be located for either language. Static vectors are concatenated with character-level representations (CNN or LSTM) and gold POS tag embeddings. The transformer-based models do not use static vectors; they generate contextualized subword representations internally.

Training protocol. Final models are trained on the full data with the best hyperparameters. Biaffine models run for up to 1,000 epochs (patience 100); Stack-Pointer models run for up to 600 epochs (patience 60). In practice, no model approached its epoch ceiling before early stopping triggered. Transformer models on high-resource languages (Romanian, Dutch, French) are additionally capped at 100 epochs to balance computational costs, as these models require fewer passes to converge due to their large volumes of data. Each configuration is trained five times with different random seeds, and we report mean UAS and LAS across runs to reduce sensitivity to initialization.

6 Results

6.1 Parsing Performance

Tables 3a and 3b report mean LAS and UAS with standard deviations across five training runs for all four architectures and twelve languages. Overall, performance scales with training set size, though the relative ordering of architectures varies by language. The transformer models outperform the Biaffine LSTM on higher-resource languages, while the LSTM is competitive with or superior to both transformers on several lower-resource ones. The Stack-Pointer parser is the weakest architecture across nearly all languages and metrics.

We quantify architecture-level differences using the *relative error rate* (RER). Let $\text{LAS}_{\text{Biaffine LSTM}}(\ell)$ and $\text{LAS}_{\text{TF}}(\ell)$ denote the labeled attachment scores of the Biaffine LSTM baseline and a transformer comparison model on language ℓ. We define:

$$RER_{\text{LAS}}(\ell) = \frac{\text{LAS}_{\text{Biaffine LSTM}}(\ell) - \text{LAS}_{\text{TF}}(\ell)}{100 - \text{LAS}_{\text{Biaffine LSTM}}(\ell)} \quad (1)$$

and likewise for unlabeled attachment (RER_{UAS}). We prefer RER over raw score differences for two reasons. First, absolute LAS or UAS gains are confounded with baseline difficulty: a 2-point gain at a baseline of 95 is not comparable to the same gain at a baseline of 70, as the remaining headroom differs substantially. Second, because attachment scores are bounded at 100, raw differences are heteroscedastic across the training-size axis, with low-resource languages exhibiting lower baselines and higher variance; normalizing against the remaining error stabilizes variance across languages. A value $RER_{\text{LAS}} > 0$ indicates that the comparison model incurs more errors than the Biaffine LSTM, while $RER_{\text{LAS}} < 0$ indicates fewer errors.

Tables 4 and 5 report RER_{LAS} and RER_{UAS} for each comparison model. Both transformer models show $RER_{\text{LAS}}, RER_{\text{UAS}} < 0$ (fewer errors than the LSTM) on French, Romanian, and Afrikaans, but $RER > 0$ on Xhosa, Wolof, and—for RemBERT—Swahili. Looking only at RER_{LAS}, the pattern of LSTM's competitiveness extends further. AfroXLMR shows positive RER_{LAS} on Swahili ($+0.049$) in addition to Xhosa ($+0.148$) and Wolof ($+0.077$). RemBERT's label-level underperformance is even more extensive: beyond Swahili ($+0.181$), Wolof ($+0.130$), and Xhosa ($+0.314$), it records positive RER_{LAS} on Yoruba ($+0.071$) and Kinyarwanda ($+0.042$). The Stack-Pointer parser consistently underperforms, with $RER > 0$ on nearly every language. Figure 1 plots RER_{LAS} and RER_{UAS} for each non-baseline model across languages ordered by training set size.

6.2 Training Data Scaling

To test whether RER is systematically related to training set size, we fit linear mixed models predicting RER_{LAS} and RER_{UAS} as a function of log training sentences, with language as a random intercept. We restrict this analysis to the two transformer models, since the Stack-Pointer parser is consistently outperformed by the LSTM baseline, showing $RER > 0$ on nearly every language.

Language	Biaffine LSTM	AfroXLMR-large	RemBERT	Stack-Pointer
French (fra)	93.51 ± 0.13	95.81 ± 0.05	95.53 ± 0.08	92.41 ± 0.05
Dutch (nld)	91.07 ± 0.12	94.55 ± 0.15	94.50 ± 0.13	90.39 ± 0.21
Romanian (ron)	87.23 ± 0.18	90.61 ± 0.12	90.51 ± 0.06	87.41 ± 0.12
Turkish (tur)	67.35 ± 0.52	72.76 ± 0.56	73.07 ± 0.49	63.88 ± 0.28
Afrikaans (afr)	85.85 ± 0.31	88.90 ± 0.28	87.75 ± 0.19	85.91 ± 0.61
Swahili (swa)	79.48 ± 0.43	78.48 ± 0.49	75.77 ± 0.53	73.10 ± 0.95
Kinyarwanda (kin)	76.16 ± 0.30	76.21 ± 0.43	75.16 ± 0.94	74.76 ± 0.35
Nigerian Pidgin (pcm)	70.56 ± 0.51	72.12 ± 0.42	71.81 ± 0.52	68.65 ± 0.77
Wolof (wol)	79.63 ± 0.28	78.06 ± 0.76	76.99 ± 0.42	76.27 ± 0.76
Yoruba (yor)	77.71 ± 0.66	77.75 ± 0.81	76.13 ± 0.42	75.33 ± 0.45
Hausa (hau)	79.66 ± 0.61	80.75 ± 0.63	79.75 ± 1.25	78.55 ± 0.60
Xhosa (xho)	70.38 ± 0.67	66.00 ± 0.80	61.08 ± 1.72	58.20 ± 1.30

(a) Mean Labeled Attachment Score (LAS) $\pm$ standard deviation.

Language	Biaffine LSTM	AfroXLMR-large	RemBERT	Stack-Pointer
French (fra)	95.53 ± 0.08	97.30 ± 0.03	97.17 ± 0.07	94.69 ± 0.08
Dutch (nld)	93.85 ± 0.16	96.34 ± 0.12	96.37 ± 0.07	93.24 ± 0.22
Romanian (ron)	92.45 ± 0.14	95.38 ± 0.08	95.26 ± 0.04	92.01 ± 0.11
Turkish (tur)	75.98 ± 0.23	82.30 ± 0.37	82.86 ± 0.30	72.40 ± 0.37
Afrikaans (afr)	88.74 ± 0.31	91.37 ± 0.33	90.42 ± 0.01	88.68 ± 0.59
Swahili (swa)	88.57 ± 0.37	89.39 ± 0.22	87.54 ± 0.57	84.92 ± 0.82
Kinyarwanda (kin)	86.44 ± 0.24	86.59 ± 0.26	86.11 ± 0.54	85.71 ± 0.32
Nigerian Pidgin (pcm)	80.41 ± 0.72	82.13 ± 0.30	81.83 ± 0.44	78.12 ± 0.48
Wolof (wol)	85.79 ± 0.16	84.88 ± 0.76	84.19 ± 0.47	82.83 ± 0.77
Yoruba (yor)	85.46 ± 0.41	85.97 ± 0.67	84.25 ± 0.39	83.69 ± 0.35
Hausa (hau)	92.30 ± 0.49	92.32 ± 0.29	91.81 ± 0.74	91.47 ± 0.74
Xhosa (xho)	81.70 ± 0.52	76.78 ± 1.07	72.78 ± 1.69	69.80 ± 1.44

(b) Mean Unlabeled Attachment Score (UAS) $\pm$ standard deviation.

Table 3: Parsing performance results across five seeds for LAS and UAS.

Language	AfroXLMR	RemBERT	Stk-Ptr
French	-0.354	-0.311	0.169
Dutch	-0.390	-0.384	0.076
Romanian	-0.265	-0.257	-0.014
Turkish	-0.166	-0.175	0.106
Afrikaans	-0.216	-0.134	-0.004
Swahili	0.049	0.181	0.311
Kinyarwanda	-0.002	0.042	0.059
Nigerian Pidgin	-0.053	-0.042	0.065
Wolof	0.077	0.130	0.165
Yoruba	-0.002	0.071	0.107
Hausa	-0.054	-0.004	0.055
Xhosa	0.148	0.314	0.411

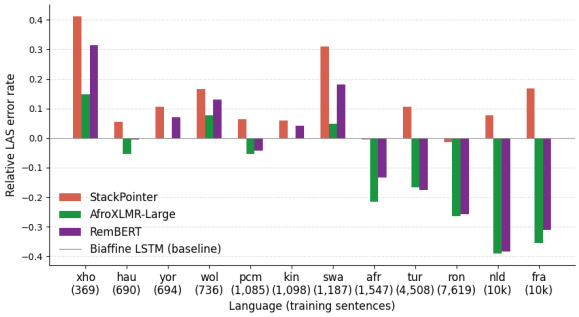
Table 4: RER_{LAS} compared to the Biaffine LSTM baseline. Values $RER_{\text{LAS}} < 0$ favor the comparison model; values $RER_{\text{LAS}} > 0$ favor the LSTM.

Language	AfroXLMR	RemBERT	Stk-Ptr
French	-0.396	-0.367	0.188
Dutch	-0.405	-0.410	0.100
Romanian	-0.388	-0.372	0.059
Turkish	-0.263	-0.286	0.149
Afrikaans	-0.234	-0.149	0.006
Swahili	-0.072	0.090	0.320
Kinyarwanda	-0.011	0.024	0.054
Nigerian Pidgin	-0.088	-0.072	0.117
Wolof	0.064	0.113	0.208
Yoruba	-0.035	0.083	0.122
Hausa	-0.003	0.064	0.108
Xhosa	0.269	0.487	0.650

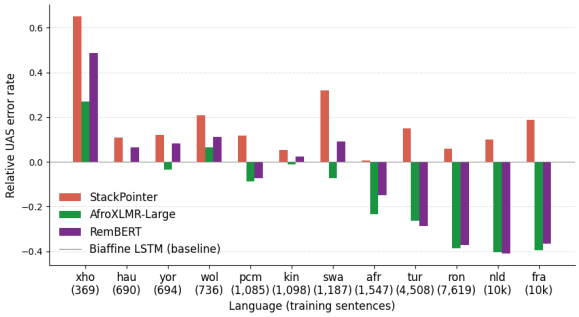
Table 5: RER_{UAS} compared to the Biaffine LSTM baseline. Values $RER_{\text{UAS}} < 0$ favor the comparison model; values $RER_{\text{UAS}} > 0$ favor the LSTM.

Both models show a statistically significant negative relationship between training size and RER—RER decreases as more data becomes available. For LAS, the slope on `log_train` is $\hat{\beta} = -0.388$ ($p < .001$) for RemBERT and $\hat{\beta} = -0.321$ ($p < .001$) for AfroXLMR-large; the corresponding UAS slopes are steeper at $\hat{\beta} = -0.496$ and $\hat{\beta} = -0.398$ (both $p < .001$). Spearman correlations between training size and mean RER are negative and significant across all model–metric combinations, corroborating this pattern under minimal distributional assumptions. For RemBERT, $\rho_{\text{LAS}} = -0.830$ ($p < .001$) and $\rho_{\text{UAS}} = -0.869$ ($p < .001$); for AfroXLMR-large, $\rho_{\text{LAS}} = -0.844$ ($p < .001$) and $\rho_{\text{UAS}} = -0.935$ ($p < .001$). Fitted curves for LAS are shown in Figure 2.

The fitted curves cross $RER = 0$—the point at which a transformer's predicted error equals the

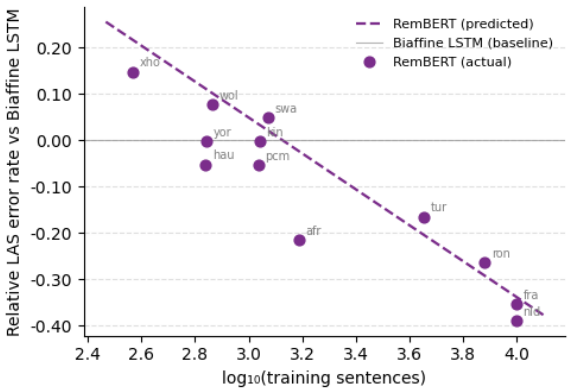


(a) RER_{LAS}

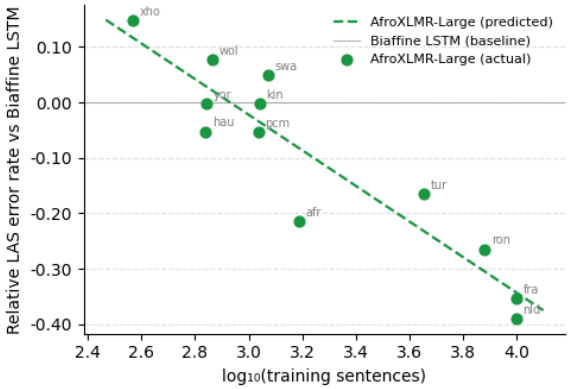


(b) RER_{UAS}

Figure 1: RER vs. sentence count. $RER < 0$ indicates fewer errors than Biaffine LSTM baseline.



(a) RER_{LAS}: RemBERT vs. Biaffine LSTM



(b) RER_{LAS}: AfroXLMR-large vs. Biaffine LSTM

Figure 2: RER_{LAS} as a function of training size. Plots for RER_{UAS} are omitted for brevity, although the trends are qualitatively identical, with RER_{UAS} decreasing as training data grows.

LSTM's—at the sentence counts summarized in Table 6. AfroXLMR-large reaches the crossover at around 830-860 sentences, while RemBERT requires approximately 1,310–1,340. Several languages in our evaluation fall close to these thresholds: Hausa (690), Yoruba (694), and Wolof (736) sit just below the AfroXLMR-large crossover, while Swahili (1,187), Nigerian Pidgin (1,085), and Kinyarwanda (1,098) fall between the estimates.

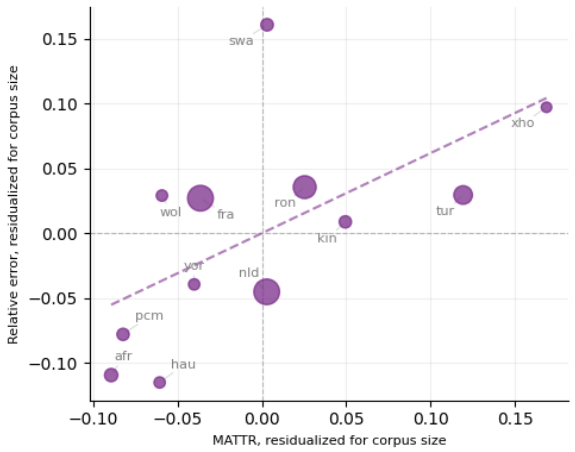
Model	Metric	$\log_{10}$(train)	Sents.
AfroXLMR-large	LAS	2.931	854
RemBERT	LAS	3.127	1,339
AfroXLMR-large	UAS	2.924	839
RemBERT	UAS	3.118	1,311

Table 6: Estimated training-set sizes at which fitted RER curve crosses $RER = 0$, i.e. where transformers' predicted error equals the Biaffine LSTM's.

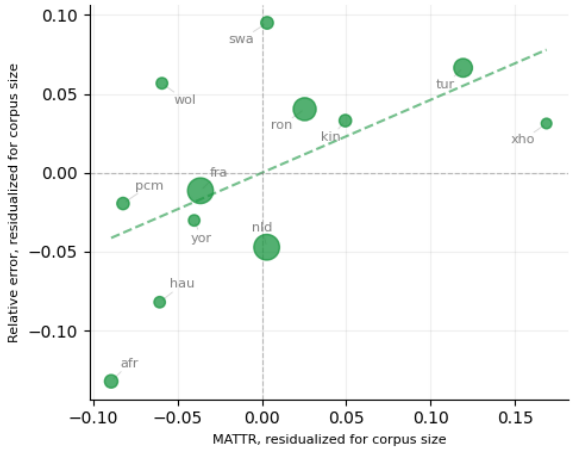
6.3 Morphological Complexity

To assess whether morphological complexity explains residual variance in architecture-level differences beyond what training size captures, we extend the mixed models by adding MATTR as a predictor alongside log training sentences, and compare model fit with and without this term.

For RemBERT, adding MATTR significantly improves fit on both LAS ($\chi^2(1) = 5.44$, $p = .020$) and UAS ($\chi^2(1) = 8.65$, $p = .003$), with positive `mattr_z` coefficients of $\hat{\beta} = 0.050$ ($p = .023$) and $\hat{\beta} = 0.065$ ($p = .002$) respectively: higher morphological complexity is associated with a positive shift in RER, even after controlling for data size. Similarly, for AfroXLMR-large, both the LAS model ($\chi^2(1) = 4.51$, $p = .034$; $\hat{\beta} = 0.037$, $p = .043$) and the UAS model ($\chi^2(1) = 9.89$, $p = .002$; $\hat{\beta} = 0.048$, $p = .001$) show a significant effect. The magnitude of the MATTR effect is somewhat smaller for AfroXLMR than for RemBERT on both metrics. It is possible that AfroXLMR's African-language pretraining partially offsets the elevated RER that complex morphology imposes on transformer models, though this was not formally tested. In all extended models, the negative effect of training size on RER remains significant. Figure 3 shows partial regression plots in which variables have been residualized for training size, isolating the complexity effect.



(a) RemBERT: RER_{LAS}



(b) AfroXLMR-large: RER_{LAS}

Figure 3: Partial regression of RER_{LAS} on MATTR for RemBERT (top) and AfroXLMR-large (bottom), after residualizing both variables for training size. Partial regression plots for RER_{UAS} are omitted for brevity, although the relationship of RER_{UAS} with morphological complexity is consistent with that of RER_{LAS} and leads to the same conclusions.

7 Discussion

7.1 Scaling Behavior Across Architectures

Across both transformer models, RER decreases as training corpus size grows, with transformers achieving $RER \leq 0$ once treebanks reach roughly 800–1,400 training sentences. This pattern is consistent with standard fine-tuning dynamics: multilingual pre-trained models carry useful cross-lingual syntactic representations, but adapting those representations to a specific treebank's annotation conventions requires sufficient supervision. With little training data, fine-tuning is underconstrained—the model lacks sufficient signal to effectively map its pre-trained representations to the specific structural conventions of the target treebank. The Biaffine LSTM, learning its representations entirely from the treebank, avoids this problem; its inductive biases toward locality and sequential composition suit the short-range dependencies that dominate treebank data. As supervision increases, transformers gain the signal needed to resolve this ambiguity, RER crosses zero, and their cross-lingual transfer becomes advantageous.

The estimated crossover points are practically significant: a substantial fraction of African UD treebanks fall within the 839–1,339 sentence range, meaning the LSTM remains the stronger choice for many under-resourced languages in practice. RER_{UAS} shows a steeper scaling trajectory than RER_{LAS}. We attribute this to structural attachment being more typologically general and therefore more transferable via pre-training, while label assignment depends more heavily on language-specific annotation conventions that neither scale nor pre-training fully resolves.

7.2 Morphological Complexity and the Low-Resource Penalty

Even after controlling for corpus size, morphological complexity predicts a consistent positive shift in RER for both transformer models. Morphologically rich languages produce a larger inventory of distinct inflected forms, compounding the difficulty of learning an accurate parser from limited data. The effect is weaker for AfroXLMR-large—its MATTR coefficient is roughly 1.35 times smaller than RemBERT's—which is consistent with its continued pre-training on African-language text providing richer coverage of inflectional paradigms, partially offsetting the morphological component of RER.

In absolute terms the MATTR effect is modest, and interpretation is complicated by Xhosa's status as the smallest treebank and most extreme MATTR outlier, giving it disproportionate leverage over the complexity estimates. We also note that MATTR captures lexical diversity rather than morphological complexity directly. More principled operationalizations—such as morpheme-per-word ratios or paradigm entropy—would provide a stronger test in future work. While lemmatizing the treebanks could theoretically mitigate this morphological penalty, robust lemmatizers for many of these languages remain unavailable.

7.3 Stack-Pointer Underperformance

The Stack-Pointer parser shows $RER > 0$ on nearly every language, suggesting that its architectural complexity is counterproductive in this data regime. Its top-down, depth-first decoding introduces error propagation, since mistakes at high nodes cascade to descendants—an effect that is exacerbated when data is scarce. The pointer network decoder also adds substantially more parameters than the Biaffine LSTM, increasing the risk of overfitting on small treebanks, and the higher-order sibling and grandparent features that distinguish this architecture depend on prior parsing decisions that may themselves be unreliable under low-resource conditions.

7.4 Broader Limitations

Several factors limit the generalizability of our conclusions. (1) Our treebanks differ in domain and genre, and some of the observed RER differences may reflect genre mismatch rather than resource level. (2) For non-transformer models, the quality of word embeddings introduces a confound that remains unquantified in our crossover estimates, since higher-resource languages benefit from embeddings trained on larger corpora. (3) Our transformer baselines are all multilingual—languages with monolingual pre-trained encoders available may show different crossover behavior. (4) Typological differences in word order and non-projectivity affect parsing difficulty independently of resource level and cannot be fully separated from it in our sample. (5) The test set for our lowest-resource language contains fewer than 50 sentences, limiting the statistical precision of reported means. A larger and more diverse language sample would provide more robust grounding for our claims.

8 Conclusion

This work examined whether transformer-based parsers maintain their advantage over recurrent architectures when training data is scarce. Across twelve typologically diverse languages spanning a wide range of resource levels, our results consistently show $RER > 0$ for both AfroXLMR-large and RemBERT on the lowest-resource languages, while RER decreases with training data until transformers achieve $RER \leq 0$ at higher resource levels. Statistical analysis confirms a significant negative relationship between training set size and RER, and morphological complexity—operationalized via MATTR—emerges as a secondary predictor of elevated RER: higher morphological richness compounds the transformer's error disadvantage, an effect that is attenuated for AfroXLMR-large, likely due to its pre-training on African-language text. These findings suggest that data scarcity and morphological richness jointly constrain the effectiveness of multilingual transformer fine-tuning in low-resource conditions.

The LAS scores on the lowest-resource languages remain lower than most downstream applications require, but they nonetheless carry practical value. Imperfect parsers may still assist with annotation, reducing the clerical burden on human annotators and accelerating treebank development. Parsing accuracy also functions as a sensitive diagnostic of the overall NLP infrastructure available for a language. Taken together, these results suggest that the Biaffine LSTM may be better suited for syntactic tool development in under-resourced languages—for which $RER > 0$ favors it over transformer-based models—until sufficient annotated data is available to leverage the representational capacity of large pre-trained models.

Future work. Several directions would strengthen and extend these findings. The precise $RER = 0$ crossover could be localized by systematically subsampling larger treebanks to produce controlled, within-language training sets of varying sizes. Whether the crossover generalizes beyond dependency parsing to structurally simpler tasks such as POS tagging and NER, and whether zero-shot or few-shot cross-lingual transfer can push it toward lower resource levels, are open questions. The evaluation should also be extended to morphologically rich agglutinative languages—Turkic languages, Japanese, Korean, and Indigenous languages of the Americas—which are not represented in the current study, and to architectures such as the TreeCRF parser (Zhang et al., 2020), which has been reported to perform well under low-resource conditions. Finally, more principled measures of morphological complexity—such as information-theoretic metrics—would allow a cleaner test of whether the elevated RER for transformers is specifically driven by inflectional disambiguation load, and evaluating these architectures on lemmatized data, where available, could test if neutralizing inflectional variance mitigates this MATTR confound.

Acknowledgments

This work was supported by the Princeton Language and Intelligence (PLI) Seed Grant Program. We thank Janet Pierrehumbert for the conversation that led to the central hypothesis on which this work is based.

References

Anne Abeillé, Lionel Clément, and Loïc Liégeois. 2019. Un corpus arboré pour le français : le French Treebank [A parsed corpus for French: the French treebank]. *Traitement Automatique des Langues*, 60(2):19–43.

David Ifeoluwa Adelani. 2022. AfricaNLP resources. https://github.com/dadelani/africanlp-resources. Accessed: 2026-04-27.

David Ifeoluwa Adelani. 2025. Natural language processing for african languages. *arXiv preprint arXiv:2507.00297*.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, Sana Al-azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, Abraham Owodunni, Nnaemeka Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeb Afolabi, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Ogbu, Chinedu Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede, Toadoum Sari Sakayo, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Kanda Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos Nigusse, Abdulmejid Johar, Shafie Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. 2023. MasakhaNEWS: News topic classification for African languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 144–159, Nusa Dua, Bali. Association for Computational Linguistics.

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba O. Alabi, Shamsuddeen H. Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Elvis Mboning, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo L. Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Adeyemi, Gilles Q. Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu Ngoli, and Dietrich Klakow. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Liesbeth Augustinus, Peter Dirix, Daniel van Niekerk, Ineke Schuurman, Vincent Vandeghinste, Frank Van Eynde, and Gerhard van Huyssteen. 2016. AfriBooms: An online treebank for Afrikaans. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 677–682, Portorož, Slovenia. European Language Resources Association (ELRA).

Verginica Barbu Mititelu and Elena Irimia. 2016. Linguistic data retrievable from a treebank. In *Proceedings of the Second International Conference on Computational Linguistics in Bulgaria (CLIB 2016)*, pages 19–27, Sofia, Bulgaria. Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences.

Gábor Bella et al. 2023. Towards bridging the digital language divide. *arXiv preprint arXiv:2307.13405*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Gosse Bouma, Gertjan van Noord, and Robert Malouf. 2001. Alpino: Wide-coverage computational analysis of dutch. In *Computational Linguistics in the Netherlands 2000*, volume 37 of *Language and Computers: Studies in Practical Linguistics*, pages 45–59, Amsterdam. Rodopi.

Ted Briscoe and John Carroll. 1993. Generalised probabilistic LR parsing of natural language (corpora)

with unification-based grammars. In *Sixth Conference of the European Chapter of the Association for Computational Linguistics*, Utrecht, The Netherlands. Association for Computational Linguistics.

Happy Buzaaba, Cheikh Mouhamadou Bamba Dione, David Ifeoluwa Adelani, Sylvain Kahane, Kim Gerdes, Bruno Guillaume, Kevin Guan, Aremu Anuoluwapo, Naome A Etori, Shamsuddeen Hassan Muhammad, et al. 2026. Afrisud: A dependency treebank collection for evaluating models on african languages. *arXiv preprint arXiv:2606.12708*.

Bernard Caron, Marine Courtin, Kim Gerdes, and Sylvain Kahane. 2019. A surface-syntactic UD treebank for Naija. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, Paris, France. Association for Computational Linguistics.

Yoeng-Jin Chu. 1965. On the shortest arborescence of a directed graph. *Scientia Sinica*, 14:1396–1400.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. *arXiv preprint arXiv:2010.12821*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Michael A. Covington and Joe D. McFall. 2010. Cutting the gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. 2018. Universal transformers. *arXiv preprint arXiv:1807.03819*. Published at ICLR 2019.

Cheikh Bamba Dione. 2019. Developing Universal Dependencies for Wolof. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 12–23, Paris, France. Association for Computational Linguistics.

Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jonathan Mukiibi, Blessing Sibanda, Bonaventure F. P. Dossou, Andiswa Bukula, Rooweither Mabuya, Allahsera Auguste Tapo, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Fatoumata Ouoba Kabore, Amelia Taylor, Godson Kalipe, Tebogo Macucwa, Vukosi Marivate, Tajuddeen Gwadabe, Mboning Tchiaze Elvis, Ikechukwu Onyenwe, Gratien Atindogbe, Tolulope Adelani, Idris Akinade, Olanrewaju Samuel, Marien Nahimana, Théogène Musabeyezu, Emile Niyomutabazi, Ester Chimhenga, Kudzai Gotosa, Patrick Mizha, Apelete Agbolo, Seydou Traore, Chinedu Uchechukwu, Aliyu Yusuf, Muhammad Abdullahi, and Dietrich Klakow. 2023. MasakhaPOS: Part-of-speech tagging for typologically diverse African languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10883–10900, Toronto, Canada. Association for Computational Linguistics.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

Bruce C. Donaldson. 1993. *A Grammar of Afrikaans*. Number 8 in Mouton Grammar Library. Mouton de Gruyter, Berlin and New York.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *Proceedings of the 5th International Conference on Learning Representations*.

Jack Edmonds. 1967. Optimum branchings. *Journal of Research of the National Bureau of Standards B*, 71(4):233–240.

Luke Gessler and Amir Zeldes. 2022. MicroBERT: Effective training of low-resource monolingual BERTs through parameter reduction and multitask learning. In *Proceedings of the 2nd Workshop on Multi-lingual Representation Learning (MRL)*, pages 86–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Stefan Grünewald, Annemarie Friedrich, and Jonas Kuhn. 2021. Applying occam's razor to transformer-based dependency parsing: What works, what doesn't, and what is really necessary. In *Proceedings of the 17th International Conference on Parsing Technologies (IWPT 2021)*, pages 131–144, Bangkok, Thailand.

Prabhat Kumar Jha, Rajesh Kumar, and Vikram Sahula. 2020. A survey of the model transfer approaches to cross-lingual dependency parsing. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 19(3):1–39.

Oluwadara Kalejaiye, Luel Hagos Beyene, David Ifeoluwa Adelani, Mmekut-mfon Gabriel Edet, Aniefon Daniel Akpan, Eno-Abasi Urua, and Anietie Andy. 2025. Ibom NLP: A step toward inclusive

natural language processing for Nigeria's minority languages. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Ying Li, Zhenghua Li, Min Zhang, Rui Wang, Sheng Li, and Luo Si. 2019. Self-attentive biaffine dependency parsing. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, pages 5067–5073, Macao.

Gabriel Lindenmaier, Sean Papay, and Sebastian Padó. 2025. Efficient language modeling for low-resource settings with hybrid RNN-Transformer architectures. *arXiv preprint arXiv:2502.00617*.

Xuezhe Ma. 2024. NeuroNLP2: Deep neural models for core NLP tasks (Pytorch version).

Xuezhe Ma, Zecong Hu, Jingzhou Liu, Nanyun Peng, Graham Neubig, and Eduard Hovy. 2018. Stack-pointer networks for dependency parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1403–1414, Melbourne, Australia. Association for Computational Linguistics.

Martin Maiden, Adina Dragomirescu, Gabriela Pană Dindelegan, Oana Uță, and Rodica Zafiu. 2021. *The Oxford History of Romanian Morphology*. Oxford University Press, Oxford.

R. Thomas McCoy, Robert Frank, and Tal Linzen. 2020. Does syntax need to grow on trees? Sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics*, 8:125–140.

John H. McWhorter. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology*, 5:125–166.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Alipio Jorge, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Baye Messelle, Hailu Beshada Balcha, Sisay Adugna Chala, Hagos Tesfahun Gebremichael, Bernard Opoku, and Stephen Arthur. 2023. AfriSenti: A Twitter sentiment analysis benchmark for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13968–13981, Singapore. Association for Computational Linguistics.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Solomon Nsumba, Benjamin Akera, Evelyn Nafula Ouma, Medadi Ssentanda, Deo Kawalya, Engineer Bainomugisha, Ernest Tonny Mwebaze, and John Quinn. 2026. SALT-31: A machine translation benchmark dataset for 31 Ugandan languages. In *Proceedings of the 7th Workshop on African Natural Language Processing (AfricaNLP 2026)*, pages 211–216. Association for Computational Linguistics.

Antoine Nzeyimana. 2024. Low-resource neural machine translation with morphological modeling. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 182–195, Mexico City, Mexico. Association for Computational Linguistics.

Kelechi Ogueji. 2022. Afriberta: Towards viable multilingual language models for low-resource languages. Master's thesis, University of Waterloo, Waterloo, Ontario, Canada, August.

Tracy N. Probert. 2019. A comparison of the early reading strategies of isiXhosa and Setswana first language learners. *South African Journal of Childhood Education*, 9(1).

Umut Sulubacak, Memduh Gokirmak, Francis Tyers, Çağrı Çöltekin, Joakim Nivre, and Gülşen Eryiğit. 2016. Universal dependencies for turkish. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3444–3454, Osaka, Japan. The COLING 2016 Organizing Committee.

Ke Tran, Arianna Bisazza, and Christof Monz. 2018. The importance of being recurrent for modeling hier-

archical structure. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4731–4736, Brussels, Belgium. Association for Computational Linguistics.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? *arXiv preprint arXiv:2005.09093*.

Yu Zhang. 2024. SuPar: State-of-the-art Parsers for Natural Language. Python package.

Yu Zhang, Zhenghua Li, and Min Zhang. 2020. Efficient second-order TreeCRF for neural dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3295–3305, Online. Association for Computational Linguistics.

He Zhou, Daniel Dakota, and Sandra Kübler. 2025. Cross-lingual dependency parsing for a language with a unique script. *Natural Language Processing*, 31(2):277–305.

A finite-state model of Innu verbal inflection

Loïc Daignault-Pichette
OLST, Université de Montréal
loic.daignault-pichette@umontreal.ca

François Lareau
OLST, Université de Montréal
francois.lareau@umontreal.ca

Abstract

This article describes a finite-state transducer (FST) of Innu verbal inflection. Unlike approaches that model word formation by concatenating unanalysed inflectional chunks to a root, the proposed model is linguistically motivated and reflects the largely agglutinative nature of Innu verbal morphology. Our implementation relies on a sequence of intermediate representations, making it modular. It covers 19,628 unique forms for 68 lemmas and it has been evaluated against data from the most comprehensive resources currently available. We argue that FSTs can serve not only as a practical natural language processing tool, but also as a means of testing the descriptive adequacy of a morphological analysis. We discuss the challenges encountered in developing this model.

1 Introduction

Finite-state transducers (FSTs) have been used for morphology in natural laage processing (NLP) since the mid-1980s and have been applied to a wide variety of languages (Karttunen and Beesley, 2005). A morphological FST represents a set of pairs of symbol strings mapping a *deep level* to a *surface level*,[1] either for synthesis or analysis. Analysis maps a word to its valid analyses. Synthesis is the inverse: it maps an abstract morphological representation to the corresponding word. These mappings can be one-to-many, thus modelling ambiguity and homography. A FST can also be made productive, meaning it can analyze or synthesize new forms, independently of its internal lexicon. We have set out to create an open-source morphological model of this type to help develop computational resources for Innu.[2] The value of this model lies in its subject and its structure.

The subject. The model targets Innu, an Algonquian language spoken in the eastern part of the Québec-Labrador peninsula with approximately 10,000 speakers (Tanner, 2025). Models have been created for other Algonquian languages (Hammerly et al., 2026; Bowers et al., 2017; Harrigan et al., 2017; Davis et al., 2021; Kazeminejad et al., 2017) and for other Indigenous languages of North America (Arppe et al., 2016; Forbes et al., 2021; Lachler et al., 2018), but to our knowledge, only two automatic verb conjugators have been developed for Innu—one of which has not been published—(Junker and Stewart, 2008), and none of them are still in use. Our model is therefore a valuable resource for the language processing of Innu.

The model presented here only deals with inflectional morphology, not with derivational morphology. In Innu, as in many other languages, derivational morphology is much more idiosyncratic and therefore difficult to generalize and implement. Also, although some of it is suitable for implementation, much of it is either not well documented or sensitive to factors such as vowel length, which are not accounted for in the linguistic resources currently available.

The structure. The structure of the model was designed to simplify its implementation from a theoretically motivated linguistic description. Computational models are often shaped by practical considerations such as ease of implementation, operational efficiency, or compatibility with available tools, rather than by theoretical considerations such as the plausibility of the description, its logical elegance, or its consistency with descriptions of other languages. In our implementation, we have paid close attention to theoretical considerations.

The expressiveness of automata theory and the set of high-level operations available in the XFST formalism (Beesley and Karttunen, 2003),[3] allow

[1] Often referred to as *upper level* and *lower level*, respectively.

[2] The source code is available at `https://codeberg.org/ldp/innu-aimuna`, under GPLv3. However, the test corpus we used is not included, as it is copyrighted.

[3] We use Foma (Hulden, 2009), a free clone of XFST.

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 80–88
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_010

for the faithful representation of a wide range of morphological phenomena and can be used to describe morphological transducers in a manner that resembles certain classical formal linguistic descriptions. This allows us to envisage that morphological descriptions can be systematically tested against a large amount of real data, and thus subjected to the quantitative comparison of their descriptive adequacy.

Since the 1990s, rule-based approaches have gradually given way to machine learning methods. In particular, deep learning has led to models that demonstrate real competence in NLP, including for morphology (Baxi and Bhatt, 2024). Our rule-based approach is motivated by theoretical, practical, and social considerations. First, since deep learning models are less transparent, they would be less suitable for testing linguistic descriptions. Second, the corpus of available texts in the Innu language is too small to train a statistical or neural model for morphological processing. In addition, some ethical issues, such as those related to the ownership of the data used for training, are particularly sensitive when it comes to Indigenous peoples (Junker, 2024). Finally, it is well known that statistical or neural models trained on small corpora make many errors, which poses a problem when attempting to describe a language for which very few reference works are available to speakers. Indeed, due to deficiencies in the transmission of the Innu language and the fact that Innu people are primarily educated in French or English, many speakers suffer from linguistic insecurity in both spoken and written forms. To avoid exacerbating this further, it is important to provide reliable resources. In this regard, since they are deterministic, rule-based models offer greater guarantees of predictable behaviour than machine learning models, at the cost of more limited coverage.

2 Innu verbal inflection

To make the description of the model easier to understand, this section provides a brief overview of Innu verb inflection. This overview is based on a morphological analysis that we conducted specifically for the purposes of the model. For this reason, the analysis is based on standardized spelling rather than on the spoken language, and we considered the meaning of linguistic units only insofar as it could inform the study of their form. Furthermore, it largely follows established Algonquianist terminology. Our analysis proved to be similar in many respects to that of Agnès (2023), of which we only became aware once our work was already well underway. Both adopt a positional approach to Innu's inflectional morphology. However, when confronted to forms that could be analyzed but only imperfectly with more affixes than less, we favoured the more radically agglutinative one, whereas Agnès opted in these cases for a slightly more fusional analysis. For example, the segment *nan* in *nuapama**nan**at* ('we.EXCL see them') is one morpheme that encodes the local persons as well as the first person plural in Agnès's analysis and two morphemes in ours, where *n* is responsible for the local persons and *an* for the first person plural. This difference of approach leads to a small difference in the number of positions we recognized. It should be noted that Agnès paid a much closer attention to meaning than we did.

Innu verbs are inflected for person, number, obviation, order, tense, and mood. The person category distinguishes the three usual persons. There are two numbers: singular and plural. The first-person plural is also marked for the inclusion (or exclusion) of the addressee. Obviation is a category characteristic of Algonquian languages that ranks third persons according to their peripheral nature (Volchok, 2019). It opposes the *proximate* (central) and the *obviative* (peripheral), and some paradigms allow up to four degrees of obviation, including the proximate. In traditional terminology, these degrees are often conflated with the person category, where the third person corresponds to the proximate, the fourth to the first degree of the obviative, the fifth to its second degree, and the sixth to its third degree. Since the categories of person and obviation are semantically and morphologically distinct, in our model we have chosen to distinguish them. The three orders are the independent, the conjunct, and the imperative. The tenses are present, past and future. There are numerous moods, and many have an evidential meaning. Not all combinations of order, tense, and mood are permitted. We have excluded from our model inflections that are expressed by an auxiliary, such as the future tense (except in the imperative, where it is synthetic).

From a lexical standpoint, verbs are classified according to their transitivity and to the animacy of their subject (for intransitive verbs) or their object (for transitive verbs—the subject of a transitive verb must be animate). The intersection of

Form	PPERS	R	DÉC	DIR	OBV	SPERS	NBLC	TIR	NBNLC
2SG→1SG	tsh	uapam				in		atak	
2PL→1SG	tsh	uapam				in	au	atak	
1SG→2PL	tsh	uapam		it		in	au	atak	
1PL.EX→3SG	n	uapam		a		n	an	atak	
3SG→1PL.EX	n	uapam		iku		n	an	atak	
3SG→2PL	tsh	uapam		ik		u	au	atak	
3PL→2PL	tsh	uapam		ik		u	au	atak	at
1PL.IN→3OBV	tsh	uapam	im	a	n		an	atak	a

Table 1: Positional analysis of different forms of the VTA *uapameu* ('to see') in the present indirect independent.

these features yields the four commonly used Algonquian verb classes: animate intransitive verb (VAI), inanimate intransitive verb (VII), transitive animate verb (VTA), and transitive inanimate verb (VTI). These classes are important because verbs within the same class share the same morphemic patterns. In most cases, the inflectional and syntactic behaviours of members of the same class align, but when they do not, it is inflection that prevails. For example, the VAI class includes a subclass called VAI+O, whose members require a direct object and are thus actually transitive. Verbs in this subclass, however, conjugate like any other VAI. All verbs take the subject's features, but VIIs only allow for the third-person form (since the first two persons cannot be inanimate), and VTAs additionally take the object's features. The VTA paradigm is defective: it does not allow all combinations of subject and object features. For a given Order-Tense-Mood combination, VIIs have at most 4 different forms, VAIs and VTIs have 8, and VTAs have 45 subject-object combinations.

The verbal morphology of Innu can be fairly well represented by a sequence of *positions*, or *slots*. Depending on the context, a position may be occupied by a morpheme or left empty. Morphemes that appear together in the same form cannot, therefore, occupy the same position. This criterion is decisive for determining the number of positions but insufficient for assigning morphemes to the various positions. Indeed, it often happens that the distribution of a morpheme allows for its assignment to several adjacent positions left empty by other morphemes. In such cases, the semantic criterion, according to which morphemes occupying the same position tend to express the same class of grammatical meanings, allows us to resolve the issue. In this way, we assign to each position the expression of specific features. We have identified 9 positions: the person prefix (PPERS),[4] the root (R), the shift (DÉC), which in VTAs is used to "shift" a third person by one degree of obviation, the directional (DIR), the obviative (OBV), the person suffix (SPERS), the local person number (NBLC), the tense or mood (TIR), the non-local person number (NBNLC), and the perceptive (PERCEP), that is only used in the perceptive mood. Not all verb classes realize the same positions, and only VTAs realize all of them. Table 1 presents the positional analysis of different forms of the VTA *uapameu* ('to see') in the present indirect independent.

Verbal roots can also be decomposed in the same way, but since they belong to derivational morphology and the lexical resources we use represent them as simple roots, we treated them as unanalyzable.

The DIR position deserves further explanation. It is found in VTAs and primarily distinguishes between *direct* forms and *inverse* forms. Its role is better understood through the following hierarchy:

$$2 > 1 > 3 > 3\text{OBV}$$

The distribution of some morphemes in VTAs, and many other phenomena in Innu, can be explained largely by this hierarchy (Drapeau, 2014). For example, the second-person prefix *tshi–* can contribute, in VTAs, to the expression of the person combinations 2→1, 1→2, 2→3, and 3→2. By default, or when the DIR position contains a direct morpheme, only combinations where the subject is higher than the object in the hierarchy are permitted (2→1 and 2→3); when the DIR position contains an inverse morpheme, the condition is reversed.

Morphemes are not entirely orthogonal. Some-

[4] We use French abbreviations in this paper, to echo the symbols used in our code. Since most native speakers of Innu speak French, more than English, having the code in French increases maintainability.

times a combination of morphemes has an unexpected meaning, and sometimes several combinations could logically express the same meaning, while only one is actually used. In fact, the two major dialects, Eastern and Western, have sometimes retained different combinations to express the same meaning. Also, especially in the conjunct order, some morphemes are fusional, which means that they span multiple positions in our analysis. For example, the morpheme *ishk*, used in the conjunct order, realizes both the direction (inverse) and the person suffix (second and third person) positions.

This positional analysis views Innu morphology as mostly concatenative. This is largely true, but certain moods of the conjunct order exhibit vowel alterations called *initial changes* (IC in the glosses below) that affect the first vowel of a verb's root. These alterations are regular and depend on vowel quality and length. For example, the forms of the iterative subjunctive conjunct differ from those of the proximate subjunctive conjunct only through the application of these alterations (Drapeau, 2014).

(1) Pipu-k-i, uemut ni-ka kushpi-n
be.winter-NLC-SBJV surely 1-FUT go.inland-LC
'When winter comes, I will surely go inland.'

(2) Pepu-k-i, nanitam ni-kushpi-n
be.winter.IC-NLC-SBJV always 1-go.inland-LC
'In winter, I always go inland.'

The form *pepuki* follows the rule where a short *i* becomes a long *e* in a form that presents initial change. However, vowel length is not consistently indicated in dictionaries and cannot be deduced from general rules. Furthermore, dialects diverge on the length of the first vowel of certain roots. This is the case with *tshitshipańu* ('to begin'), whose *i* is long in eastern dialects and short in western ones (Drapeau, 2014). For this reason, the inflected forms of the roots have not been standardized.

We found two patterns that do not fit well into our analysis because they present the apparent duplication of a morpheme, which we discuss here using the VTA *uapameu* ('to see') as an example. In the past indicative independent form 2PL→3OBV, *tshuapamimat**au**aua* is used instead of the expected *tshuapamimataua*, with the apparent duplication of the second-person plural morpheme *au*. Similarly, in the western dialect, the present indicative conjunct 3OBV→1PL.EX form, *uapam**im**imiti*, appears to duplicate the shift morpheme *im*. In both cases, the two identical segments could be treated either as a single special morpheme or as two instances of the same morpheme. Our implementation adopts the latter analysis, although the former would be more compatible with a strictly positional account.

3 Implementation

As discussed by Lachler et al. (2018), Muradoglu et al. (2020) and Hammerly et al. (2026), early in the design of a morphological transducer, one has to choose between favouring a chunk-based approach, which treats word formation using unanalysed blocks affixed to a root, a concatenative one, where morphemes are independently concatenated following the most general rules possible in a manner that can be similar to that used in classical morphological descriptions, or a mix of both. In a way, the chunk-based approach treats a language's morphology as if it were completely fusional. Both approaches can be used to model any language's morphology. In agglutinative languages, where the differences are the most salient, there are examples of implementations using chunks (Hammerly et al., 2026) or concatenation (Bowers et al., 2017) (note that both models target Ojibwe dialects).

A chunk-based approach simplifies the treatment of irregularity because it enumerates inflectional forms in relatively large units, thus treating both regular and irregular forms in the same way. By contrast, a concatenative approach requires separate rules for both regular patterns and exceptions, which increases implementation complexity but also improves generalization and linguistic transparency. However, the broader the scope of a rule, the greater the likelihood that it will overgenerate. All things considered, and given our interest in linguistic plausibility and model explainability, we have opted for a concatenative approach.

To address the inherent complexity of describing the morphology of an agglutinative language, we decided not to directly map the deep level to the surface level, but rather to create four intermediate levels with well-defined structures and roles. Each level isolates a distinct descriptive problem: feature ordering, positional structure, morpheme selection, canonical exponent insertion, allomorphic realization, and orthographic normalization. This sequence of levels of analysis is loosely inspired by the Meaning-Text Theory (Mel'čuk, 1993) and is made natural by the composition operation, sup-

ported by any finite transducer. For example, let T_1 be a transducer that removes all circumflex accents, which are used in the definition of our model to mark vowel length but are not used in Innu orthography, and T_2 a transducer that lowers the superscript of *u* (^{u}) to *u* when it is not at the end of a word, as per spelling rules. It is possible to construct a transducer T_3 equivalent to the successive application of T_1 and T_2.

$$T_1(\text{tûtam}^{\text{u}}\text{at}) = \text{tutam}^{\text{u}}\text{at}$$
$$T_2(\text{tûtam}^{\text{u}}\text{at}) = \text{tûtamuat}$$
$$T_3 = T_1 \circ T_2$$
$$T_3(\text{tûtam}^{\text{u}}\text{at}) = \text{tutamuat}$$

The composition operation makes it possible to construct abstractions and to modularize the implementation without affecting the model's runtime performance. The four intermediate levels are therefore only relevant for the description of the model and are merged at compile-time into one final transducer, which directly maps the deep level to the surface level.

3.1 Defining the lexicon

The internal lexicon of the transducer is defined in a file that lists each lemma, along with any necessary lexical information.

```
tshimuan+impers+VII
tau+IdTau+Râ+VAI
ashtueim^u+VTI
uitshieu+VTA
```

The verb class is specified for each verb using the features `+VII`, `+VAI`, `+VTI` and `+VTA`. Some VIIs are impersonal; they are marked with `+impers`, which informs the model that it has to block the realization of all non-third-person singular forms for this verb. Moreover, since it is not always possible to determine the root of a verb based solely on its citation form, we sometimes have to encode information about the root. For instance, in the excerpt above, `+Râ` indicates that the root ends in /aː/ (the final *u* of the citation form *tau* being a third-person suffix).

In the overwhelming majority of cases, knowing the verb class is sufficient to derive the inflectional form from a root. It therefore seems that, by directly specifying roots rather than citation forms, we could do without the features that allow us to recover a verb's root. However, since we want to productively generate forms from a citation form (because all the resources of the language refer to it), there must be a way to provide the information necessary to discover the root. Features such as `+Râ` are therefore unavoidable, as is the infrastructure that makes them work. It then seemed simpler to us to use the same mechanism to define the internal lexicon as to specify the productive lexicon.

The uniformity of the definition of the internal lexicon and of the productive one goes even further. *Tau* ('to be (somewhere)') belongs to a small group of irregular verbs in Innu. It would have been possible to account for its irregularity using rules targeting the string *tau* (or *tâ*, in cases where it occurred after the recovery of the roots). Instead, we decided that these rules would target a symbol specific to the lemma, `+IdTau`. This method has three advantages. First, it makes the irregularity of verbs bearing such features explicit. Second, it opens the door to the productive processing of irregular verbs and some of their derivatives that would not be in the internal lexicon. Finally, since productive forms are generated in the same way as non-productive ones, by verifying the correctness of the non-productive forms, we also verify a portion of the productive forms.

Toward the end of the transducer's compilation, features such as `+Râ` and `+IdTau` are removed from the deep level of non-productive forms, so that knowledge of the idiosyncrasies of a verb is not required to synthesize its non-productive form. Those features need to be provided explicitly to synthesize productive forms, as otherwise the model would have no way to recover roots with certainty, which is necessary to produce correct forms. However, for cases where the root can be deducted unambiguously from the citation form, the user is not required to provide a root extraction feature.

3.2 Levels

Deep level. The deep level is the starting point of synthesis and the endpoint of analysis. Its elements are divided into three classes: lemma, lexical features, and inflectional features. They are ordered as follows: the citation form must precede the lexical features, which must precede the inflectional features. Within a class, the elements must also follow a specific order. The order of lexical features is: `+Prod` (which triggers productive mode), root finding feature, `+impers`, length of the first vowel, and verb class; the order of inflectional features

Level	Structure
Deep	nipau+Râ+VAI+Est+Indép+Ind+Prés+1+Pl+Ex
Positional	[PPers]nipâ[Déc][Dir][Obv][SPers][NbLc][Tir][NbNLc][Percep]+VAI+Râ+Est+Indép+Ind+Prés+1+Pl+Ex
Morphological	<1>nipâ<Radical><Lc><1Pl>+VAI+Râ+Est
Morphemic	nit<1>nipâ<Radical>n<Lc>an<1Pl>+VAI+Râ+Est
Morphic	ni<1>nipâ<Radical>n<Lc>an<1Pl>+VAI+Râ+Est
Surface	ninipanan

Table 2: All levels of representation for *ninipanan*, the first-person exclusive plural of the independent present indicative of the VAI *nipau* ('to stand').

is: order, mood, tense, person, number, inclusivity, and obviation. This strict ordering is a consequence of the choice of FSTs for implementation.

The two main dialects of the Innu language are distinguished in the deep level by two mandatory features: `+Est` for the Eastern dialect and `+Ouest` for the Western dialect. Most often, they map to the same surface form, but they are required as the two dialects don't share all forms. In general, a deep form maps to only one surface form, but one surface form can map to many deep ones.

To indicate the length of the first vowel of a root, two features are used: `+VICourte` (for a short vowel) and `+VILongue` (for a long vowel). They are used for the application of initial change. As this length is not well described nor consistent across dialects, these features cannot be set in the lexicon but have to be supplied explicitly at runtime to conjugate a form that triggers initial change. In fact, for these forms, the model will generate nothing unless one of them is given. We had to resort to this mechanism because of the absence of information on vowel length in available resources.

For non-productive forms, lexical features are limited to the verb class and the dialect; for productive forms, they also include the features necessary to identify the root, as well as the features `+impers`, `+Prod`, etc.[5] Inflectional features specify the order (`+Indép` for *independent*, `+Conj` for *conjunct*, or `+Imp` for *imperative*), the mood (`+Ind` for *indicative*, `+Indir` for *indirect*, etc.), tense (`+Prés` or `+Pas`), person (`+2`, `+1`, or `+3`), number (`+Sing` or `+Pl`), inclusivity (`+In` or `+Ex`) and obviation (`+Ntr`, `+Obv`, `+Obv2` or `+Obv3`).

[5] As mentioned above, non-productive forms bear the same features as productive forms (except `+Prod`) until very late in the model's compilation. This explanation concerns runtime.

Positional level. At this stage, roots are identified using features such as `+Râ` or via default rules. This representation introduces the positions we discussed earlier. Lexical and inflectional features are preserved and moved to the right.

```
{a} -> {â} || _ \Lettre* "+Râ"
```

Sample rule. *Replace `a` with `â` when only non-letters separate it from `+Râ`.*

Morphological level. At this level, a position is either filled with a symbol representing a morpheme or left permanently empty. For example, at the morphological level in table 2, the position `[PPers]` is filled with the morpheme `<1>` (first person), `[SPers]` with `<Lc>` (local person) and `[NbLc]` with `<1Pl>` (first person plural). `<Radical>` is the root, which can be treated like any other morpheme. These morphemes are selected based on lexical and inflectional features. All the information necessary for the realization of the morphemes must be conveyed either by the identity of these morphemes or by the lexical features inherited from the deep level, since inflectional features are not carried over to the morphemic level.

```
"[PPers]" -> "<1>" || \
    _ [$["+1"] - $["+In"]] \
    & $["+Indép"]
```

Sample rule. *Fill `[PPers]` position with morpheme `<1>` in non-inclusive independent first-person forms.*

Morphemic level. Here, the canonical forms of morphemes are placed to the left of the morpheme symbol. These symbols are retained because they uniquely identify the morphemes and can therefore resolve ambiguities between homographic forms that would otherwise not be targeted by the same morphophonological rules. For this reason, we

generally prefer to work with symbols rather than character strings, and we delay the introduction of strings until as late as possible.

```
"<1>" -> {nit} "<1>"
```

Sample rule. *Add the canonical form of the first person morpheme symbol before its symbol.*

Morphic level. The morphic level is similar to the morphemic level, but it uses fully realized morphemes rather than canonical forms.

```
{t} -> [] || \
     _ ["<2>"|"<1>"] \Voy
```

Sample rule. *Remove the final* t *of the* `<2>` *or* `<1>` *morphemes before a non-vowel.*

Surface level. The surface representation is a grammatically correct inflected form of a verb. Compared to the morphic level, morpheme symbols and lexical features no longer appear. In addition, some general spelling rules are applied. For example, the superscript *u* (ᵘ) is lowered *(u)* when it is not at the end of a word; the acute accent on *ń* is removed for eastern dialects; and the circumflex accent indicating a long vowel is removed, since standard spelling does not represent vowel length.

3.3 Coverage

The non-productive subset of the model currently contains 68 roots and generates 113,682 forms, of which 19,628 are distinct. This discrepancy is largely artificial and has to do with the structure of the deep level. Indeed, the presence of a feature marking the dialect (`+East` or `+West`) results in nearly doubling the number of feature-form mappings without adding many distinct forms. This is also true for `+VICourte` and `+VILongue`, which are only useful with initial change. Also, some systematic syncretisms are not treated as such at the deep level. For example, the plural obviative marker is systematically conflated in the language with the proximate plural marker for inanimate nouns and with the singular obviative marker for animate nouns. We chose to distinguish these forms at the deep level for simplicity, but one could argue that this distinction simply does not exist even at a semantic level. Overall, only a small portion of this discrepancy is attributable to actual homography.

Throughout the model's development, we tested it against the forms contained in the conjugation tables available on `verbe.innu-aimun.ca`, as well as in the grammar by Drapeau (2014). These resources contain 12,068 distinct forms (27,070 in total)[6] distributed across 68 lemmas and thus contain 61% of the 19,628 distinct forms produced by the model (24% of all 113,682 forms). Our system is able to generate every form in this evaluation set, fully covering the available reference paradigms. As for the generated forms that are not in the evaluation set, they are not necessarily false positives. Indeed, the reference material does not exhaustively list all correct forms for the 68 lemmas they contain. For example, the VTA *natutueu* is only listed for the imperative mood in the resources we employed. It thus contributes 1,808 forms to the model but only 72 to the test corpus. For the other moods, *petueu* is used to exemplify its paradigm. Upon closer inspection, we confirmed that the generalizations encoded in our transducer are consistent with the available paradigmatic reference data.

4 Discussion

Since our model is rule-based, we were able to refine it iteratively until it covered all forms in the available reference dataset. This result is meaningful because the conjugation tables from which the evaluation set is drawn were designed specifically to present and summarize the diversity of the language's regular verbs. It should nevertheless be interpreted as coverage of the attested reference paradigms rather than as an absolute measure of language-wide accuracy. As for irregular verbs, we have implemented only one, *tau* ('to be (somewhere)'). Thus, we know that our infrastructure supports them, but many have not yet been described by linguists, making it impossible to integrate them into the model.

Among North American Indigenous languages, Innu is relatively well documented: it has a grammar spanning several hundred pages (Drapeau, 2014), a substantial and actively maintained online dictionary (`dictionnaire.innu-aimun.ca`), and a widely adopted standard orthography. Nevertheless, some aspects of the language remain less thoroughly documented. Thus, our task was made somewhat more difficult by the lack of standardization regarding changed forms and by the fact that the length of the first vowel is not systematically specified. Similarly, difficulties arise from the fact

[6] We temporarily excluded 123 forms from our test corpus, as the two resources do not agree on them. We are in the process of checking if either of them contains errors, or if we need to include the forms from both.

that the conjugation of irregular verbs has not been officially established, including for some very common verbs such as *tau* ('to be (somewhere)'). This verb is, in fact, the only irregular verb included in our model, and we were able to include it only because its forms were specifically provided to us.

We also aimed to implement a linguistically plausible description. Of course, our analysis of Innu verbal inflection has several limitations. In particular, being based exclusively on written data, it ignores vowel length, which standard orthography does not capture. Furthermore, the two exceptions to our positional analysis suggest that some parts of the verbal paradigms are more diachronically relevant than synchronically. It should be noted, however, that our analysis was conducted primarily for the purposes of this model, and we consider that it proved adequate for this purpose. It contributes to showing that a theoretically motivated model is viable even for a morphologically rich language such as Innu. Also, we showed that the use of multiple clearly defined intermediate levels of description can be very valuable to manage the inherent complexity of an agglutinative language.

We should point out that this project, although inspired by the many high-level operations offered by the XFST formalism, has revealed the limitations of this formalism. In particular, the limited means of abstraction it offers and its relatively strict adherence to finite-state calculus at compile-time led us to hope for a formalism that combines a task-tailored syntax with more powerful means of abstraction, allowing for the use of Turing machines at compile-time. For example, despite the obvious operationalization of the person hierarchy presented above, we were forced to define a large part of its implications by enumeration.

Finally, one of the main challenges in developing digital resources for the Indigenous languages of North America is the lack of resources—both in terms of funding and personnel. We do not yet know what impact our approach will have on this issue. On the one hand, our model is undoubtedly more complex than it would have been had we used a fusional approach. On the other hand, the abstractions on which it is built allow it to be modified (by adding a root, a morpheme, etc.) in a much more localized and productive manner. This modularity should make the model easier to maintain and extend, provided that its implementation is accompanied by clear documentation.

5 Conclusion

We presented a FST for the synthesis and analysis of Innu verbal inflection. This model departs from chunk-based implementations by encoding a linguistically motivated concatenative analysis of Innu verbal morphology. Its architecture relies on positional structure and on a sequence of intermediate representations that separate feature organization, morpheme selection, allomorphic realization, and orthographic normalization.

The current implementation covers the regular verbal paradigms represented in the available reference materials and includes one irregular verb, *tau* ('to be (somewhere)'). Remaining irregular verbs require further description by linguists before they can be fully incorporated.

A morphological transducer lends itself to an array of practical applications. It enables, for instance, the dynamic generation of conjugation tables. Current practical resources on Innu verb conjugation present it using a small set of model verbs, but Innu's verbal inflection is very rich, features many paradigms, and includes irregular verbs, so it would be useful to have an interface that allows users to look up the inflected forms of any verb in the language. We are currently working on such an interface. Also, the rich inflectional system of Innu often makes it difficult for a learner to isolate a verb's root and, consequently, to figure out its lookup form in a dictionary. Adding a morphological analysis engine to an electronic dictionary would allow users to access an entry via any of its inflected forms. Our model could eventually be expanded to include forms from other parts of speech, which could serve as the basis for a lemmatizer or a spell checker. Finally, Innu, like many other Indigenous languages, has very few educational resources. Morphological analysis and synthesis could play an important role in the creation of interactive educational exercises.

Beyond its immediate applications, the model shows that finite-state implementation can function as an empirical testbed for evaluating the descriptive adequacy of morphological analyses in a low-resource language context.

Acknowledgements

We would like to thank Yvette Mollen for her invaluable feedback and Marie-Odile Junker for sharing lexical data with us.

References

Sauvane Agnès. 2023. *La morphosyntaxe algonquienne: vers un nouveau modèle de description à partir des données de l'innu-aimun, langue algonquienne du Québec*. Ph.D. thesis, Sorbonne Université.

Antti Arppe, Christopher Cox, Mans Hulden, Jordan Lachler, Sjur N. Moshagen, Miikka Silfverberg, and Trond Trosterud. 2016. Computational modeling of verbs in dene languages: The case of tsuut'ina. In *Working Papers in Dene Languages 2016*, pages 51–69. Alaska Native Language Center.

Jatayu Baxi and Brijesh S. Bhatt. 2024. Recent advancements in computational morphology: A comprehensive survey. *ArXiv*, abs/2406.05424.

K.R. Beesley and L. Karttunen. 2003. *Finite State Morphology*. CSLI studies in computational linguistics. CSLI Publications, Stanford.

Dustin Bowers, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trond Trosterud. 2017. A morphological parser for odawa. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 1–9, Honolulu. ACL.

Fineen Davis, Eddie A. Santos, and Heather Souter. 2021. On the computational modelling of Michif verbal morphology. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2631–2636, Online. ACL.

Lynn Drapeau. 2014. *Grammaire de la langue innue*. Presses de l'Université du Québec, Québec.

Clarissa Forbes, Garrett Nicolai, and Miikka Silfverberg. 2021. An FST morphological analyzer for the gitksan language. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 188–197, Online. ACL.

Christopher Hammerly, Nora Livesay, Antti Arppe, Anna Stacey, and Miikka Silfverberg. 2026. Ojibwemorph: an approachable finite-state transducer for ojibwe (and beyond). *Lang Resources & Evaluation*, 60(27).

Atticus G. Harrigan, Katherine Schmirler, Antti Arppe, Lene Antonsen, Trond Trosterud, and Arok Wolvengrey. 2017. Learning from the computational modelling of plains cree verbs. *Morphology*, 27:565–598.

Mans Hulden. 2009. Foma: a finite-state compiler and library. In *Proceedings of the Demonstrations Session at EACL 2009*, pages 29–32, Athens, Greece. ACL.

Marie-Odile Junker. 2024. Data-mining and extraction: the gold rush of AI on indigenous languages. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 52–57, St. Julians, Malta. ACL.

Marie-Odile Junker and Terry Stewart. 2008. Building search engines for algonquian languages. *Algonquian Papers - Archive*, 39.

Lauri Karttunen and Kenneth R. Beesley. 2005. Twenty-five years of finite-state morphology. In Antti Arppe, Lauri Carlson, Krister Lindén, Jussi Piitulainen, Mickael Suominen, Martti Vainio, Hanna-Maria Westerlund, Anssi Yli-Jyrä, Juho Tupakka, and Björn Gamb, editors, *Inquiries into words, constraints and contexts: Festschrift in the honour of Kimmo Koskenniemi on his 60th birthday*, pages 71–83. CSLI Publications, Stanford.

Ghazaleh Kazeminejad, Andrew Cowell, and Mans Hulden. 2017. Creating lexical resources for polysynthetic languages—the case of Arapaho. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 10–18, Honolulu. ACL.

Jordan Lachler, Lene Antonsen, Trond Trosterud, Sjur Moshagen, and Antti Arppe. 2018. Modeling Northern Haida verb morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. ELRA.

Igor Mel'čuk. 1993. *Cours de morphologie générale*, volume 1. Les Presses de l'Université de Montréal/ CNRS Éditions, Montréal/Paris.

Saliha Muradoglu, Nicholas Evans, and Hanna Suominen. 2020. To compress or not to compress? a finite-state approach to Nen verbal morphology. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 207–213, Online. ACL.

Adrian Tanner. 2025. Innu (montagnais-naskapi). L'Encyclopédie canadienne.

Irina Volchok. 2019. A synthesis of obviation in algonquian languages. Master's thesis, University of Manitoba.

Morphological Operations in the Persian Verbal System

Marzieh Rabiei Max Silberztein
Marie and Louis Pasteur University

Abstract

Persian is a morphologically rich language whose verbal system presents several challenges for computational modeling. In particular, the interaction between lexical preverbs, grammatical prefixes, and orthographic conventions often obscures the internal morphological boundaries of compound verbs. Additionally, prefixation in verbs with stems beginning with ا or آ (alef-initial stems) may trigger systematic morphophonological alternations that are not directly visible in the written form.

This paper proposes a formal computational description of these processes using electronic dictionaries and morphological grammars implemented in the NooJ linguistic development environment. We formalize two morphological operations designed to capture key structural properties of Persian verbal morphology and implement the corresponding operators: operator <F> identifies internal lexical boundaries and dynamically positions grammatical prefixes within compound verbs involving a restricted set of lexical preverbs, and operator <G> models conditional stem alternations in prefixed contexts through a controlled glide insertion mechanism.

The resulting description has been implemented and applied to a corpus of Persian verbal forms, producing an accurate representation of morpho-orthographic interactions in Persian verbs.

1 Introduction

This study falls within the field of Computational Linguistics, whose objective is to provide explicit and formal descriptions of natural languages through electronic dictionaries and structured morphological and syntactic grammars. In rule-based approaches, linguistic knowledge is encoded in a transparent and formally controlled way, allowing automatic parsing, annotation, and generation [1]. Formalization is therefore both a computational and theoretical commitment to making linguistic structure explicit.

In this context, the present work focuses on the Persian verbal system, which constitutes the morphologically richest component of the language (Windfuhr, 2009; Lazard, 1992). Persian verbs encode tense, aspect, mood, polarity, and person agreement, and combine verbal stems with both grammatical prefixes (mi-, ne-, be-, etc.) and lexical preverbs (bar-, dar-, farâ-, zir-, etc.). This interaction generates complex morphosyntactic configurations in which lexical structure, prefixation, and orthographic constraints must be respected simultaneously. Lexical preverbs form semantically integrated units, while grammatical prefixes must be inserted within these structures without violating the lexical integrity.

From a computational perspective, these properties raise a central question: How can the morpho-orthographic structure of Persian verbs be formally modeled in a way that preserves lexical structure, correctly positions grammatical prefixes, and accounts for prefix-triggered stem alternations? To address this question, we propose a formal modeling approach based on electronic dictionaries and morphological grammars. The study introduces mechanisms for dynamic boundary detection in prefixed verbs and conditional stem mutation in alef-initial verbs. The model is evaluated on a corpus of literary texts in order to assess both descriptive adequacy and computational robustness.

2 Persian Verbal Morphology and Orthography

2.1 Verbal Prefixes and Lexical Preverbs

Persian verbal morphology is characterized by the interaction between two types of preverbal elements: grammatical prefixes and lexical preverbs. Although both precede the verbal stem, they differ in their morphological status, semantic contribution, and computational treatment. Grammatical prefixes express inflectional categories such as aspect, mood, or polarity. The most frequent are *mi-* (imperfective), *ne-* / *na-* (negation), and *be-* (sub-

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 89–95
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_011

junctive or imperative). These prefixes are productive and combine systematically with verbal stems across paradigms without modifying lexical meaning (Windfuhr, 2009). In contrast, lexical preverbs form part of the lexical identity of the verb. Elements such as بر-, در-, فرا-, باز-, پس-, and پیش- combine with a verbal base to create semantically distinct units that cannot be reduced to simple compositional combinations (Samvelian, 2012; Karimi-Doostan, 1997).
This distinction is illustrated in Table 1.

The addition of a lexical preverb creates a new lexical unit with its own semantic identity. These forms must therefore be treated as lexical entries, whereas grammatical prefixes operate at the inflectional level. Failing to distinguish these two layers leads to incorrect segmentation and prefix placement.

2.2 Orthographic Constraints

Persian orthography introduces constraints that directly affect morphological analysis. These concern (i) the absence of short vowels, (ii) obligatory attachment of lexical preverbs, and (iii) the use of the Zero Width Non-Joiner (ZWNJ).[1]

Standard Persian writing does not represent short vowels, leaving vowel interpretation to the reader (Windfuhr, 2009; Lazard, 1992). As a result, a single orthographic form may correspond to multiple morphological analyzes, increasing the ambiguity in stem identification and inflection.

Many lexical preverbs are obligatorily attached to the verbal base. Forms such as برداشتن [bar dâštan] and درآوردن [dar âvardan] are written as single units, concealing internal morphological boundaries that remain relevant for inflectional processes.

The ZWNJ (*nim-fâsele*) prevents incorrect ligatures while preserving morphological cohesion. For example, پیش‌رفتن requires a ZWNJ between the preverb and the base, whereas پیشرفتن is orthographically incorrect and پیش رفتن incorrectly separates the lexical unit (Academy of Persian Language and Literature, 2016).

These constraints show that orthography is not merely superficial. The absence of vowels, the obligatory attachment, and the usage of ZWNJ directly impact segmentation and the application of rules. Therefore, any formal model of Persian verbal morphology must integrate orthographic structure as part of the morphological system.

3 Computational Approaches to Persian Morphology

Computational resources for Persian have been developed within several frameworks. Corpus-based resources, including the Bijankhan corpus (Bijankhan et al., 2004), the Hamshahri corpus (AleAhmad et al., 2003), and the Persian Dependency Treebank (Seraji et al., 2016), provide annotated data for tasks such as part-of-speech tagging and syntactic parsing. Morphological resources have also been developed specifically for Persian. For example, PerLex 2 provides a linguistically motivated morphological lexicon, used together with the MEltfa morphosyntactic tagger (Sagot et al., 2011). These resources constitute important lexical and annotation infrastructures, although their primary objective differs from the explicit generation and analysis of the morpho-orthographic processes considered here.

Beyond these corpus and lexical resources, Persian morphology has also been investigated from both descriptive and computational perspectives. Mahmoodi-Bakhtiari (2018) provides a comprehensive description of Persian morphology, including verbal inflection, derivation, and compounding. From a computational perspective, Shamsfard et al. (2010) integrate morphological analysis into the STeP-1 Persian text-processing toolkit, while also addressing orthographic normalization and spacing variation. More recently, Haghdoost et al. (2019) developed a morphological network for Persian on the basis of a morpheme-segmented lexicon. These studies illustrate the diversity of computational and linguistic approaches that have been proposed to model Persian morphology.

A particularly relevant point of comparison is the finite-state morphological analyzer proposed by Megerdoomian (2004). Using Xerox finite-state technology, this system provides a two-level analysis of Persian morphology and addresses complex verbal conjugation, morphological dependencies, orthographic variation, tokenization, and phonological alternations at morpheme boundaries. It therefore demonstrates that many of the linguistic phenomena addressed in the present study are well suited to explicit rule-based modeling.

The present work does not claim that these

[1]The Zero Width Non-Joiner (ZWNJ) is a non-printing character used in Persian orthography to prevent the cursive connection between characters while preserving morphological boundaries.

Simple Verb	Meaning	Compound Verb	Meaning	Structural Effect
داشتن [dâštan]	to have	برداشتن [bar dâštan]	to pick up	lexical reanalysis
آوردن [âvardan]	to bring	درآوردن [dar âvardan]	to extract / bring out	semantic specialization
دادن [dâdan]	to give	پس دادن [pas dâdan]	to return	directional reinterpretation
رفتن [raftan]	to go	پیش رفتن [pish raftan]	to progress	metaphorical extension

Table 1: Lexical preverbs and semantic differentiation in Persian compound verbs

underlying linguistic phenomena are themselves novel. Rather, its contribution lies in their formalization within the NooJ framework through two structure-sensitive operators. Operator <F> dynamically locates internal lexical boundaries in verbs containing lexical preverbs in order to position grammatical prefixes, whereas operator <G> implements context-dependent glide insertion in prefixed alef-initial stems. Whereas finite-state implementations encode morphotactic and phonological constraints through finite-state lexicons, continuation paths, boundary markers, and composed rules, the present implementation expresses the targeted operations directly within NooJ morphological grammars through operator-based transformations.

The aim is therefore not to establish superiority over existing Persian morphological analyzers, but to provide an explicit NooJ implementation of these interactions and to evaluate its behavior on Persian textual data. A direct empirical comparison with existing finite-state systems on a shared gold-standard dataset remains an important direction for future work.

4 Formalizing Persian Morphology With NooJ

To formalize the morpho-orthographic structure described in the previous sections, we have used the NooJ linguistic development environment (Silberztein, 2016), a rule-based platform dedicated to the construction of electronic dictionaries and formal morphological, syntactic, and semantic grammars. This framework has allowed us to explicitly formalize morphosyntactic constraints using controlled rules and guaranties that our morphological description will be integrated in a future syntactic formalization of Persian syntax.

The following sections focus on the operator-based mechanisms implemented within this framework to model Persian-specific morphological phenomena. In NooJ, inflectional and derivational grammars are formalized as enhanced context-free grammars, in which rules associate sets of suffixes to be concatenated to a lexical entry with their corresponding linguistic properties.

For example, to produce the inflected forms of the lexical entry *drink*, we need to delete its final letters before adding the suffix *unk*. In NooJ, this is done using the operator <B> (backspace):

```
DRINK = <E>/STD | s/PR+3+s
| ing/G | <B3>unk/PP |
<B3>ank/PT ;
```

By adding the suffix <B3>unk to the lexical entry *drink* (resp. *stink*), one produces the word form *drunk* (resp. *stunk*), and then associates it with the property code PP (Past Participle).

NooJ offers several other generic operators, such as <L> (move cursor to the left), <P> (move to the previous wordform), and <C> (change case). Besides these generic operators available for any language, NooJ also provides operators that are specific to certain languages. We will discuss in Sections 5 and 6 the addition of two specific operators for Persian.

5 Prefix Handling and Boundary Detection

5.1 Linguistic Motivation

Persian includes a class of lexical preverbs such as بر-, در-, فرا-, باز-, پیش, and پس, which attach to a verbal base and form semantically integrated lexical units. Constructions such as برداشتن [bar dâštan] and درآوردن [dar âvardan] are therefore not syntactic sequences but lexicalized verbs (Lazard, 1992; Samvelian, 2012).

Grammatical prefixes (*mi-*, *na-*, *be-*) must be inserted between the preverb and the base without breaking the lexical unit.

For example:

برداشتن [*bar dâštan*, ‘to pick up’] → برنداشته [*bar-na-dâšte*, ‘not picked up’]

where negation applies to the base داشتن [*dâštan*] rather than to the entire form. The incorrect form:

نبرداشته [**na-bar-dâšte*]

illustrates the consequence of applying the prefixation at the word level instead of at the level of the verbal base.

This reveals a structural asymmetry: although the verb appears as a single orthographic unit, it contains an internal boundary that must be accessed for inflection.

5.2 Implementation

We introduce the operator <F> as a structure-sensitive positioning operator for internal boundary detection. When a lexical preverb is detected, <F> relocates the rule cursor to the boundary between the preverb and the verbal base, while preserving the lexical preverb:

- if a lexical preverb is detected, <F> shifts the cursor after the preverb;
- otherwise, <F> applies at the beginning of the word.

This operator relies on structure-sensitive positioning rather than linear insertion. For example, <F> correctly produces the negative past participle form برنداشته [*bar-na-dâšte*, 'not picked up'] from the lexical entry برداشتن [*bar dâštan*, 'to pick up'].

This mechanism provides a formal solution to prefix positioning in compound verbs and constitutes a necessary component of any computational model of Persian verbal morphology.

6 Glide Insertion in Prefixal Contexts

6.1 Linguistic Motivation

A second difficulty in modeling Persian verbal morphology concerns verbs beginning with آ or ا (alef-initial stems) in prefixal contexts.

Consider the verb آوردن [âvardan]. When the prefix ب is added, simple concatenation yields:

آوردن [*âvardan*] + ب [*be-*] →
بآورد [**be-âvarad*]

However, the correct form is as follows:

بیاورد [*bi-âvarad*]

In this configuration, the initial آ is replaced by a glide sequence یا. This corresponds to glide insertion in prefixal contexts, whereby Persian avoids the adjacency of two vowel nuclei.

This behavior is not uniform across all alef-initial verbs. For example:

ایستادن [*istâdan*, 'to stand'] →
بایست [*be-ist*, 'stand'] /
بایستم [*be-istam*, 'I stand']

Here, the initial alef is preserved, and no glide insertion occurs. The phenomenon is therefore lexically conditioned and cannot be reduced to a general rule.

6.2 Phonological Characterization

The prefixal context may create a hiatus between the vowel of the prefix and the initial vowel of the verbal stem. Persian resolves this configuration through glide insertion. However, this process is not uniform and must be distinguished into two distinct patterns depending on the phonological nature of the stem-initial element.

Case 1: Alef-madda-initial stems (آ) In this class of verbs, the stem begins with آ, which represents a long vowel. In prefixal contexts, this vowel is not deleted; instead, it is preserved and realized through a glide structure:

آ → یا

For example:

آوردن [*âvardan*] → بیاورد [*bi-âvarad*]

In this case, the vocalic quality is maintained and surfaces as the sequence یا.

Case 2: Alef-initial stems (ا) In this second case, the stem begins with ا, which functions as a vowel carrier rather than a fully specified vowel. The actual vowel quality is determined by its phonological context (e.g., /a/, /o/).

In prefixal contexts, this initial ا is deleted and replaced by the glide ی, while the associated vocalic quality is preserved and transferred onto the resulting structure.

For example:

اُفتادن [*oftâdan*] → بیُفتد [*bi-oftad*]

Here, the initial carrier ا disappears, the glide ی is inserted, and the original vowel quality is maintained within the new structure.

These two patterns demonstrate that glide insertion in Persian is not a uniform process but is conditioned by the phonological nature of the stem-initial element. Consequently, any computational model must explicitly distinguish between آ-initial and ا-initial stems in order to correctly model prefix-triggered stem alternations.

6.3 Implementation

We introduce the operator <G> as a context-sensitive transformation operator for glide insertion in prefixed alef-initial verbs. When the relevant prefixed context is detected, <G> triggers the insertion of the glide ی and the corresponding stem-initial alternation. Its application is lexically conditioned by the phonological class of the stem and therefore does not apply uniformly to all alef-initial verbs.

7 Interaction

Operators <F> and <G> address complementary aspects of Persian verbal morphology: boundary detection and stem alternation. In many cases, both mechanisms must apply simultaneously. For example, in the verb درآوردن [dar âvardan], the negative form درنیاوردم [*dar-na-yâvardam*] requires:

- <F> to detect the lexical preverb -در [*dar*] and correctly position the prefix;
- <G> to trigger the insertion of the glide that affects the stem آورد [*âvard*] in the prefixal context, producing forms such as نیاوردم [*na-yâvardam*].

Without <F>, prefixation would incorrectly target the entire word; without <G>, the resulting form would be phonologically ill-formed. The correct output therefore depends on the interaction of both operators.

This interaction shows that Persian prefixation requires access to internal structure and conditional transformation, highlighting the necessity of structure-sensitive morphological modeling.

8 Evaluation

We used the NooJ software to apply our linguistic resources to a corpus of Persian texts. In natural data, most words are written without short vowels, and the space character is often used instead of the correct ZWNJ (Zero Width Non-Joiner) character.

In order to construct a linguistically controlled evaluation setting, we explicitly encoded vowels in our dictionaries and grammars and systematically used the correct ZWNJ character. However, when applied to real texts, NooJ processes both vowel absence and orthographic variation, including inconsistent use of spacing and ZWNJ.

The evaluation shows that the proposed operators correctly handle prefix positioning and stem alternation across a wide range of verbal forms. In particular, <F> successfully detects internal boundaries in compound verbs, while <G> accurately models glide insertion in prefixal contexts.

These results demonstrate that a rule-based, structure-sensitive approach can effectively capture morpho-orthographic interactions in Persian verbal morphology, even in the presence of orthographic variability.

The results show that the combination of <F> and <G> allows correct generation and analysis of prefixed forms while preserving lexical integrity and handling orthographic constraints such as attachment and ZWNJ.

The evaluation was conducted on a contemporary Persian literary corpus containing 172,548 word forms (230,085 tokens in total).

The system identified 16,253 verbal forms using the implemented morphological resources.

Among these, 361 forms required the application of the <F> operator for correct prefix positioning, while 183 forms involved the <G> operator, corresponding to the alef-initial stems undergoing glide insertion. These counts provide a targeted evaluation of the two phenomena addressed by the proposed operators, namely internal prefix positioning in lexical-preverb constructions and glide insertion in prefixed alef-initial stems.

The system did not recognize a total of 192 verbal forms, corresponding to a recall of 98%. These unrecognized forms mainly reflect gaps in the current lexical coverage and indicate the need for further extension of the verbal lexicon. For example, forms such as تجمع کنند [*tajammo' konand*] and باز نیافته بود [*bâz na-yâfte bud*] correspond to compound verb constructions composed of a nominal element and a simple verb. These compound verbs are not yet fully included in the current lexical resources. Although approximately 2000 compound verbs have already been encoded in the dictionary, some constructions remain uncovered. In such cases, NooJ correctly identifies the simple verbal component, but does not fully analyze the entire compound structure.

In addition, 134 verbal forms were incorrectly analyzed, corresponding to a precision of 99%. These errors are mainly due to orthographic variation, especially the use of space instead of the Zero Width Non-Joiner (ZWNJ), a problem that can be addressed within the NooJ framework.

Table 2 summarizes the main quantitative re-

Corpus size (word forms)	172548
Total tokens	230085
Verbal forms identified	16253
Forms using <F>	361
Forms using <G>	183
Forms not recognized	192 (recall = 98%)
Incorrect analyses	134 (precision = 99%)

Table 2: Evaluation results on the Persian literary corpus

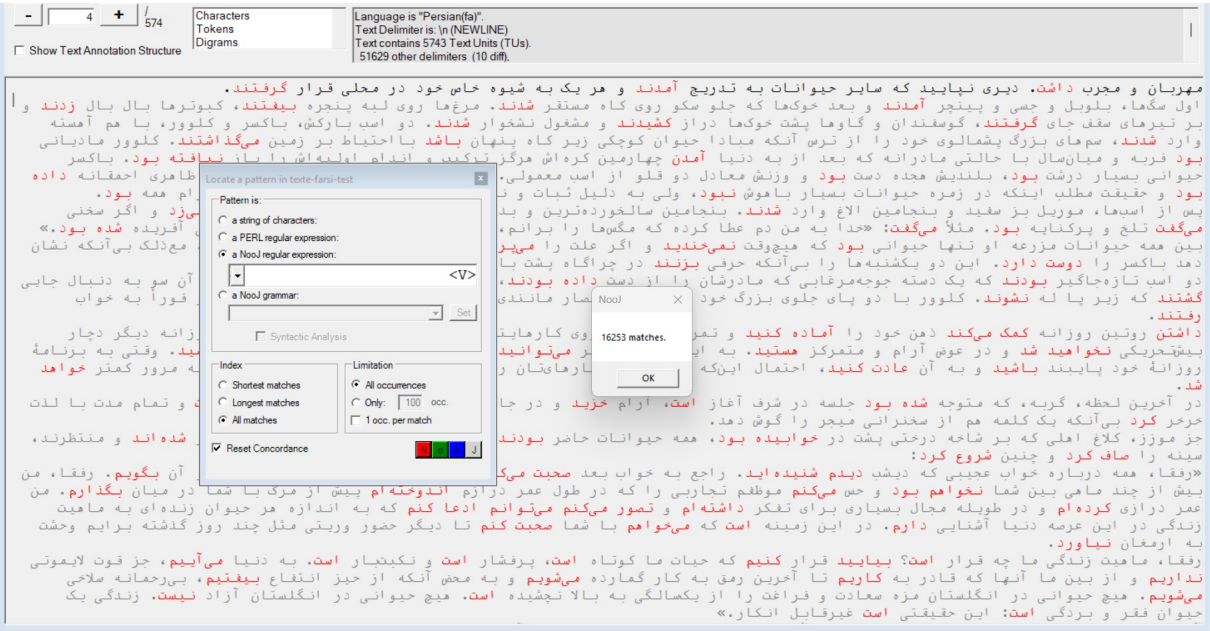


Figure 1: Example of morphological analysis in NooJ showing detected verbal forms

sults of the evaluation.

These results confirm that the proposed morphological resources and operators effectively capture structural and morphophonological constraints, with a recall of 98% and a precision of 99%. At the same time, the remaining unrecognized and incorrectly analyzed forms highlight the importance of extending lexical coverage and addressing orthographic variation to further improve the robustness of the system.

9 Conclusion

This paper has presented a formal, structure-sensitive model of Persian verbal morphology, centered on two operators designed to handle internal boundary detection and prefix-triggered stem alternations. In particular, operator <F> captures the internal structure of compound verbs by identifying lexical preverbs and correctly positioning grammatical prefixes, while operator <G> models glide insertion in alef-initial stems, accounting for context-dependent morphophonological transformations.

The evaluation conducted on a contemporary Persian literary corpus demonstrates that these mechanisms effectively capture key morpho-orthographic interactions that are not accessible through purely concatenative approaches. The results show that the proposed model can accurately analyze and generate prefixed verbal forms while preserving lexical integrity and handling orthographic constraints such as attachment and ZWNJ usage.

At the same time, the observed limitations, including a small number of unrecognized or incorrectly analyzed forms, highlight the need to extend lexical coverage and refine the treatment of complex compound constructions.

Given the close morphological and orthographic relationship between Persian and Dari, the proposed approach could also provide a basis for extending these resources to Dari, although such an

extension would require adaptation to its specific lexical and morphophonological properties.

Overall, this work underscores the importance of structure-sensitive modeling for morphologically rich and low-resource languages. It provides a robust foundation for future developments, including the integration of richer lexical resources and the extension of the model toward syntactic and semantic processing.

References

Academy of Persian Language and Literature. 2016. Persian orthography guidelines. Tehran.

Abolfazl AleAhmad, Hadi Amiri, Ehsan Darrudi, Masoud Rahgozar, and Farhad Oroumchian. 2003. Hamshahri: A standard persian text collection. In *Proceedings of the ACL Workshop on Computational Approaches to Arabic Script-based Languages*.

Mahmood Bijankhan, Javad Sheykhzadegan, Mohammad Bahrani, and Masood Ghayoomi. 2004. Lessons from building a persian written corpus: Peykare. *International Journal of Corpus Linguistics*, 9(2):195–213.

Hamid Haghdoost, Ebrahim Ansari, Zdeněk Žabokrtský, and Mahshid Nikravesh. 2019. Building a morphological network for Persian on top of a morpheme-segmented lexicon. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pages 91–100, Prague, Czechia. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics.

Gholamhossein Karimi-Doostan. 1997. Light verb constructions in persian. *Studies in the Linguistic Sciences*, 27(2):149–173.

Gilbert Lazard. 1992. *A Grammar of Contemporary Persian*. Mazda Publishers.

Behrooz Mahmoodi-Bakhtiari. 2018. Morphology. In Anousha Sedighi and Pouneh Shabani-Jadidi, editors, *The Oxford Handbook of Persian Linguistics*, pages 273–299. Oxford University Press.

Karine Megerdoomian. 2004. Finite-state morphological analysis of Persian. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, pages 35–41, Geneva, Switzerland. COLING.

Benoît Sagot, Géraldine Walther, Pegah Faghiri, and Pollet Samvelian. 2011. Développement de ressources pour le persan : PerLex 2, nouveau lexique morphologique et MEltfa, étiqueteur morphosyntaxique. In *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 73–78, Montpellier, France. ATALA.

Pollet Samvelian. 2012. Auxiliation in persian. *Lingua*, 122(5):514–535.

Mojgan Seraji, Beata Megyesi, and Joakim Nivre. 2016. Building a persian dependency treebank. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*.

Mehrnoush Shamsfard, Hoda Sadat Jafari, and Mahdi Ilbeygi. 2010. STeP-1: A set of fundamental tools for Persian text processing. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Max Silberztein. 2016. *Formalizing Natural Languages: The NooJ Approach*. Wiley.

Gernot Windfuhr. 2009. *The Iranian Languages*. Routledge.

A Massive Open-Source Corpus for Valencian: Over 4.7 Billion Tokens for Low-Resource Language Modelling

Yoan Gutiérrez[*] **Juan Pablo Consuegra-Ayala**[†]
Robiert Sepúlveda-Torres[‡] **Rafael Muñoz**[§]
Language Processing and Information Systems Group,
University Institute for Computing Research / University of Alicante
{ygutierrez,juan.consuegra,robiert.sepulveda,rafael.munoz}@ua.es

Abstract

Valencian is a severely underrepresented language in Natural Language Processing (NLP). Building robust foundation models requires massive, high-quality datasets that are currently unavailable for this Romance language variant. In this paper, we present the **AITANA Valencian Corpus**, to the best of our knowledge the largest open-source dataset curated for Valencian to date. The resource comprises 3.01 billion non-parallel tokens and 1.72 billion parallel tokens (Valencian–Spanish/English), for a total of 4.73 billion tokens across institutional, legislative, and media registers. We detail the acquisition, Markdown-based structuring, and parallel-alignment pipelines, and report corpus analytics covering lexical-diversity metrics and a gender-bias audit. The corpus is publicly released to advance equitable NLP research.

1 Introduction

The development of Large Language Models (LLMs) requires vast amounts of high-quality pre-training data. For high-resource languages, internet-scale crawling provides sufficient volume; however, for low-resource regional languages, data acquisition remains a critical bottleneck. Valencian, a variant of the Catalan language (Acadèmia Valenciana de la Llengua, 2022), co-official in the Valencian Community (Spain), suffers from acute data scarcity in modern NLP (García Cerdá et al., 2025).

Valencian, the traditional and official denomination used in the Valencian Community for the same linguistic system also spoken in Catalonia, the Balearic Islands, and Andorra (Acadèmia Valenciana de la Llengua, 2005), co-official in the Valencian Community (Spain), suffers from acute data scarcity in modern NLP (García Cerdá et al., 2025).

To address this gap, we introduce the **AITANA Valencian Corpus**, a multi-source collection designed specifically for pre-training and aligning language models. By aggregating institutional records, regional news, and parallel administrative data, we have compiled over 4.7 billion tokens.

The contributions are threefold: (i) to the best of our knowledge, the largest open corpus curated for Valencian to date, comprising 3.01B non-parallel and 1.72B document-aligned parallel tokens; (ii) a pipeline for Markdown-based document structuring and multilingual alignment at the document, paragraph, and sentence levels for bilingual institutional sources; and (iii) a corpus-analytics suite covering lexical-diversity metrics and a gender-bias audit.

2 Related Work

Recent efforts have begun to narrow the resource gap for regional and low-resource languages, particularly in the Iberian and European context. Open multilingual suites such as Salamandra (González-Agirre et al., 2025) cover European languages including Catalan, while language-specific initiatives such as Latxa for Basque (Etxaniz et al., 2024) pair dedicated corpora with evaluation suites. Continual pre-training on curated regional data has proven effective for closely related Romance varieties, as demonstrated for Galician–Portuguese (Gamallo et al., 2024). For the Catalan–Valencian continuum specifically, prior work documents both the digital under-resourcing of Catalan (European Language Equality (ELE), 2022) and the difficulty of automatically distinguishing the two varieties (García Cerdá et al., 2025). Our work complements these lines by contributing a large, curated, Valencian-focused corpus together with an aligned parallel component for machine

[*]Design, writing and supervision.
[†]Data collection and processing.
[‡]Language model training and evaluation.
[§]Data management coordination and quality supervision.

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 96–101
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_012

translation.

3 Data Acquisition Strategy

To ensure high linguistic quality and domain diversity, we targeted established institutional and media repositories rather than relying solely on generic web scraping.

Institutional and Legislative Data. We processed the full historical archives of the *Diari Oficial de la Generalitat Valenciana* (DOGV) and the *Boletín Oficial de la Universidad de Alicante* (BOUA). For parliamentary discourse, we extracted transcripts from *Les Corts Valencianes*, separating plenary sessions (DSCV) from commission debates (DSCCV) to capture formal parliamentary discourse.

Media and Web-scale Data. To capture contemporary, everyday Valencian, we ingested the AMIC repository (regional news outlets and blogs) and the news archives of the *Universitat Jaume I* (UJI).

For each source, we collected the full historical archive available at collection time: AMIC (up to May 2024), Les Corts (up to March 2025), UJI (up to December 2025), DOGV (1980–December 2025), and BOUA (up to March 2026).

4 Processing and Alignment Pipeline

4.1 Markdown Transformation

Raw PDFs, HTML, and EPUB files were uniformly converted into structured Markdown. This approach preserves essential document hierarchy (headers, lists, tables), which can support the layout understanding and reasoning of downstream LLMs. Extraneous metadata and boilerplate navigation text were stripped using heuristic rules. Beyond boilerplate removal, we applied a traditional cleaning stage that discards redundant and non-informative content (e.g., repeated headers and footers, empty or near-empty records).

4.2 Parallel Alignment

The parallel subset derives from sources whose records are published in more than one language version; in particular, regional regulations mandate that official publications such as those of the DOGV and the BOUA be issued in both Valencian and Spanish, providing a naturally parallel corpus. After cleaning, paired publications were stored as a document-aligned corpus by exploiting the formal structure of the documents, which is identical across language versions, to match each record and its sections with their counterparts. However, these pairs are aligned only at the document level. We employed the **Translation Alignment Tool**[1] to extract sentence-level and paragraph-level alignments. Specifically, we utilized an embeddings-based alignment approach using the `distiluse-base-multilingual-cased-v2` model, setting a semantic similarity threshold of 0.7 to discard misaligned segments. To further ensure alignment quality, we applied additional automated filtering based on a minimum length ratio of 0.67 between segment pairs, alongside Named Entity Recognition (NER) matching. This multi-stage filtering pipeline yielded 1.72 billion tokens of parallel data for cross-lingual transfer and machine-translation tasks.

4.3 Language Labelling

The language labels reported in Table 1 were not assigned by an automatic language-identification classifier. The language label follows the official classification under which each institutional source publishes by regulation (i.e., records issued as Valencian or as Spanish by the issuing body).

5 Corpus Analytics and Quality

5.1 Basic Statistics and Distribution

Table 1 presents the volumetric breakdown of the corpus and reports Valencian and Spanish token counts separately, derived from the record-level language labels described in Section 4.3. Throughout the paper, token counts are computed by whitespace segmentation; they are therefore not directly comparable to counts produced by subword tokenisers used in modern LLMs. The non-parallel subset is dominated by the DOGV (1.75B tokens), providing a highly reliable, grammatically consistent foundation, followed by AMIC with 932.2M tokens from regional media. Overall, Valencian accounts for 54.7% of the non-parallel tokens that carry a language label (1,548.9M out of 2,833.2M), the remainder being Spanish; the mixed-language Les Corts transcripts are excluded from this figure, as they cannot be split by language.

The non-parallel and parallel subsets are *not* disjoint and should not be read as 4.73B unique tokens. The non-parallel subset is the raw corpus intended for continual pre-training of base models, whereas

[1] https://github.com/gplsi/translation-alignment

the parallel subset is a selection of the same underlying material reformatted into aligned language pairs for machine translation. The headline figure therefore aggregates two task-specific views of the data; in particular, the DOGV contributes to both the non-parallel and the parallel rows of Table 1.

5.2 Lexical Richness

To quantify lexical diversity, we compute the standardised Type-Token Ratio (STTR) and the Measure of Textual Lexical Diversity (MTLD) (McCarthy and Jarvis, 2010) across subsets.

Table 2 reports both metrics for the non-parallel sources, while Table 3 reports the same metrics for the parallel datasets divided by language. In the non-parallel subset, Les Corts has the highest lexical diversity, reflecting parliamentary debate, followed by BOUA and DOGV. AMIC shows lower diversity, consistent with recurrent regional-news formats, while UJI News has the lowest MTLD.

The substantially higher MTLD values in the AMIC and UJI parallel subsets (125.0 and 127.8, respectively) than in their non-parallel counterparts (62.2 and 44.9) are best understood as a selection effect rather than a document-length artefact, to which MTLD is comparatively robust. The alignment pipeline retains only bilingual records with structurally and semantically matched segments and filters pairs with low semantic similarity, discrepant lengths, or inconsistent named entities. It therefore disproportionately removes noisy, fragmented, and repetitive web content, leaving a smaller collection of coherent prose with a denser lexical distribution.

5.3 Downstream Use

The corpus has already been used as continual pre-training and fine-tuning data for Valencian language models built on open multilingual foundations. Its institutional, parliamentary, and media subsets have been used to continually pre-train decoder-only models on top of both FLOR (Da Dalt et al., 2024) and Salamandra (González-Agirre et al., 2025), yielding gains over the respective base models across Valencian, Catalan, and Spanish benchmarks, with notable improvements on summarisation and reading comprehension. The parallel subset has additionally been used to instruction-tune the Spanish–Valencian translation model Aitana-TA-2B-S (Sepúlveda-Torres et al., 2025), built on top of SalamandraTA-2b-instruct[2], improving translation quality over the multilingual baseline (e.g., from 75.5 to 81.2 BLEU on va–es). Translation quality was evaluated on the ES–VA Translation Test benchmark,[3] following the task specification defined in LM Evaluation Harness for the corresponding Spanish Bench task.[4] Crucially, this benchmark shares no material with the corpus presented in this work: its source utterances originate from Common Voice—a crowd-sourced read-speech dataset that is not part of any of our five sources (DOGV, Les Corts, AMIC, BOUA, and UJI)—and belong to a conversational spoken-language domain distinct from the institutional, legislative, and media registers of our data. Moreover, the Spanish and Valencian references were produced by independent translation of those English utterances rather than drawn from any text in the corpus. Train–test overlap is therefore precluded by construction, which rules out data contamination in the reported figures.

5.4 Ethics, Privacy, and Bias

GDPR Compliance. As the datasets stem from public repositories, personal data is generally public record. Nonetheless, we applied an automated anonymisation pipeline[5] to sensitive fields to ensure GDPR compliance prior to open release. Specifically, the system detects and anonymises the following entity types: EMAIL, FINANCIAL, ID, TELEPHONE, VEHICLE, ZIP, and CODE.

Gender-Bias Audit. Legislative texts often encode historical biases. Recent work has shown that LLMs can be leveraged to quantify gender representation bias in gendered-language corpora (Derner et al., 2025). A quantitative audit of the DOGV corpus revealed a 78%-to-22% ratio of masculine to feminine professional nouns (e.g., *director* vs. *directora*). Although this ratio faithfully reflects the historical source texts, documenting the skew is important for downstream LLM training, as it enables practitioners to apply de-biasing techniques during instruction tuning.

[2] https://huggingface.co/BSC-LT/salamandraTA-2b-instruct

[3] https://huggingface.co/datasets/gplsi/ES-VA_translation_test

[4] https://github.com/EleutherAI/lm-evaluation-harness/tree/main/lm_eval/tasks/spanish_bench/phrases_es

[5] https://github.com/gplsi/AnonymizationPipeline

Cat.	Dataset	Lang.	Docs	Avg. len.	Tok. (M)	VA (M)	ES (M)
	DOGV	VA/ES	772,630	2,259	1,745.4	962.4	782.9
	Les Corts	VA/ES^{mixed}	3,628	48,666	176.6	*N/A*	*N/A*
NP	AMIC	VA/ES	136,076	6,850	932.2	501.6	430.6
	BOUA	VA/ES	17,417	4,710	82.0	45.1	36.9
	UJI News	VA/ES	26,659	2,759	73.6	39.8	33.8
	DOGV Para.	VA-ES	354,582	4,311	1,528.7	820.1	708.5
	BOUA Para.	VA-ES	6,548	8,012	52.5	27.9	24.6
P	AMIC Para.	VA-ES	72,215	1,569	113.3	60.9	52.3
	UJI Para.	VA-ES	9,145	1,462	13.4	7.2	6.1
	UJI Para.	VA-EN	8,214	1,019	8.4	6.6	1.8

Table 1: Volumetric summary of the AITANA Valencian Corpus. Cat.: NP = non-parallel, P = parallel. Avg. len.: average document length in tokens. Tok. (M): total tokens in millions; VA (M) and ES (M): Valencian and Spanish token counts in millions. NP sources contain both Valencian and Spanish records; "mixed" indicates that both languages can occur within a single document. Accordingly, the VA and ES token counts for Les Corts are reported as N/A because its mixed-language documents cannot be separated using the record-level language labels. ES is not applicable to the VA–EN row, whose cell includes English tokens instead. For parallel rows, document counts denote aligned document pairs and average length sums tokens across both sides of the pair. The parallel subset is a reformatted selection of the same underlying sources as the non-parallel subset; the two are not disjoint.

Dataset	STTR	MTLD
DOGV	0.355	84.0
Les Corts	0.403	95.3
AMIC	0.300	62.2
BOUA	0.370	88.2
UJI News	0.336	44.9

Table 2: Lexical-diversity metrics for the non-parallel sources. STTR denotes standardised Type-Token Ratio; MTLD denotes Measure of Textual Lexical Diversity.

Dataset	Pair	Lang.	STTR	MTLD
DOGV Para.	VA-ES	VA	0.342	86.5
		ES	0.363	80.7
BOUA Para.	VA-ES	VA	0.355	85.6
		ES	0.372	79.0
AMIC Para.	VA-ES	VA	0.412	125.0
		ES	0.448	122.3
UJI Para.	VA-ES	VA	0.421	127.8
		ES	0.459	126.2
	VA-EN	VA	0.421	127.7
		EN	0.472	96.7

Table 3: Lexical-diversity metrics for the parallel datasets, reported separately for each language side of the aligned document pairs.

6 Conclusions

The AITANA Valencian Corpus resolves a critical resource gap for Valencian NLP. By aggregating and aligning over 4.7 billion tokens of non-parallel and parallel data, we provide foundational infrastructure for training regional Large Language Models and advancing translation technologies. The resource has already enabled continual pre-training of Valencian language models on top of open multilingual foundations and the construction of a dedicated Spanish–Valencian machine-translation system. The dataset is released under open licenses to foster equitable AI development.

Data Availability. The corpus is released under the CC BY-SA 4.0 license and hosted on Hugging Face. The non-parallel subsets are available for DOGV,[6] Les Corts,[7] AMIC,[8] BOUA,[9] and UJI News.[10] The parallel corpus variants can be ac-

[6] https://huggingface.co/datasets/gplsi/alia_dogv

[7] https://huggingface.co/datasets/gplsi/alia_les_corts

[8] https://huggingface.co/datasets/gplsi/alia_amic

[9] https://huggingface.co/datasets/gplsi/alia_boua

[10] https://huggingface.co/datasets/gplsi/alia_uji

cessed via DOGV Parallel,[11] BOUA Parallel,[12] AMIC Parallel,[13] UJI Parallel va-en,[14] and UJI Parallel va-es.[15]

Limitations

Coverage and representativeness. The corpus draws on institutional, legislative, and regional-media registers, which over-represent formal and administrative language and under-represent spontaneous, conversational, and dialectally diverse Valencian; models trained on it may inherit this register skew. Relatedly, the gender skew reported above reflects the source texts and is not corrected in the released data.

Language labelling. Language labels are assigned from each issuing body's official publication language rather than by per-document language identification. Within-document language mixing is therefore not separated, which affects the Les Corts transcripts in particular and prevents us from reporting a Valencian–Spanish split for that source.

Interpreting the reported counts. The non-parallel and parallel subsets are not disjoint: the parallel data is a reformatted selection of the same sources, so the 4.73B headline figure should be read as the sum of two task-specific views rather than as a count of unique tokens. Token counts are computed by whitespace segmentation and are therefore not directly comparable to those produced by subword tokenisers.

Processing and alignment quality. Cleaning is heuristic and does not include hash-based near-duplicate deduplication, so some redundancy may remain. The alignment pipeline applies several automatic quality-control filters—an embedding-based semantic-similarity threshold, a minimum inter-segment length ratio, and NER-based consistency checks—but its output has not been validated against human judgments on a held-out annotated sample. We therefore do not report precision, recall, or alignment error rate (AER), and the parallel token counts should be read as the result of an automatic, conservatively filtered process rather than as human-verified alignments.

Future work. We plan to complement the automatic pipeline with a manual evaluation of alignment quality on a stratified held-out sample of 100–200 segment pairs, reporting the metrics above and disaggregating results by domain (legislative vs. news) to characterise where automatic alignment is most and least reliable. Beyond validation, these annotations will serve to calibrate the semantic-similarity and length-ratio thresholds of the pipeline and to release a small gold-standard alignment set alongside the corpus.

[11] https://huggingface.co/datasets/gplsi/dogv_parallel

[12] https://huggingface.co/datasets/gplsi/boua_parallel

[13] https://huggingface.co/datasets/gplsi/amic_parallel

[14] https://huggingface.co/datasets/gplsi/uji_parallel_va_en

[15] https://huggingface.co/datasets/gplsi/uji_parallel_va_es

Acknowledgments

This work was supported by the Spanish Ministry of Science, Innovation and Universities (MICIU) and the State Research Agency (AEI/10.13039/501100011033) through the projects HEART-NLP (PID2024-156263OB-C22), co-funded by the European Regional Development Fund (ERDF/EU), and AIA2025-163322-C63, co-funded by the European Social Fund Plus (ESF+). Additional support was provided by the Vice-Rectorate for Research of the University of Alicante through grants ACIE25-08, UAUSTI25-12, and ATI2025. This work also received funding from the Ministerio para la Transformación Digital y de la Función Pública, co-financed by the European Union – NextGenerationEU, under the project "Desarrollo de Modelos ALIA".

References

Acadèmia Valenciana de la Llengua. 2022. Statement on Valencian as a variant of Catalan. https://avl.gva.es.

Acadèmia Valenciana de la Llengua. 2005. Resolution of the acadèmia valenciana de la llengua concerning principles and criteria for protecting the name and identity of valencian.

Severino Da Dalt, Joan Llop, Irene Baucells, Marc Pàmies, Yishi Xu, Aitor Gonzalez-Agirre, and Marta Villegas. 2024. FLOR: On the effectiveness of language adaptation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7377–7388, Torino, Italia. ELRA and ICCL.

Erik Derner, Sara Sansalvador De La Fuente, Yoan Gutierrez, Paloma Moreda Pozo, and Nuria M Oliver.

2025. Leveraging large language models to measure gender representation bias in gendered language corpora. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 468–483, Vienna, Austria. Association for Computational Linguistics.

Julen Etxaniz, Oscar Sainz, Naiara Perez, Itziar Aldabe, Germán Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. 2024. Latxa: An open language model and evaluation suite for Basque. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, pages 14952–14972.

European Language Equality (ELE). 2022. Report on the Catalan language in the digital age. Technical report, ELE Project, Deliverable D1.6.

Pablo Gamallo, Pablo Rodríguez, Daniel Santos, Susana Sotelo, Nuno Miquelina, Silvia Paniagua, Daniela Schmidt, Iria de Dios-Flores, Paulo Quaresma, Daniel Bardanca, José Ramom Pichel, Vítor Nogueira, and Senén Barro. 2024. A Galician-Portuguese generative model. In *Progress in Artificial Intelligence: 23rd EPIA Conference on Artificial Intelligence, EPIA 2024, Viana Do Castelo, Portugal, September 3–6, 2024, Proceedings, Part III*, page 292–304, Berlin, Heidelberg. Springer-Verlag.

Raúl García Cerdá, María Miró Maestre, and Miquel Canal. 2025. Building a lightweight classifier to distinguish closely related language varieties with limited supervision: The case of Catalan vs Valencian. In *Proceedings of the First Workshop on Advancing NLP for Low-Resource Languages*, pages 7–11, Varna, Bulgaria. Association for Computational Linguistics.

Aitor González-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, and 1 others. 2025. Salamandra: An open-source multilingual LLM suite for European languages. *arXiv preprint arXiv:2502.08489*.

Philip M. McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.

Robiert Sepúlveda-Torres, Santiago Galeano, María Miró Maestre, Iván Martínez-Murillo, Eduardo Grande, Miquel Canal-Esteve, Ernesto L. Estevanell-Valladares, Fabio Yáñez-Romero, Juan Pablo Consuegra-Ayala, Mar Bonora, Yoan Gutierrez, José Ignacio Abreu Salas, Elena Lloret, Andrés Montoyo, Muñoz-Guillena, and Manuel Palomar. 2025. Aitana-ta-2b-s: Translation model for spanish and valencian. https://huggingface.co/gplsi/Aitana-TA-2B-S. Accessed: 2025-10-03.

LuxIT: A Luxembourgish Instruction Tuning Dataset from Monolingual Seed Data

Julian Valline, Cedric Lothritz, Siwen Guo, Jordi Cabot
Luxembourg Institute of Science and Technology
5 Av. des Hauts-Fourneaux
L-4362 Esch-sur-Alzette
{julian.valline, cedric.lothritz,
siwen.guo, jordi.cabot}@list.lu

Abstract

The effectiveness of instruction-tuned Large Language Models (LLMs) is often limited in low-resource linguistic settings due to a lack of high-quality training data. We introduce LuxIT, a monolingual instruction tuning dataset for Luxembourgish developed to mitigate this challenge. We synthesize the dataset from a corpus of native Luxembourgish texts, utilizing DeepSeek-R1-0528, chosen for its shown proficiency in Luxembourgish. Following generation, we apply a quality assurance process, employing an LLM-as-a-judge approach, retaining 227,507 high-quality instruction-answer pairs. To investigate the practical utility of the dataset, we fine-tune 14 smaller-scale LLMs ($\leq$15B parameters) on LuxIT and evaluate them on standardized Luxembourgish proficiency exams and five downstream NLP tasks. Training on LuxIT yields a mean accuracy change of +5.37 percentage points on language exams across all 14 models, with 12 of 14 showing improvement. On NLP downstream tasks, 9 of 14 models improve in macro-averaged F1, though gains on the two benchmarks do not systematically correlate. These results underscore the feasibility of leveraging monolingual synthetic data to improve LLM capabilities in low-resource languages, while highlighting the multi-faceted nature of language proficiency.

1 Introduction

In recent years, Large Language Models (LLMs) have demonstrated remarkable proficiency across a diverse range of natural language tasks (Zhang et al., 2025; Brown et al., 2020; Touvron et al., 2023b; Yin et al., 2023; Peng et al., 2023; Li et al., 2023a). Their rapid and widespread adoption as virtual assistants (Shu et al., 2023) has been largely driven by their accessibility through intuitive chat-based interfaces (Weber et al., 2024; Touvron et al., 2023b). The training of these assistants involves multiple steps, beginning with a pre-training stage where the model learns from a vast text corpus using a self-supervised objective such as next-word prediction (Weber et al., 2024; Touvron et al., 2023b; Zhou et al., 2023). The subsequent, critical step is instruction tuning (IT) (Wei et al., 2022) which refines the model's ability to follow user instructions and engage in a conversational manner (Weber et al., 2024; Zhang et al., 2025). This process involves fine-tuning an LLM on an extensive dataset, which typically consists of instruction-answer pairs. By training on these pairs, the model learns to generate appropriate responses to user instructions, resulting in more anticipated and manageable outputs (Zhang et al., 2025). A notable benefit of this approach is that it can enhance the model's ability to generalize and respond to previously unseen instructions (Nayak et al., 2024; Zhou et al., 2023; Sanh et al., 2022).

A significant issue at present is the scarcity of open-source, instruction tuning datasets for underrepresented languages. Existing datasets are predominantly in English, marginalizing other languages and leading to inferior model performance and higher deployment costs in these settings (Weber et al., 2024). In the context of this work, "low-resource" refers to a scarcity of NLP tooling, open-source datasets, and specialized instruction-tuning resources. Translating an existing English instruction set, or pairing target-language content with a higher-resource language both depend on resources that are themselves scarce in these settings (Köksal et al., 2025). Luxembourgish, a West Germanic language spoken by about 600,000 people[1] primarily in Luxembourg, exemplifies this challenge as adequate training data is scarce.

In this paper, we introduce LuxIT, a monolingual instruction tuning dataset in Luxembourgish,

[1] https://cursus.edu/en/23040/luxembourgish-at-its-best, as of 11.08.2026

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 102–120
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_013

synthesized from articles of the news website RTL[2] and Wikipedia entries. Following the work of Lothritz et al. (2026), we select DeepSeek-R1-0528, as the top-performing model for Luxembourgish generation and comprehension to synthetically produce a final dataset of 227,507 high-quality instruction-answer pairs. To validate the quality of our dataset, we assess the Luxembourgish capabilities of several LLMs before and after they have been fine-tuned on LuxIT.

To our knowledge, LuxIT is the first Luxembourgish instruction tuning dataset generated entirely within the target language, requiring neither a machine translation system nor parallel data.

Our primary contributions are twofold:

- We introduce LuxIT[3], a synthetically generated instruction tuning dataset in Luxembourgish.
- We evaluate the utility of LuxIT by fine-tuning 14 widely-used small-scale LLMs ($\leq$ 15B parameters) on LuxIT and evaluating them on standardized Luxembourgish proficiency exams and five downstream NLP tasks, demonstrating improvements in language proficiency for 12 of 14 models and NLP task performance for 9 of 14 models.

To validate the pipeline's ability to produce high-quality data, we assess the generated outputs against four custom metrics (see Section 3.4), quantifying the proportion of retained samples. We then address the following research questions:

RQ1: How does fine-tuning on LuxIT affect performance on Luxembourgish language exams?

i) Does fine-tuning on LuxIT improve overall accuracy across CEFR proficiency levels? We fine-tune 14 LLMs ($\leq$ 15B parameters) on LuxIT and benchmark their performance in Luxembourgish against their corresponding base instruct models, following the evaluation framework by Lothritz et al. (2026).

ii) How does the impact vary across individual linguistic categories? We examine the results from RQ1 further and analyze them over the individual categorical exam levels, akin to Lothritz et al. (2026).

RQ2: Does fine-tuning on LuxIT improve performance on downstream Luxembourgish NLP tasks? We benchmark our fine-tuned models against their respective base instruct models on 5 well-known NLP tasks in Luxembourgish.

[2] https://www.rtl.lu/

[3] We provide the Wikipedia subset, RTL subset and code

2 Related work

2.1 Multilingual instruction tuning datasets

xP3 (Muennighoff et al., 2023) is a multilingual, human-crafted instruction tuning dataset, combining data from existing datasets, such as P3 (Sanh et al., 2022) into a unified format across 46 languages. The same work additionally introduces xP3mt, by applying machine translation to the English prompt templates. Bactrian-X (Li et al., 2023b) machine-translates English instructions from Alpaca (Taori et al., 2023) and Dolly (Conover et al., 2023) into 51 other languages. Responses are then generated by GPT-3.5-Turbo. LIMA (Zhou et al., 2023) is a fine-tuned version of Llama (Touvron et al., 2023a), trained on 1000 attentively selected instruction-answer pairs. Weber et al. (2024) extend this resource to Lima-X by translating the LIMA instructions into 4 languages. Comparative overviews of these and related resources are provided by Zhang et al. (2025) and Weber et al. (2024). However, none of these resources covers Luxembourgish.

2.2 Luxembourgish language resources

Recently, there have been various contributions to the creation of Luxembourgish language resources. Lothritz et al. (2022) developed a pretraining dataset in Luxembourgish through a data-augmentation technique where they partially translated text from German into Luxembourgish, resulting in LuxemBERT. Plum et al. (2025) presented LuxGen, a benchmark for evaluating data generation in Luxembourgish and LuxT5, an mT5-based (Xue et al., 2021) text generation model in Luxembourgish, pre-trained on a German, French and Luxembourgish text corpus, where the latter is obtained through transfer learning from German and French. Most recently, Philippy et al. (2025b) introduced a hand-crafted cross-lingual dataset for training Luxembourgish sentence embedding models, which led to the LuxEmbedder model.

With the increasing ubiquity of LLMs, there is a need for a high-quality instruction tuning dataset in Luxembourgish to guide weaker models in improving their Luxembourgish capabilities. Two recent approaches generate Luxembourgish instruction data from native texts. Köksal et al. (2025) ap-

ply reverse instructions to entries from Wikipedia and CulturaX (Nguyen et al., 2024) across 200 languages, including Luxembourgish. Their pipeline however relies on machine translating the article into English and the generated instruction back into the target language. The cross-lingual dataset from Philippy et al. (2025a) demonstrates an effective strategy for generating instruction data in Luxembourgish without machine translation, pairing Luxembourgish with English, German, or French content. Yet, this approach is contingent on the availability of parallel seed data, which is often a bottleneck for low-resource languages. Our work complements both approaches by performing generation directly in Luxembourgish, requiring neither machine translation nor parallel data.

3 LuxIT

This section details the methodology employed for the creation of our Luxembourgish instruction tuning dataset. The data generation process for LuxIT, illustrated in Figure 1 is executed as a five-step pipeline: The initial step (1) involves the extraction of raw data from two distinct sources (see Section 3.1). In the second step (2), we subject the data to a series of heuristic filters (see Section 3.2). We subsequently provide the refined data to DeepSeek-R1-0528, which generates instruction-answer pairs in Luxembourgish (3). Following this, we implement an LLM-as-a-judge approach to evaluate the quality of our synthetic data (4). In parallel, a subset of these pairs undergoes manual human evaluation. In the final step, we apply a post-filtering process to remove all samples that received poor scores (5), resulting in the final LuxIT dataset.

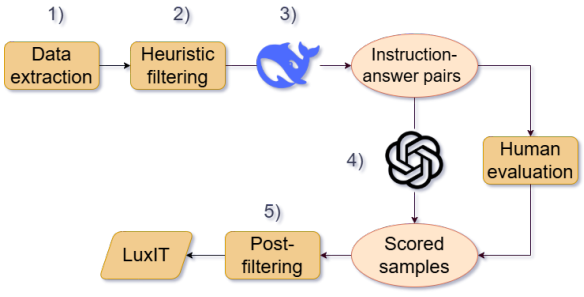


Figure 1: LuxIT Data Generation Pipeline.

3.1 Data

Our methodology utilizes two primary data sources: a complete dump of the Luxembourgish Wikipedia and a collection of all news articles and comments from RTL up to May 2024. We obtained the most recent dump of the Luxembourgish Wikipedia[4] on August 5th, 2025 and subsequently use Wikiextractor (Attardi, 2015) for extraction and cleaning. We convert the processed data from both sources into a JSON format for easier handling. Further details on the data structure are available in Appendix A.1. The initial collection comprises approximately 1.2 million RTL comments, 303,000 RTL news articles, and 80,000 Wikipedia articles, summing to approximately 1.6 million entries. We choose to exclude RTL user comments from our seed data due to their brevity and the potential for biased content, typographical mistakes, and grammatical inaccuracies.

3.2 Heuristic filtering

Before the generation phase, we implement a set of heuristic filters to remove low-quality samples. We retain only articles containing at least 750 characters. The threshold was set heuristically, to ensure each seed carries enough information. For Wikipedia articles, we exclude list articles, disambiguation pages, and pages marked as stubs. We filter out all non-Luxembourgish RTL news articles based on the language tags provided directly by RTL and apply additional filters, including boilerplate removal and semantic format filtering, which identifies keywords indicating a non-prose structure. This filtering process yields 16,558 Wikipedia articles and 100,340 RTL news articles.

3.3 Instruction-Answer pair generation

We employ DeepSeek-R1-0528[5] to synthetically create instruction-answer pairs in Luxembourgish from the filtered RTL news and Wikipedia articles. We select all 16,558 Wikipedia articles and randomly sample 66,776 RTL news articles. We then prompt the model to generate three instruction-answer pairs for each seed, yielding a total of 245,624 synthetic pairs[6].

To ensure the generation of high-quality data, we engineer a comprehensive prompt designed to process information from both data sources effectively. The prompt instructs the model to embed all necessary context within the instruction to facilitate a self-contained answer. For news articles, we

[4] https://dumps.wikimedia.org/lbwiki/latest/

[5] While Lothritz et al. (2026) used the original Deep-Seek-R1 in their experiments, we use the latest DeepSeek-R1-0528

[6] There were some parsing errors for the generated json strings, hence we exclude those samples.

direct the model to handle temporal context with care. We also guide the model to produce a diverse set of instruction types, including summarization, question answering, information extraction, and explanation, where the latter task consists of providing a justification or reasoning alongside the response. For summarization tasks specifically, the model is required to include the original source text within the instruction itself. The complete data generation prompt is available in Appendix A.2.

The final LuxIT dataset is structured with four columns: `instruction`, `response`, `messages` and `source_type` (RTL/Wiki). The `instruction` and `response` columns store the generated instructions and their corresponding answers. The `messages` column consolidates these two into ChatML format. We show the structure of our prompt template in Appendix A.4 and provide a representative sample of an instruction-answer pair from LuxIT in Table 4 in Appendix A.3.

3.4 LLM-as-a-judge scoring

Following generation, we apply an LLM-as-a-judge approach for quality control using GPT-5-mini (OpenAI, 2025). We select GPT-5-mini over GPT-4o-mini based on a comparison reported in Table 5 (Appendix A.5) which shows that despite a slight underperformance at lower Common European Framework of Reference for Languages (Council of Europe, 2001) (CEFR) levels, GPT-5-mini outperforms GPT-4o-mini by up to 24 points at higher levels (+9.7 pp macro average). A small-scale pilot comparing GPT-5 and GPT-5-mini judging found retention rates differing by only a few percent, justifying the cost-performance tradeoff.

For the judging task, we prompt GPT-5-mini to assign scores to each pair based on four criteria: linguistic quality, factual accuracy, instruction adherence (or instruction following), and helpfulness & relevance. Scoring is performed on the instruction–response pair in isolation. We use a three-point scoring system, with 1 representing poor quality and 3 indicating excellent quality. The linguistic quality score aims to check for the overall quality of the Luxembourgish language in the text, such as grammar, spelling and phrasing. Factual accuracy simply checks whether the answer factually answers the instruction without contradictions. Instruction adherence aims to verify whether the response correctly answers the instruction as intended and helpfulness relevance examines the general usefulness of the instruction. The full breakdown of our scoring system and the prompt used for this evaluation are provided in Appendices A.6 and A.7. We retain only samples that achieve a score of at least 2 across all four metrics, discarding all others to form the final dataset.

3.4.1 Human evaluation

To validate our automated quality control, we conduct a manual human evaluation on a random subset of 100 generated samples. This sample size is selected to provide a meaningful preliminary audit of the dataset's quality while remaining feasible within the standard resource constraints of low-resource language research, where recruiting qualified native-speaker annotators is both challenging and cost-intensive. The evaluation was performed independently by three native Luxembourgish speakers (two male, one female; ages 26, 32, and 35) who hold academic backgrounds in computer science and history.

Each annotator evaluated the samples using the identical 3-point scoring rubric applied by the automated LLM judge. Annotators were instructed to assess each pair using only the information contained in the instruction and response, without consulting the seed article. All scoring disagreements are resolved through a strict majority vote to determine the final aggregated label for each sample.

The automated judge retains a higher proportion of samples (92.6%) than the human evaluation (69.0%), indicating a leniency bias in the LLM judge. Inter-annotator agreement (Fleiss' Kappa) is fair for linguistic quality ($\kappa = 0.243$) but poor across the remaining metrics, including below-chance agreement for factual accuracy ($\kappa = -0.019$). This reflects the inherent difficulty of consistently evaluating synthetic Luxembourgish text and suggests that future annotation efforts would benefit from more granular rubrics and calibration sessions. Despite this low agreement, majority-vote aggregation over retained samples shows 94.2% rated "Excellent" for factual accuracy, pointing to generally high quality while underscoring the subjectivity of the individual judgments. We report the full breakdown in Table 7 in Appendix A.8.1.

To directly evaluate the alignment between our automated judge and human insight, Table 1 presents a per-item agreement analysis across the 100-sample subset. Exact agreement is low

Evaluation Metric	Exact Agreement	Cohen's κ	Human Mean (Std)	LLM Mean (Std)
Linguistic Quality	57.0%	0.121	2.24 (0.57)	2.19 (0.46)
Factual Accuracy	45.0%	-0.058	2.84 (0.39)	2.40 (0.65)
Instruction Adherence	83.0%	0.224	2.93 (0.26)	2.80 (0.45)
Helpfulness Relevance	51.0%	0.076	2.29 (0.88)	2.74 (0.44)

Table 1: Comparison of human and LLM evaluator (GPT-5-mini) agreement across four quality metrics on a random subset of 100 samples. Exact Agreement reports the percentage of matching scores, Cohen's κ measures inter-rater reliability, and Mean (Std) values are on a 1–3 scale.

and Cohen's Kappa reveals poor reliability overall, highlighted by a negative coefficient for Factual Accuracy ($\kappa = -0.058$). This below-chance agreement is consistent with a mismatch visible in the metric means. For Factual Accuracy, human reviewers heavily favored perfect marks, resulting in a highly skewed mean of 2.84. Conversely, the LLM judge was significantly more conservative, frequently downgrading those same pairs to acceptable, leading to a lower mean of 2.40. By contrast, a reverse trend is visible in Helpfulness Relevance, where the LLM displays a leniency bias, overestimating value compared to humans (2.74 vs. 2.29).

3.5 Post-filtering

After filtering out all samples with at least one score of 1, the final version of our dataset contains 227,507 instruction-answer pairs in Luxembourgish. We provide a detailed overview in Table 6 (Appendix A.8) and show the score distribution on LuxIT from our post-filtering step. The original dataset consists of 245,624 instruction-answer pairs. We reject 18,117 samples with low scores, observing a retention rate of 92.62%. Among all samples, 32,215 (13.12%) have perfect scores (score 3) across all score types. We observe that `instruction adherence` scores best among all metrics, with 223,065 (90.8%) obtaining a score of 3. Furthermore, among the rejected samples, `factual accuracy` has the highest rejection rate, with 11,745 entries having a score of 1. For the rejection rate, we observe 6,924, 11,745, 1,039 and 246 entries of score 1 (lowest) for linguistic quality, factual accuracy, instruction adherence and helpfulness relevance respectively.

3.6 Dataset composition

To characterize the task-type distribution of LuxIT and evaluate whether our open-ended generation prompt produces balanced coverage, we conduct a classification audit. While categorizing the entire 227,507-pair dataset is computationally prohibitive, we sample 100 random pairs and classify their instruction types using GPT-5.5 (OpenAI, 2026). The distribution of the sampled dataset is as follows: Summarization (32%), Information Extraction (30%), Question Answering (23%), and Explanation (15%). These results indicate that the generation prompt yields a relatively balanced mix of task types, though it exhibits a slight skew toward summarization and extraction tasks.

4 Experimental setup

In this section, we outline and break down our experiments in detail.

4.1 Fine-tuning on LuxIT

We fine-tune a variety of small models ($\leq$15B parameters) on LuxIT. We split the dataset into train, evaluation and test split with ratios 0.92, 0.05 and 0.03 respectively, training on 209,307 and evaluating on 11,375 samples. We select the following models: Llama-3.1-8B-Instruct, Llama-3.2-1B-Instruct (Grattafiori et al., 2024), Gemma-3-1B-IT, Gemma-3-12B-IT (Team et al., 2025), GLM-4-9B-0414 (GLM et al., 2024), Qwen2.5-0.5B-Instruct, Qwen2.5-1.5B-Instruct, Qwen2.5-7B-Instruct (Qwen et al., 2025), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Ministral-3-3B-Instruct-2512 (Liu et al., 2026), Phi-4 (Abdin et al., 2024), Apertus-8B-Instruct-2509 (Hernández-Cano et al., 2026), EuroLLM-1.7B-Instruct (Martins et al., 2025), Olmo-3-7B-Instruct (Olmo et al., 2025). The models are fine-tuned with LoRA (Hu et al., 2022) using the Unsloth[7] framework. Unless otherwise specified, we apply a standard configuration across models: a LoRA rank (r) of 16, a LoRA alpha (α) of 32, a learning rate of 1×10^{-4} with a cosine learning rate scheduler, and a warmup ratio of 0.03. Models are trained for 2 epochs using

[7] https://docs.unsloth.ai/

the AdamW 8-bit optimizer. Due to computational constraints, hardware setups differ slightly between runs. A comprehensive table detailing the hyperparameters, exceptions, and GPU hardware used for each specific model can be found in Appendix B.1.

4.2 Evaluation on language exams

We investigate the impact of LuxIT on the Luxembourgish language proficiency of small LLMs, following the same approach proposed by Lothritz et al. (2026). After fine-tuning, we evaluate the LLMs on Luxembourgish language proficiency exams, instructing the models to solve exams consisting of multiple-choice questions, testing for vocabulary, grammar, reading comprehension, and conversational comprehension (done through transcripts). The language exams stem from the Luxembourgish language institute *Institut National des Langues Luxembourg (INLL)*[8] and are divided into 6 skill levels defined in the CEFR, ranging from A1 (basic level) to C2 (native level), each exam level consisting of fill-in-the-blank and multiple choice questions (Lothritz et al., 2026). The full dataset consists of 629 such questions, with slightly more than 100 questions per difficulty level and a nearly equal distribution of testing categories (vocabulary, grammar, reading comprehension, conversational comprehension). We measure the models' performance using accuracy as a metric, akin to Lothritz et al. (2026). We compare instruct models to our fine-tuned models.

The INLL exams are proprietary and not publicly accessible. This substantially reduces the risk of data contamination within our Wikipedia and RTL news seed corpora, ensuring that the models' performance reflects genuine language comprehension rather than test-set memorization.

4.3 Evaluation on Luxembourgish downstream tasks

We evaluate our 14 fine-tuned models on 5 well-known NLP downstream tasks in Luxembourgish, namely Intent Classification (IC) (Lothritz et al., 2021), Winograd Natural Language Inference (WNLI) (Lothritz et al., 2022), Recognizing Textual Entailment (RTE), Sentiment Analysis and Stanford Sentiment Treebank (SST-2), taken from Lothritz et al. (2023). According to the authors, the data in these task-benchmarks was not taken from Wikipedia or RTL articles. The dataset consists of 3,823 samples. We report the class distribution in Table 9 in Appendix B.2.

[8] https://www.inll.lu/en/

5 Results

5.1 RQ1 i): Does fine-tuning on LuxIT improve overall accuracy across CEFR proficiency levels?

Table 2 shows the comparison between our fine-tuned models and their respective base models evaluated on Luxembourgish language exams, reporting total accuracy on all language exam levels. Fine-tuning on LuxIT leads to a macro average accuracy increase in 12 out of 14 models, with macro average gains ranging from +2.5 percentage points (pp) (Gemma-3-12B-IT) to +21.0 pp (Ministral-3-3B-Instruct). On average, the models see a macro average accuracy increase of 5.37 pp. The effectiveness of instruction tuning on LuxIT is most evident for Ministral-3-3B-Instruct, with its macro average increasing from 25.8% to 46.8% (+21.0 pp). Similarly, for Llama-3.2-1B-Instruct, Phi-4 and Qwen2.5-7B-Instruct, the macro average increases by 14.4 pp, 14.2 pp and 13.6 pp respectively. For 2 of the 14 models tested, Gemma-3-1B and Qwen2.5-0.5B, we observe a degradation. We observe a decrease in macro average accuracy of -22.0 pp and -13.6 pp, respectively. With regard to the individual language exam levels, we observe the most consistent increase for B1/B2 level across all models except our 2 overall decreasing models, Gemma-3-1B-IT and Qwen2.5-0.5B-Instruct. C2 level shows the most inconsistent gains, with 6 out of 14 models declining.

Out of all models, Gemma-3-12B-IT has the strongest baseline before fine-tuning, with a baseline macro average of 56.9%. We identify GLM-4-9B-0414 as the top performing model in terms of absolute performance, with a final macro average of 60.6%, followed by Gemma-3-12B-Instruct with an average of 59.4%. Appendix C.1 contains a more detailed visual breakdown.

5.2 RQ1 ii): How does the impact vary across individual linguistic categories?

Table 3 shows macro-averaged accuracy changes across the four skill categories introduced in section 4.2. The largest gains are observed in vocabulary and conversational comprehension, where multiple models improve substantially — notably Llama-3.2-1B-Instruct (+24.4 pp vocabulary) and Phi-4 (+25.8 pp conversational comprehension). Read-

Model	A1 LuxIT (Δbase)	A2 LuxIT (Δbase)	B1 LuxIT (Δbase)	B2 LuxIT (Δbase)	C1 LuxIT (Δbase)	C2 LuxIT (Δbase)	Macro Avg LuxIT (Δbase)
∼0.5–1.5B parameter models							
qwen2.5-0.5b	20.2 (−22.1)	21.2 (−9.6)	12.6 (−16.5)	23.7 (−8.8)	15.4 (−3.9)	6.9 (−20.8)	16.7 (−13.6)
qwen2.5-1.5b	41.3 (+1.0)	40.4 (+1.0)	34.0 (+10.7)	38.6 (+15.8)	29.8 (+1.0)	27.7 (+6.9)	35.3 (+6.0)
llama-3.2-1b	36.5 (+17.3)	32.7 (+21.2)	24.3 (+10.7)	26.3 (+14.0)	21.2 (+11.5)	22.8 (+11.9)	27.3 (+14.4)
gemma-3-1b	23.1 (−22.1)	21.2 (−18.3)	13.6 (−24.3)	10.5 (−25.4)	7.7 (−18.3)	5.9 (−23.8)	13.7 (−22.0)
eurollm-1.7b	26.9 (±0.0)	26.9 (+3.8)	22.3 (+2.9)	27.2 (+7.9)	14.4 (±0.0)	21.8 (+8.9)	23.3 (+3.9)
ministral-3-3b	61.5 (+25.0)	48.1 (+16.3)	45.6 (+22.3)	48.2 (+29.8)	38.5 (+13.5)	38.6 (+18.8)	46.8 (+21.0)
∼7–9B parameter models							
mistral-7b	51.9 (+3.8)	45.2 (+5.8)	41.7 (+13.6)	44.7 (+10.5)	29.8 (+12.5)	26.7 (+4.9)	40.0 (+8.5)
olmo-3-7b	46.2 (+14.4)	48.1 (+23.1)	38.8 (+10.7)	33.3 (+2.6)	35.6 (+12.5)	22.8 (−4.0)	37.5 (+9.9)
qwen2.5-7b	73.1 (+21.2)	59.6 (+14.4)	56.3 (+16.5)	61.4 (+15.8)	48.1 (+14.4)	36.6 (−1.0)	55.9 (+13.6)
llama-3.1-8b	49.0 (±0.0)	46.2 (+1.9)	41.7 (+10.7)	40.4 (+1.8)	39.4 (+7.7)	25.7 (−4.0)	40.4 (+3.0)
apertus-8b	56.7 (+3.8)	51.0 (−1.0)	48.5 (+13.6)	46.5 (+10.5)	36.5 (+2.9)	32.7 (+3.0)	45.3 (+5.5)
glm-4-9b	74.0 (+11.5)	65.4 (+2.9)	59.2 (+11.6)	59.6 (+7.0)	53.8 (+7.7)	51.5 (+8.9)	60.6 (+8.3)
∼10–14B parameter models							
gemma-3-12b	68.3 (±0.0)	64.4 (±0.0)	64.1 (+8.7)	60.5 (+3.5)	58.7 (+5.8)	40.6 (−3.0)	59.4 (+2.5)
phi-4	67.3 (+7.7)	65.4 (+16.3)	58.3 (+20.4)	52.6 (+7.0)	64.4 (+24.0)	46.5 (+9.9)	59.1 (+14.2)

Table 2: Accuracy (%) of LuxIT fine-tuned models on Luxembourgish language exams, with change relative to the base model (pre-fine-tuning instruct model) in brackets. Green indicates improvement over the base, red indicates degradation. Δbase = macro-averaged accuracy of the LuxIT fine-tuned model minus that of the base model (higher is better).

Model	Vocabulary		Grammar		Reading		Conversation	
	LuxIT	ΔBase	LuxIT	ΔBase	LuxIT	ΔBase	LuxIT	ΔBase
∼0.5–1.5B parameter models								
qwen2.5-0.5b	26.2	(−4.1)	26.2	(−5.1)	5.1	(−25.6)	9.1	(−19.9)
qwen2.5-1.5b	36.8	(+9.6)	33.3	(+2.4)	33.3	(+2.6)	38.0	(+9.0)
llama-3.2-1b	37.5	(+24.4)	32.6	(+8.8)	16.7	(+7.7)	21.9	(+16.8)
gemma-3-1b	14.8	(−24.6)	14.0	(−24.1)	10.9	(−21.8)	14.8	(−17.4)
eurollm-1.7b	29.0	(+5.8)	29.9	(+2.0)	14.1	(+3.2)	19.3	(+4.5)
ministral-3-3b	54.7	(+15.9)	44.9	(+11.8)	43.6	(+30.8)	44.5	(+25.9)
∼7–9B parameter models								
mistral-7b	47.6	(+20.1)	33.5	(+1.9)	41.0	(+8.3)	38.6	(+4.6)
olmo-3-7b	41.6	(+14.0)	29.6	(+1.0)	39.1	(+13.5)	40.0	(+10.9)
qwen2.5-7b	65.8	(+14.1)	40.0	(+8.2)	59.0	(+13.5)	59.3	(+18.2)
llama-3.1-8b	48.5	(+7.9)	33.4	(+1.8)	39.7	(+4.5)	40.7	(−1.8)
apertus-8b	53.4	(−2.2)	33.9	(+8.5)	44.2	(+4.5)	50.9	(+11.0)
glm-4-9b	74.6	(+3.1)	42.8	(+7.1)	56.4	(+7.7)	69.6	(+15.5)
∼10–14B parameter models								
gemma-3-12b	70.5	(−1.1)	41.1	(+1.9)	59.6	(+4.5)	67.7	(+5.1)
phi-4	68.6	(+18.3)	43.4	(+6.0)	54.5	(+7.7)	71.6	(+25.8)

Table 3: Macro-averaged accuracy (%) per skill category for each LuxIT fine-tuned model, together with the change over the respective base model. Green = improvement; red = degradation. ΔBase = LuxIT accuracy minus base accuracy (higher is better).

ing comprehension shows more variable results, with large gains for some models (Ministral-3-3B-Instruct: +30.8 pp) but modest improvements elsewhere. Grammar scores show the most modest changes, with gains rarely exceeding 10 pp, suggesting that the synthetic data is less effective at improving structural linguistic knowledge. The two degrading models (Qwen2.5-0.5B-Instruct and Gemma-3-1B-IT) show uniform drops across all four categories, consistent with their overall regression on language exams in Table 2.

5.3 RQ2: Does fine-tuning on LuxIT improve performance on downstream Luxembourgish NLP tasks?

Figure 2 shows the average change in performance on the Luxembourgish NLP tasks (Section 4.3), comparing baseline and LuxIT fine-tuned models (see Appendix C.2 for a breakdown of the results). Fine-tuning on LuxIT leads to improvements in Luxembourgish NLP task performance for 9 out of 14 models, as measured by macro-averaged F1 (Figure 2) across intent classification, RTE, sentiment analysis, SST-2, and WNLI. The most substantial gains are observed for Phi-4 (ΔF1-macro = +0.325), Olmo-3-7b (+0.211), and qwen2.5-7b (+0.172), while the largest regressions occur for Apertus-8b (-0.206) and Gemma-3-12b (-0.128). Models that improve on language exams (RQ1) do not systematically improve on NLP tasks, and vice versa with notable divergences for Ministral-3-3B, which shows the largest language exam gain (+21.0pp) alongside near-zero NLP task change (-0.003), and Gemma-3-1b, which shows a modest NLP task gain (+0.064) despite a substantial language exam regression (-22.0pp).

It is to note that there are substantial outliers that influence the average performance change (see Appendix C.2, Figure 4). Most notably, Apertus-8b performs poorly on the RTE task after fine-tuning on LuxIT, with a change in performance of -0.677 and Gemma3-12b's performance on IC changes by

-0.522. On the other hand, we observe a considerable gain of +0.802 for Phi-4 in the SST task. While those outliers explain the extreme average changes we observe for Phi-4, Gemma3-12b, and Apertus-8b, it is unclear why those outliers occur.

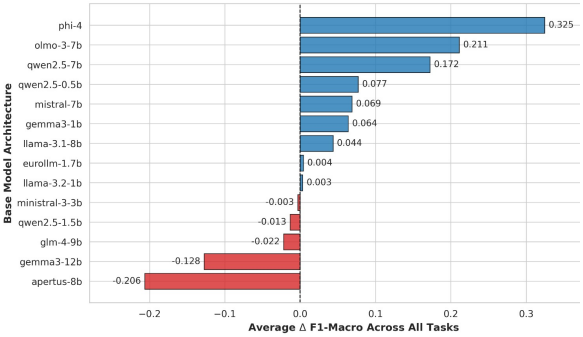


Figure 2: Average Impact of LuxIT Fine-Tuning per Model across all tasks.

6 Discussion

LuxIT represents an important step toward addressing the scarcity of instruction-tuning resources for Luxembourgish. Our pipeline effectively synthesizes a large dataset from monolingual seed data. Furthermore, our dataset composition audit demonstrates that the open-ended generation prompt yields a balanced coverage of diverse instruction types (summarization, extraction, Q&A, and explanation). The human evaluation underscores the subjectivity of manually evaluating synthetic Luxembourgish text. We observe low human inter-annotator agreement (Table 7) alongside metric-specific biases between the automated judge and native speakers (Table 1). While the LLM judge shows a leniency bias regarding Helpfulness Relevance, human evaluators are more forgiving regarding Factual Accuracy. This divergence highlights the need for more objective evaluation frameworks for future iterations.

Regarding language proficiency, fine-tuning generally enhanced formal Luxembourgish comprehension. However, the divergent outcomes among the smallest models suggest that fine-tuning receptivity depends on more than model size alone. Models with limited prior task alignment saw the most benefit, whereas regressions in other sub-1.5B models with stronger baselines likely point to catastrophic forgetting of pre-trained representations.

A deeper analysis of the language exam levels reveals that gains from LuxIT are not uniformly distributed across proficiency levels. Improvements were most consistent at intermediate proficiency levels (B1/B2) and in vocabulary and conversational comprehension. Contrarily, structural grammar and advanced native-like nuances (C2) proved difficult to improve. This suggests that the synthetic data, constrained by the generator model's capabilities and the formal nature of the seed corpora, effectively expands the models' lexicon but skews toward intermediate proficiency.

Furthermore, the limited correlation between language exam improvements and downstream NLP task performance highlights the specific nature of our dataset. The language exams feature standardized, well-formed Luxembourgish, while the NLP tasks contain noisy, non-standardized text. Because LuxIT is derived from formal Wikipedia and news articles, the fine-tuned models appear to align better with the language style of the formal exams. This divergence suggests that format alignment and linguistic improvement are separable effects that instruction tuning on LuxIT does not uniformly induce across all models.

Despite these nuances, in terms of absolute capability, fine-tuning medium-sized models on LuxIT successfully yielded highly proficient Luxembourgish models.

7 Conclusion

This work addresses the scarcity of instruction tuning resources for low-resource languages by introducing LuxIT, a high-quality dataset synthesized from monolingual Luxembourgish seed texts entirely within the target language. By fine-tuning fourteen small-scale LLMs, we demonstrate that LuxIT broadly improves formal language comprehension, particularly in expanding vocabulary and enhancing intermediate-level proficiency, even if performance on noisy, structured downstream NLP tasks remains less consistent. Ultimately, our findings validate that leveraging monolingual synthetic data is an effective and feasible strategy for enhancing LLM capabilities in low-resource linguistic settings. Future work will explore open-ended conversational evaluations and multi-turn data integration.

Limitations

We acknowledge several limitations. First, LuxIT is derived from Wikipedia and RTL news articles and therefore skewed toward formal, standardized, and objective language registers. Consequently, LuxIT may lack the conversational flow, regional

dialectal variations, or informal creative language styles of everyday Luxembourgish. Future data synthesis efforts should incorporate conversational transcripts or informal text.

Second, our human evaluation, while crucial for validation, was conducted on a relatively small subset of 100 samples. A larger-scale human evaluation by multiple native speakers would lend greater statistical power to our quality claims and provide deeper insights into the dataset's strengths and weaknesses.

Third, we acknowledge the observed leniency bias of our single LLM judge (GPT-5-mini) and human evaluators. As shown by the low Exact Agreement and Cohen's Kappa scores (Table 1), biases are metric-dependent. The LLM overestimates Helpfulness Relevance compared to humans, suggesting the final dataset may contain instructions that are marginally less useful than their automated scores imply. Furthermore, the human audit reveals a negative inter-annotator agreement among the native speakers themselves (Fleiss' κ = -0.019, Table 7). Combined with their skewed leniency compared to the LLM (human mean of 2.84 vs. the LLM's 2.40), this indicates our evaluation guidelines allowed for highly subjective interpretations. This emphasizes the difficulty of manually evaluating synthetic text in Luxembourgish. Future iterations should implement more granular rubrics and annotator calibration to improve human consistency, alongside stricter threshold prompting or multi-model LLM ensembles to enhance automated filtering. Similarly, our factual accuracy rubric should not be interpreted as a faithfulness metric. With the exception of the cases where the generation model reproduced the seed text within the instruction, neither the automated judge nor the human annotators were shown the source article. A response that is fluent and plausible but untrue with respect to the seed article would therefore not reliably be filtered out.

Fourth, while our approach is strictly monolingual in its seed data, the generation model, DeepSeek-R1-0528, is a multilingual model. Its internal knowledge is shaped by the vast multilingual corpus it was pre-trained on, which could subtly influence the style and structure of the generated Luxembourgish. More fundamentally, our pipeline requires a generator model already proficient in the target language.

Fifth, our downstream evaluation is restricted to multiple-choice language proficiency exams and classification tasks. While this provides a rigorous and objective measure of grammatical, lexical, and comprehension skills, it does not evaluate the models' open-ended text generation, multi-turn conversational abilities, or specific instruction-following nuances. Consequently, the true conversational proficiency of the fine-tuned models remains to be evaluated.

Finally, Apertus-8B-Instruct is constrained to a maximum sequence length of 512 tokens due to hardware limitations, versus 2048 tokens for all other models. As some LuxIT pairs may exceed this length, its results may underestimate the model's true capacity under standard training conditions.

Ethical considerations and licensing

The Luxembourgish Wikipedia seed data is available under the CC BY-SA 4.0 license, and our use and distribution of RTL news articles is governed by a formal data-sharing agreement with RTL. As DeepSeek-R1-0528 places no restrictions on its outputs, we release the LuxIT Wikipedia subset under CC BY-SA 4.0 and the RTL subset under CC BY-NC 4.0. The INLL proficiency exams (Section 4.2) are proprietary and not publicly distributed; we obtained explicit permission from INLL to use them solely for the automated evaluation of our models. The five downstream NLP benchmarks are used under their original academic terms.

Synthesizing data with LLMs risks propagating biases present in the seed texts or the generator model. We therefore excluded user-generated RTL comments, which are more prone to biased and toxic content, and applied heuristic filtering favoring formal, objective language.

Acknowledgments

This research was supported by LLMs4EU, co-funded by the Digital Europe Programme under GA 101198470 , and by the BESSER project through the Luxembourg National Research Fund (FNR) PEARL programme under the grant agreement 16544475. Our gratitude goes out to RTL for supplying the Luxembourgish news articles and to INLL for providing the Luxembourgish language exams. We thank Aline Fries and Aaron Conrardy for helping us with the human evaluation part and Bernhard Voggenberger for advising us on the fine-tuning experiments.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, et al. 2024. Phi-4 technical report. Preprint, arXiv:2412.08905.

Giuseppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, et al. 2023. Free Dolly: Introducing the world's first truly open instruction-tuned LLM. Accessed on July 31, 2026.

Council of Europe. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press, Cambridge, UK.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. Preprint, arXiv:2406.12793.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. 2024. The llama 3 herd of models. Preprint, arXiv:2407.21783.

Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pásztor, et al. 2026. Apertus: Democratizing open and compliant LLMs for global language environments. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46877–46955, San Diego, California, United States. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, et al. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, et al. 2023. Mistral 7b. Preprint, arXiv:2310.06825.

Abdullatif Köksal, Marion Thaler, Ayyoob Imani, Ahmet Üstün, Anna Korhonen, and Hinrich Schütze. 2025. MURI: High-quality instruction tuning datasets for low-resource languages via reverse instructions. *Transactions of the Association for Computational Linguistics*, 13:1032–1055.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, et al. 2023a. Mimic-it: Multi-modal in-context instruction tuning. Preprint, arXiv:2306.05425.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023b. Bactrian-x: Multilingual replicable instruction-following models with low-rank adaptation. Preprint, arXiv:2305.15011.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, et al. 2026. Ministral 3. Preprint, arXiv:2601.08584.

Cedric Lothritz, Kevin Allix, Bertrand Lebichot, Lisa Veiber, Tegawendé F. Bissyandé, and Jacques Klein. 2021. Comparing multilingual and multiple monolingual models for intent classification and slot filling. In *Natural Language Processing and Information Systems*, pages 367–375, Cham. Springer International Publishing.

Cedric Lothritz, Jordi Cabot, and Laura Bernardy. 2026. Testing low-resource language support in LLMs using language proficiency exams: the case of Luxembourgish. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 2453–2476, Rabat, Morocco. Association for Computational Linguistics.

Cedric Lothritz, Saad Ezzini, Christoph Purschke, Tegawendé Bissyandé, Jacques Klein, Isabella Olariu, et al. 2023. Comparing pre-training schemes for Luxembourgish BERT models. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 17–27, Ingolstadt, Germany. Association for Computational Lingustics.

Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Lisa Veiber, Tegawende Bissyande, Jacques Klein, et al. 2022. LuxemBERT: Simple and practical data augmentation in language model pre-training for Luxembourgish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5080–5089, Marseille, France. European Language Resources Association.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, et al. 2025. Eurollm: Multilingual language models for europe. *Procedia Computer Science*, 255:53–62. Proceedings of the Second EuroHPC user day.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, et al. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Nihal Nayak, Yiyang Nan, Avi Trost, and Stephen Bach. 2024. Learning to generate instruction tuning datasets for zero-shot task adaptation. In *Findings of the Association for Computational Linguistics: ACL*

2024, pages 12585–12611, Bangkok, Thailand. Association for Computational Linguistics.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, et al. 2024. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, et al. 2025. Olmo 3. Preprint, arXiv:2512.13961.

OpenAI. 2025. Gpt-5 system card. https://openai.com/index/gpt-5-system-card/. Accessed on August 18th, 2025.

OpenAI. 2026. Gpt-5.5 system card. https://deploymentsafety.openai.com/gpt-5-5/gpt-5-5.pdf. Accessed on March 6th, 2026.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. Preprint, arXiv:2304.03277.

Fred Philippy, Laura Bernardy, Siwen Guo, Jacques Klein, and Tegawendé F. Bissyandé. 2025a. Luxinstruct: A cross-lingual instruction tuning dataset for luxembourgish. Preprint, arXiv:2510.07074.

Fred Philippy, Siwen Guo, Jacques Klein, and Tegawende Bissyande. 2025b. LuxEmbedder: A cross-lingual approach to enhanced Luxembourgish sentence embeddings. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11369–11379, Abu Dhabi, UAE. Association for Computational Linguistics.

Alistair Plum, Tharindu Ranasinghe, and Christoph Purschke. 2025. Text generation models for Luxembourgish with limited data: A balanced multilingual strategy. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 93–104, Abu Dhabi, UAE. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. 2025. Qwen2.5 technical report. Preprint, arXiv:2412.15115.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Manli Shu, Jiongxiao Wang, Chen Zhu, Jonas Geiping, Chaowei Xiao, and Tom Goldstein. 2023. On the exploitability of instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 61836–61856. Curran Associates, Inc.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, et al. 2023. Alpaca: A strong, replicable instruction-following model. Accessed on September 16, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, et al. 2025. Gemma 3 technical report. Preprint, arXiv:2503.19786.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, et al. 2023a. Llama: Open and efficient foundation language models. Preprint, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. Preprint, arXiv:2307.09288.

Alexander Arno Weber, Klaudia Thellmann, Jan Ebert, Nicolas Flores-Herr, Jens Lehmann, Michael Fromm, et al. 2024. Investigating multilingual instruction-tuning: Do polyglot models demand for multilingual instructions? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20829–20855, Miami, Florida, USA. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, et al. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, et al. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, et al. 2023. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. In *Advances in Neural Information Processing Systems*, volume 36, pages 26650–26685. Curran Associates, Inc.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, et al. 2025. Instruction tuning for large language models: A survey. Preprint, arXiv:2308.10792.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, et al. 2023. Lima: Less is more for alignment. In *Advances in Neural Information Processing Systems*, volume 36, pages 55006–55021. Curran Associates, Inc.

A LuxIT

A.1 Data structure

Both data sources are contained in a JSON file. The following shows an example entry of an

RTL news article[9] and a Wikipedia article. We only extract `public_date`, `title`, `header` and `text` from the RTL news article and extract `title` and `text` from the Wikipedia article.

RTL news article

```
{
 'category_name': 'International',
 'category_id': 5,
 'article_id': 2190645,
 'type': 'news',
 'public_date': '2024-04-28 14:04:54',
 'title': '<article title>',
 'header': '<article header>',
 'text': '<article text>',
 'tags': ['international'],
 'text_id': 301779,
 'lang_id': 'lb'
}
```

Wikipedia article

```
{
  "id": '25',
  "revid": '580',
  "url": "https://lb.wikipedia.org/
    wiki?curid=25",
  "title": "Matthew Perry",
  "text": "De Matthew Langford Perry,
    gebuer den 19. August 1969 zu
    Williamstown am Massachusetts, a
    gestuerwen den 28. Oktober 2023
    zu Los Angeles, war en
    US-amerikanesch-kanadesche
    Schauspiller, dee virun allem
    duerch seng Roll als Chandler
    Muriel Bing an der
    Televisiounsserie Friends bekannt
    ginn ass..."
}
```

A.2 Data generation model prompt

The data generation is performed by DeepSeek-R1-0528. We instruct the model to return the data in JSON format, to ensure compatible formatting and to make it easily accessible. The prompt is formulated as follows:

Data Generation Model Prompt

```
You are an expert in the
    Luxembourgish language tasked
    with creating high-quality
    synthetic training data for
    language models.

OBJECTIVE:
Generate 3 instruction-response
    pair(s) in authentic
    Luxembourgish based on the
    provided text. These pairs will
    be used for instruction
    fine-tuning of language models.

REQUIREMENTS:
1. LANGUAGE: All content MUST be in
    fluent, natural Luxembourgish
   - Use proper Luxembourgish
    grammar, spelling, and idioms
   - Avoid unnecessary German or
    French loan words
   - Ensure the language sounds
    natural to native speakers

2. QUALITY STANDARDS:
   - Instructions should be clear,
    specific, and answerable based on
    the provided text
   - Responses should be
    comprehensive, accurate, and
    well-structured
   - Include ALL necessary context in
    the instruction for a complete
    answer
   - If insufficient information
    exists, indicate that more
    details are needed

3. SUMMARIZATION INSTRUCTIONS:
   - When creating summarization
    tasks, ALWAYS include the
    original seed-text unchanged in
    the instruction for reference
   - The instruction should present
    the source text and ask for a
    summary
   - This ensures the training data
    contains both the source material
    and the summary

4. TEMPORAL CONTEXT:
   - When a date is provided,
    incorporate it appropriately
   - Add temporal context to maintain
    relevance when applicable
   - Consider whether dates belong in
    the instruction, response, or both

5. DIVERSITY:
   - Create varied types of
    instructions (e.g.,
    summarization, Q&A, information
    extraction, explanation)
   - Vary complexity levels
    appropriately
   - Ensure each pair is unique and
    adds value

6. OUTPUT FORMAT:
   Return ONLY a valid JSON array
    with the following structure:
   [
     {{
       "instruction": "Clear
    instruction in Luxembourgish",
```

[9] We do not show title, header and text for the RTL news article as we do not redistribute raw articles in this paper

```
        "response": "Detailed response
    in Luxembourgish"
      }}
    ]

SOURCE TEXT:
{source_context}

Generate 3 high-quality
    instruction-response pair(s)
    based on the above text.
```

A.3 LuxIT Sample

Here we show a sample of LuxIT taken from the Wikipedia subset.

Instruction	Response
Wien dierft bei de Gemengewalen wielen (aktiivt Walrecht), a wat sinn d'Konditioune fir Net-Lëtzebuerger?	Fir ze wielen, muss een 18 Joer al sinn, bierger-lech a politesch Rechter hunn an zu Lëtzebuerg wunnen. Net-Lëtzebuerger mussen hire Wunnsëtz zu Lëtzebuerg hunn, op der Gemeng an d'Wielerlëscht ageschriwwen sinn, a bei Bierger vu Länner baussent EU/Schengen e gültegt Openthaltsdokument hunn.

Table 4: Example instruction-response pair from the LuxIT dataset, taken from the Wikipedia subset.

A.4 Prompt Template

Here we show the prompt template in more detail that unifies columns from LuxIT (Table 4). The `Instruction` and `Response` columns are brought together in the ChatML format:

Prompt Template

```
messages =
[
    {
        "role": "user",
        "content": <instruction>
    },
    {
        "role": "assistant",
        "content": <response>
    }
]
```

A.5 LLM-as-a-judge

Here we break down the reasoning for using GPT-5-mini as a judge (Section 3.4). Table 5 shows the evaluation results from GPT-5 and GPT-5-mini compared to GPT-4o and GPT-4o-mini, where the accuracy scores for the latter 2 were taken from Lothritz et al. (2026) and we simply added the macro average.

Model	A1	A2	B1	B2	C1	C2	Avg
GPT-4o	92.3	85.6	**78.6**	84.2	74.0	60.4	79.2
GPT-4o-mini	90.4	90.4	59.2	59.6	54.8	43.6	66.3
GPT-5	**94.2**	**92.3**	**78.6**	**87.7**	**87.5**	**64.4**	**84.1**
GPT-5-mini	83.7	82.7	76.7	78.9	78.8	55.4	76.0
Δ *(5-mini vs. 4o-mini)*	−6.7	−7.7	+17.5	+19.3	+24.0	+11.8	+9.7

Table 5: Accuracy (%) on Luxembourgish CEFR-level proficiency exams (A1–C2). **Bold** denotes the best score per column. Shading marks our chosen judge model. The Δ row shows the gain/degradation of GPT-5-mini over GPT-4o-mini.

A.6 Custom scoring metric

We fully break down our custom scoring metric introduced in section 3.4:

Linguistic Quality score

Score 1: Contains significant grammatical errors, spelling mistakes, or unnatural phrasing in Luxembourgish. Text that is actually German or French instead of proper Luxembourgish should receive this score.

Score 2: Mostly correct Luxembourgish, but has minor errors or sounds slightly robotic/unnatural. May mix in too many loan words unnecessarily.

Score 3: Fluent, idiomatic, and grammatically perfect Luxembourgish. Natural-sounding text that a native speaker would produce.

Factual Accuracy score

Score 1: Contains factual errors that contradict the source text or general knowledge.

Score 2: Mostly accurate but might have minor inaccuracies or omissions.

Score 3: Completely accurate according to the source text and factual knowledge.

Instruction Adherence score

Score 1: Fails to follow the core instruction (e.g., provides a summary when asked for a list).

Score 2: Follows the main instruction but misses a constraint (e.g., writes 4 bullet points when asked for 3, wrong format, or incorrect tone).

Score 3: Perfectly follows all parts of the instruction, including constraints like length, format, and tone.

Helpfulness Relevance score

Score 1: The instruction is nonsensical, irrelevant to any reasonable context, or the response is unhelpful/off-topic.

Score 2: The instruction is plausible but not very insightful or creative. The response addresses the instruction but in a basic way.

Score 3: A genuinely useful, interesting, or creative instruction that elicits a helpful, comprehensive response.

A.7 Post-filtering model prompt

We post filter the synthetic data with GPT-5-mini. Again, we want the scores to be in JSON format. The prompt is as follows:

Post-Filtering Model Prompt

```
You are an expert evaluator of
    Luxembourgish text quality. Your
    task is to evaluate the following
    instruction-response pair written
    in Luxembourgish based on four
    specific criteria.

IMPORTANT: The texts below are in
    Luxembourgish. You must evaluate
    them as Luxembourgish texts, NOT
    as German, French, or any other
    language. Luxembourgish has its
    own distinct grammar, vocabulary,
    and spelling conventions.

INSTRUCTION (in Luxembourgish):
{instruction}

RESPONSE (in Luxembourgish):
{response}

EVALUATION CRITERIA:

1. linguistic_quality (Linguistic
    Quality):
   - Score 1 (Poor): Contains
    significant grammatical errors,
    spelling mistakes, or unnatural
    phrasing in Luxembourgish. Text
    that is actually German or French
    instead of proper Luxembourgish
    should receive this score.
   - Score 2 (Acceptable): Mostly
    correct Luxembourgish, but has
    minor errors or sounds slightly
    robotic/unnatural. May mix in too
    many loan words unnecessarily.
   - Score 3 (Excellent): Fluent,
    idiomatic, and grammatically
    perfect Luxembourgish.
    Natural-sounding text that a
    native speaker would produce.

2. factual_accuracy (Factual
    Accuracy):
   - Score 1 (Incorrect): Contains
    factual errors that contradict
    the source text or general
    knowledge.
   - Score 2 (Mostly Correct): Mostly
    accurate but might have minor
    inaccuracies or omissions.
   - Score 3 (Perfect): Completely
    accurate according to the source
    text and factual knowledge.

3. instruction_adherence (Instruction
    Following):
   - Score 1 (Not Followed): Fails to
    follow the core instruction
    (e.g., provides a summary when
    asked for a list).
   - Score 2 (Partially Followed):
    Follows the main instruction but
    misses a constraint (e.g., writes
    4 bullet points when asked for 3,
    wrong format, or incorrect tone).
   - Score 3 (Fully Followed):
    Perfectly follows all parts of
    the instruction, including
    constraints like length, format,
    and tone.

4. helpfulness_relevance (Helpfulness
    and Relevance):
   - Score 1 (Not Helpful): The
    instruction is nonsensical,
    irrelevant to any reasonable
    context, or the response is
    unhelpful/off-topic.
   - Score 2 (Somewhat Helpful): The
    instruction is plausible but not
    very insightful or creative. The
    response addresses the
    instruction but in a basic way.
   - Score 3 (Very Helpful): A
    genuinely useful, interesting, or
    creative instruction that elicits
    a helpful, comprehensive response.

CRITICAL INSTRUCTIONS:
- Respond ONLY with a JSON object
    containing the four scores.
- Each score must be an integer: 1,
    2, or 3.
- Do NOT include any explanations,
    comments, or additional text
    outside the JSON.
- Evaluate the text AS LUXEMBOURGISH,
    not as any other language.

JSON FORMAT:
{{
    "linguistic_quality": <score 1-3>,
    "factual_accuracy": <score 1-3>,
```

```
    "instruction_adherence": <score
    1-3>,
    "helpfulness_relevance": <score
    1-3>
}}
```

A.8 Dataset Quality

In Table 6 we provide detailed statistics on the LuxIT score distribution introduced in Section 3.5.

A.8.1 Human Evaluation

We provide details for our human evaluation results in Table 7.

B Experimental Setup

Here we provide more details accompanying the Experimental Setup Section (Section 4).

B.1 Fine-Tuning on LuxIT

We provide more details for the hyperparameter settings in Table 8. (Introduced in Section 4). Several hyperparameters and configurations are held constant across all models unless otherwise specified.

All models are fine-tuned for 2 epochs using the 8-bit AdamW optimizer and a cosine learning rate scheduler. We consistently apply a warmup ratio of 0.03, a weight decay of 0.01, and keep the default value for max_grad_norm at 1.0. For the Low-Rank Adaptation (LoRA) setup, the rank (r) is set to 16, the scaling factor (α) to 32, and the dropout rate to 0.05. Gradient checkpointing is enabled across all training runs to optimize memory consumption. During training, logging occurs every 200 steps, while evaluation and model checkpoint saving are performed every 500 steps. For the dataset, 5% is strictly held out for validation and 3% for testing across all runs.

Notably, while a maximum sequence length of 2048 tokens is utilized for the vast majority of the models, we constrain the Apertus-8B-Instruct model to a sequence length of 512 tokens due to hardware memory limitation requirements. Furthermore, all models trained via the Unsloth framework utilize a fixed random seed of 3407 to ensure reproducibility, whereas the Apertus-8B-Instruct model, which we train using standard Hugging Face pipelines, is initialized with a random seed of 42.

B.2 Evaluation on Luxembourgish downstream tasks

We break down the dataset class distribution in Table 9 introduced in Section 4.3.

Task	Samples
Intent Classification (IC)	159
Recognizing Textual Entailment (RTE)	801
Sentiment Analysis (SA)	367
Stanford Sentiment Treebank (SST-2)	2,360
Winograd NLI (WNLI)	136

Table 9: Sample distribution across Luxembourgish downstream tasks.

C Results

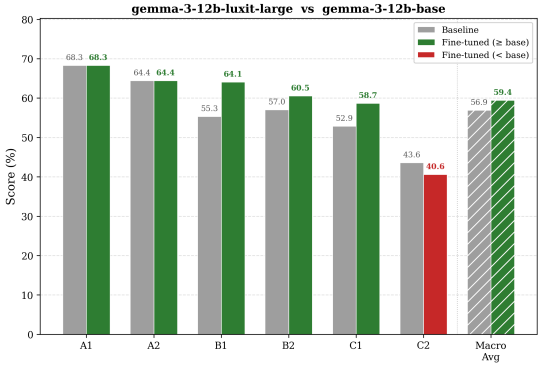
Here we provide more details accompanying the Results section (Section 5).

C.1 RQ1 i)

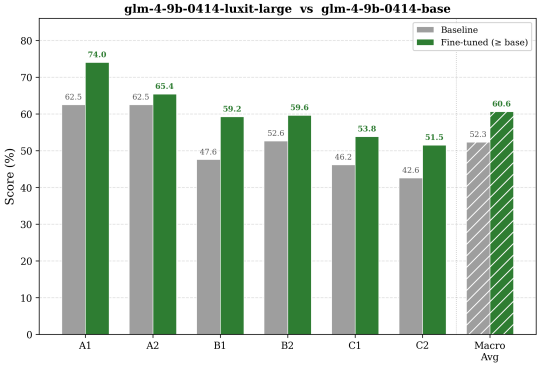
We provide detailed visual comparison between our fine-tuned models and their respective baseline in Figure 3 (Section 5.1).

C.2 RQ2

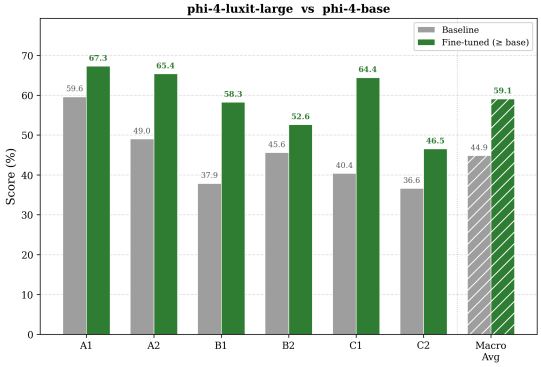
We report the results of benchmarking our fine-tuned models against their respective baseline models (Section 5.3) on the Luxembourgish downstream tasks in Table 10 (introduced in Section 4.3) and show the per-task changes and the average change per model in Figure 4 and Figure 2 respectively.



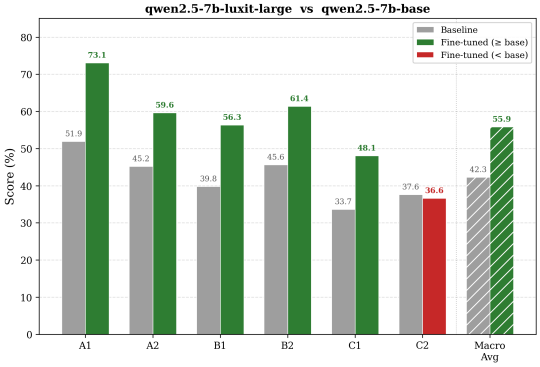
(a) gemma-3-12b



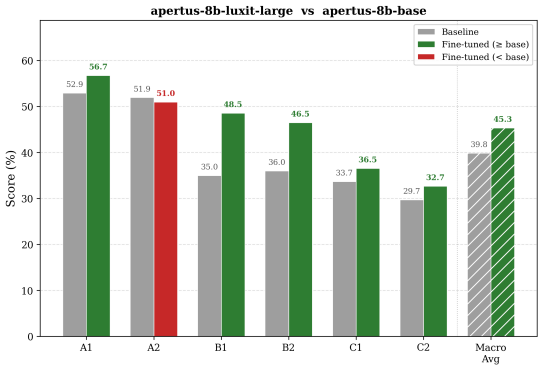
(b) glm-4-9b-0414



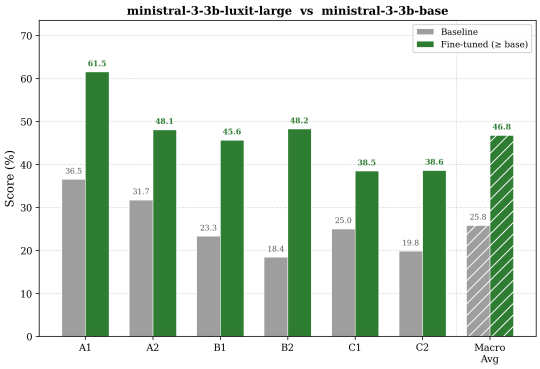
(c) phi-4



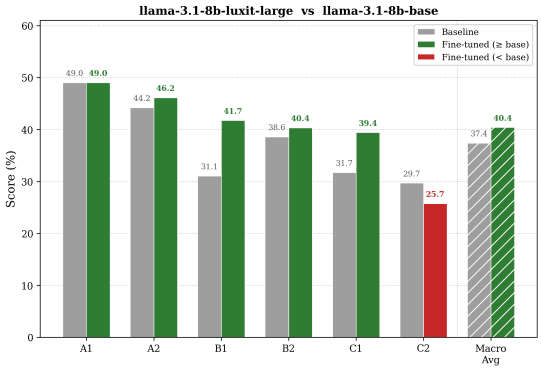
(d) qwen2.5-7b



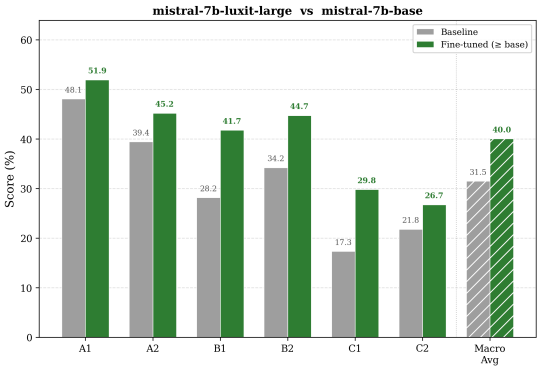
(e) apertus-8b



(f) ministral-3-3b

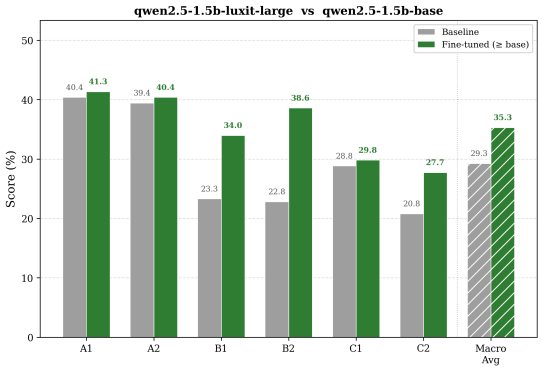


(g) llama-3.1-8b

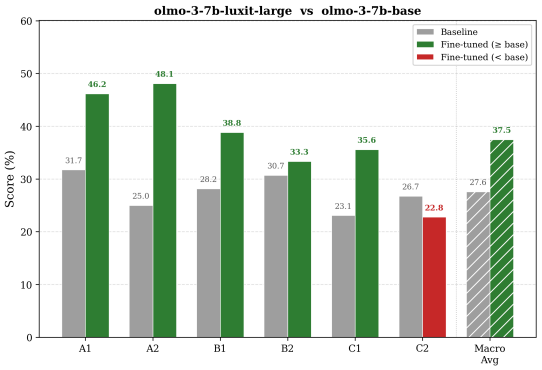


(h) mistral-7b

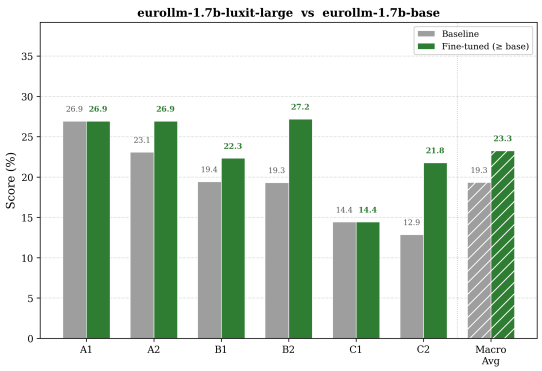
Figure 3: Total accuracy of fine-tuned models (labeled luxit-large) compared to their baselines on Luxembourgish language exams (A1–C2). Green bars indicate improvement over baseline; red bars indicate regression. Models are ordered roughly by macro-average performance of the fine-tuned variant (best to worst, left-to-right, top-to-bottom).



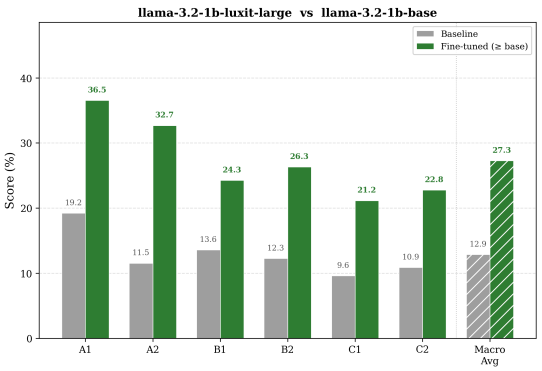
(i) qwen2.5-1.5b



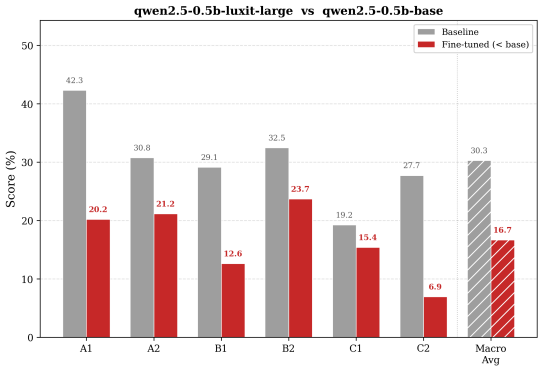
(j) olmo-3-7b



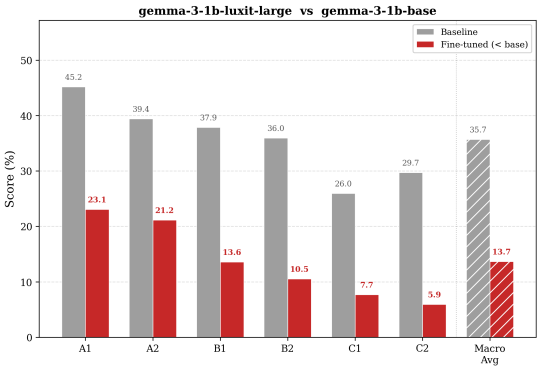
(k) eurollm-1.7b



(l) llama-3.2-1b



(m) qwen2.5-0.5b



(n) gemma-3-1b

Figure 3: (Continued) Total accuracy of fine-tuned models compared to their baselines on Luxembourgish language exams (A1–C2).

Score type	Score 1	Score 2	Score 3	Mean	Median
Original Dataset (245,624 entries)					
LQ	6,924 (2.8%)	189,858 (77.3%)	48,842 (19.9%)	2.171	2.0
FA	11,745 (4.8%)	101,693 (41.4%)	132,186 (53.8%)	2.490	3.0
IA	1,039 (0.4%)	21,520 (8.8%)	223,065 (90.8%)	2.904	3.0
HR	246 (0.1%)	44,755 (18.2%)	200,623 (81.7%)	2.816	3.0
Filtered Dataset (227,507 entries)					
LQ	N/A	180,075 (79.2%)	47,432 (20.8%)	2.208	2.0
FA	N/A	97,068 (42.7%)	130,439 (57.3%)	2.573	3.0
IA	N/A	16,343 (7.2%)	211,164 (92.8%)	2.928	3.0
HR	N/A	29,305 (12.9%)	198,202 (87.1%)	2.871	3.0

Table 6: LuxIT Score Distribution. We report scores from post-filtering on LuxIT with scores 1 (low), 2 (acceptable) and 3 (excellent) for Linguistic Quality (LQ), Factual Accuracy (FA), Instruction Adherence (IA) and Helpfulness Relevance (HR). We show the score distribution on the original dataset compared with the filtered dataset after rejecting 18,117 samples with low scores.

Evaluation Metric	Fleiss' κ	Aggregated Mean	% Acceptable (Score 2)	% Excellent (Score 3)
Linguistic Quality	0.243	2.33	66.7%	33.3%
Factual Accuracy	-0.019	2.94	5.8%	94.2%
Instruction Adherence	0.102	2.93	7.2%	92.8%
Helpfulness Relevance	0.090	2.80	20.3%	79.7%

Table 7: Human evaluation results on a random subset of 100 samples. Inter-annotator agreement is reported using Fleiss' κ across three independent evaluators. The final aggregated scores (Majority Vote) are reported for the 69 retained samples (scores of 1 "Low" are discarded).

Model	GPU	4-bit Load	BS	GAS	LR	Target Modules
Gemma-3-12B-IT	RTX 4090	Yes	2	16	1×10^{-4}	Gemma-specific[1]
Olmo-3-7B-Instruct	RTX 4090	No	8	4	2×10^{-4}	Standard[2]
EuroLLM-1.7B-Instruct	RTX 4090	No	16	3	1×10^{-4}	Standard[2]
Gemma-3-1B-IT	RTX 4090	No	16	3	1×10^{-4}	Gemma-specific[1]
Apertus-8B-Instruct[3]	RTX 4090	No	2	16	2×10^{-4}	all-linear
Qwen2.5-7B-Instruct	V100	No	2	16	1×10^{-4}	Standard[2]
Qwen2.5-1.5B-Instruct	V100	No	2	16	2×10^{-4}	Standard[2]
Qwen2.5-0.5B-Instruct	V100	No	2	16	2×10^{-4}	Standard[2]
Phi-4	V100	Yes	2	16	1×10^{-4}	Standard[2]
Mistral-7B-Instruct-v0.3	V100	No	4	8	1×10^{-4}	Standard[2]
Ministral-3-3B	V100	No	4	10	2×10^{-4}	Standard[2]
Llama-3.2-1B-Instruct	V100	No	8	6	2×10^{-4}	Standard[2]
Llama-3.1-8B-Instruct	V100	No	2	16	1×10^{-4}	Standard[2]
GLM-4-9B-0414	V100	No	2	16	1×10^{-4}	Standard[2]

Table 8: Hyperparameter Settings by Model. Constants across all models include 2 training epochs, AdamW 8-bit optimizer, cosine learning rate scheduler, a warmup ratio of 0.03, weight decay of 0.01, and LoRA parameters ($r = 16, \alpha = 32$, dropout = 0.05). BS = Train Batch Size per Device, GAS = Gradient Accumulation Steps, LR = Learning Rate. [1]Gemma-specific: vision_layers=False, language_layers=True, attention_modules=True, mlp_modules=True. [2]Standard: q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj. [3]Apertus-8B-Instruct was the only model fine-tuned utilizing standard Hugging Face pipelines rather than Unsloth.

	IC			RTE			SA			SST-2			WNLI			
Base Model	Base	LuxIT	Δ	Base	LuxIT	Δ	Base	LuxIT	Δ	Base	LuxIT	Δ	Base	LuxIT	Δ	**Avg. Δ**
phi-4	0.647	0.695	+0.048	0.449	0.775	+0.326	0.017	0.430	+0.413	0.015	0.816	+0.802	0.293	0.327	+0.034	**+0.325**
olmo-3-7b	0.150	0.245	+0.096	0.142	0.666	+0.524	0.249	0.218	−0.031	0.300	0.787	+0.488	0.285	0.265	−0.020	**+0.211**
qwen2.5-7b	0.491	0.702	+0.210	0.658	0.774	+0.116	0.323	0.349	+0.026	0.285	0.770	+0.485	0.511	0.536	+0.025	**+0.172**
qwen2.5-0.5b	0.008	0.000	−0.008	0.276	0.392	+0.115	0.065	0.076	+0.011	0.297	0.504	+0.207	0.331	0.390	+0.059	**+0.077**
mistral-7b	0.412	0.419	+0.008	0.548	0.716	+0.168	0.354	0.383	+0.029	0.652	0.767	+0.115	0.265	0.290	+0.025	**+0.069**
gemma-3-1b	0.150	0.186	+0.036	0.530	0.560	+0.031	0.246	0.251	+0.005	0.636	0.694	+0.058	0.293	0.482	+0.189	**+0.064**
llama-3.1-8b	0.518	0.545	+0.027	0.332	0.623	+0.291	0.188	0.348	+0.160	0.711	0.651	−0.060	0.520	0.320	−0.200	**+0.044**
eurollm-1.7b	0.006	0.019	+0.013	0.139	0.054	−0.085	0.052	0.213	+0.161	0.267	0.289	+0.023	0.318	0.227	−0.090	**+0.004**
llama-3.2-1b	0.010	0.009	−0.001	0.251	0.324	+0.072	0.106	0.082	−0.024	0.175	0.338	+0.163	0.459	0.265	−0.194	**+0.003**
ministral-3-3b	0.398	0.322	−0.076	0.392	0.551	+0.159	0.425	0.192	−0.233	0.545	0.684	+0.139	0.329	0.324	−0.005	−0.003
qwen2.5-1.5b	0.232	0.142	−0.090	0.310	0.414	+0.105	0.263	0.240	−0.023	0.565	0.501	−0.064	0.385	0.390	+0.006	−0.013
glm-4-9b	0.684	0.743	+0.059	0.775	0.735	−0.040	0.307	0.328	+0.021	0.823	0.838	+0.016	0.478	0.311	−0.167	−0.022
gemma-3-12b	0.865	0.342	−0.522	0.744	0.770	+0.025	0.524	0.455	−0.068	0.783	0.716	−0.067	0.569	0.564	−0.005	−0.128
apertus-8b	0.576	0.460	−0.116	0.686	0.010	−0.677	0.505	0.234	−0.271	0.706	0.459	−0.247	0.265	0.543	+0.278	−0.206

Table 10: Complete evaluation results comparing **Baseline** and **LuxIT** fine-tuned F1-Macro scores across five Luxembourgish NLP tasks: Intent Classification (IC), Recognizing Textual Entailment (RTE), Sentiment Analysis (SA), Stanford Sentiment Treebank (SST-2), and Winograd Natural Language Inference (WNLI). Δ = LuxIT − Baseline; bold values in the *Avg.* Δ column indicate net improvement. Models are sorted by average Δ in descending order; the horizontal rule separates positive from negative average gains.

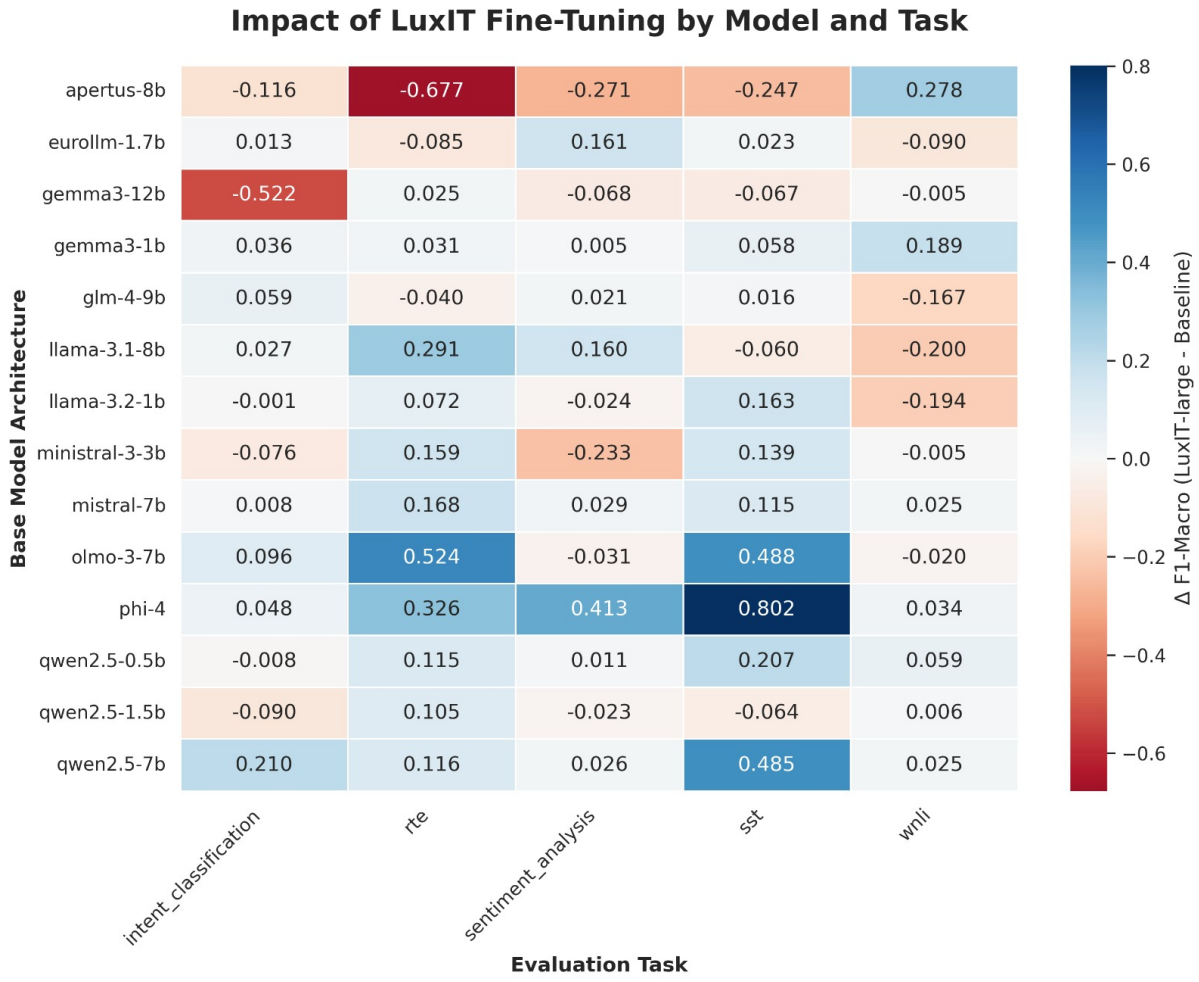


Figure 4: Impact of LuxIT Fine-Tuning by Model and Task (ΔF1-Macro).

Topic Modeling for Moroccan Darija: A Comparative Study of Classical Machine Learning Approaches

Salma MEKAOUI[1] Ilham CHAKER[2] Arsalane ZARGHILI[2] Nikola S. Nikolov[1]
[1]University of Limerick, Department of CSIS, Ireland
[2]Faculty of Sciences and Technology, USMBA, Morocco
mekaoui.salma@ul.ie

Abstract

Topic modeling for low-resource dialects such as Moroccan Arabic (Darija) remains largely underexplored, mainly due to the scarcity of linguistic resources and standardized datasets. In this work, we present a comparative study of lightweight topic modeling methods, including Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), and Latent Semantic Analysis (LSA), applied to two datasets of different sizes derived from Moroccan Darija Wikipedia corpora. To identify the optimal number of topics and ensure robust evaluation, we employ multiple complementary metrics, namely coherence C_v, normalized pointwise mutual information (NPMI), and Topic Diversity. Our results highlight the importance of jointly considering these metrics, as relying on a single measure may lead to suboptimal conclusions. Across all experiments, NMF consistently achieves the best overall performance, reaching a coherence C_v score of 0.76, an NPMI score of 0.20, and high Topic Diversity across both small and large corpora. To the best of our knowledge, this is among the first studies to systematically compare multiple classical topic modeling methods on native Moroccan Darija text. These findings provide a baseline for future work on dialectal Arabic and pave the way for exploring more advanced neural topic modeling approaches.

1 Introduction

Natural Language Processing (NLP) has experienced significant advancements in recent years, particularly for high-resource languages such as English and Modern Standard Arabic (MSA) (Jannani et al., 2025; Abuzayed and Al-Khalifa, 2021; Kumari et al., 2026). However, many dialects and low-resource language varieties remain underexplored. Moroccan Arabic (Darija), a widely spoken Arabic dialect, is one such example. Despite its extensive use in informal communication, Darija lacks standardized orthography, large annotated corpora, and well-developed linguistic resources, making computational processing particularly challenging.

Topic modeling is a key task in NLP that aims to uncover latent thematic structures within collections of documents (Blei et al., 2003). It has been widely applied to extract meaningful information from large corpora across various domains (Vayansky and Kumar, 2020). However, most existing work focuses on high-resource languages, while the application of topic modeling to dialectal Arabic, including Darija, remains limited.

In this paper, we present an empirical study of topic modeling techniques applied to Darija. We focus on classical machine learning approaches, namely LDA, NMF, and LSA. These methods are evaluated on a Darija corpus to assess their effectiveness in a low-resource setting. Given the noisy and non-standard nature of Darija text, we propose a preprocessing pipeline that includes normalization and stopword removal in order to improve the quality of the extracted topics. To ensure a robust evaluation, we employ multiple complementary metrics, including coherence C_v, NPMI, and Topic Diversity.

The contributions of this work are threefold:

- To the best of our knowledge, this study provides one of the first empirical evaluations of topic modeling techniques for Darija using a Wikipedia derived dataset, and investigates the impact of the number of topics on model performance.
- We propose a pre-processing pipeline tailored to the orthographic and lexical characteristics of Darija, including a custom normalization scheme and a dialect-specific stopword list.

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 121–130
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_014

- We demonstrate the importance of combining multiple evaluation metrics for a more reliable assessment of topic quality.

The results offer insights into the behavior of topic modeling methods when applied to noisy, low-resource dialectal data and highlight the challenges and opportunities for future research in this area.

2 Related Works

Moroccan Darija has recently attracted increasing attention from the NLP community despite its status as a low-resource dialect. Research efforts have focused on developing linguistic resources, language models, and task-specific applications to address challenges related to limited annotated data and the lack of standardized orthography. In the field of machine translation, Rosca et al. (2025) investigated methods for improving Neural Machine Translation (NMT) systems for translation into English. Their work explored data augmentation strategies such as back-translation and paraphrasing, as well as transfer learning using multilingual and bilingual models, demonstrating the effectiveness of these approaches in low-resource settings. Recent studies have also examined the application of Large Language Models (LLMs) to dialectal Arabic text processing. Aftiss et al. (2025) conducted an empirical evaluation of LLMs for abstractive summarization of Moroccan news articles written in this variety, comparing Arabic-specific, multilingual, and causal language models under both fine-tuning and zero-shot settings. Similarly, Ennasri et al. (2025) developed a chatbot for the Moroccan fiscal administration that combines LLM-based translation with semantic similarity techniques to facilitate citizen interactions. Several works have focused on enriching lexical and semantic resources for this dialect. Belbachir et al. (2025) proposed an automated methodology for identifying semantic relations such as synonyms, antonyms, hypernyms, and hyponyms by leveraging bilingual resources and WordNet. Extending this effort, Belbachir et al. (2026) introduced a methodology for constructing a structured lexical-semantic resource through WordNet and SUMO ontology alignment, enhanced with contextual word embeddings to improve semantic accuracy.

Text classification has also received considerable attention. Rachidi et al. (2025) studied the impact of data augmentation techniques on emotion detection, showing that Easy Data Augmentation (EDA) can alleviate data scarcity and improve BERT-based classifiers. In the banking domain, Skiredj et al. (2025) introduced the DarijaBanking dataset and proposed BERTouch, a specialized model for intent classification, achieving competitive results compared to retrieval-based and LLM-based approaches.

The development of language resources and pretrained models has further accelerated research in this area, Gaanoun et al. (2025) introduced DarijaBERT, the first family of BERT models trained specifically for Moroccan Arabic in both Arabic and Latin scripts. These models were evaluated on several downstream tasks, including sentiment analysis, dialect identification, and topic classification, highlighting the advantages of monodialectal pretraining. In the speech processing domain, Labied et al. (2024) developed DARIJA-C, a crowdsourced speech corpus designed to support automatic translation from spoken forms into Modern Standard Arabic (MSA).

Beyond traditional NLP tasks, this dialect has also been explored in educational and conversational systems. El Alaoui and Cavalli-Sforza (2024) proposed DarijaGenie, a multimodal chatbot that assists non-native speakers in learning Moroccan Arabic through interactive simulations. The system relies on a dedicated conversational corpus and Natural Language Understanding techniques for intent recognition and dialogue management.

On the other hand, Habbat (2022) analyzed Moroccan tweets by first translating Arabic and French content into English and then applying LDA topic modeling to sentiment based categories. After labeling tweets as positive, negative, or neutral using machine learning classifiers, the authors identified latent topics within each sentiment class. These studies demonstrate the growing maturity of the NLP ecosystem for Moroccan Arabic, spanning machine translation, summarization, lexical resource construction, text classification, conversational agents, and language resource development. However, comparatively limited attention has been devoted to topic modeling and thematic discovery in native corpora. As noted earlier, Habbat (2022) performed topic modeling only after translating tweets into English, meaning that thematic extraction was not conducted on the original texts, and the number of topics was not systematically investi-

gated. This gap motivates the present work, which investigates topic modeling techniques directly on Moroccan Darija data to better understand latent thematic structures in dialectal text collections.

Most closely related to the present work, Amzil et al. (2026) conducted a large-scale comparison of classical (LDA, NMF) and embedding-based (BERTopic, Top2Vec) topic modeling methods on over 17,000 short Moroccan Darija texts collected from online/social media sources spanning ten topic domains, using a fixed number of ten topics throughout. While this study provides a valuable benchmark for classical-vs-embedding performance on short, informal Darija text, it does not include LSA, does not investigate how the number of topics affects model performance, and its findings are specific to the short-text, social-media regime. The present work instead focuses on longer-form, encyclopedic Darija text (Wikipedia-derived corpora), evaluates LSA alongside LDA and NMF, and systematically investigates how corpus size and topic count affect classical method performance, questions left unaddressed by Amzil et al. (2026).

3 Methodology

This study follows a topic modeling pipeline consisting of preprocessing, feature extraction, topic modeling, and evaluation as shown in Figure 1.

The preprocessing stage includes text normalization and stopwords removal to reduce noise in dialectal Arabic data. The processed text is then transformed into numerical representations using both bag-of-words and TF-IDF (Manning, 2008) vectorization techniques.

Three classical topic modeling methods are evaluated: LDA (Blei et al., 2003), NMF (Shahnaz et al., 2006), and LSA (Dumais et al., 2004). These methods are applied to uncover latent thematic structures within the corpus. To investigate the impact of the number of topics on model performance, experiments were conducted using different topic counts. This analysis allows us to assess the behavior of each method under varying configurations and to identify the topic cardinality that yields the most coherent and meaningful topics according to the evaluation metrics.

To assess topic quality, we employ three complementary evaluation metrics: topic coherence C_v (Röder et al., 2015), which measures the semantic relatedness of the top words within a topic, NPMI (Aletras and Stevenson, 2013), which evaluates word co-occurrence patterns, and Topic Diversity (Dieng et al., 2020), which quantifies the uniqueness of words across topics.

The metrics were first analyzed independently to examine their individual behavior and their ability to characterize topic quality. Subsequently, a composite score was computed to facilitate the comparison of topic modeling methods. This score combines the three metrics through a weighted aggregation scheme, allowing multiple aspects of topic quality to be considered simultaneously.

Greater emphasis was placed on coherence C_v and NPMI measures, as they directly reflect the semantic interpretability of the generated topics. Topic Diversity was also incorporated to account for topic distinctiveness, although with a relatively lower contribution in order to avoid excessively penalizing meaningful overlaps that may naturally occur across related topics. The weighting strategy adopted in this study is detailed in the experimental setup section. The Final Score is computed as follows:

$$\text{Score} = \alpha \cdot C_v + \beta \cdot \text{NPMI} + \gamma \cdot \text{TD}$$

where C_v denotes topic coherence C_v, NPMI denotes Normalized Pointwise Mutual Information, TD denotes Topic Diversity, and α, β, and γ are the corresponding weighting coefficients.

4 Experiments and Results

In this section, we provide a detailed description of the datasets used for topic modeling in Darija. We also present the experimental setup, including the parameter configurations used in our experiments, as well as the obtained results and their discussion. The full code and detailed experimental results are available in a public repository[1].

4.1 Dataset Description

The experiments were conducted on two Darija corpora: the Moroccan Darija Wiki Audio Dataset[2] and Arywiki (B., 2025). Both datasets originate from the Moroccan Darija Wikipedia[3] and were selected due to their suitability for topic modeling and other NLP classifcation tasks.

[1] https://github.com/SalmaMekaoui/Darija-topic-modeling

[2] https://huggingface.co/datasets/atlasia/Moroccan-Darija-Wiki-Audio-Dataset

[3] https://ary.wikipedia.org/wiki

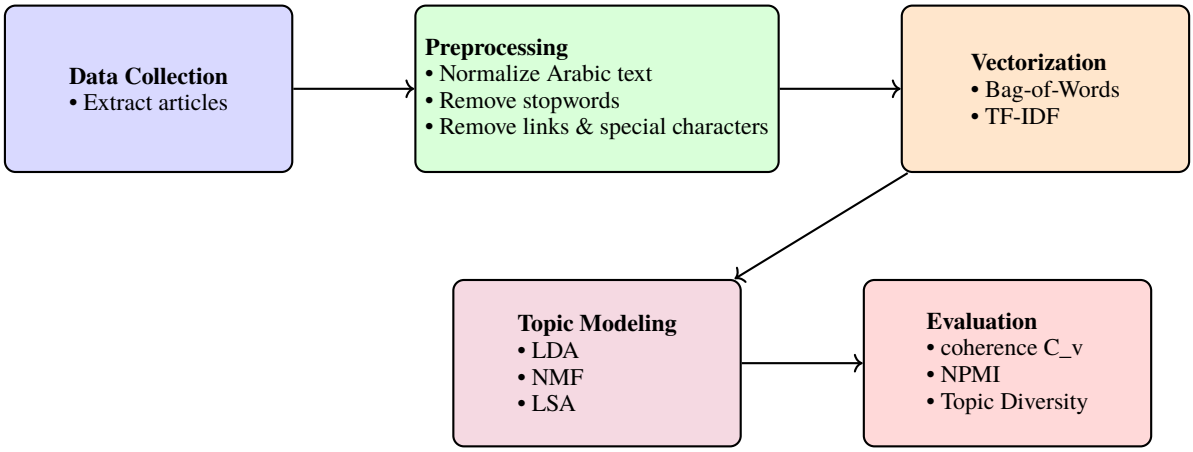


Figure 1: Overview of the proposed topic modeling methodology. The pipeline includes preprocessing, feature extraction, application of topic modeling methods and evaluation.

The primary motivation for choosing these datasets is their common source domain combined with their substantial difference in size. This characteristic enables an analysis of how different topic modeling approaches behave under varying corpus sizes while minimizing the influence of domain-specific variations. Since both datasets are derived from Wikipedia articles written in Darija, they provide relatively homogeneous content in terms of style and topic coverage, allowing a more controlled comparison of topic modeling methods.

Furthermore, both corpora have undergone preprocessing and cleaning procedures prior to their release, making them suitable for downstream NLP applications. The Moroccan Darija Wiki Audio Dataset consists of 551 Wikipedia articles and includes audio recordings aligned with the textual content. In addition to its multimodal nature, the dataset benefits from curation efforts that contribute to the quality of the textual resources. In contrast, Arywiki is a larger corpus containing more than 10,000 articles extracted from a cleaned and processed snapshot of the Moroccan Darija Wikipedia. The corpus has been filtered and normalized to provide a high-quality textual resource for NLP research involving the Darija dialect. Table 1 summarizes the main characteristics of the two datasets.

4.2 Preprocessing

Although both datasets were released in a preprocessed and cleaned form, an additional preprocessing step was necessary for our experiments. During preliminary analysis, we observed the persistence of a significant number of dialectal stopwords, as well as noise in the form of inconsistent character usage and orthographic variations, which can negatively affect topic modeling performance.

To address these issues, we applied a normalization procedure tailored to Moroccan Darija. This process included the unification of variant Arabic characters (e.g., different forms of *alif*, *ya*, and *ta marbuta*) and the elimination of non-Arabic symbols and URLs. Furthermore, extra whitespace and redundant characters were cleaned to ensure a more consistent textual representation.

In addition, a custom stopword list was constructed, combining standard Arabic stopwords with frequently occurring Darija-specific terms (e.g., *had*, *dial*, *bzaaf*, *daba*), which were identified as non-informative for topic extraction. Tokens shorter than a predefined threshold were also removed to reduce noise. The complete stopword list and preprocessing code are available in the project repository[4].

Finally, documents with very limited textual content were filtered out to ensure sufficient semantic richness for topic modeling. These preprocessing steps resulted in a cleaner and more coherent corpus, improving the quality and interpretability of the extracted topics.

4.3 Vectorization

For document vectorization, two different representations were employed. For LDA, a Bag-of-Words (BoW) representation was used, as LDA is a probabilistic generative model that operates directly on term frequencies. In contrast, TF-IDF

[4]https://github.com/SalmaMekaoui/Darija-topic-modeling

Table 1: Statistics of the datasets used in this study.

Statistic	Arywiki	Moroccan Darija Wiki Audio Dataset
Documents	10,726	551
Total Words	1,098,517	97,254
Vocabulary Size	137,163	23,501
Average Document Length	102.42	176.50

weighting was applied for LSA and NMF, since both methods are based on matrix factorization and generally benefit from weighting schemes that emphasize informative terms while reducing the influence of highly frequent words.

4.4 Results and Discussions

For the evaluation, we employ three metrics: topic coherence C_v, NPMI, and Topic Diversity. In addition, a composite score is reported, combining the three metrics as described in Section 3. The weighting coefficients α, β, and γ are set to 0.4, 0.4, and 0.2, respectively, giving equal and higher priority to coherence C_v and NPMI as direct measures of interpretability, with a lower weight assigned to diversity to avoid over-penalizing meaningful topic overlap (see Section 3).

To ensure a fair and consistent comparison across methods while identifying the best-performing approach, we varied the number of topics from 10 to 100 for each dataset. This analysis aims to determine the topic cardinality that yields the most informative representation of the corpus based on the selected evaluation metrics.

In this section, we present the results obtained by applying the three topic modeling methods, LDA, LSA and NMF, separately to the Moroccan Darija Wiki Audio Dataset and the Arywiki dataset. The performance of each method is analyzed using both the individual evaluation metrics and the composite score.

4.4.1 Topic Modeling on the Moroccan Darija Wiki Audio Dataset

For the Darija Wiki Audio Dataset, only the textual content was considered. The three topic modeling methods were applied and evaluated using the proposed metrics. As shown in Figure 2, NMF consistently outperformed both LDA and LSA across all evaluation metrics. In particular, NMF achieved the highest composite score and produced highly coherent topics, with coherence C_v values exceeding 0.7 when the number of topics was set to 10.

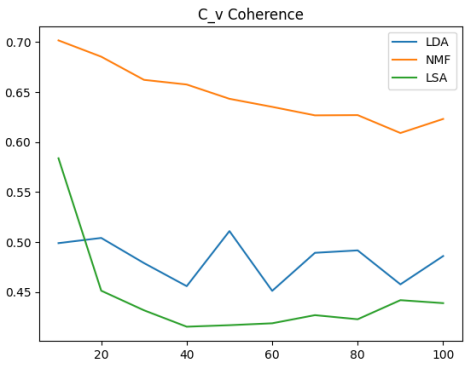
Another notable observation is the stability of LDA across different topic counts. Although its performance remained below that of NMF, the evaluation metrics exhibited relatively small variations as the number of topics increased, suggesting that LDA is less sensitive to changes in topic cardinality. In contrast, the performance of LSA deteriorated as the number of topics increased, resulting in lower coherence C_v, NPMI, diversity, and overall scores.

While the coherence C_v scores obtained by NMF and LSA generally fall within the range of 0.4 to 0.6, suggesting a moderate degree of semantic consistency among the generated topics, the NPMI values remained negative across most experimental settings. Negative NPMI values indicate limited co-occurrence among the top words within topics, suggesting that the discovered topics are not strongly supported by word co-occurrence statistics. This behavior may be attributed to the relatively small size of the corpus, the sparsity of the vocabulary, or linguistic characteristics of Darija.

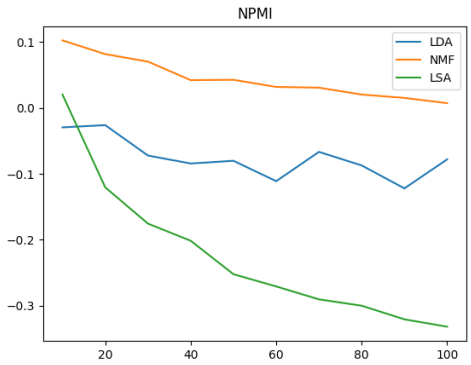
These observations highlight the importance of evaluating topic quality using multiple complementary metrics rather than relying solely on coherence C_v. Consequently, the composite weighted score provides a more balanced assessment by jointly considering coherence C_v, NPMI, and Topic Diversity, thereby enabling a more comprehensive comparison of the evaluated topic modeling methods. The resulting topics, along with their translations, are reported in Table 2.

4.4.2 Topic Modeling on Arywiki Dataset

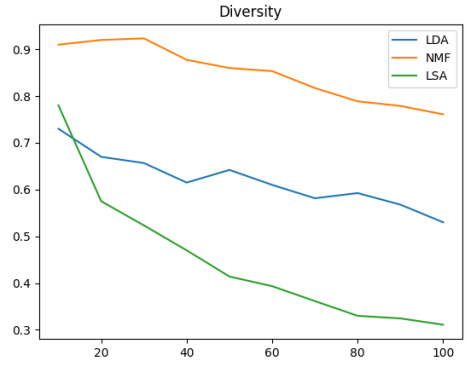
The main objective of repeating the experiments on a different dataset is to evaluate the robustness of the topic modeling methods under varying corpus sizes. The Moroccan Darija Wiki Audio Dataset is relatively small, with only 551 articles, which is insufficient to draw generalizable conclusions. For this reason, we also use a larger dataset derived from the Moroccan Wikipedia, which contains more than 10,000 articles, as shown in Table 1.



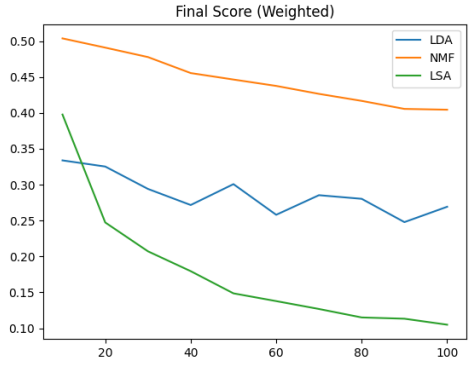
(a) coherence C_v



(b) NPMI



(c) Topic Diversity



(d) Final Score

Figure 2: Evaluation metrics for topic modeling methods across different numbers of topics on The Moroccan Darija Wiki Audio Dataset.

The results again show that NMF performs best on this dataset, achieving a coherence C_v score of approximately 0.75, as shown in figure 3. LDA exhibits a similar behavior to that observed in the smaller dataset, remaining relatively stable across different numbers of topics, and this time achieving positive NPMI values and a competitive overall score.

In contrast, LSA consistently underperforms and appears particularly sensitive to the dataset characteristics. It is the only method that produces negative NPMI values on this dataset. While LSA performs reasonably well for small numbers of topics (e.g., 10 or 20 topics), its performance degrades significantly as the number of topics increases, making it less suitable for Moroccan Darija corpora compared to NMF and LDA.

Another observation is that both NMF and LDA achieve their best coherence C_v and NPMI scores when the number of topics is around 20. However, Topic Diversity is highest when the number of topics is 10 and decreases noticeably as the number of topics increases. This highlights a trade-off between topic interpretability and redundancy, as topic modeling aims to produce compact and non-overlapping representations of large corpora. These results further motivate the use of the proposed composite score for selecting the most appropriate model configuration. The resulting topics and their translations are reported in Table 3.

Finally, it is well known that LDA tends to perform better on larger corpora (Mifrah and Benlahmar, 2020; Nanyonga et al., 2023). This expectation is partially confirmed, as performance improves on the larger dataset compared to the smaller one, although NMF remains superior overall. This can be partly explained by the characteristics of the corpus: although the dataset is larger, the vocabulary size remains moderate, and the average document length is relatively short with approximately 102 tokens (table 1). This limits LDA's ability to effectively learn stable topic distributions, which typically benefit from longer documents and richer word co-occurrence patterns.

Topic ID	Label	Top Words
0	جغرافيا (Geography)	امريكا، لمحيط، لكاريبي، لبلاد، الشماليه، لجزر، بلاد، ليزيه، الصبليونيه، لهادي "America, ocean, Caribbean, country, northern, islands, country, money, Spanish, Pacific"
1	شخصيات مغربية (Moroccan figures)	مغربي، ناس، لمغريب، محمد، تولد، زياده، فاس، مات، عبد، تزادو "Moroccan, people, Morocco, Mohammed, born, increase, Fes, died, Abd, were born"
2	شركات وصناعة (Companies & industry)	شركه، الشركه، الطوموبيلات، جاپونيه، لماطر، المانيه، لكاميويات، تاسسات، لمقر، لمعروفه "company, the company, cars, Japanese, airports, German, trucks, established, headquarters, known"
3	عواصم ومدن (Capitals & cities)	اكبر، لعاصيما، مدينه، عواصم، لمدينه، مطار، لبلاد، جات، سان، ليونيسكو "largest, capital, city, capitals, city, airport, country, located, Saint/San, UNESCO"
4	سياسة أمريكية (US politics)	لميريكان, الريييس, ميريكان, ميريكاني, حكم, ويكيميديا, رايس, رييس, كومنز, لمزاود "American, president, American, American (adj), rule, Wikimedia, president, head, Commons, hot peppers"
5	قارات ومناطق (Continents & regions)	اسيا، لبلاد، لجمهوريه، لجنوب، بلاد، وروپا، الشمال، اوروپا، لبحر، لمتحده "Asia, country, republic, south, country, Europe, north, Europe, sea, united"
6	اللغات (Languages)	لوغه، لوغات، واحد، رسميه، لفرانساويه، علي، ليزيه، لاتينيه، اللوغه، لغه "language, languages, one, official, French, uses, money, Latin, language, language"
7	دول إفريقيا (African countries)	افريقيا، لبلاد، غينيا، السودان، لكون، جنوب، جمهوريه، لغرب، لفرانساويه، بلاد "Africa, country, Guinea, Sudan, Universe, south, republic, west, French language, country"
8	كرة القدم (Football)	كاس، لكوره، كوره، فرقه، لفرقه، الطاليان، لعاب، التيران، لعالم، ميلانو "cup, football, ball, team, club, Italian, players, stadium, world, Milan"
9	جهات المغرب (Moroccan regions)	لجيهه، جهه، اقليم، لوطنيه، فلجيهه، مكناس، لمغريب، مطار، والطريق، مراكش "region, region, province, national, in the region, Meknes, Morocco, airport, road, Marrakesh"

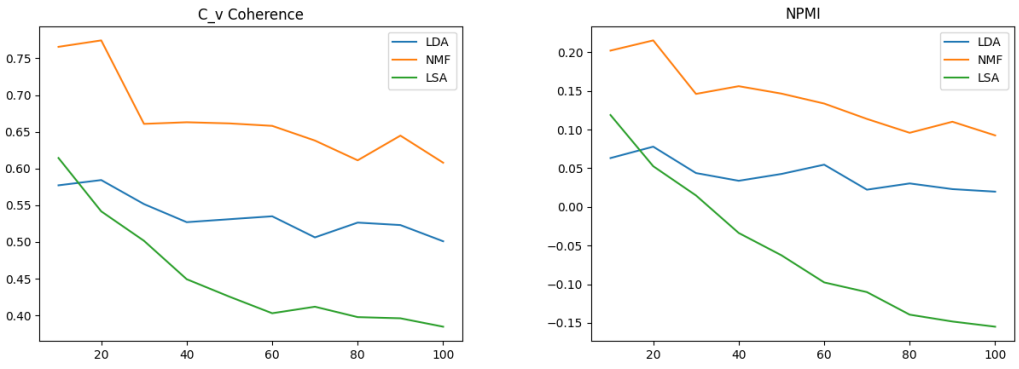
Table 2: Top words for each discovered topic for the Moroccan Darija Wiki Audio Dataset. Arabic terms are supplemented with English translations for readability. Topic labels (in Darija/Arabic script and English) were assigned manually by the authors based on the top words to facilitate interpretation.

5 Conclusion

This work presented a comparative study of lightweight statistical topic modeling methods applied to Moroccan Arabic (Darija) datasets. We conducted experiments on two corpora of different sizes, both derived from the Moroccan Darija Wikipedia dump. To determine the optimal number of topics, we employed multiple evaluation metrics commonly used in topic modeling, namely coherence C_v, Topic Diversity, and NPMI. Our results highlight the importance of combining these complementary metrics rather than relying on a single measure to obtain reliable and interpretable results.

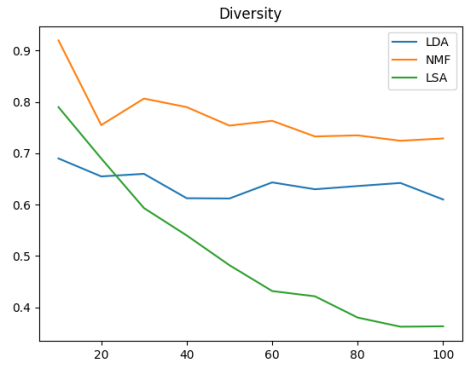
We evaluated three traditional topic modeling approaches NMF, LSA, and LDA. Across all experimental settings, NMF consistently outperformed the other methods, achieving a coherence C_v score of 0.76, an NPMI score of 0.20, and a Topic Diversity of 0.92. Furthermore, NMF effectively identified 10 meaningful topics in both corpora (comprising 551 and 10,000 articles, respectively), demonstrating its robustness across different dataset sizes.

To the best of our knowledge, this is the first study to investigate topic modeling techniques for Moroccan Darija, a low-resource dialect, and to systematically compare the performance of classical methods on such data. Nevertheless, this work should not be considered a definitive answer to topic modeling in Darija. Several limitations remain, including the reliance on traditional models and the absence of human evaluation or gold-standard benchmarks, which are crucial for validating topic quality in unsupervised settings.

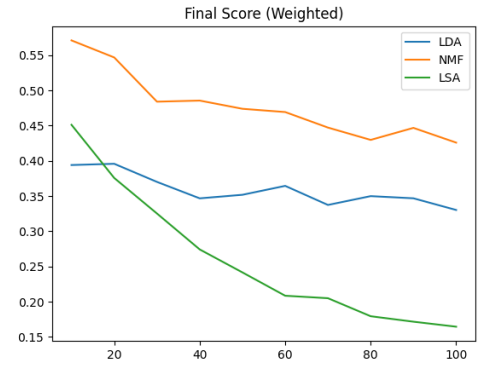


(a) coherence C_v

(b) NPMI



(c) Topic Diversity



(d) Final Score

Figure 3: Evaluation metrics for topic modeling methods across different numbers of topics on Arywiki dataset.

This study has several limitations. The comparison is restricted to classical methods, while embedding-based approaches such as BERTopic and Top2Vec have recently shown strong performance on short, informal Darija text; whether this advantage extends to longer, encyclopedic text such as ours remains untested. Both corpora are also drawn from Wikipedia, which offers a controlled, relatively formal register but may not reflect noisier domains such as social media or conversational speech. Finally, our evaluation relies entirely on automatic metrics, without human judgment of topic interpretability.

Future work should explore neural topic modeling approaches and leverage pre-trained language models to better capture the linguistic characteristics of Darija. Additionally, expanding the analysis to larger and more diverse datasets beyond Wikipedia, as well as incorporating human evaluation, would provide a more comprehensive understanding of topic modeling performance in this context.

6 Acknowledgements

This publication has emanated from research conducted with the financial support of Taighde Éireann - Research Ireland under Grant number 18/CRT/6223.

References

Abeer Abuzayed and Hend Al-Khalifa. 2021. Bert for arabic topic modeling: An experimental study on bertopic technique. *Procedia computer science*, 189:191–194.

Azzedine Aftiss, Salima Lamsiyah, Christoph Schommer, and Said Ouatik El Alaoui. 2025. Empirical evaluation of pre-trained language models for summarizing moroccan darija news articles. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 77–85.

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th interna-*

Topic ID	Label	Top Words
0	ديموغرافيا (Demographics)	واصله, نسبه, الجواج, عدد, التسكويل, احصاييات, دوار, لخصوبه, لعيالات, تزاد "rate, percentage, marriage, number, schooling, statistics, village, fertility, women, increase"
1	التقويم (Calendar systems)	تقويم, ثمره, نهار, لميلاد, لهولوسيني, كيوافق, لهيجري, لامازيغي, لايرا, لعامه "calendar, number/date, day, Gregorian, Holocene, corresponds, Hijri, Amazigh, era, year"
2	التقسيم الترابي (Territorial divisions)	ترابيه, حسب, ساكنين, قرويه, كاينه, لعام, اقليم, جماعه, جهه, لاحصاء "territorial, according to, inhabitants, rural, existing, year, province, municipality, region, census"
3	دواوير وقرى (Villages & settlements)	دوار, تارودانت, ماسه, سوس, مجمع, مشيخه, كتضم, دواور, كينتامي, نوطات "village, Taroudant, Massa, Souss, cluster, district, includes, villages, belongs to, notes"
4	السنة والتقويم (Days & years)	نهار, لعام, مكبس, يلا, لميلادي, باقين, موراه, لخر, شهر, يوم "day, year, leap, if, Gregorian, remaining, after, last, month, day"
5	التسلسل الزمني (Chronology)	عقد, كرونولوجيا, مقالات, شوف, الزمان, دجنبر, يناير, سالا, لعقد, عوام "decade, chronology, articles, see, time, December, January, ended, decade, years"
6	جهة مراكش آسفي (Marrakesh-Safi region)	اسفي, مراكش, لحوز, دوار, لاحصاء, الصويره, ايت, اسني, تحناوت, جهه "Safi, Marrakesh, Al Haouz, village, census, Essaouira, Ait, Asni, Tahanaout, region"
7	العناصر الكيميائية (Chemical elements)	العنصر, عنصر, مرتب, دريه, كيماوي, لعمود, طابلو, السطر, لمعياريه, دوري "element, element, ordered, particle, chemical, column, table, row, standard, periodic"
8	شخصيات وأماكن (Figures & places)	علي, واحد, لبلاد, مدينه, بين, محمد, لعالم, بعد, قبل, اكبر "Ali, one, country/village, city, between, Mohammed, world, after, before, biggest"
9	جهة فاس مكناس (Fez-Meknes region)	مكناس, فاس, عماله, تاونات, عين, تازه, لاحصاء, احواز, بولمان, لعام "Meknes, Fes, prefecture, Taounate, Ain, Taza, census, surroundings, Boulemane, year"

Table 3: Top words for each discovered topic for Arywiki dataset. Arabic terms are supplemented with English translations for readability. Topic labels (in Darija/Arabic script and English) were assigned manually by the authors based on the top words to facilitate interpretation.

tional conference on computational semantics (IWCS 2013)–Long Papers, pages 13–22.

Taoufik Amzil, Ayoub Jannani, Rachida Ait Abdelouahid, and Nawal Sael. 2026. Short text topic modeling for moroccan darija online content: A comparative study between classical and embedding-based approaches. In *2026 6th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, pages 1–6.

Safouane B. 2025. Darija wiki (arywiki) cleaned dataset. https://huggingface.co/datasets/safouaneb/arywiki.

Said Belbachir, Nadia Khlif, Mohamed Chahhou, Mohammed El Mohajir, Azzeddine Mazroui, and Ouafae Nahli. 2025. A proposed approach for extracting semantic and lexical relations for low-resource languages: A case study of darija. In *2025 IEEE 8th Congress on Information Science and Technology (CiSt)*, pages 153–160. IEEE.

Said Belbachir, Mohammed El Mohajir, Mohamed Chahhou, and Ouafae Nahli. 2026. Building a semantic resource for the moroccan dialect: a hybrid approach with leoni and semantic similarity. *Language Resources and Evaluation*, 60(2):43.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Susan Dumais et al. 2004. Latent semantic analysis. *ARIST (Annual Review of Information Science Technology)*, 38:189–230.

Hamza El Alaoui and Violetta Cavalli-Sforza. 2024. Darijagenie: Learning moroccan arabic through a multimodal chatbot. In *International Conference on Arabic Language Processing*, pages 74–89. Springer.

Oumaima Ennasri, Brahim El Bhiri, and Yann Ben Maissa. 2025. Moroccan institutional chatbots: A hybrid approach with llms, semantic matching, and dialect adaptation for darija. *Engineering Proceedings*, 112(1):41.

Kamel Gaanoun, Abdou Mohamed Naira, Anass Allak, and Imade Benelallam. 2025. Darijabert: a step forward in nlp for the written moroccan dialect. *International Journal of Data Science and Analytics*, 20(2):917–929.

Nassera Habbat. 2022. Sentiment analysis and topic modeling on arabic twitter data during covid-19 pandemic. *Indonesian Journal of Innovation and Applied Sciences (IJIAS)*.

Ayoub Jannani, Soukaina Bouhsissin, Nawal Sael, and Faouzia Benabbou. 2025. Topic modeling and sentiment analysis of arabic news headlines for a societal well-being scoring and monitoring system: Moroccan use case. *IEEE Access*, 13:26345–26364.

Pratima Kumari, Sachin Kadian, Mukund Vora, Pradyumna Vemuri, Sumit Kumar, Om Prakash Verma, and Roger J Narayan. 2026. Recent advancements in topic modeling techniques for healthcare, bioinformatics, and other potential applications. *Advanced Intelligent Systems*, 8(1):2400528.

Maria Labied, Abdessamad Belangour, and Mouad Banane. 2024. Darija-c: a crowdsourced corpus for moroccan darija speech-to-text translation. *Int. J. Eng. Trends Technol*, 72(10):257–266.

Christopher D Manning. 2008. *Introduction to information retrieval*. Syngress Publishing,.

Sara Mifrah and EH Benlahmar. 2020. Topic modeling coherence: A comparative study between lda and nmf models using covid'19 corpus. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(4):5756–5761.

Aziida Nanyonga, Hassan Wasswa, and Graham Wild. 2023. Topic modeling analysis of aviation accident reports: A comparative study between lda and nmf models. In *2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, pages 1–2. IEEE.

Rabia Rachidi, Wassima Moutaouakil, Othmane Daanouni, Mounir Omari, Bouchaib Cherradi, and Hassan Silkan. 2025. Augmenting moroccan darija text data for improved classification: A comparative study of augmentation techniques. In *2025 5th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, pages 1–7. IEEE.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.

Alexei Rosca, Abderrahmane Issam, and Gerasimos Spanakis. 2025. Low-resource machine translation for moroccan arabic. In *Proceedings of the First Workshop on Advancing NLP for Low-Resource Languages*, pages 32–38.

Farial Shahnaz, Michael W Berry, V Paul Pauca, and Robert J Plemmons. 2006. Document clustering using nonnegative matrix factorization. *Information Processing & Management*, 42(2):373–386.

Abderrahman Skiredj, Ferdaous Azhari, Ismail Berrada, and Saad Ezzini. 2025. Darijabanking: A new resource for overcoming language barriers in banking intent detection for moroccan arabic speakers. *Natural Language Processing*, 31(5):1234–1264.

Ike Vayansky and Sathish AP Kumar. 2020. A review of topic modeling methods. *Information Systems*, 94:101582.

Distil and Evolve: Compact Adaptive Document Classification with Continual Learning for Low-Resource Settings

Mehdi Soufiane[1,2,3] **Mourad Hammale**[1] **Kawtar Zerhouni**[1]
Adil Ahidar-Coutrix[1] **Rachid Benouini**[2,3]
[1]Laboratory of Complex Systems, École Centrale Casablanca, Morocco
[2]LEYTON, Casablanca, Morocco [3]LEYTON CognitX
{mehdi.soufiane, mourad.hammale}@centrale-casablanca.ma
{adil.ahidar, kawtar.zerhouni}@centrale-casablanca.ma

Abstract

Document image classification is the entry point of document-processing pipelines, yet the institutions that need them most operate under tight compute and annotation budgets. Knowledge distillation (KD) compresses large document Transformers for such deployment, but whether a heavily distilled model can still undergo continual learning (CL) without catastrophic forgetting, and which CL strategy suits it, is unclear. We distil a DiT-Large teacher (307 M parameters) into a ViT-Tiny student (5.5 M, 21.2 MiB) and evaluate four CL methods in a two-phase class-split protocol on RVL-CDIP across five random seeds. Replay-based methods significantly outperform regularisation-based ones: DER++ retains $90.4 \pm 1.2\%$ of prior accuracy against $75.9 \pm 2.1\%$ for the best regulariser, a 14.5 pp paired difference ($p < 10^{-4}$). Every pairwise ordering is significant and holds at every seed. Because hyperparameters were selected on one seed, we also measure the resulting optimism: the selection seed overstates retention by up to 7.4 pp, and most for the weakest method. Per-class analysis shows macro retention conceals concentrated damage, with two of five target classes absorbing most of the loss. We profile each strategy against deployment metrics, including instrumented GPU energy and single-thread CPU latency, and find that counting method preparation shrinks the time advantage of regularisation over replay from 9.1 to 3.3 minutes. A same-size non-distilled control starts near chance and therefore cannot isolate a distillation effect; we report it as a diagnostic only. We release the distilled model, the harness, and all seed-level artifacts.

1 Introduction

Automated classification of document images, including invoices, contracts, forms, and scientific reports, is a core component of enterprise information systems. Transformer-based document understanding models such as DiT (Li et al., 2022) and LayoutLMv3 (Huang et al., 2022) achieve strong accuracy but contain hundreds of millions of parameters, making them impractical for latency-sensitive or resource-constrained deployment. This barrier is most acute in low-resource contexts, where digitisation pipelines must run on limited infrastructure and large annotated corpora are unavailable. Document image classification is largely script-agnostic, operating on visual layout rather than textual content, so compact adaptive classifiers are a transferable building block for document processing in low-resource languages, including the multilingual administrative archives common in North African institutions. Knowledge distillation (KD) (Hinton et al., 2015) addresses the compute side of this gap by transferring a large teacher's representations into a compact student, with modest accuracy loss (Gou et al., 2021).

Compression is only part of the deployment challenge. Once in production, a document classifier must adapt as organisational needs evolve. Naïvely fine-tuning on new classes causes catastrophic forgetting (McCloskey and Cohen, 1989; French, 1999): the model loses previously learned classes. Continual learning (CL) methods (Wang et al., 2024) mitigate this, but have been studied primarily on full-capacity models. A distilled student has substantially reduced representational redundancy compared with its teacher: fewer layers, a narrower hidden dimension, fewer attention heads. This motivates a practical question. Parameter-regularisation methods rely on spare weight-space directions where new knowledge can be absorbed without disturbing old representations, and a compact model offers fewer of them; replay methods instead inject direct gradient signal for past tasks at every step. Which family should a practitioner deploying a distilled doc-

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 131–139
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_015

ument classifier actually choose?

We stress at the outset what this study can and cannot settle. Our experiments hold capacity fixed at ViT-Tiny. They therefore establish which family wins *on a compact distilled document classifier*, with quantified uncertainty, but they cannot attribute that outcome to compactness: separating "regularisation fails because the model is small" from "regularisation underperforms replay here as it commonly does at full capacity" requires a capacity sweep that we did not run. We treat the capacity argument as motivation, not as a finding.

Two families offer contrasting mechanisms. Regularisation-based methods penalise changes to parameters deemed important for previous tasks. Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017) measures importance via the diagonal Fisher information matrix, while Memory Aware Synapses (MAS) (Aljundi et al., 2018) uses the sensitivity of the network's output norm to parameter perturbations. Related approaches include Synaptic Intelligence (Zenke et al., 2017) and Learning without Forgetting (LwF) (Li and Hoiem, 2017), which constrains the output distribution rather than the parameters. Replay-based methods store and re-expose past examples. Experience Replay (ER) (Chaudhry et al., 2019) replays raw samples from a memory buffer, while Dark Experience Replay++ (DER++) (Buzzega et al., 2020) additionally preserves the previous model's logits for buffered samples. Among replay variants, iCaRL (Rebuffi et al., 2017) and GEM (Lopez-Paz and Ranzato, 2017) further constrain gradient updates.

Our contributions are:

- a five-seed evaluation of four CL methods on a distilled ViT-Tiny document classifier, with paired significance tests, establishing that the replay advantage is a stable ordering rather than a single-run artifact;
- a direct measurement of how much single-seed hyperparameter selection inflates reported retention, and of which methods it flatters most;
- a per-class analysis showing that macro retention hides concentrated degradation, and a corrected reading of the non-distilled control, which we show cannot support a distillation claim;
- an instrumented deployment profile including GPU energy, method preparation cost, and single-thread CPU latency, released with all seed-level artifacts.

2 Related Work

Document representations. LayoutLM introduced joint text-and-layout pre-training (Xu et al., 2020); LayoutLMv2 extended it multimodally (Xu et al., 2021); LayoutLMv3 unifies text and image masking (Huang et al., 2022); DiT studies self-supervised pre-training directly on document images (Li et al., 2022). These concern representation learning and backbone design. We instead fix one compact student and study its post-deployment adaptation.

Continual learning. Wang et al. (2024) survey CL across five families but do not address how compression modulates their relative effectiveness. CL has been studied for general image recognition (Rebuffi et al., 2017) and for language (d'Autume et al., 2019), yet CL on distilled document classifiers remains little explored.

Replay-free alternatives. Recent parameter-efficient work offers an important alternative to the four methods compared here. L2P learns a prompt pool over a frozen backbone (Wang et al., 2022), and CODA-Prompt learns input-conditioned prompt components without rehearsal data (Smith et al., 2023). These avoid retaining a document buffer entirely, which matters where page retention is restricted by governance rather than by storage. We did not implement them: both assume a frozen, high-capacity pre-trained backbone, and their behaviour on a 5.5 M-parameter student that must itself keep adapting is an open question we flag as the most relevant extension to our comparison.

3 Methodology

3.1 Knowledge Distillation Pipeline

We employ offline logit-based distillation on RVL-CDIP (Harley et al., 2015), a benchmark of 16 document classes. The original dataset contains approximately 400,000 images; our experiments use a curated subset[1] of 30,400 training and 3,200 test images. The subset was adopted for tractability under our compute budget; its sampling proce-

[1] `sitloboi2012/rvl_cdip_large_dataset` on Hugging Face Datasets, pinned at revision `05a84f4e`.

dure is not documented by the source, so our absolute accuracies are not directly comparable to published full-split RVL-CDIP numbers, and we do not claim they are. All comparisons in this paper are internal to the subset. The student minimises

$$\mathcal{L}_{\text{KD}} = \alpha \cdot \mathcal{L}_{\text{CE}}(y, \sigma(z_S)) + (1-\alpha) \cdot T^2 \cdot \text{KL}(\sigma(z_T/T) \| \sigma(z_S/T)) \quad (1)$$

where z_T and z_S are teacher and student logits, T is the temperature, α balances hard and soft targets, and σ is softmax.

The teacher is DiT-Large (Li et al., 2022) (307 M parameters, 24 layers). The student is ViT-Tiny (5,527,504 parameters, 12 layers, hidden size 192, 3 heads, 21.2 MiB on disk, 1.26 GMACs per image), a 55.5× parameter reduction. The distilled student is the starting point θ_{orig} for all CL experiments.

3.2 Two-Phase Protocol and Metrics

We evaluate CL using a two-phase protocol (Figure 1). In Phase 1 (Task A, specialisation), θ_{orig} is fine-tuned without CL protection on 5 target classes, producing a specialist θ_{naive}. In Phase 2 (Task B), each CL method starts from θ_{naive} and fine-tunes on the 11 remaining classes. All CL anchors are computed from θ_{naive} on Task-A data only. We also evaluate a No-CL Phase-2 condition as a lower bound.

Because θ_{orig} already predicts all 16 classes, Task B is a *recovery* group rather than a set of globally unseen classes. We therefore use "two-phase class-split adaptation" rather than "class-incremental learning". This matches a deployed classifier that is temporarily specialised and later broadened; it does not cover the case where genuinely new labels arrive after deployment.

Let $a_c(\theta)$ be the per-class accuracy of θ, $\mathcal{T}_A$ the 5 target classes and $\mathcal{T}_B$ the 11 remaining. Task-A retention measures the fraction of θ_{naive}'s target-class accuracy preserved:

$$\text{Ret} = \frac{1}{|\mathcal{T}_A|} \sum_{c \in \mathcal{T}_A} \frac{a_c(\theta_{\text{CL}})}{a_c(\theta_{\text{naive}})} \quad (2)$$

Task-B efficiency measures recovery of θ_{orig}'s accuracy on the remaining classes:

$$\text{Eff} = \frac{\frac{1}{|\mathcal{T}_B|} \sum_{c \in \mathcal{T}_B} a_c(\theta_{\text{CL}})}{\frac{1}{|\mathcal{T}_B|} \sum_{c \in \mathcal{T}_B} a_c(\theta_{\text{orig}})} \quad (3)$$

Retention is a mean of per-class ratios, so values above 100% are possible when a class improves. We also report global 16-class accuracy, because retention alone rewards a method that simply resists learning Task B.

We complement these with deployment metrics recorded by the instrumented pipeline: parameter count, on-disk size, inference latency, wall-clock adaptation time separated into method preparation and training, peak GPU memory, GPU energy, and auxiliary adaptation storage S_{aux}, the persistent memory a method needs beyond the weights (importance and anchor matrices, or the replay buffer).

3.3 Continual Learning Methods

EWC (Kirkpatrick et al., 2017) adds $\frac{\lambda}{2} \sum_i F_i(\theta_i - \theta_{\text{naive},i})^2$ to the loss, with F_i the Fisher diagonal computed on θ_{naive} from Task-A samples. MAS (Aljundi et al., 2018) shares this quadratic form but estimates importance as the mean absolute gradient of the output L_2 norm, an output-based rather than loss-curvature signal. ER (Chaudhry et al., 2019) maintains a class-balanced buffer of Task-A samples and adds a replay cross-entropy term at each Phase-2 step. DER++ (Buzzega et al., 2020) additionally stores θ_{naive}'s logits for buffered samples and adds an MSE term constraining outputs to stay close to them. Note that ER weights replay cross-entropy by 1 whereas DER++ weights it by β and adds α-weighted logit MSE, so their comparison reflects the combined configuration rather than isolating the stored-logit term.

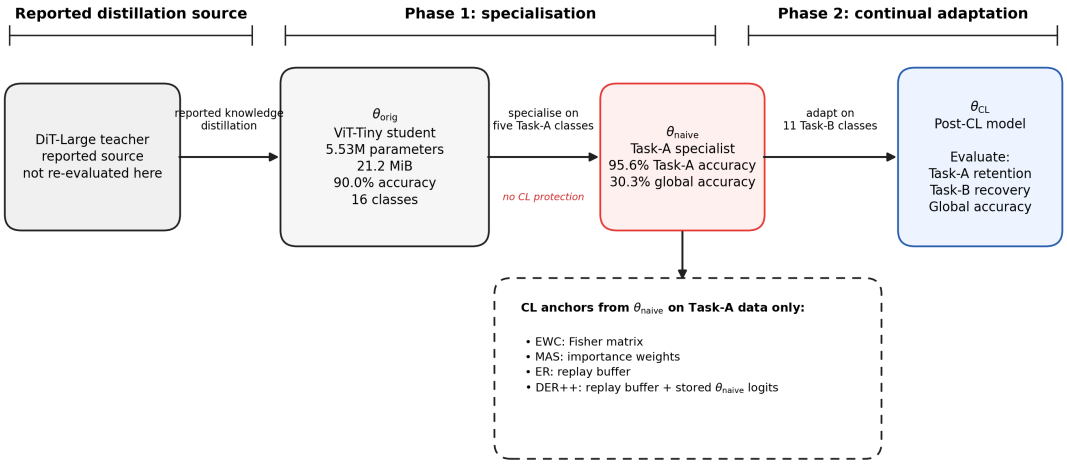


Figure 1: Two-phase class-split continual-learning protocol. The distilled ViT-Tiny student θ_{orig} (5.5 M parameters, obtained by KD from DiT-Large, 307 M) is the single starting point. Phase 1 fine-tunes it without CL protection on 5 target classes (Task A), producing the specialist θ_{naive}. Phase 2 starts from θ_{naive} and fine-tunes on the remaining 11 classes (Task B) under a CL method. All anchors (Fisher, importance weights, replay buffers, stored logits) derive from θ_{naive} on Task-A data only. The whole protocol is repeated independently for each of five seeds.

4 Experimental Setup

We use the 16-class RVL-CDIP subset[1] with 30,400 training and 3,200 test images (200 per class). Task A comprises 5 semantically diverse classes: email (2), handwritten (3), specification (7), invoice (11), presentation (12). The remaining 11 form Task B. All experiments use AdamW with learning rate 5×10^{-5}, weight decay 0.01, batch size 8, and gradient clipping at 1.0. Phase 1 uses a constant learning rate over 10 epochs; Phase 2 uses cosine annealing with 5% linear warmup over 10 epochs. Each phase draws up to 8,000 training samples. Fisher and importance matrices are computed on all available Task-A samples ($\sim$9,500), which is common to all compared methods but is not a strict online protocol.

Seeds, and the separation of selection from evaluation. A grid search was run on seed 42, selecting per method the configuration maximising Task-A retention: EWC $\lambda \in \{5{,}000, 10{,}000, \mathbf{50{,}000}\}$; MAS $\lambda \in \{\mathbf{5}, 10, 20\}$; ER buffer $\in \{200, 500, \mathbf{1{,}000}\}$; DER++ buffer $\in \{200, 500, \mathbf{1{,}000}\}$ with $\alpha = 0.3, \beta = 0.5$. We then *froze* those choices and ran the complete protocol on four further seeds (43–46), which are therefore held-out evaluations of a selection made elsewhere. Unless stated otherwise, reported values are means $\pm$ sample standard deviations over all five seeds, and comparisons between methods are paired across seeds. Section 5.3 uses the held-out seeds to quantify what the single-seed selection cost us.

LwF (Li and Hoiem, 2017) was evaluated in the sweep but excluded from the main comparison: its retention-selected configuration reaches 100.2% retention at only 21.6% Task-B efficiency and 43.3% global accuracy, i.e. it preserves Task A by declining to learn Task B. We report it as evidence that retention alone is an inadequate selection criterion.

5 Results

5.1 Main Comparison

The distilled student reaches 90.0% accuracy on the 16-class test set against 93.7% for the DiT-Large teacher, a 3.7 pp drop at 55.5$\times$ fewer parameters. After Phase 1, θ_{naive} reaches $96.3 \pm 0.5\%$ on Task-A classes but falls to $30.4 \pm 0.2\%$ globally, confirming catastrophic forgetting of Task B. Unprotected Phase-2 fine-tuning retains only $23.5 \pm 5.5\%$ of Task-A accuracy while recovering Task B efficiently ($97.2 \pm 0.8\%$).

All four CL methods improve substantially on that lower bound, and the ordering is identical at every seed (Table 1, Figure 2). DER++ retains $90.4 \pm 1.2\%$, ahead of ER (84.1 ± 1.8), MAS (75.9 ± 2.1) and EWC (63.3 ± 3.8). DER++ also

achieves the highest global accuracy ($85.6 \pm 0.4\%$), while ER is the strongest Task-B learner ($96.5 \pm 0.5\%$). Notably, the standard deviation shrinks monotonically with retention: EWC varies three times as much across seeds as DER++, so the weakest method is also the least predictable.

Table 1: Main results, mean ± sd over five seeds. A-Ret. = Task-A retention; B-Eff. = Task-B recovery efficiency; Global = 16-class accuracy. Best CL result in bold.

Method	A-Ret.↑	B-Eff.↑	Global↑
θ_{orig} (reference)	—	—	90.0
θ_{naive} (Phase 1)	—	—	30.4 ± 0.2
No-CL Phase-2	23.5 ± 5.5	**97.2** ± 0.8	67.1 ± 1.8
EWC (λ=50k)	63.3 ± 3.8	89.5 ± 1.2	74.3 ± 1.4
MAS (λ=5.0)	75.9 ± 2.1	94.3 ± 0.7	81.0 ± 0.9
ER (buffer=1k)	84.1 ± 1.8	96.5 ± 0.5	84.9 ± 0.7
DER++ (buffer=1k)	**90.4** ± 1.2	94.6 ± 0.4	**85.6** ± 0.4

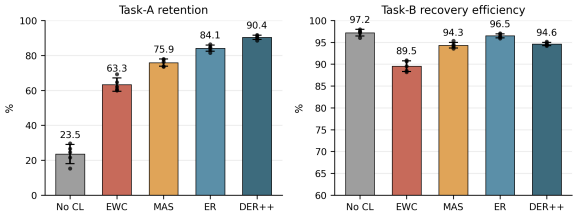


Figure 2: Task-A retention and Task-B recovery efficiency, mean ± sd over five seeds. Individual seed values are overlaid as points. The replay advantage on retention holds at every seed with no crossovers.

5.2 Statistical Significance

With five paired seeds we test each pairwise difference using a paired t-test (Table 2). Every retention comparison is significant at $p < 10^{-3}$, including the DER++-over-ER margin of 6.3 ± 1.1 pp that a single run could not have separated from noise. On global accuracy the ordering is likewise significant, though the DER++-over-ER margin is much smaller (0.7 ± 0.4 pp, $p = 0.018$): DER++ buys most of its retention advantage back in Task-B efficiency, so the two replay methods are far closer overall than the retention column alone suggests.

We report t-tests rather than the Wilcoxon signed-rank test because with $n = 5$ paired observations the smallest attainable two-sided Wilcoxon p-value is $2/2^5 = 0.0625$; the test cannot reach $\alpha = 0.05$ regardless of effect size, and would have declared every comparison here non-significant. We note this because it is an easy trap at small seed counts.

Table 2: Paired differences in Task-A retention across five seeds (percentage points), with paired t-test statistics.

Comparison	Δ Ret.	t	p
DER++ − ER	$+6.3 \pm 1.1$	12.9	2×10^{-4}
DER++ − MAS	$+14.5 \pm 1.2$	26.3	1×10^{-5}
DER++ − EWC	$+27.1 \pm 3.3$	18.7	5×10^{-5}
ER − MAS	$+8.3 \pm 1.0$	18.6	5×10^{-5}
ER − EWC	$+20.9 \pm 3.0$	15.8	9×10^{-5}
MAS − EWC	$+12.6 \pm 2.6$	10.9	4×10^{-4}

5.3 The Cost of Selecting on One Seed

Because the configurations were chosen to maximise retention on seed 42, that seed is expected to be optimistic. Holding the selection fixed and comparing seed 42 against the mean of the four held-out seeds quantifies the effect, and it is not uniform. EWC is inflated by $+7.4$ pp (69.2 versus 61.8), MAS by $+2.6$, ER by $+2.8$, and DER++ by only $+1.1$. The method with the widest search grid and the weakest performance is flattered most.

This has a practical consequence beyond our own numbers. Reporting a CL comparison at each method's retention-selected operating point, on the same run used for selection, systematically compresses the gap between strong and weak methods: on seed 42 the DER++-over-EWC margin reads 22.1 pp, whereas on the held-out seeds it is 28.4 pp. Selection-conditioned single-run comparisons should be read as lower bounds on the spread between methods, not as unbiased estimates.

5.4 Where Forgetting Concentrates

Macro retention conceals substantial heterogeneity (Table 3). Email is essentially untouched under every method, and handwritten and specification stay high, while invoice and presentation absorb most of the damage. Under EWC, invoice retains $21.8 \pm 6.0\%$ and presentation $25.1 \pm 9.9\%$ while email retains $99.6 \pm 0.6\%$ — a single macro number of 63.3% describes neither. The methods differ mainly in how far they rescue these two classes: DER++ lifts invoice to $87.3 \pm 1.6\%$ and presentation to $72.7 \pm 4.3\%$.

Per-class variance is also concentrated in the same place. EWC's specification retention varies

by ± 12.6 pp across seeds, an order of magnitude more than its email retention. A practitioner reading only the macro mean would neither know which document types are at risk nor how reproducible that risk is. The retained artifacts contain per-class accuracies but not per-image predictions, so we cannot determine whether visual similarity between invoices, presentations and the Task-B classes drives the pattern; that would require a confusion-level analysis we did not archive.

Table 3: Task-A per-class retention (%), mean $\pm$ sd over five seeds. Two classes absorb most of the loss under every method.

Method	Email	Hand.	Spec.	Invoice	Pres.
EWC	99.6 ± 0.6	91.6 ± 2.8	78.4 ± 12.6	21.8 ± 6.0	25.1 ± 9.9
MAS	99.4 ± 0.7	82.9 ± 3.4	88.4 ± 3.5	63.3 ± 4.1	45.3 ± 4.7
ER	99.7 ± 1.0	91.6 ± 2.7	88.7 ± 4.0	77.3 ± 3.3	63.4 ± 3.4
DER++	100.7 ± 0.7	95.8 ± 2.2	95.5 ± 2.5	87.3 ± 1.6	72.7 ± 4.3

5.5 Buffer Size and Label Budget

Both sweeps below were run on seed 42 only and are reported as trends, without the uncertainty attached to Table 1.

Retention improves monotonically with buffer size, DER++ ahead of ER at every size (Table 4). DER++ with only 200 buffered samples (40 per class, 9.6 MiB) already reaches 81.9% retention, exceeding the best regularisation configuration at under a quarter of the 42.2 MiB those regularisers spend on importance matrices. At small buffers replay is thus both stronger and cheaper; at 1,000 samples (47.9 MiB) it remains stronger but no longer cheaper, trading data storage for parameter statistics. The efficiency column shows the cost: DER++'s Task-B recovery declines from 96.7% to 94.9% as the buffer grows, so a larger buffer buys stability by conceding plasticity.

Table 4: Replay buffer sweep (seed 42). Retention rises with buffer size while DER++'s Task-B efficiency falls.

	ER		DER++	
Buffer	Ret.	Eff.	Ret.	Eff.
200	72.6	97.0	81.9	96.7
500	80.2	97.3	88.4	96.1
1,000	86.4	96.9	91.3	94.9

Label scarcity tells a mirror-image story (Figure 3). Re-running Phase 2 across labelled Task-B budgets from 200 to 8,000 samples, replay dominates retention at every budget, and at 200 labels all four methods retain at least 97.4% of Task-A accuracy. That apparent success is misleading: Task-B recovery at that budget spans only 49.0–72.8%, so high retention under extreme scarcity indicates limited adaptation rather than a good balance. ER is the most label-efficient new-task learner throughout, whereas DER++ is the most conservative, its logit matching over-regularising when data is scarce. Increasing the budget monotonically raises recovery and global accuracy while lowering retention, for every method.

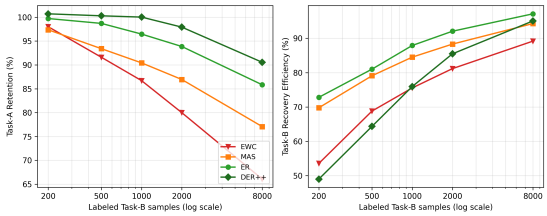


Figure 3: Task-A retention (left) and Task-B efficiency (right) versus labelled Task-B budget (200–8,000, log scale), seed 42. Near-total retention at the smallest budget reflects limited adaptation, not a favourable trade-off.

5.6 The Non-Distilled Control Cannot Isolate Distillation

The submitted version of this work compared the distilled student against a same-size non-distilled ViT-Tiny (ImageNet-initialised `WinKawaks/vit-tiny-patch16-224`) and read the resulting gaps as a distillation effect. That reading does not hold, and we correct it here. The non-distilled model begins Phase 1 at 6.6% accuracy on the 16 classes — essentially chance — against 90.0% for the distilled student. An 83.4 pp gap in starting accuracy confounds optimisation, representation quality and calibration simultaneously, so the comparison cannot attribute anything to distillation specifically.

What the control does show is narrower but still useful: the family ordering reproduces under a drastically weaker initialisation, with DER++ and ER retaining 82.5% and 72.7% against 51.5% and 32.7% for EWC and MAS (Table 5). The replay advantage is therefore not an artifact of the distilled starting point. Establishing a distillation effect would require a non-distilled ViT-Tiny trained to comparable 16-class accuracy before Phase 1; we did not run that control, and we make no distillation claim without it.

Table 5: Distilled versus non-distilled ViT-Tiny of identical size, single seed. The 6.6% versus 90.0% starting-accuracy gap makes this a diagnostic, not an accuracy-matched control.

	Task-A Ret. (%)		Global Acc. (%)	
Method	Distil.	Raw	Distil.	Raw
Start	—	—	90.0	6.6
No-CL	26.5	8.4	68.1	55.4
EWC	69.2	51.5	75.7	49.2
MAS	78.0	32.7	81.6	56.0
ER	86.4	72.7	85.5	73.6
DER++	91.3	82.5	85.8	72.6

5.7 Deployment Cost

Table 6 reports Phase-2 cost on a single NVIDIA T4, averaged over the four held-out seeds, with GPU energy read from the on-board NVML counter. Two points matter for practitioners.

First, once method *preparation* is counted, the time advantage of regularisation largely disappears. Estimating a Fisher or importance matrix over $\sim$9,500 Task-A samples costs 7.3–7.5 minutes, whereas constructing a replay buffer costs 1.4–1.5. Comparing training loops alone puts EWC 9.1 minutes ahead of ER; comparing total adaptation time narrows this to 3.3 minutes. Prior reports that time only the optimisation loop overstate how much cheaper regularisation is.

Second, the absolute costs are small. Every method completes an adaptation round in under 37 minutes, below 600 MiB of GPU memory, and under 26 Wh — roughly the energy of running a laptop for two hours. Distillation is performed once; each subsequent adaptation is cheap enough for institutions where annotation arrives incrementally and compute budgets are fixed.

For serving rather than adaptation, the student runs at 6.3 ms per image (batch 1) on the T4. Since the motivating deployments may lack a GPU entirely, we also measured a controlled single-thread CPU benchmark on an Intel Core i7-13700H: median model-forward latency is 61.6 ms (mean 64.9; 10th–90th percentile 45.5–89.6) over 200 timed passes after 30 warm-ups. This is one machine and one timing session, and it excludes data loading and application logic, but it establishes that CPU-only serving of the 21.2 MiB student is feasible at roughly 16 pages per second per core.

Table 6: Phase-2 deployment cost on a single NVIDIA T4, mean over four held-out seeds. Prep = method preparation (Fisher/importance estimation or buffer construction); Train = optimisation loop; S_{aux} = persistent auxiliary storage; Energy from NVML.

Method	Prep (min)	Train (min)	GPU (MiB)	S_{aux} (MiB)	Energy (Wh)
Naive FT (ref.)	0.0	18.7	358	0	13.6
EWC (λ=50k)	7.3	26.1	421	42.2	18.0
MAS (λ=5.0)	7.5	26.9	421	42.2	18.5
ER (buffer=1k)	1.4	35.2	591	47.9	25.8
DER++ (buffer=1k)	1.5	33.3	586	47.9	25.6

6 Discussion

The consistent advantage of replay over regularisation — 27 pp between DER++ and EWC, significant at every pairwise step — is consistent with how each family uses limited capacity. Regularisation penalises deviation from θ_{naive} but supplies no direct signal for the target classes during Task-B learning; the model must satisfy the Task-B loss and the penalty simultaneously. Replay instead includes Task-A samples in every batch, and DER++'s logit matching additionally preserves the output distribution. We emphasise that this is an interpretation of the observed behaviour, not an isolated mechanism: with capacity held fixed at ViT-Tiny throughout, we cannot show that compactness is what makes the indirect mechanism fail. A Tiny/Small/Base sweep would be required, and is the single most informative extension of this work.

The 12.6 pp advantage of MAS over EWC, despite their shared quadratic penalty, is notable and now significant. EWC's Fisher measures loss curvature; MAS measures sensitivity of the output norm. For a specialist confident on only 5 classes, output-based importance appears to identify the parameters essential to those predictions better than loss-curvature signals do.

For practitioners deploying distilled document classifiers that must adapt over time, our results support four guidelines. First, prefer replay: DER++ retains 90% of prior knowledge at an auxiliary storage cost ($\sim$48 MiB) comparable to what regularisation already requires, and the ordering is stable across seeds. Second, choose between the replay variants by what you need: DER++ for stability, ER if Task-B recovery matters more, since their global-accuracy difference is under a point. Third, if a replay buffer is infeasible — and page

retention may be prohibited by governance rather than by storage — MAS is clearly preferable to EWC. Fourth, do not read a macro retention number as uniform protection: budget for the two or three document types that will absorb most of the loss, and validate on those specifically.

7 Limitations

Several boundaries qualify these findings. The evaluation uses a single benchmark subset, a single student architecture, and a single fixed task partition and order; CL results are known to be sensitive to task composition, and multiple partitions would be needed to show the ordering is not specific to this split. The subset's sampling procedure is undocumented, so absolute accuracies should not be compared with published full-split RVL-CDIP results. Although the headline comparison now rests on five seeds with paired tests, the buffer, label-budget and non-distilled analyses remain single-seed and are reported as trends. Hyperparameters were selected on seed 42 and frozen rather than re-tuned per seed, which is what makes seeds 43–46 valid held-out evaluations but also means each method is reported at a configuration tuned on one draw; Section 5.3 quantifies the resulting optimism rather than removing it. The capacity argument that motivates the study is untested, and the non-distilled control cannot support a distillation claim. Finally, the CPU measurement is one machine and one session, and no genuine low-resource-language document collection was evaluated: our claim of transferability rests on document classification being largely layout-driven, which is an argument, not evidence.

8 Conclusion

We studied the interaction between knowledge distillation and continual learning for document classification. Distilling a DiT-Large teacher (307 M) into a ViT-Tiny student (5.5 M, 21.2 MiB, 90.0% accuracy) and evaluating four CL methods across five seeds, we found that replay consistently beats regularisation — DER++ $90.4 \pm 1.2\%$ retention against $75.9 \pm 2.1\%$ for the best regulariser — with every pairwise ordering significant and stable across seeds. Measuring the cost of single-seed selection showed it inflates retention by up to 7.4 pp and flatters the weakest method most. Per-class analysis showed that macro retention hides concentrated damage in two of five target classes. Counting method preparation reduced regularisation's apparent speed advantage from 9.1 to 3.3 minutes per round, with every method finishing under 37 minutes and 26 Wh on one commodity GPU.

Future work should run the capacity sweep the compactness argument requires, add an accuracy-matched non-distilled control, evaluate multiple class partitions and orders, and validate on a genuine low-resource-language document collection. Code, the distilled model, the benchmarking harness, and all seed-level result artifacts are available at `https://github.com/Mehdi-Soufiane/distil-and-evolve`.

References

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. 2018. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 139–154. Springer.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. 2020. Dark experience for general continual learning: a strong, simple baseline. In *Advances in Neural Information Processing Systems*, volume 33, pages 15920–15930.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Tal Hassner, and Louis-Philippe Morency. 2019. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*.

Cyprien de Masson d'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic memory in lifelong language learning. In *Advances in Neural Information Processing Systems*, volume 32.

Robert M. French. 1999. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819.

Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis. 2015. Evaluation of deep convolutional nets for document image classification and retrieval. In *Proceedings of the 2015 International Conference on Document Analysis and Recognition (ICDAR)*, pages 991–995. IEEE.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. LayoutLMv3: Pre-training for document AI with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091. ACM.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. 2022. DiT: Self-supervised pre-training for document image transformer. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3530–3539. ACM.

Zhizhong Li and Derek Hoiem. 2017. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947.

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. 2017. iCaRL: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2001–2010. IEEE.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. 2023. CODA-Prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11909–11919.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. 2024. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5362–5383.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. 2022. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 139–149.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. 2021. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 2579–2591. ACL.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. LayoutLM: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200. ACM.

Friedemann Zenke, Ben Poole, and Surya Ganguli. 2017. Continual learning through synaptic intelligence. *Proceedings of the 34th International Conference on Machine Learning*, 70:3987–3995.

Do Large Language Models Style-Shift? Register-Conditioned Stylistic Fronting in AI-Generated Icelandic

Anton Karl Ingason
University of Iceland
antoni@hi.is

Johanna Mechler
University of Iceland
mechler@hi.is

Lilja Björk Stefánsdóttir
University of Iceland
lbs@hi.is

Abstract

We examine whether large language models (LLM) encode sociolinguistic knowledge of formal and informal register in a low-resource language, Icelandic, finding that register sensitivity is present in the two systems tested (GPT-5.5 and Gemini 3.1 Pro). Our findings are based on a 2×2 persona design with elicited production (n=1,200), manually coded for Stylistic Fronting, a property of formal language. The direction and size of their style shift are not detectably different, but they differ sharply in overall rate; against a matched human comparator, only GPT-5.5 produces a persona-appropriate rate. Neither system reproduces the gender pattern documented for human speakers. Thus, register sensitivity does not guarantee that output rates are appropriate to the persona the model was instructed to adopt, an issue that may reflect limited training data for low-resource languages.

1 Introduction

In recent research, Large Language Models (LLMs) are increasingly used as proxies for human speakers (e.g., Argyle et al., 2023), and their sociolinguistic properties are under ongoing investigation (e.g., Hofmann et al., 2024). A key open question in this respect is whether and how such systems encode sociolinguistic knowledge of languages that are underrepresented in their training data. Icelandic is an interesting test case for exploring this issue, since it is low-resource in terms of Language Technology and has a well-developed sociolinguistic literature on syntactic variation. This combination allows us to examine LLM-generated Icelandic in the context of what we know about human language variation.

We examine Stylistic Fronting (SF) (Ingason and Mechler, 2025; Mechler et al., 2025), a word order phenomenon that is well documented and sensitive to register. By controlling the syntactic context, we get a binary contrast in SF that is easy for an expert human to annotate, thus being ideal for controlled elicitation.

While LLM performance has been studied on certain tasks in Icelandic (Ármannsson et al., 2025b,a; Ingimundarson et al., 2026), no previous study has examined whether and how LLMs encode sociolinguistic knowledge about Icelandic that allows them to distinguish between formal and informal registers. We test two independently developed systems (GPT-5.5 and Gemini 3.1 Pro) and show that they produce significantly more SF in the formal conditions than the informal ones, thus register sensitivity is present, and not vendor-specific. The direction and size of the shift are not detectably different between the systems, but their overall usage rate differs sharply. Against a matched human comparator averaging 90.4% SF, only GPT-5.5 falls within that speaker's attested range (81.8%), while Gemini 3.1 Pro produces 34.0%, so register sensitivity does not guarantee a persona-appropriate rate. We find that neither system reproduces the gender pattern documented for human speakers (see Section 2.1).

The paper is structured as follows. Section 2 outlines the background, then Section 3 discusses our methods. We present our results in Section 4. Section 5 provides discussion and Section 6 concludes.

2 Background

2.1 Sociolinguistic variable: Stylistic Fronting

A sociolinguistic variable is defined as two or more ways of expressing the same meaning, where choice between variants is conditioned by social and linguistic factors, not by truth conditions (Labov, 1972). Here, we focus on SF as a syntactic sociolinguistic variable. SF is a word-order phenomenon in Icelandic clauses with subject gaps.

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 140–149
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_016

This includes some relative clauses where a non-finite main verb may optionally precede a finite auxiliary (Maling, 1980; Holmberg, 2006; Thráinsson, 2007; Angantýsson, 2011). The following examples show the two word orders in such a clause. Example (1) shows the auxiliary-first order (no SF); Example (2) shows the non-finite-first order (SF). Using either variant does not affect the truth condition of the sentence.

(1) *Varðandi* *það* *[$_{CP}$* *sem* ***var*** ***sagt*** *hér]*
regarding it that **was** **said** here
'Regarding what was said here' [no SF]

(2) *Varðandi* *það* *[$_{CP}$* *sem* ***sagt*** ***var*** *hér]*
regarding it that **said** **was** here
'Regarding what was said here' [SF]

Note here that the type of auxiliary may modulate SF use, where construction with *vera* 'be' passives show the highest rates (Ingason and Mechler, 2025). To isolate the register effect, the present study holds auxiliary type constant at *vera*.

SF indexes formality (Wood, 2011) and is genre-sensitive, e.g., higher rates of SF were documented for a newspaper corpus compared to parliament speeches (Ingason and Mechler, 2025). SF rates may also vary for individual speakers: Considering the same large corpus of parliamentary speeches, SF rates have been found to range from approximately 20–80% depending on the individual (Stefánsdóttir and Ingason, 2024; Stefánsdóttir et al., 2025a,b, 2026) (about 31–96% for the subset of *vera* constructions examined in this study). The variable use of SF is conditioned by social factors, e.g., by gender. Female politicians produce significantly more SF than male ones (Mechler et al., 2025), consistent with research on formal language use and professional identity (Labov, 2001, *inter alia*). The professional background of the speaker also matters, as a female parliament member who was previously an Icelandic teacher produces notably high rates of SF (about 62–82%; about 75–96% in *vera* subset), aiming to construct a formal and professional political identity (Stefánsdóttir et al., 2025b). We use this case study as a direct human comparator (Condition 1) and test whether LLMs reproduce sociolinguistic patterns, in particular for gender and professional background.

2.2 LLMs as sociolinguistic informants

In this study, we apply LLMs as synthetic informants, where LLMs function as stand-ins for human participants (e.g., Argyle et al., 2023). Previous work reveals that LLMs can encode complex sociocultural contexts and attitudes, yet not without limits and mistakes if social factors are not explicitly incorporated (e.g., Hovy and Yang, 2021). Evidence on the sociolinguistic competence of LLMs suggests that they do encode aspects of social variation and consequently racist and sexist biases (Holliday, 2026, 2025; Holliday and Reed, 2025; Hofmann et al., 2024). Yet, "the use of LLMs for modelling and analysing variation in speech appears to be a largely unexplored area" (Nguyen, 2025, p. 4) and more work is needed to understand the complex interplay of multiple social factors in constructing speaker personas linguistically. Here, we address this gap and explore how two LLMs employ sociolinguistic knowledge to construct different personas.

We also need more studies on languages like Icelandic, which is considered low-resource, despite a recent Language Technology Program which developed key resources (e.g., Nikulásdóttir et al., 2020). Prior work has examined LLM performance on some Icelandic tasks (Ármannsson et al., 2025b,a; Ingimundarson et al., 2026). We add a sociolinguistic dimension to this line of work.

3 Method

3.1 Experimental design

The study is designed to elicit differences in the generation of formal and informal language depending on whether an LLM is instructed to write on behalf of a specific gender, female/male, and depending on a contrast between language-sensitive and non-language-sensitive occupation. For each condition, the LLM produces sentences that may include SF. Previous work demonstrates that the importance of language in a person's life affects their use of formal/standard variants and women also generally use more formal/standard variants, including the use of SF (Sankoff and Laberge, 1978; Mechler et al., 2025). We set up a 2×2 design crossing gender and occupation type, yielding four persona conditions. We acknowledge that treating gender as binary is a simplification of a nuanced variable (see Limitations) and we aim to explore more aspects of gender in future work.

In Condition 1, (formal/language-sensitive) the LLM is instructed to be a female parliament member and former Icelandic teacher. Being a parliament member is already a language-sensitive social

context associated with formal language and this is reinforced by a background as a language teacher, as educators have a special status in the use of standard language (Grama et al., 2023). Condition 2 is the male equivalent of Condition 1. Equivalent, except the LLM is instructed to be a man, rather than a woman. Condition 3 introduces the informal female persona. Here, the LLM is instructed to be a woman who is an electrician. This is a non-language-sensitive occupation, expected to be linked to informal language. Finally, Condition 4 is the informal male persona, the male equivalent of Condition 3, a man who is an electrician. The electrician profession was chosen on the basis that it is not language-sensitive as compared to the profession used in the formal conditions and language sensitivity has been previously found to be linked to use of formal/standard variants as described in the theory of the Linguistic Market (Sankoff and Laberge, 1978; Grama et al., 2023). This choice is not based on corpus evidence on language use of electricians, we do not have such data.

In addition to stating the occupation, some extra text in the prompt is used to emphasize the prevalence of formal vs. informal language use for each persona. Age of the target speaker is held constant at mid-forties across conditions and occupation is identical within gender pairs so only gender varies between C1–C2 and C3–C4. It should be noted that while we are focusing on language-sensitivity of the occupation, occupation type does co-vary with socioeconomic status which is also a potential predictor of formal/standard language use (e.g., Sankoff and Wagner, 2011, see also Limitations).

Two LLMs were tested: GPT-5.5 (OpenAI) and Gemini 3.1 Pro (Google).[1] We study how LLMs encode sociolinguistic knowledge in a less-resourced language and how they apply that knowledge when producing syntactic structures that human speakers use systematically in relation to the social context within which they speak and write. Testing both GPT and Gemini allows us to examine whether we find the same register effect in independently developed systems. We selected these particular LLMs because they were widely used by consumers at the time when we collected the data and we submitted prompts via the user interfaces that face non-specialists rather than API access because this is how such users encounter the systems.

[1]A third system, Claude Sonnet 4.6, was also tested but is not reported here. After data collection we discovered that those data were affected by memory contamination: account-level memory had been silently activated without our knowledge, so our protocol of using a fresh chat session for each prompt did not deliver independent observations. We found no sign of a comparable issue for GPT-5.5 or Gemini 3.1 Pro, where the memory setting was confirmed to be off. A clean re-collection was not possible, as the model version we had used was no longer available in the user interface.

3.2 Prompt and data collection

The experiment is carried out by submitting prompts to the LLMs systems that are being tested. The prompt starts with a persona description including the variables of interest, gender, and language-sensitivity of the individual's occupation. This is followed by a constant task instruction, designed to elicit sentences with potential for producing SF.

The instruction asks the model to generate 15 Icelandic sentences, each of which ends in a relative clause with a subject gap. Every relative clause should start with the relative complementizer *sem* 'that', a single finite form of the auxiliary *vera* 'be', and a non-finite main verb in the form of a passive past participle. The finite auxiliary and the non-finite main verb can appear in either of the two possible word orders and it is explicitly noted in the prompt that both orders are grammatical. While the term "Stylistic Fronting" is not used, the order in which the non-finite main verb precedes the finite auxiliary is exactly the type of SF that we are interested in.

It is important to note that it was necessary to instruct the models that both variants are grammatical because pilot testing revealed that without this instruction, the LLMs will generally default to never using SF, opting for the less marked word order. The task instruction was accompanied by two example sentences illustrating the target construction, one with each word order, to demonstrate the required syntactic form and to further emphasize that both orders are grammatical.

The following is an example of a full prompt for Condition 1:

> You are a woman in her mid-forties who serves as a member of the Icelandic parliament (Alþingi). Before entering politics, you spent many years as a secondary school teacher of Icelandic language and literature. You are careful and precise in your use of language in both formal and informal settings, and colleagues often remark on the clarity and polish of your expression.
>
> Generate 15 sentences in Icelandic as this person would naturally speak or write. Each sentence must end with a relative clause with the following properties:
>
> - The relative clause has a subject gap: the head noun it modifies is understood as the subject of the pas-

sive verb inside the relative clause, and no overt noun phrase or pronoun appears in subject position within the relative clause

- The relative clause begins with the complementizer sem
- The relative clause contains exactly one auxiliary verb — a finite form of vera — and a passive past participle as the main verb. The subject of the passive is always a nominative noun phrase. Do not use quirky passive constructions where the subject is dative or accusative (e.g., avoid sem henni var sýnt). Do not use perfect passive constructions combining hafa and vera (e.g., avoid sem hefur verið byggð; use sem var byggð instead)

Note that in Icelandic, the finite form of vera and the passive past participle may appear in either order within the relative clause. Both orders are grammatically correct, as shown in the examples below.

Use common, contemporary Icelandic vocabulary.

Make sure the sentences cover a range of different topics. Present your output as a list in plain text with one sentence per line and no numbering.

Examples of correct sentences:

- Við ræddum um skýrsluna sem send var til nefndarinnar.
- Hann las bókina sem var gefin út af þessum höfundi.

Prompts were nearly identical between systems. The only minor difference involved formatting the output as either a numbered list or a list in plain text with no numbering. The only motivation for this detail was to make it easy to copy and paste the output of each system to a spreadsheet. In all cases, persona descriptions and other instructions were given in English, apart from the two Icelandic examples. We needed to choose one main prompt language and we selected English because we suspected the systems would perform better when the prompt is mostly in their primary language, but we have not tested whether this is the case for generating Icelandic SF under our experimental conditions. We leave it for future work to compare outputs in this task depending on prompt language.

All prompts excluded constructions with non-nominative subjects since these complicate the generation of subject gaps as well as perfect passives to keep the word order variants simple. For data collection, we generated 10 independent batches of 15 sentences per condition per system, yielding 150 target sentences per condition (n=1,200 total). Each batch was run in a fresh session with no prior conversational context to collect independent observations; the memory setting in the user interface of both systems was confirmed to be off.

It is worth noting that when generating sentences using the Gemini 3.1 Pro LLM, the system expressed its reasoning steps (in the surrounding user interface, not as part of its main output) and they sometimes included explicit sociolinguistic framing such as "defining persona voice". When asked to explain such steps, the system made references to choice of appropriate vocabulary, but there was no explicit mention of phenomena such as word order choice.

3.3 Annotation

All 1,200 generated sentences were annotated manually by the first author, a native Icelandic speaker with expert knowledge of Stylistic Fronting. We tested automatic coding of SF but this approach was abandoned after pilot testing revealed that the output of the automatic methods was very unreliable. In the context of the tightly controlled syntactic structure used for our experiment, determining whether SF was applied or not is a quick and easy task for an expert human annotator.

The manual annotation includes two codes per sentence. For the Valid column, 1 means the relative clause is grammatical and contains a target structure that can be coded for SF; 0 means that these conditions are not met and the sentence will be excluded. For the SF column, sentences that were already coded as Valid are coded as 1 if the non-finite main verb precedes the finite auxiliary and 0 if the finite auxiliary comes first. Exclusions were limited to GPT-5.5, which sometimes generated adjectival passives that often involved further complications.

4 Results

4.1 Register effect

The primary issue of interest here is whether and how the LLM systems produce SF variation that reflects the type of sociolinguistic knowledge that human speakers of Icelandic apply in their use of language. We find that both models, GPT-5.5 and Gemini 3.1 Pro, produce significantly more SF in formal conditions than informal conditions. This can be seen in Table 1 and Figure 1 where SF rates are reported across systems and conditions. This is interesting since it means that at least in this tightly controlled LLM prompting setting, the persona descriptions that vary between language-sensitive and non-language-sensitive conditions consistently lead to a formality contrast in the direction that we

would expect for human speakers. Specifically, for GPT, SF is produced at a rate of 81.9% in the formal condition (across both genders) as opposed to 71.0% in the informal one ($\chi^2(1) = 9.5$, $p = .002$), whereas Gemini produced SF in 34.7% of the sentences in the formal condition vs. 23.7% in the informal condition ($\chi^2(1) = 8.8$, $p = .003$).

System	Condition	Valid	SF rate
GPT-5.5	C1 F formal	148	81.8%
	C2 M formal	150	82.0%
	C3 F informal	138	65.2%
	C4 M informal	138	76.8%
Gemini 3.1 Pro	C1 F formal	150	34.0%
	C2 M formal	150	35.3%
	C3 F informal	150	16.7%
	C4 M informal	150	30.7%

Table 1: SF rates by system and condition (Valid = sentences meeting target-structure criteria; C1/C2 = language-sensitive persona; C3/C4 = non-language-sensitive persona; F/M = female/male).

Both systems style-shift, producing more SF in the formal conditions, and the degree to which they do so is statistically indistinguishable. They differ in terms of their overall SF levels, with GPT at 76.7% SF as opposed to Gemini at 29.2%. The shift is 10.9 pp in the case of GPT and 11.0 pp for Gemini. As described below, a System × Formality interaction was not needed in our regression analysis. These similar style-shift levels mean that the visual display of Gemini's reasoning steps described in Section 3.2 do not coincide with a larger register effect.

Table 2 presents the final results of our mixed-effects logistic regression. A fully saturated model including a System × Formality × Gender interaction was fitted first, but this interaction did not significantly improve the fit of the model and was dropped ($\chi^2(1) = 0.04$, $p = .833$). Neither interaction involving System was required. Likelihood-ratio tests showed that System × Formality could be removed without a significant loss of fit ($\chi^2(1) = 0.004$, $p = .948$), as could System × Gender ($\chi^2(1) = 0.088$, $p = .767$), and the two could also be removed jointly ($\chi^2(2) = 0.092$, $p = .955$). The reported model includes Formality × Gender, an interaction that is marginal ($\chi^2(1) = 2.99$, $p = .084$). Including this interaction is slightly favored by AIC (1311.5 vs 1312.5). Detecting a potential effect of this type was part of the design of the study, which motivates including it in the reported model as a trend, despite this not being a significant effect.

Predictor	OR	SE	z	p
(Intercept)	0.20	0.218	-7.49	**<.001**
System [GPT-5.5]	9.79	0.194	11.77	**<.001**
Formality [formal]	2.53	0.268	3.47	**<.001**
Gender [male]	2.05	0.269	2.67	**<.01**
Formality × Gender	0.52	0.377	-1.75	0.080
Random effects: (1|batch_id)				
τ_{00} batch	0.31			
Observations	1,174			
Marginal R^2	0.289			
Conditional R^2	0.351			

Table 2: Mixed-effects logistic regression: SF as response variable. Reference levels: Gemini 3.1 Pro (system), informal (formality), female (gender). The model includes a Formality × Gender interaction, so Formality [formal] is the register effect for female personas and Gender [male] is the gender effect in the informal condition.

The model shows a large effect of System; holding register and persona gender constant, the odds of GPT-5.5 producing SF are nearly ten times Gemini's (OR 9.79). Formality, as a simple effect for female personas, is also a strong predictor (OR 2.53). The random effect of batch is modest (τ_{00} = 0.31). Fixed effects account for 28.9% of variance; including the batch random effect raises this to 35.1%, reflecting that batch-level variability is a limited source of variance.

The effect of gender is not constant across registers. The significant Gender [male] predictor reflects different SF production by gender in the informal condition. The CI of the Formality × Gender term ranges from a substantial reduction of the gender effect in the formal register down to essentially nothing (OR = 0.52, 95% CI [0.25, 1.08], $\chi^2(1) = 2.99$, $p = .084$). While the interaction is not significant, we retain it in the model and interpret it as a trend. The style shift is asymmetric by persona gender. While the female persona shows a clear register contrast (OR = 2.53, $z = 3.47$, $p < .001$), the male persona starts from a higher baseline in the informal condition and does not show significant style shift, despite the raw proportions moving slightly in the expected direction (OR = 1.31, $z = 1.01$, $p = .311$). This difference is the marginal interaction in the model and it should therefore be read as suggestive rather than established. The proportions reported in Table 1 break the gender contrast down by condition, i.e., system and register. Applying chi-square tests to these proportions, we do not find evidence of a significant gender difference in the formal conditions, for either GPT-5.5 ($\chi^2(1)$

= 0.00, p = .957) or Gemini 3.1 Pro ($\chi^2(1)$ = 0.06, p = .808). In contrast, male personas produce more SF in the informal conditions for both GPT-5.5 ($\chi^2(1)$ = 4.51, p = .034) and Gemini 3.1 Pro ($\chi^2(1)$ = 8.14, p = .004). Recall that the three-way System × Formality × Gender interaction was not needed, so the gender-by-register pattern does not differ detectably between the two systems. These chi-square tests are reported here with the caveat that they treat sentences as independent, but sentences come in batches of 15 from a single prompt. This clustering is accounted for by the mixed-effects model via the random intercept.

The gender pattern reported here can be set against human data. As noted in Section 2.1, women are generally found to use formal/standard variants at higher rates than men, whereas the LLM outputs show the opposite pattern in the informal condition (see also Section 5).

4.2 Comparison with human data

Having presented the output of the LLMs, we now consider how the AI-generated sociolinguistic variation compares with human language behavior. First, let us focus on a specific human speaker, a Member of Parliament (MP) who is a woman and has an extensive background as an educator, exactly like the persona described in the prompt for Condition 1. We extracted 595 relative clauses with the target structure (*vera*) from her parliament speeches in the Icelandic Parliament Corpus (Steingrímsson et al., 2020) and found that her average SF use is 90.4%, with annual rates ranging between 75.0-95.7% (see also Figure 1). In this condition, the SF rates produced by the AI systems are: GPT-5.5 81.8% and Gemini 3.1 Pro 34.0% (Table 1). GPT is clearly within the range of the human comparator. In contrast, Gemini produces SF at a substantially lower rate than the MP.

To put this into a broader human context of the Icelandic Parliament in general, we extracted further data from the Icelandic Parliament Corpus (Steingrímsson et al., 2020). We selected data (n=177,255) from all MPs born in 1930 or later, who produced at least 100 tokens of relative clauses with a subject gap and a *vera*-passive of the type produced by the LLMs in our study. This gave us 249 MPs and the range of SF rates attested in this group was between 31.8%–96.0% with a median of 78.6%. While both systems style-shift by a similar amount, their overall SF rates compare differently with human MPs, as GPT is at the level of a typical MP whereas Gemini is similar to the MPs who use the least amount of SF.

Lastly, while corpora for informal Icelandic exist (e.g., Hilmisdóttir, 2024), SF rates from these have not been published yet, and they are also not straightforwardly comparable to the electrician persona in our study. Future research is warranted to provide further insight into this matter.

4.3 Output quality and target-structure compliance

Finally, we test the validity and quality of the sentences which the respective LLMs produced as part of the study design. The measure that we use is compliance with the requested target structure, and failing it is not the same as being ungrammatical. While most non-compliant sentences are indeed degraded or ungrammatical, a minority are well-formed Icelandic outside the specification. Figure 2 illustrates the proportion of sentences meeting target-structure criteria for the two tested systems. Gemini 3.1 Pro has a 100% rate across all conditions, outperforming GPT-5.5 in target-structure compliance. There were only two sentences with minor errors, but these did not affect the target construction, so they were included in the final dataset. GPT-5.5 reaches an overall compliance rate of 95.7%. The results on the quality of the sentences reveal modest differences between the two systems, in particular for the informal conditions, yet with no indication of a gender effect.

The dominant error type for GPT concerned producing adjectival (stative) passives, often with corresponding verbs requiring dative subjects (17 instances, all in informal conditions). These adjectival passives appear to be a repair strategy used by GPT. The informal persona's topic domain is more likely to elicit dative-assigning verbs, whose passives require a dative subject, a pattern forbidden by the prompt. The model resolves the conflict by switching from a verbal to an adjectival passive, which satisfies the nominative requirement, but produces unnatural output that also violates the requirement for a main verb. In rare cases, GPT did produce grammatical passives with dative-assigning verbs; these are coded as non-compliant because they do not meet the criteria of the target structure.

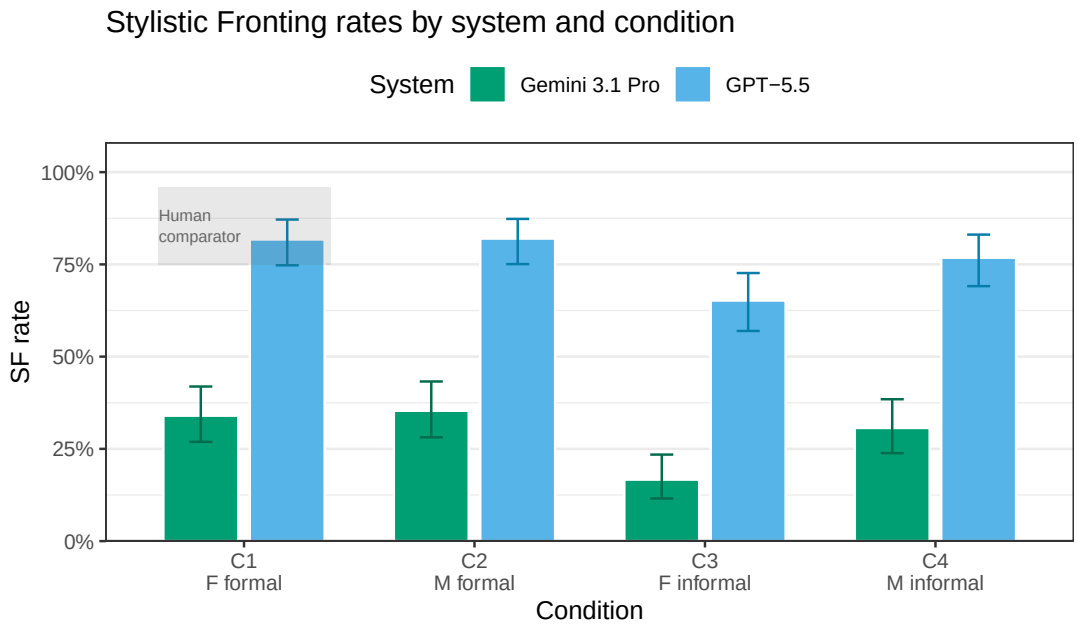


Figure 1: SF rates by system and condition with 95% Wilson confidence intervals. Shaded band: documented human rate range for a comparable speaker (Condition 1 only). C1 = female formal; C2 = male formal; C3 = female informal; C4 = male informal.

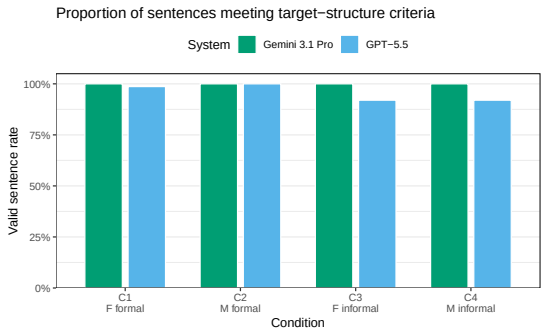


Figure 2: Proportion of sentences meeting target-structure criteria. C1 = female formal; C2 = male formal; C3 = female informal; C4 = male informal.

5 Discussion

Both LLMs that we tested do indeed style-shift: they produce more SF in the formal conditions than the informal ones. Two independently developed systems encode the register effect, which makes it unlikely to be an artifact of one company's training or alignment choices. While the size of the style shift is statistically indistinguishable between the systems, the overall level differs by an odds ratio of 9.79. The systems agree about the stylistic constraint but they disagree enormously about the overall rate at which they generate the variable. Human speakers also differ in SF rate: in our parliament speech data from 249 MPs, the range of SF use spans 31.8%–96.0%. Therefore, differing in base rate does not in itself make the systems unlike human speakers. What is notable is that they agree on the constraint while landing at opposite ends of the attested human range. The SF rate of GPT in Condition 1 (81.8%) is just above the human median of 78.6%, while Gemini's rate (34.0%) is just above the human minimum of 31.8%.

These findings may reflect that the style constraint is invariant to differences in corpus composition in a way the overall rate is not. Inferring that formal contexts favor SF needs relatively few contrasting examples, but getting the base rate right is more sensitive to the composition of the limited data available for a low-resource language. While our data show this difference in the output of the two systems under these specific experimental conditions, our study design does not identify which aspects of these LLMs produce the wide gap between them in their SF usage rates, and testing two systems cannot establish whether or how the findings would generalize to LLMs more broadly.

We must consider one key limitation for interpreting the result, namely that there is a chance of contamination in how the LLMs reach the attested

levels of SF production. The most interesting mechanism for how the LLM would acquire the relevant sociolinguistic knowledge is that it picks it up from the distribution of SF across social contexts attested in the (limited) Icelandic training data it encounters. This is indeed the scenario that we are most interested in. However, it is also possible that the systems have been exposed to linguistic publications that report on human rates in the production of SF, and this might affect the output. While other studies have published a range of SF rates, the exact envelope of variation as defined for this study has not been used in previous work. These studies have included relative clauses with other finite auxiliaries than we use here (*hafa* 'have', *munu* 'will', *skulu* 'shall'), and these are associated with lower SF rates than *vera* 'be', or they have made other decisions about which SF contexts to include (e.g., Wood, 2011; Stefánsdóttir et al., 2025a). Therefore, the rates produced by the systems tested here are not simply reproducing a directly comparable published rate. The systems could have acquired knowledge from the literature that SF is register-sensitive, but not exact rates for this specific context. Note that the *vera*-specific human rates that we report in Sections 2.1 and 4.2 were computed specifically for this paper. It should also be noted that these two mechanisms are not mutually exclusive; a given system may be affected by both the distribution of SF in attested social contexts and potential exposure to sociolinguistic literature.

Our study reveals a null gender effect within the language-sensitive condition across both models, i.e., neither model differentiates the female and male parliament member persona. In the informal condition, we find an effect in the unexpected direction: males produce more SF than females (see Sections 2.1 and 4.1). Alternatively, we can describe this distribution of SF by gender as style shift that is asymmetric by persona gender: the female persona produces more SF in the formal condition than in the informal one whereas the male persona does not shift detectably. Both descriptions follow from a single observation: the male persona's informal output stays close to its formal output, so the models do not produce a distinct informal register for that persona. Neither system reproduces the gender effect documented for human speakers; if anything, the informal condition reverses it. One explanation for this pattern may be that training data reflects biases or an absence of sufficient data on gendered register variation in Icelandic, which seems highly plausible for a low-resource language. More generally, these findings suggest that while these LLMs are able to handle one aspect of sociolinguistic competence, register-sensitivity, they cannot realize multiple aspects of sociolinguistic identity, at least not simultaneously (see also Holliday, 2025; Holliday and Reed, 2025).

Here, we have focused on only two aspects of sociolinguistic competence (register, gender), and human speakers integrate much more complex interactions between different social factors in their sociolinguistic identities. At least within the framework of our study, the two tested systems were able to replicate only one of these aspects.

6 Conclusion

Our study tests two systems, Gemini 3.1 Pro and GPT-5.5, in their sociolinguistic competence, in particular their ability to style-shift when generating Icelandic, and they both produce more SF in the formal conditions than the informal ones, a novel finding for a low-resource language. The systems agree on the stylistic constraint, as the direction and the size of the shift are not detectably different between them, but they disagree on the base rate at which they generate SF.

The style shift acquired by both systems suggests that LLMs can encode sociolinguistic knowledge of low-resource languages, but comparing the SF rate of the systems with the human rate in our corpus that matches Condition 1 (an MP who is a woman and a former Icelandic teacher) shows that only GPT-5.5 is at a similar level as the human speaker, while Gemini's output is near the bottom of the attested range for the MPs in our corpus, a failure that may reflect how acquiring persona-appropriate stylistic usage rates is sensitive to limited data availability in a low-resource language. Neither system reproduces the gender effect documented for human speakers, and they fail in the same way: the male persona's informal output stays close to its formal output.

Future research could test SF rates for other types of auxiliaries (*hafa*/modal) that were excluded here and interactions with other social factors such as age. The current study represents an important building block for setting up such future studies, highlighting how LLMs (un)successfully encode sociolinguistic knowledge.

Limitations

While our study presents certain novel and valuable findings, it is subject to some limitations.

Explicit availability: The prompt informs the models that both word orders are grammatical and therefore we are measuring register-condition variant selection under explicit availability rather than in spontaneous production. This means that absolute SF rates from the LLMs should be interpreted carefully, but even so, the relative ordering of conditions is informative.

Contamination: It is possible that the models have been trained on linguistics publications that report SF rates from studies on human speakers; this cannot be ruled out (see Discussion).

Auxiliary type: Here, we only study *vera* 'be' passives. It is unclear whether our results will generalize to constructions with other auxiliaries like *hafa* 'have' or modals.

Design scope: The study only includes two occupation types and two genders that are treated as binary while real world gender is much more nuanced. We also face a confound in that occupation type and socioeconomic status cannot be disentangled by our experimental setup.

Human comparison: When comparing with corpus data from humans, we only have formal-register speakers. We do not currently have studies that report SF rates for the informal-register conditions.

Single coding: The manual annotation was carried out by only one expert annotator and therefore we do not have any information on inter-annotator agreement. Annotation of the SF word order is a straightforward task and not a likely source of many errors.

Excluded sentences: The GPT-5.5 sentences that were excluded fall almost entirely in the informal conditions, and they show SF word order more often than the sentences that were retained there. Excluding them may understate GPT-5.5's informal SF rate by about one percentage point, so the effect on the reported rates is negligible. Gemini 3.1 Pro produced no sentences that had to be excluded.

Ethical Considerations

In this study, all of the data we use are generated by AI systems and no human subjects are involved. The systems that we analyze are publicly available and no sensitive data are used.

Acknowledgments

The project is supported by a grant from the European Research Council (ERC), project ID 101117824. We would also like to thank the reviewers for helpful comments which improved this paper.

References

Ásgrímur Angantýsson. 2011. *The Syntax of Embedded Clauses in Icelandic and Related Languages*. Ph.D. thesis, University of Iceland.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Bjarki Ármannsson, Hinrik Hafsteinsson, Jóhannes B. Sigtryggsson, Atli Jasonarson, Einar Freyr Sigurðsson, and Steinþór Steingrímsson. 2025a. Playing by the rules: A benchmark set for standardized Icelandic orthography. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 28–36.

Bjarki Ármannsson, Finnur Ágúst Ingimundarson, and Einar Freyr Sigurðsson. 2025b. An Icelandic linguistic benchmark for large language models. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 37–47.

James Grama, Johanna Mechler, Lea Bauernfeind, Mirjam E. Eiswirth, and Isabelle Buchstaller. 2023. Post-educator relaxation in the u-shaped curve: Evidence from a panel study of Tyneside (ing). *Language Variation and Change*, 35(3):325–350.

Helga Hilmisdóttir. 2024. Young Icelanders and popular culture: An empirical study of Anglicisms in conversation. *Sociolinguistica*, 38(2):215–236.

Valentin Hofmann, Pratyusha Kalluri, Dan Jurafsky, and Sharese King. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature*, 633:147–154.

Nicole Holliday. 2026. Unrealistic feedback in socially prescriptive speech technologies. *International Journal of Applied Linguistics*, pages 1–12.

Nicole R. Holliday. 2025. Socially prescriptive speech technologies: Linguistic, technical, and ethical issues. *The Journal of the Acoustical Society of America*, 158(6):4361–4369.

Nicole R. Holliday and Paul E. Reed. 2025. Gender and racial bias issues in a commercial "tone of voice" analysis system. *PLOS ONE*, 20(2):e0314470.

Anders Holmberg. 2006. Stylistic Fronting. In Martin Everaert and Henk van Riemsdijk, editors, *The Blackwell Companion to Syntax*, volume IV, pages 532–565. Blackwell, Oxford.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Anton Karl Ingason and Johanna Mechler. 2025. Stability and change in Icelandic corpora: The case of Stylistic Fronting. *Journal of Data Mining & Digital Humanities*, pages 1–9.

Finnur Ágúst Ingimundarson, Steinunn Rut Friðriksdóttir, Bjarki Ármannsson, Iris Nowenstein, and Steinþór Steingrímsson. 2026. Who benchmarks the benchmarks? A case study of LLM evaluation in Icelandic. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 4702–4715, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

William Labov. 1972. *Sociolinguistic Patterns*. University of Pennsylvania Press.

William Labov. 2001. *Principles of Linguistic Change: Volume 2: Social Factors*. Blackwell, Malden & Oxford.

Joan Maling. 1980. Inversion in embedded clauses in Modern Icelandic. *Íslenskt mál*, 2:175–193.

Johanna Mechler, Lilja Björk Stefánsdóttir, and Anton Ingason. 2025. Language use of political parties over time: Stylistic Fronting in the Icelandic Gigaword Corpus. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 313–318.

Dong Nguyen. 2025. Collaborative growth: When large language models meet sociolinguistics. *Language and Linguistics Compass*, 19(2):e70010.

Anna Nikulásdóttir, Jón Guðnason, Anton Karl Ingason, Hrafn Loftsson, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, and Steinþór Steingrímsson. 2020. Language technology programme for Icelandic 2019-2023. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3414–3422, Marseille, France. European Language Resources Association.

David Sankoff and Suzanne Laberge. 1978. The linguistic market and the statistical explanation of variability. In *Linguistic Variation: Models and Methods*, pages 239–250. Academic Press.

Gillian Sankoff and Suzanne Evans Wagner. 2011. Age grading in the Montréal French inflected future. *Language Variation and Change*, 23:275–313.

Lilja Björk Stefánsdóttir, Johanna Mechler, and Anton Karl Ingason. 2025a. Constructing a liberal identity via political speech: Tracking lifespan change in the Icelandic Gigaword Corpus. In *Proceedings of the 5th Conference on Language, Data and Knowledge*, pages 312–317, Naples, Italy. Unior Press.

Lilja Björk Stefánsdóttir, Johanna Mechler, and Anton Karl Ingason. 2025b. Examining the dynamics of individual lifespan change through sociolinguistic interviews: The case of MP Ragnheiður Ríkharðsdóttir. Paper presented at New Ways of Analyzing Variation 53 (NWAV 53), University of Michigan.

Lilja Björk Stefánsdóttir, Johanna Mechler, and Anton Karl Ingason. 2026. The stylistic aftermath of the financial crisis in the Icelandic parliament: Reinventing identity at the level of political parties. In *Proceedings of the 49th Annual Penn Linguistics Conference*, pages 312–317. University of Pennsylvania.

Lilja Björk Stefánsdóttir and Anton Karl Ingason. 2024. Wiggly lifespan change in a crisis: Contrasting reactive and proactive identity construction. *U. Penn Working Papers in Linguistics*, 30(2):119–125.

Steinþór Steingrímsson, Starkaður Barkarson, and Gunnar Thor Örnólfsson. 2020. IGC-Parl: Icelandic corpus of parliamentary proceedings. In *Proceedings of the Second ParlaCLARIN Workshop*, pages 11–17.

Höskuldur Thráinsson. 2007. *The Syntax of Icelandic*. Cambridge University Press, Cambridge.

Jim Wood. 2011. Stylistic Fronting in spoken Icelandic relatives. *Nordic Journal of Linguistics*, 34(1):29–60.

Benchmarking Large Language Models on Mandarin Proverb Explanation and Contextual Matching

Xiaojing Zhao[1], Salima Lamsiyah[2], Emmanuele Chersoni[1], Han Xu[1]
[1]The Hong Kong Polytechnic University
[2]University of Luxembourg
xiaojing.zhao@connect.polyu.hk, salima.lamsiyah@uni.lu
{emmanuele.chersoni, han12.xu}@polyu.edu.hk

Abstract

Mandarin proverbs condense historical allusions, figurative imagery, and conventionalized pragmatic functions into short expressions, making them a challenging test of culturally grounded language understanding. We evaluate four large language models (LLMs) on the Mandarin subset of the WISDOM dataset through two complementary tasks under zero-shot and few-shot prompting: situation-to-proverb selection, assessed by accuracy, and bilingual proverb explanation, assessed with automatic metrics and human judgments of semantic correctness, cultural faithfulness, clarity, and learner usefulness. The results reveal a clear gap between recognition and explanation. Models achieve high selection accuracy, with GPT-5.4 and DeepSeek-V4-Flash above 98%. Yet, human evaluation shows that fluent explanations often fail to preserve culturally conventionalized meanings, particularly for proverbs that rest on historical allusions, and are correspondingly less useful for learners. Few-shot demonstrations benefit some weaker models in open-ended explanation but do not consistently help stronger ones. These findings suggest that selection accuracy alone does not capture proverb understanding, and that Mandarin proverbs remain a demanding task for evaluating how LLMs handle figurative and culturally embedded language.

Note: The appendices referenced throughout this paper are provided in the accompanying GitHub repository, together with the dataset and supplementary materials: https://github.com/zcocozz/Mandarin-Proverb.

1 Introduction

Large language models (LLMs) have achieved remarkable performance across a wide range of multilingual NLP tasks, from translation to open-ended question answering (Xu et al., 2020; Yue et al., 2024; Lamsiyah, 2026). However, strong performance on standard benchmarks does not imply cultural competence: a growing body of work shows that LLMs often default to English knowledge and struggle with culturally grounded reasoning, particularly when the relevant knowledge is tied to a specific language community (Haagsma et al., 2020; Alwajih et al., 2025; Chang et al., 2025; Magdy et al., 2025).

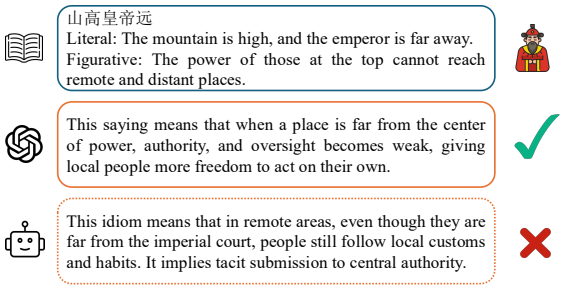


Figure 1: Example from the benchmark. For the proverb "山高皇帝远" (*the mountain is high, and the emperor is far away*), one model captures the figurative meaning of weakened central authority, while another gives a fluent but incorrect interpretation.

Proverbs are a particularly demanding test bed for cultural understanding. As condensed carriers of folk wisdom, their meanings depend on metaphorical mappings, historical narratives, social norms, and pragmatic conventions that are not recoverable from literal wording alone (Ghosh and Srivastava, 2022; Liu et al., 2024; Wang et al., 2025). Therefore, understanding a proverb requires more than lexical semantics: a model should recover the intended lesson, the situations in which the proverb is appropriately used, and the attitude it conveys. Prior work on figurative language has shown that LLMs frequently fall back on literal or culturally generic interpretations when cultural grounding is required (Magdy et al., 2025; Zhao et al., 2026a), yet most existing evaluations frame

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 150–160
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_017

the problem as multiple-choice inference over English or translated data, leaving open whether models can actively *use* and *explain* proverbs.

Mandarin Chinese offers an especially informative setting for this question. Although Mandarin is a high-resource language in general NLP, its proverbial expressions, including *yanyu* (proverb), *chengyu* (four-character idioms), and idiomatic allusions, are sparsely represented in existing benchmarks and evaluation data, which rarely go beyond cloze-style or selection-based formats (Zheng et al., 2019). These expressions often encode references to classical literature and historical events (Yue et al., 2024; Zhao et al., 2025), so understanding them requires cultural and pragmatic knowledge that general-domain text abundance does not guarantee. This combination makes Mandarin a useful test case: when a model fails on Mandarin proverbs, the failure cannot be attributed to overall data scarcity for the language, but may instead be attributed to a gap in cultural and pragmatic understanding. Existing cloze-style benchmarks are useful for evaluating contextual appropriateness but may not fully assess whether models can explain culturally conventionalized meanings. As Figure 1 illustrates, a model may produce a fluent explanation that nonetheless misreads the proverb's figurative core: given "山高皇帝远" (*the mountain is high, and the emperor is far away*), one model correctly identifies the weakening of central authority with distance, while another generates a fluent but culturally wrong answer.

To bridge this gap, we introduce a Mandarin proverb understanding benchmark derived from the Mandarin subset of WISDOM, targeting two complementary abilities. The first task, **situation-to-proverb selection**, presents a short situation that implicitly expresses a proverb's meaning and requires the model to identify the proverb that best summarizes it, testing pragmatic abstraction and appropriate usage in context. The second task, **bilingual proverb explanation**, requires the model to generate the intended pragmatic meaning of a proverb in both Mandarin and English, testing whether it can move beyond word-level translation to recover the underlying lesson and implied attitude. Together, the two tasks separate *knowing when to use a proverb* from *knowing what a proverb means*.

Our main contributions are as follows:

- We develop a two-task evaluation framework for culturally grounded Mandarin proverb understanding, covering both situation-to-proverb selection and bilingual explanation.
- We evaluate four contemporary LLMs under zero-shot and few-shot prompting settings, combining accuracy, automatic metrics, and human judgments along four dimensions: semantic correctness, cultural faithfulness, clarity, and learner usefulness.
- Our analysis reveals a gap between recognition and explanation: models that perform strongly in selecting the correct proverb do not necessarily explain it with the same level of cultural faithfulness or learner usefulness. Beyond LLM evaluation, this framework also has implications for language-learning applications, where learners need culturally grounded explanations rather than literal translations.

2 Related Work

2.1 Figurative Language and Proverbs

Figurative language has long posed difficulties for language models because its meaning often goes beyond the literal meanings of individual words. Recent studies show that LLMs still struggle with figurative language understanding and reasoning (Liu et al., 2022, 2024; Kim et al., 2025), and these challenges become more pronounced when expressions are grounded in non-Western cultural contexts (Kabra et al., 2023; Hu et al., 2026; Zhao et al., 2026b), where interpretation depends on cultural knowledge rather than lexical meaning alone.

Proverbs are a particularly demanding form of figurative language because they compress social norms, practical wisdom, historical allusions, and culturally conventionalized attitudes into short expressions. Therefore, understanding proverbs requires recognizing their figurative meanings, knowing when they are appropriately used, and identifying the stance they convey. For Chinese, existing resources have focused primarily on *chengyu*, the four-character idiomatic expressions (Fu et al., 2025; Qiu et al., 2025). ChID formulates idiom understanding as a cloze-style selection task (Zheng et al., 2019), while later datasets extend Chinese idiom and proverb evaluation to settings such as comprehension and translation (Qiang et al., 2023; Li et al., 2024; Wang et al., 2025). However, Mandarin proverbs with historical and cultural allusions

remain less explored for whether LLMs can produce culturally faithful explanations rather than only select plausible answers.

2.2 Cultural Knowledge and Evaluation of LLMs

A growing line of work examines whether LLMs possess culturally grounded knowledge and whether they tend to default to English-centric perspectives (Xu et al., 2020; Kabra et al., 2023; Alwajih et al., 2025; Lamsiyah et al., 2025). Existing cultural benchmarks probe everyday norms, commonsense reasoning, social conventions, and region-specific knowledge across languages and cultures (Myung et al., 2024; Chang et al., 2025; Chiu et al., 2025). Results consistently find that performance degrades for underrepresented languages and regions. Much of this work, however, centers on declarative cultural knowledge, typically assessed through recognition-based formats such as multiple choice.

In contrast, cultural competence also involves a pragmatic dimension: knowing how culturally meaningful expressions are used in context and how their meanings can be conveyed to others (Hu et al., 2023; Yue et al., 2024). Proverbs sit precisely at this intersection, encoding social knowledge and conventionalized communicative functions in compact expressions. This motivates an evaluation that connects recognition with explanation rather than treating proverb understanding as only a closed-form inference problem.

3 Methodology

3.1 Research Questions

The experimental design is guided by the following research questions:

- **RQ1:** Can LLMs infer the correct proverb from an implicit real-life situation?
- **RQ2:** Can LLMs explain Mandarin proverbs accurately in both Mandarin and English?
- **RQ3:** Do LLMs preserve cultural and pragmatic meaning when explaining proverbs, or do they rely mainly on literal translation?

3.2 Dataset

This study uses the Mandarin proverb subset of the WISDOM dataset, a cross-regional multilingual benchmark for proverbs and figurative language. The full WISDOM dataset contains human-annotated proverbs from a wide range of languages and regional varieties. Each entry preserves the proverb in its original form and is accompanied by explanations in the original language and in English, together with semantic labels and relevant metadata. The dataset was constructed through a multilingual annotation process involving native-speaking contributors, with entries collected from written collections and online sources. Native speakers validated the language-specific annotations to ensure that the explanations reflected both the literal and culturally situated meanings of the proverbs. However, in WISDOM's evaluation, given a proverb, models identify its meaning, semantic category, or cross-lingual equivalent. A model may succeed at all three yet still not know when a proverb is appropriately used, or be unable to explain its cultural allusion to a learner. Our evaluation extends WISDOM with a situation-to-proverb selection task and a bilingual explanation task to probe these abilities.

For the present study, we use the Mandarin subset, which contains 1,152 proverbs. Each entry includes a unique proverb ID, the Mandarin proverb, a Mandarin reference explanation, an English reference explanation, and a broad semantic category. This subset is suitable for evaluating LLMs because Mandarin proverbs often encode abstract social, moral, and pragmatic meanings through compact figurative expressions. Correct interpretation therefore requires more than literal semantic matching, as it also involves recognizing the proverb's implied stance, usage context, and culturally conventionalized meaning. Further details on dataset construction are provided in Appendix A.

3.3 Pipeline

Figure 2 illustrates the construction and evaluation pipeline used in this study. Starting from the Mandarin subset, we first normalized the entries into a structured format. From these 1,152 proverbs, three fixed examples were reserved as few-shot demonstrations and excluded from model evaluation to avoid leakage. The remaining 1,149 proverbs formed the evaluation set for both tasks.

We then constructed two tasks. The first task evaluates situation-to-proverb reasoning by pairing each proverb with a Chinese situation that expresses its pragmatic meaning without explicitly revealing the proverb. Task 1 situations were gen-

erated using *Claude-Opus-4.8* [1] and were automatically validated for format and leakage. The validation checked that each situation was non-empty, did not contain the full target proverb, and did not include explicit explanatory cue phrases that could reveal the intended answer. We also manually spot-checked a small sample for correspondence between the situation and target proverb. An example of Task 1 is provided in Appendix B. The second task evaluates bilingual proverb explanation by asking models to explain each proverb in Mandarin and English.

3.4 Tasks

The experiment consists of two complementary tasks. Task 1 evaluates whether models can recognize the appropriate proverb from an implicit usage scenario. Task 2 evaluates whether models can explicitly explain the meaning and use of a Mandarin proverb. Both tasks were conducted under zero-shot and fixed few-shot prompting conditions. All prompts were written in Simplified Chinese, and model responses were required to follow a JSON format. The full prompt templates used for Task 1 and Task 2 are provided in Appendix B.

Task 1 Situation-to-Proverb Selection In Task 1, the input is a Chinese situation and four candidate proverbs. The candidate set contains the target proverb and three distractors drawn from the same Mandarin proverb dataset. The situation is designed to express the pragmatic meaning of the target proverb without directly mentioning the proverb itself. The model is asked to select the best-matching proverb and return a JSON object in the format `"answer":"..."`. The selected answer must exactly match one of the candidate proverbs. This task is formulated as multiple-choice selection rather than open-ended generation, allowing model performance to be evaluated using accuracy.

Task 2 Bilingual Proverb Explanation In Task 2, the input is a Mandarin proverb. The model is asked to generate a bilingual explanation in the following JSON format: `"zh":"...","en":"..."`. The prompt instructs the model to explain the proverb's core meaning, usage context, and implied attitude rather than merely provide a literal translation. This task is open-ended: a valid explanation may differ from the reference wording as long as it captures the intended meaning and culturally appropriate use.

3.5 Experimental Setup

We evaluate four LLMs covering proprietary and open-weight systems. The proprietary model is *GPT-5.4* (Singh et al., 2025). The open-weight models are *DeepSeek-V4-Flash* (Xu et al., 2026), *Qwen-2.5-Instruct-7B* (Yang et al., 2024) and *Llama-3.1-Instruct-8B* (Grattafiori et al., 2024). These models were selected to compare systems across different parameter scales and multilingual capabilities: Qwen-2.5 and Llama-3.1 allow a controlled comparison between similarly sized dense models (7B–8B), DeepSeek-V4-Flash represents large-scale open-weight MoE systems, and GPT-5.4 represents frontier proprietary systems. All models were accessed through hosted API routing services, including OpenRouter[2] and Infron[3], and generation temperature was set to 0.2 for all runs.

Each model was evaluated under two prompting settings: zero-shot and fixed few-shot. In the zero-shot setting, the model received only the task instruction and the current test item. In the few-shot setting, the model additionally received a fixed set of demonstration examples. This was to ensure comparability across models.

3.6 Evaluation Metrics

Task 1 was evaluated using accuracy. A prediction was counted as correct when the answer exactly matched the gold proverb. Task 2 was evaluated using both automatic metrics and human judgments. Proverb explanations are open-ended, and semantically valid outputs may differ substantially from the reference wording. Therefore, human evaluation was adopted alongside automatic metrics.

Automatic Evaluation For automatic evaluation of Task 2, we report three metrics on both Mandarin and English explanations: BLEU (Papineni et al., 2002) for surface-level lexical overlap with the reference, BERTScore-F1 (Zhang et al., 2019) (computed with *bert-base-multilingual-cased* [4]) for token-level semantic similarity, and sentence-embedding cosine similarity (Reimers and Gurevych, 2020) (computed with *paraphrase-*

[1] https://www.anthropic.com/news/claude-opus-4-8

[2] https://openrouter.ai/

[3] https://infron.ai/

[4] https://huggingface.co/google-bert/bert-base-multilingual-cased

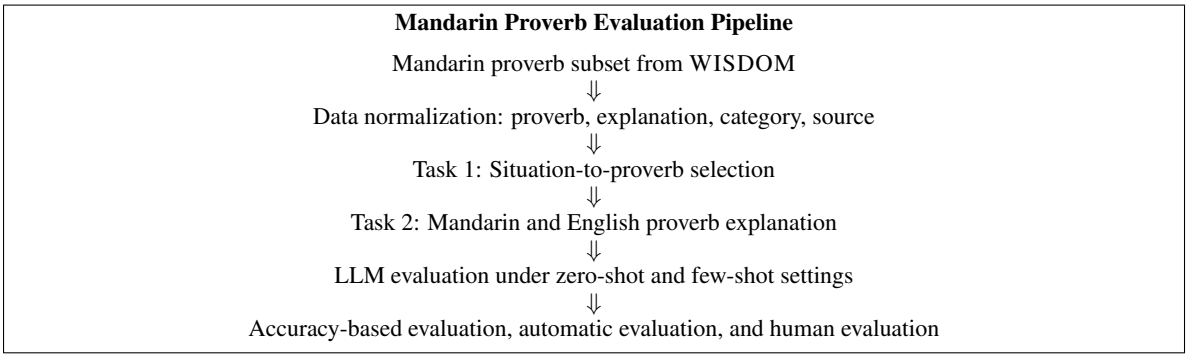


Figure 2: Pipeline for constructing and evaluating the Mandarin proverb understanding benchmark. The benchmark evaluates both situation-to-proverb reasoning and proverb explanation.

multilingual-MiniLM-L12-v2[5]) for sentence-level semantic similarity. For BLEU, Mandarin explanations are tokenized at the character level to avoid reliance on Chinese word segmentation, which can be unstable for compact proverb expressions and culturally specific phrases.

Human Evaluation For human evaluation of Task 2, two native Chinese speakers with master's degree in translation and C1-level English proficiency independently annotated a sample of 160 bilingual explanation outputs, with 20 outputs sampled from each model-setting condition. Four 5-point Likert scale dimensions were used: ***semantic correctness***, ***cultural faithfulness***, ***clarity***, and ***learner usefulness***.

Semantic correctness measures whether the explanation captures the proverb's intended meaning. **Cultural faithfulness** measures whether the explanation preserves the proverb's cultural, metaphorical, and pragmatic implications. **Clarity** measures whether the explanation is understandable and well expressed. **Learner usefulness** measures whether the explanation would help Mandarin learners or cross-cultural readers understand when and why the proverb is used. Final human evaluation scores were computed by averaging the two annotators' ratings.

We calculated 95% confidence intervals for condition means using 10,000 nonparametric bootstrap resamples. Because the zero-shot and few-shot conditions contained different proverb items, their overall scores were compared within each model using independent-samples Welch tests. The resulting p-values were adjusted using the Holm procedure across the four model-wise comparisons. We additionally assessed the association between cultural faithfulness and learner usefulness across the 160 outputs using Spearman's ρ, with a 95% confidence interval based on 10,000 output-level bootstrap resamples.

Annotator Agreement Inter-annotator agreement was assessed on the 160 sampled Task 2 outputs using quadratic weighted Cohen's kappa. Agreement was substantial for semantic correctness and cultural faithfulness, moderate for learner usefulness, and lower for clarity. We retained clarity because it measures a conceptually distinct and practically important property: an explanation can be readable while remaining semantically or culturally inaccurate. The lower agreement on clarity may indicate greater subjectivity in how annotators weighed fluency, concision, and unnecessary elaboration. The complete human evaluation guideline is provided in Appendix C.

Rating Dimension	Cohen's kappa
Semantic Correctness	0.757
Cultural Faithfulness	0.706
Clarity	0.360
Learner Usefulness	0.637

Table 1: Inter-annotator agreement for Task 2 human evaluation.

4 Results and Discussion

This section reports model performance on the two tasks and discusses the gap between proverb recognition and culturally faithful explanation.

4.1 Task 1: Situation-to-Proverb Selection

This task evaluates whether models can identify the proverb that best matches an implicit real-life

[5] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

Model	Zero-shot	Few-shot
Qwen-2.5-7B	0.950	0.941
Llama-3.1-8B	0.757	0.742
DeepSeek-V4-Flash	0.987	0.985
GPT-5.4	0.991	0.992

Table 2: Task 1 situation-to-proverb selection performance. Each model was evaluated under zero-shot and few-shot settings.

situation from a set of candidate proverbs.

4.1.1 Overall Selection Performance

Table 2 reports Task 1 selection performance. Overall, models perform substantially better on situation-to-proverb selection than on open-ended bilingual explanation. GPT-5.4 achieves the highest accuracy in both settings, reaching 99.1% in zero-shot and 99.2% in few-shot. DeepSeek-V4-Flash follows closely, with accuracy above 98% in both settings. Qwen-2.5-7B also performs strongly, with accuracy above 94%, while Llama-3.1-8B lags behind the other models.

The overall high accuracy suggests that Task 1 is easier for strong LLMs. Unlike open-ended explanation, the selection task provides both a concrete situation and a closed set of candidate proverbs, allowing models to rely on recognition, semantic matching, and candidate elimination. These results should be interpreted as evidence of strong recognition performance rather than complete mastery of proverb explanation or culturally grounded use. In other words, selecting the correct proverb from a candidate set does not necessarily mean that the model can explain the proverb accurately, preserve its cultural implications, or use it appropriately in open-ended contexts.

4.1.2 Effect of Few-shot Prompting

In addition, few-shot prompting does not lead to consistent improvements. Few-shot prompting yields only a slight improvement for GPT-5.4, while Qwen-2.5-7B, Llama-3.1-8B, and DeepSeek-V4-Flash perform slightly better in the zero-shot setting. This indicates that fixed demonstrations may be less important when the task already provides strong contextual and candidate cues. For the selection task, the situation and the four candidate proverbs already define a constrained reasoning space, making additional examples less beneficial. These results also suggest that few-shot prompting is not uniformly advantageous for Mandarin proverb understanding. Its effect depends on the task format and model strength. For multiple-choice selection, where recognition and elimination strategies are already available, few-shot examples appear to provide limited additional benefit.

4.2 Task 2: Bilingual Proverb Explanation

This task evaluates whether models can generate accurate, culturally faithful, clear, and learner-oriented explanations of Mandarin proverbs in both Mandarin and English.

4.2.1 Reference-Based Evaluation

Table 3 reports the reference-based evaluation results for Task 2. BLEU scores are generally low across models, reflecting the open-ended nature of proverb explanation. A valid explanation may differ substantially from the reference wording. Therefore, BLEU is a surface-overlap indicator rather than a direct measure of explanation quality. BERTScore-F1 and embedding cosine provide complementary semantic similarity scores, but reference-based metrics cannot fully capture culturally appropriate proverb interpretation.

Among the evaluated models, DeepSeek-V4-Flash obtains the highest automatic scores in both zero-shot and few-shot settings across most metrics. GPT-5.4 and Qwen-2.5-7B receive lower BLEU scores, but their BERTScores and embedding cosine scores remain relatively competitive, suggesting that surface-level overlap and broader semantic similarity do not always align.

Across prompting settings, few-shot prompting does not consistently improve automatic scores. For DeepSeek-V4-Flash and Llama-3.1-8B, the zero-shot setting outperforms the few-shot setting across most automatic metrics. This may indicate that the fixed few-shot examples influence response style or wording without necessarily increasing similarity to the reference explanations. Given these limitations, we complement the automatic evaluation with human judgments.

4.2.2 Human Evaluation of Explanation Quality

Table 4 presents the human evaluation results for Task 2. Overall, DeepSeek-V4-Flash achieves the strongest human-rated performance in both prompting settings, with overall scores of 4.49 in the zero-shot setting and 4.44 in the few-shot setting. This pattern is consistent with the automatic evaluation results, where DeepSeek-V4-Flash also obtains the highest scores across most metrics. GPT-5.4 also

Model	Setting	BLEU (ZH)	BLEU (EN)	BERTScore (ZH)	BERTScore (EN)	Emb. (ZH)	Emb. (EN)
Qwen-2.5-7B	Zero-shot	2.712	2.404	0.686	0.697	0.602	0.522
Llama-3.1-8B	Zero-shot	3.299	2.875	0.694	0.699	0.588	0.465
DeepSeek-V4-Flash	Zero-shot	**4.812**	**4.152**	**0.712**	**0.716**	**0.655**	**0.593**
GPT-5.4	Zero-shot	2.286	1.916	0.682	0.686	0.607	0.534
Qwen-2.5-7B	Few-shot	2.455	2.427	0.684	0.692	0.583	0.506
Llama-3.1-8B	Few-shot	2.412	2.174	0.682	0.689	0.536	0.449
DeepSeek-V4-Flash	Few-shot	**3.489**	**3.001**	**0.696**	**0.698**	**0.614**	**0.541**
GPT-5.4	Few-shot	2.312	2.174	0.684	0.690	0.594	0.519

Table 3: Reference-based evaluation results for Task 2 bilingual proverb explanation. BLEU measures surface-level lexical overlap, BERTScore measures token-level semantic similarity, and Emb. denotes sentence-transformer embedding cosine similarity.

Model	Setting	Semantic Correctness	Cultural Faithfulness	Clarity	Learner Usefulness	Overall
Qwen-2.5-7B	Zero-shot	3.55 ± 1.40	2.98 ± 1.25	4.03 ± 0.83	3.27 ± 1.09	3.46 ± 1.10
Llama-3.1-8B	Zero-shot	2.67 ± 1.30	2.40 ± 1.11	3.35 ± 0.78	2.70 ± 1.02	2.78 ± 1.00
DeepSeek-V4-Flash	Zero-shot	4.83 ± 0.34	4.30 ± 0.55	4.67 ± 0.41	4.17 ± 0.34	$\mathbf{4.49 \pm 0.31}$
GPT-5.4	Zero-shot	4.65 ± 0.49	4.12 ± 0.54	4.10 ± 0.50	3.90 ± 0.45	4.19 ± 0.35
Qwen-2.5-7B	Few-shot	4.15 ± 1.19	3.80 ± 1.06	4.65 ± 0.46	3.92 ± 0.91	4.13 ± 0.86
Llama-3.1-8B	Few-shot	3.25 ± 1.30	2.88 ± 1.15	3.73 ± 0.77	2.98 ± 1.04	3.21 ± 1.03
DeepSeek-V4-Flash	Few-shot	4.70 ± 0.57	4.20 ± 0.57	4.75 ± 0.30	4.10 ± 0.58	$\mathbf{4.44 \pm 0.44}$
GPT-5.4	Few-shot	4.20 ± 1.06	3.80 ± 0.86	4.03 ± 0.44	3.45 ± 0.81	3.87 ± 0.65

Table 4: Human evaluation results for Task 2 bilingual proverb explanation. Scores are reported as mean $\pm$ standard deviation. Final scores were obtained by averaging the two annotators' ratings. Bootstrap 95% confidence intervals are reported in Appendix D.

performs strongly in the zero-shot setting, obtaining an overall score of 4.19.

These results suggest that the stronger models generally produce accurate and culturally appropriate bilingual explanations of Mandarin proverbs. DeepSeek's advantage is also reflected in its smaller standard deviations, indicating consistently high ratings across items, whereas the other models exhibit greater variation in explanation quality.

The effect of few-shot prompting varies across models. Qwen-2.5-7B and Llama-3.1-8B receive higher mean human-evaluation scores in the few-shot setting, suggesting that demonstration examples may help lower-performing models generate more complete and useful explanations. By contrast, DeepSeek-V4-Flash performs comparably across the two settings, while GPT-5.4 obtains a higher mean score in the zero-shot setting. Independent-samples Welch tests showed that the Qwen comparison was nominally significant before correction ($p = .04$), whereas the comparisons for Llama ($p = .19$), DeepSeek ($p = .64$), and GPT-5.4 ($p = .06$) were not. However, none of the four differences remained statistically significant after Holm correction (adjusted ps = .15, .39, .64, and .18, respectively).

Taken together, these results do not establish a statistically reliable benefit of the fixed demonstrations, although the observed mean differences suggest that their effects may vary by model. Therefor, the effectiveness of few-shot prompting may depend on both model capability and demonstration design, including the number, selection, and cultural relevance of the examples.

Across the four evaluation dimensions, the results reveal a consistent pattern: models handle surface-level demands better than culturally grounded ones. The following discussion illustrates each dimension in turn.

Semantic Correctness Semantic correctness scores are generally higher than both cultural faithfulness and learner usefulness, indicating that models can usually recover at least the broad or literal meaning of a proverb. Even when other dimensions receive low scores, the core propositional content is often recognizable in the output. This suggests that the models' linguistic competence in Mandarin is sufficient to parse proverb wording and produce a semantically related paraphrase. The challenge lies in moving beyond paraphrase toward culturally accurate interpretation.

Cultural Faithfulness Cultural faithfulness scores are noticeably less stable, particularly

for proverbs that involve historical figures, culturally specific roles, or institutionally grounded metaphors. These examples suggest that background-dependent proverbs may be challenging. Mandarin proverbs draw heavily on classical literature, dynastic history, and folk narrative, and models frequently fail to recover these layers. As Figure 1 shows, "山高皇帝远" (*the mountain is high, and the emperor is far away*) illustrates this well. The proverb encodes political common sense: central authority cannot reach remote regions, and it is conventionally invoked to explain local defiance or unchecked local power. However, model outputs ranged from a purely literal reading (the mountains are "far away and majestic") to an inverted one that framed the proverb as a call for obedience to central authority. Both readings are fluent, but both miss what the proverb is actually used for. A similar failure appears with *Wang Tu* ("Butcher Wang"), whom one model promoted to a "leader figure", turning a proverb about an indispensable skilled tradesman into a generic statement about leadership.

Clarity Clarity scores are generally the highest among the four dimensions, given that clarity measures readability and coherence rather than factual or cultural accuracy. This indicates that models can produce fluent and well-organized explanations even when the underlying interpretation is incomplete or culturally off-target. The gap between clarity and learner usefulness is especially visible for Qwen-2.5-7B and Llama-3.1-8B, whose readable explanations do not consistently translate into pedagogically useful interpretations.

Learner Usefulness Learner usefulness was strongly positively associated with cultural faithfulness across the 160 evaluated outputs (Spearman's $\rho = .83$, 95% bootstrap CI $[.76, .89]$, $p < .001$). This association suggests that explanations judged to preserve culturally conventionalized meanings and pragmatic implications were also generally judged more useful for learners. A fluent explanation that omits these elements may be easy to read but offer limited help in understanding when and why the proverb is used.

LLM-assisted Language Learning The two tasks provide implications for LLM-assisted language learning. Situation-to-proverb selection could support contextual practice, while bilingual explanations could help learners connect literal imagery with conventional meaning and appropriate use. However, the gap between clarity and cultural faithfulness indicates that fluent model output should not be presented as authoritative without safeguards. Practical systems could combine generation with retrieval from curated proverb dictionaries, historical sources, or cultural knowledge bases, and route culturally opaque or low-confidence cases for teacher or expert review.

5 Conclusion

This study evaluated four large language models on Mandarin proverb understanding through two complementary tasks, situation-to-proverb selection and bilingual proverb explanation, using a Mandarin subset of the WISDOM dataset under zero-shot and few-shot prompting conditions.

Results show that proverb recognition and proverb explanation present different levels of difficulty. Models achieved high accuracy on the multiple-choice selection task, especially GPT-5.4 and DeepSeek-V4-Flash, suggesting strong recognition and semantic matching ability when candidate proverbs are provided. Open-ended explanation proved considerably more demanding. Human evaluation showed that models often produced fluent and readable explanations, yet cultural faithfulness and learner usefulness lagged behind, particularly for proverbs that rest on historical allusions or conventionalized pragmatic functions. In short, surface fluency does not guarantee culturally grounded understanding.

The comparison between prompting presents a nuanced picture. Under the fixed few-shot examples, mean scores increased for the two lower-performing models in the open-ended explanation task, whereas the stronger models showed little improvement or obtained higher scores in the zero-shot setting. However, none of the zero-shot–few-shot differences remained significant after multiple-comparison correction. The effectiveness of few-shot prompting may depend on the model and demonstration design.

Overall, Mandarin proverb understanding and explanation remain challenging for culturally grounded LLM evaluation. By combining accuracy-based evaluation, automatic metrics, and human judgment, this study shows that current models handle figurative language reliably at the level of recognition and paraphrase, but far less reliably at the level of allusion, pragmatics, and cul-

tural convention. Future evaluation should move beyond recognition-based tasks and place greater weight on explanation quality, cultural faithfulness, and pedagogical usefulness.

Limitations

The main limitations of this study concern the evaluation setup. Task 1 was formulated as multiple-choice selection, which enables accuracy-based evaluation at scale but may understate the difficulty of the underlying skill. As models receive both the situation and a closed set of candidate proverbs, they can succeed through recognition, semantic matching, and elimination, and high accuracy does not necessarily indicate mastery of culturally appropriate proverb use in open-ended settings. In addition, the three distractors were drawn from the same dataset but were not systematically constructed as category-matched or semantically nearest hard negatives. More controlled distractor construction could provide a more demanding test. Due to limited resources, human spot-checking in Task 1 and human evaluation in Task 2 was conducted on a sampled subset rather than the full set.

Relatedly, the few-shot condition used a fixed set of demonstration examples. This choice improves comparability across models and avoids variation from dynamic example selection, but the chosen examples may shape response style and cannot represent the full diversity of Mandarin proverbs. Future work could consider alternative retrieval-based or semantically or culturally matched example-selection strategies.

Moreover, the scope of the study is limited to a Mandarin subset of the broader proverb dataset. The dataset includes broad semantic categories but was not annotated by further historical or allusion details. Future work could develop such metadata with explicit annotation guidelines. The scope of this study is also limited to Mandarin. Future work could extend the evaluation to Cantonese and other regional varieties spoken in China, as well as to additional languages. It could also test open-ended proverb use and generation, and examine whether retrieval-based support improves cultural faithfulness and learner-oriented explanation quality.

Acknowledgments

We express our sincere thanks to the annotators for their efforts and thorough work. We also thank the reviewers for their constructive comments.

References

Fakhraddin Alwajih, Abdellah El Mekki, Samar Mohamed Magdy, AbdelRahim A. Elmadany, Omer Nacar, El Moatez Billah Nagoudi, Reem Abdel-Salam, Hanin Atwany, Youssef Nafea, Abdulfattah Mohammed Yahya, Rahaf Alhamouri, Hamzah A. Alsayadi, Hiba Zayed, Sara Shatnawi, Serry Sibaee, Yasir Ech-chammakhy, Walid Al-Dhabyani, Marwa Mohamed Ali, Imen Jarraya, Ahmed Oumar El-Shangiti, Aisha Alraeesi, Mohammed Anwar AL-Ghrawi, Abdulrahman S. Al-Batati, Elgizouli Mohamed, Noha Taha Elgindi, Muhammed Saeed, Houdaifa Atou, Issam Ait Yahia, Abdelhak Bouayad, Mohammed Machrouh, Amal Makouar, Dania Alkawi, Mukhtar Mohamed, Safaa Taher Abdelfadil, Amine Ziad Ounnoughene, Anfel Rouabhia, Rwaa Assi, Ahmed Sorkatti, Mohamedou Cheikh Tourad, Anis Koubaa, Ismail Berrada, Mustafa Jarrar, Shady Shehata, and Muhammad Abdul-Mageed. 2025. Palm: A culturally inclusive and linguistically diverse dataset for Arabic LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32871–32894, Vienna, Austria. Association for Computational Linguistics.

Tyler A Chang, Catherine Arnett, Abdelrahman Eldesokey, Abdelrahman Sadallah, Abeer Kashar, Abolade Daud, Abosede Grace Olanihun, Adamu Labaran Mohammed, Adeyemi Praise, Adhikarinayum Meerajita Sharma, et al. 2025. Global piqa: Evaluating physical commonsense reasoning across 100+ languages and cultures. *arXiv preprint arXiv:2510.24081*.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.

Yicheng Fu, Zhemin Huang, Liuxin Yang, Yumeng Lu, and Zhongdongming Dai. 2025. CHENGYU-BENCH: Benchmarking large language models for Chinese idiom understanding and use. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2355–2366, Suzhou, China. Association for Computational Linguistics.

Sayan Ghosh and Shashank Srivastava. 2022. ePiC: Employing proverbs in context as a benchmark for abstract language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3989–4004, Dublin, Ireland. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle,

Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Hai Hu, Siyuan Song, Chongtian Shao, Kejia Zhang, Tianjian Zhu, and Xiaojing Zhao. 2026. Reasoning or memorization: Can llms understand and generate chinese xiehouyu riddles? *arXiv preprint arXiv:2607.23440*.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284, Toronto, Canada. Association for Computational Linguistics.

Jisu Kim, Youngwoo Shin, Uiji Hwang, Jihun Choi, Richeng Xuan, and Taeuk Kim. 2025. Memorization or reasoning? exploring the idiom understanding of LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21678–21699, Suzhou, China. Association for Computational Linguistics.

Salima Lamsiyah. 2026. Responsible nlp: Why performance is no longer enough. *Natural Language Processing*, page 1–7.

Salima Lamsiyah, Kamyar Zeinalipour, Samir El amrany, Matthias Brust, Marco Maggini, Pascal Bouvry, and Christoph Schommer. 2025. ArabicSense: A benchmark for evaluating commonsense reasoning in Arabic with large language models. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 1–11, Abu Dhabi, UAE. Association for Computational Linguistics.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2024. Translate meanings, not just words: Idiomkb's role in optimizing idiomatic translation with language models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.

Samar Mohamed Magdy, Sang Yun Kwon, Fakhraddin Alwajih, Safaa Taher Abdelfadil, Shady Shehata, and Muhammad Abdul-Mageed. 2025. JAWAHER: A multidialectal dataset of Arabic proverbs for LLM benchmarking. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12320–12341, Albuquerque, New Mexico. Association for Computational Linguistics.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki A Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew A Ayele, et al. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jipeng Qiang, Yang Li, Chaowei Zhang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2023. Chinese idiom paraphrasing. *Transactions of the Association for Computational Linguistics*, 11:740–754.

Le Qiu, Emmanuele Chersoni, and Aline Villavicencio. 2025. ChengyuSTS: An intrinsic perspective on Mandarin idiom representation. In *Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 1–12, Suzhou, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Minghan Wang, Viet Thanh Pham, Farhad Moghimifar, and Thuy-Trang Vu. 2025. Proverbs run in pairs: Evaluating proverb translation capability of large language model. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1646–1662, Vienna, Austria. Association for Computational Linguistics.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv e-prints*, pages arXiv–2412.

Shisen Yue, Siyuan Song, Xinyuan Cheng, and Hai Hu. 2024. Do large language models understand conversational implicature- a case study with a Chinese sitcom. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pages 1270–1285, Taiyuan, China. Chinese Information Processing Society of China.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Xiaojing Zhao, Emmanuele Chersoni, Chu-Ren Huang, and Han Xu. 2026a. How do language models handle emotional content in video game localization? a computational linguistics approach. *Research Methods in Applied Linguistics*, 5(1):100294.

Xiaojing Zhao, Emmanuele Chersoni, Chu-Ren Huang, and Han Xu. 2026b. Translating vulgar language in video game localization: a case study of black myth: Wukong. *Perspectives*, 34(3):493–512.

Xiaojing Zhao, Han Xu, Huacheng Song, Emmanuele Chersoni, and Chu-Ren Huang. 2025. Can LLMs help sun wukong in his journey to the west? a case study of language models in video game localization. In *Proceedings of the First Workshop on Natural Language Processing and Language Models for Digital Humanities*, pages 164–173, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. ChID: A large-scale Chinese IDiom dataset for cloze test. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 778–787, Florence, Italy. Association for Computational Linguistics.

LuxDiag-RAG: Diagnostic Evaluation of Retrieval-Augmented Generation for Luxembourgish Reading Comprehension

Keerthana Murugaraj[*†] **, Hedi Tebourbi**[†]**,**
Christophe Friezas Gonçalves, Salima Lamsiyah

Faculty of Science, Technology and Medicine (FSTM),
University of Luxembourg, Luxembourg

keerthana.murugaraj@uni.lu, hedi.tebourbi.001@student.uni.lu,
christophe.friezas@uni.lu, salima.lamsiyah@uni.lu

[†] Equal contribution [*] Corresponding author

Abstract

Retrieval-augmented generation (RAG) is widely used to ground large language model (LLM) outputs in external evidence, but its evaluation remains concentrated on high-resource languages. We present LuxDiag-RAG, a diagnostic evaluation framework for Luxembourgish retrieval-augmented reading comprehension using LuxDiagRC, a corpus of 640 multiple-choice questions over 16 annotated texts. We compare closed-book, full-text oracle, text-restricted RAG, and collection-wide RAG settings to separate answerability without context, comprehension with gold context, evidence selection, and document-plus-evidence retrieval. We evaluate lexical, dense, and hybrid retrievers with four LLMs, and report answer accuracy, source-text recall, span-level evidence recall, distractor-span retrieval, diagnostic annotation-based analyses, and targeted human evaluation of answer correctness and evidence sufficiency. Results show that LLMs can use Luxembourgish context effectively when the full passage is provided, with oracle accuracy between 0.81 and 0.85. RAG accuracy improves with larger retrieved contexts and approaches oracle performance in the text-restricted setting, while collection-wide retrieval remains a bottleneck. Character-level lexical and weighted hybrid retrieval outperform off-the-shelf multilingual dense retrieval for collection-wide evidence selection. Overall, LuxDiag-RAG[1] provides a reproducible protocol for diagnosing retrieval and comprehension failures in low-resource RAG evaluation.

1 Introduction

RAG has become a standard approach for grounding LLM outputs in external evidence (Lewis et al., 2020b). By combining retrieval with generation, RAG systems reduce reliance on parametric memory and allow answers to be conditioned on explicit textual context. This is especially important for question answering and reading comprehension, where correctness depends on locating and using relevant evidence. However, RAG evaluation remains concentrated on high-resource languages and English-centric benchmarks. Low-resource and smaller European languages are less explored, even though they often have a strong need for evidence-grounded language technologies. Luxembourgish is one such case: it is socially and educationally important in Luxembourg but has comparatively limited NLP resources, benchmark datasets, and evaluation studies. Recent work has introduced Luxembourgish language models and representations, including LuxemBERT (Lothritz et al., 2022) and LuxEmbedder (Philippy et al., 2025), and has begun evaluating LLM support for Luxembourgish using proficiency-exam-style tasks (Lothritz et al., 2026). Yet the ability of RAG systems to retrieve and use Luxembourgish evidence for context-grounded reading comprehension remains insufficiently studied.

We study this problem using LuxDiagRC (Gonçalves et al., 2026), a diagnostic Luxembourgish story-based reading-comprehension corpus containing 640 multiple-choice questions over 16 annotated texts. We use LuxDiagRC as an evaluation benchmark rather than a training corpus: the source texts form the retrieval collection, the questions and answer options define the multiple-choice task, and the dataset annotations support diagnostic error analysis. Because the task is passage-based, closed-book answering is not the primary setting; we use it only as a calibration and contamination control. Our main comparison is between *full-text oracle answering*, which removes retrieval failure by providing the complete gold source text, and *RAG systems*, which must retrieve

[1] GitHub Repository

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 161–171
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_018

evidence chunks before answering. We evaluate lexical, dense, and hybrid retrieval methods, including word-level BM25, character n-gram BM25, multilingual dense retrieval, and score-fusion hybrids. We consider two retrieval settings. In *text-restricted retrieval*, candidate chunks are limited to the gold source text, isolating evidence selection within a known passage. In *collection-wide retrieval*, candidate chunks come from all texts, requiring both source-text selection and evidence selection. This design allows us to distinguish retrieval errors from comprehension errors.

The main contribution of this work is **LUXDIAG-RAG**, a diagnostic evaluation protocol for Luxembourgish RAG. Beyond answer accuracy, we use critical evidence spans, distractor spans, typed answer options, cognitive labels, linguistic annotations, and targeted human evaluation to analyze where systems fail. This enables us to distinguish cases where the relevant evidence is not retrieved, where misleading distractor evidence is retrieved, and where the model receives useful evidence but still answers incorrectly. Our contributions are as follows:

1. We present a diagnostic evaluation of RAG for Luxembourgish reading comprehension using LuxDiagRC.

2. We compare closed-book control, full-text oracle answering, text-restricted RAG, and collection-wide RAG to separate answerability, comprehension, evidence selection, and document retrieval.

3. We evaluate lexical, dense, and hybrid retrieval methods for Luxembourgish evidence retrieval.

4. We use span-level evidence annotations, typed distractors, cognitive labels, linguistic annotations, and targeted human evaluation to diagnose retrieval traps and comprehension failures.

2 Related Work

RAG evaluation. Prior work shows that answer accuracy alone is insufficient, because a correct or incorrect answer may depend on retrieval quality, context relevance, faithfulness, and the model's use of evidence (Es et al., 2024; Saad-Falcon et al., 2024; Murugaraj et al., 2026). These frameworks motivate diagnostic evaluation of the retrieval-generation pipeline. Our work adopts this perspective but studies it in a controlled Luxembourgish reading-comprehension setting where gold evidence spans and distractor spans allow retrieval and comprehension failures to be analyzed separately.

Multilingual and low-resource QA and retrieval. Multilingual QA and reading-comprehension benchmarks show that model performance varies substantially across languages, motivating targeted evaluation rather than assuming transfer from English or other high-resource languages (Clark et al., 2020; Lewis et al., 2020a; Longpre et al., 2021; Bandarkar et al., 2024). Retrieval adds a second bottleneck. Multilingual retrieval benchmarks such as Mr. TyDi (Zhang et al., 2021) and MIRACL (Zhang et al., 2023) show that ranking quality differs across languages and that dense retrieval is not uniformly superior to lexical or hybrid retrieval in non-English settings. This is particularly relevant for low-resource RAG, where retrieval may depend on language-specific morphology, spelling variation, preprocessing, and embedding coverage.

Low-resource and language-specific RAG. Recent work on multilingual and low-resource RAG further shows that language choice affects retrieval, evidence selection, and answer generation. Multilingual RAG studies report inequalities across languages and selection biases toward high-resource languages (Wu et al., 2024). Language-specific RAG studies for Persian and Arabic highlight the importance of morphology-aware processing, retrieval model choice, embedding quality, and evaluation design (Hosseini et al., 2024; Hosseinbeigi et al., 2025; El-Beltagy and Abdallah, 2024; Alsubhi et al., 2025). These studies motivate language-specific RAG evaluation. In contrast to work focused mainly on system construction or benchmark-level performance, we use Luxembourgish reading-comprehension annotations to separate document retrieval, evidence selection, and comprehension errors.

Existing diagnostic reading-comprehension datasets support evaluation beyond aggregate accuracy by linking model errors to question types, linguistic difficulty, and the evidence required to answer each question. LuxDiagRC provides such annotations for Luxembourgish (Gonçalves et al., 2026). Unlike prior work, which primarily uses reading-comprehension datasets to evaluate answer

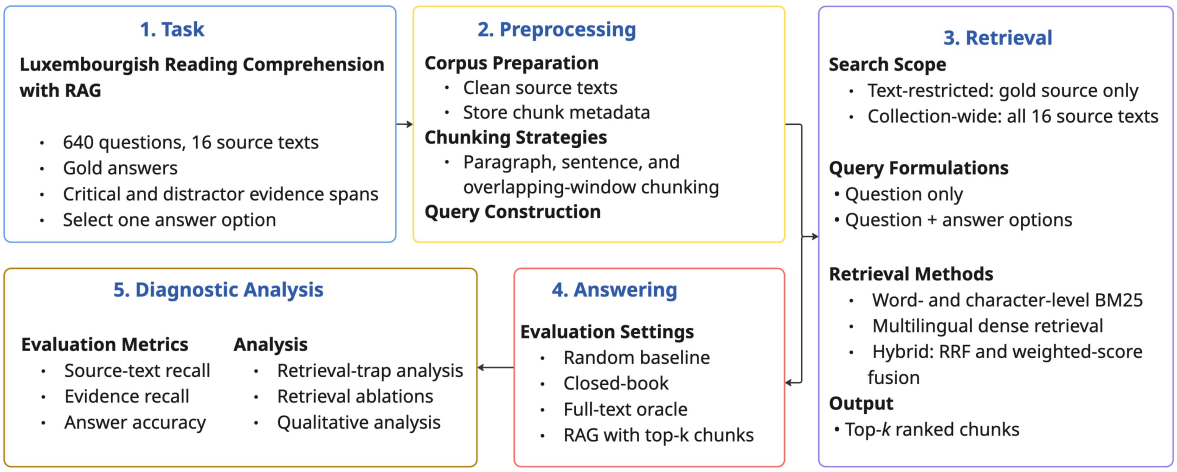


Figure 1: Overview of the LuxDiag-RAG evaluation pipeline. The framework proceeds through task formulation, preprocessing, retrieval, answering, and diagnostic analysis. Retrieval is evaluated under text-restricted and collection-wide settings using lexical, dense, and hybrid methods, followed by answering under multiple evaluation settings and analysis of retrieval and downstream performance.

accuracy, we leverage these annotations to assess RAG systems diagnostically by distinguishing errors arising from document retrieval, evidence selection, distractor retrieval, and answer selection.

3 Methodology

Figure 1 summarizes the full evaluation pipeline.

Task formulation. We study multiple-choice Luxembourgish reading comprehension with RAG. Each LuxDiagRC instance contains a question, four answer options, a gold answer, and a gold source text (Gonçalves et al., 2026). The system must select exactly one answer option. We treat the LuxDiagRC source texts as the retrieval collection and evaluate whether a system can retrieve relevant evidence and use it to answer correctly across controlled retrieval and answering settings. This formulation separates document retrieval, evidence selection, and answer prediction.

Preprocessing and query construction. The retrieval corpus consists of the cleaned LuxDiagRC source texts; we do not use external documents. Each text is split into chunks that store the source-text identifier, character offsets, token count, and chunk text. The character offsets follow the same clean-text coordinate system as the annotated critical and distractor spans, allowing retrieved chunks to be aligned with gold evidence during evaluation. We compare paragraph-based chunks, sentence-based chunks, and fixed-size overlapping chunks, with the overlapping strategy using approximately 150-token windows and 50% overlap to reduce boundary effects. For retrieval queries, we compare *question-only* queries with *question+options* queries, where the latter concatenates the question with the four answer options. This can improve lexical matching but may also introduce distractor terms.

Retrieval. We evaluate retrieval under two controlled settings. In the *text-restricted* setting, retrieval is limited to chunks from the gold source text, isolating evidence selection within a known passage. In the *collection-wide* setting, retrieval is performed over chunks from all source texts, requiring both source-text retrieval and evidence selection. Across both settings, we compare lexical, dense, and hybrid retrievers. Lexical retrieval uses BM25 with either word-level or character n-gram analysis. Dense retrieval uses multilingual sentence-transformer embeddings, with L2-normalized query and chunk vectors scored by cosine similarity:

$$s_{\text{dense}}(q, c) = \mathbf{e}_q^\top \mathbf{e}_c.$$

Hybrid retrieval combines BM25 and dense retrieval using reciprocal rank fusion (Cormack et al., 2009) or weighted min-max score fusion:

$$s_{\text{hybrid}}(q, c) = \alpha\, \tilde{s}_{\text{BM25}}(q, c) + (1-\alpha)\, \tilde{s}_{\text{dense}}(q, c).$$

We evaluate retrieval at multiple top-k values to measure how evidence coverage and answer accuracy change as more context is provided.

Answering. We evaluate four answering settings. A *random baseline* selects one of the four answer options without using an LLM. In the *closed-book* setting, the model receives only the question and answer options, serving as a calibration and possible contamination control. In the *full-text oracle* setting, the model receives the complete gold source text, estimating performance when retrieval failure is removed. In the *RAG* settings, the model receives the question, answer options, and the top-k retrieved chunks. All LLM-based settings use the same multiple-choice prompt format and require a single option label.

Diagnostic analysis. We organize our analysis into two levels. In the main results, we first use closed-book and full-text oracle controls to calibrate answerability and comprehension when retrieval is removed. We then evaluate retrieval quality using source-text recall and span-level evidence recall, followed by RAG answer accuracy under text-restricted and collection-wide settings. These results show how document retrieval, evidence selection, and answer prediction contribute to overall performance. Finally, targeted human evaluation assesses whether selected model outputs are correct and whether the retrieved context provides sufficient evidence. In the appendix, we report retrieval ablations over retriever choice, chunking strategy, query formulation, and distractor span retrieval analysis, and we further analyze retrieval traps by checking whether the retrieved context contains the critical span, the distractor span, both, or neither.

4 Experiments and Results

4.1 Experimental Setup

Dataset. We use LuxDiagRC, which contains 640 multiple-choice questions over 16 annotated Luxembourgish texts (Gonçalves et al., 2026). Each question is associated with a gold answer, a source text, cognitive and linguistic annotations, and evidence-related annotations. LuxDiagRC distinguishes four cognitive question types: direct-fact retrieval, inferential reasoning, interpretive comprehension, and evaluative reasoning. These categories respectively assess whether a model can locate explicitly stated information, connect clues across a passage, interpret meanings or implications, and make evidence-based judgments or comparisons. The dataset also includes word-level annotations of Luxembourgish-specific linguistic phenomena using tags of the form `<CATEGORY-SUBTYPE>`. These cover lexical (`LEX`), syntactic (`SYN`), morphological (`MORPH`), orthographic (`ORTHO`), and discourse-level (`DISC`) phenomena. We use these annotations to report accuracy by linguistic category and to examine how Luxembourgish variation, particularly orthographic and morphological variation, affects retrieval performance. We convert the data into a question-level evaluation format containing the question, shuffled answer options, gold label, source-text identifier, critical evidence spans, distractor spans, and annotation labels. The answer options are presented in a fixed, seeded shuffle to avoid relying on positional regularities. Critical spans indicate the textual evidence needed to answer the question, while distractor spans mark text that may support an incorrect option. These annotations are used only for evaluation and diagnostic analysis, not as input to the retriever or the LLM.

Retrieval models. Lexical retrieval uses a Python implementation of Okapi BM25, rather than an external BM25 package. We use the standard BM25 parameters $k_1 = 1.5$ and $b = 0.75$, with a Lucene-style non-negative IDF variant. We evaluate two BM25 analyzers: a word-level analyzer and a character 3–5-gram analyzer. Dense retrieval uses the *sentence-transformers* library with two multilingual models: `intfloat/multilingual-e5-base` and `BAAI/bge-m3`. Dense scores are computed as cosine similarities between L2-normalized query and chunk embeddings. Hybrid retrieval combines BM25 and dense retrieval using two implemented fusion strategies: reciprocal rank fusion and weighted min-max score fusion. Weighted score fusion is evaluated with $\alpha \in \{0.3, 0.5, 0.7\}$, with $\alpha = 0.5$ used for the main hybrid configuration. We evaluate retrieval at top-k values of $k \in \{1, 3, 5, 10\}$.

Answering models and prompting. We evaluate four instruction-tuned LLMs: GPT-5.5, DeepSeek-V4-Pro, Claude Opus 4.8, and Claude Sonnet 4.6. Each model is evaluated under three settings: closed-book, full-text oracle, and RAG. All LLM-based systems use the same multiple-choice prompt template and are instructed to return only a single answer-option label. Because the available decoding controls differ across model providers, we use

the most deterministic configuration supported by each API and record the effective decoding parameters for every run. Specifically, DeepSeek-V4-Pro and Claude Sonnet 4.6 are evaluated with the temperature set to 0. GPT-5.5 is evaluated using its default temperature because its API does not accept a value of 0. In contrast, Claude Opus 4.8 is evaluated with its default decoding configuration because the temperature parameter is deprecated.

Evaluation metrics. We evaluate retrieval and answering separately. For retrieval, we report source-text recall, span-level evidence recall, and distractor-span retrieval rate. For answering, we report accuracy overall and by system configuration, retrieval setting, top-k, cognitive question type, and linguistic category. We also compute diagnostic tables that link retrieval success to answer correctness, allowing us to distinguish missing evidence, misleading retrieved evidence, and cases where the model receives the relevant evidence but still selects the wrong answer.

4.2 Control Results: Closed-Book and Oracle Answering

Before evaluating RAG systems, we use three controls to calibrate the task. The random baseline estimates chance performance, the closed-book setting measures how much can be answered from the question and options alone, and the full-text oracle estimates performance when retrieval failure is removed. Together, these controls separate chance, possible prior knowledge or memorization, and comprehension with complete gold context.

Table 1 shows that the evaluated LLMs can use Luxembourgish passage context effectively when retrieval is removed. Full-text oracle accuracy ranges from 0.805 to 0.853 across models, with Claude Opus 4.8 obtaining the highest oracle score. This justifies the RAG setting: the task is not blocked by a complete inability to process the Luxembourgish context. The closed-book results are also well above the random baseline, ranging from 0.581 to 0.686. This confirms the need for a closed-book control. Some items may be answerable from option plausibility or general knowledge, and possible exposure to published Luxembourgish literary material cannot be ruled out. However, the oracle improves over closed-book by 0.158–0.235 absolute accuracy across models, showing that the source passage remains essential. We therefore treat closed-book performance as a calibration and contamination check, while using the oracle as the upper reference for context-grounded answering. Table 2 shows that the oracle setting preserves the diagnostic structure of LuxDiagRC. Factual retrieval questions are easiest, while evaluative, inferential, and interpretive questions are substantially harder. Thus, even when retrieval failure is removed, the benchmark distinguishes simple evidence lookup from deeper comprehension.

Control Input	Answering Model	Accuracy (95% CI)
Chance Baseline		
Random label	No LLM	0.208 [0.177, 0.241]
Closed-Book Setting (Question Only)		
Question only	DeepSeek v4-pro	0.581 [0.544, 0.620]
Question only	Claude Sonnet 4.6	0.595 [0.558, 0.634]
Question only	GPT-5.5	0.675 [0.639, 0.711]
Question only	Claude Opus 4.8	**0.686** [0.650, 0.722]
Oracle Setting (Gold Full Text)		
Gold full text	DeepSeek v4-pro	0.805 [0.773, 0.834]
Gold full text	Claude Sonnet 4.6	0.830 [0.800, 0.859]
Gold full text	GPT-5.5	0.833 [0.803, 0.861]
Gold full text	Claude Opus 4.8	**0.853** [0.825, 0.880]

Table 1: Control results for chance, closed-book, and full-text oracle answering. Best performance within each LLM-based control condition is shown in bold.

Cognitive Type	Oracle Accuracy
Retrieve: direct fact	**0.930**
Evaluative reasoning	0.789
Inferential reasoning	0.776
Interpretive comprehension	0.760

Table 2: Oracle accuracy by cognitive type for the representative DeepSeek v4-pro oracle run. The gap between factual retrieval and deeper comprehension shows that LuxDiagRC remains diagnostic even when retrieval is removed.

4.3 Retrieval Quality: Source Selection and Evidence Recall

We next evaluate retrieval independently of answer generation. This allows us to identify whether downstream errors are caused by the answering model or by the retrieval stage failing to supply the relevant evidence. We focus on the collection-wide setting, where the retriever must search across all source texts and solve both source-text selection and evidence selection. Based on the ablation

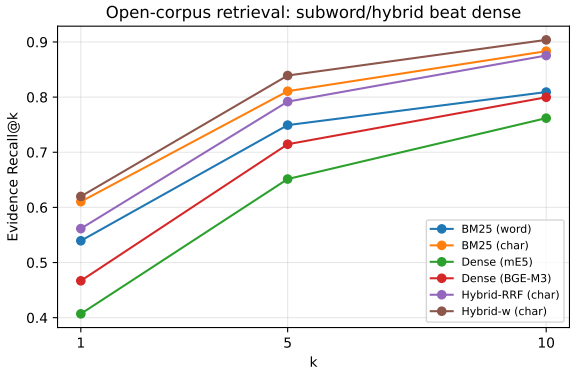


Figure 2: collection-wide evidence recall across retrieval methods.

Retriever	SrcR@10	EvidR@1	EvidR@10
BM25 (word)	0.912	0.539	0.809
BM25 (char)	0.950	0.610	0.883
Dense (mE5)	0.888	0.407	0.762
Dense (BGE-M3)	0.912	0.467	0.800
Hybrid-RRF (char)	0.948	0.562	0.875
Hybrid-w (char)	**0.956**	**0.620**	**0.904**

Table 3: collection-wide retrieval performance with overlapping chunks and question+options queries. SrcR@10 measures whether the gold source text is retrieved in the top 10; EvidR@k measures whether the annotated critical evidence span is covered by the top-k retrieved chunks. Bold marks the best score in each column.

results[2], we used overlapping chunks and question+options queries as the main retrieval configuration. Figure 2 compares evidence recall across retrieval depths. Character-level lexical retrieval and weighted hybrid fusion outperform dense-only retrieval, showing the importance of subword lexical matching for Luxembourgish evidence selection.

Table 3 shows that character-level lexical matching is particularly important for Luxembourgish retrieval. BM25 with character 3–5-grams improves over word-level BM25, increasing Evidence Recall@10 from 0.809 to 0.883 and Source Recall@10 from 0.912 to 0.950. This suggests that subword lexical overlap helps bridge spelling and word-form variation between questions and source texts. Dense retrieval alone performs worse than the strongest lexical methods in the collection-wide setting. BGE-M3 improves over multilingual E5, but both dense retrievers remain below character BM25, especially at rank 1. For example, Dense

[2]The complete ablation results are available in our GitHub repository

Retriever	TR-Acc	CW-Acc.
BM25 (word)	0.814	0.766
BM25 (char)	0.816	0.795
Dense (BGE-M3)	**0.833**	0.777
Hybrid-RRF (char)	0.813	0.784
Hybrid-w (char)	0.830	**0.798**

Table 4: Downstream RAG answer accuracy at $k = 5$ for Claude Opus 4.8 under different retrieval methods. Bold marks the best score in each column. TR-Acc and CW-Acc denote text-restricted RAG and collection-wide RAG accuracy respectively.

BGE-M3 reaches Evidence Recall@1 of 0.467, compared with 0.610 for character BM25. The best retrieval result is obtained by weighted hybrid fusion over character BM25 and dense retrieval. This method reaches 0.956 Source Recall@10 and 0.904 Evidence Recall@10, outperforming both lexical-only and dense-only retrievers. Figure 2 shows that this advantage is consistent across retrieval depths, not only at a single value of k. The gap between source-text recall and evidence recall is also important. Even when the correct source text is retrieved, the retriever may still fail to rank the exact critical evidence span highly enough for the answering model to use it. This motivates the next section, where we evaluate whether improved evidence retrieval translates into higher RAG answer accuracy.

Table 4 shows how retrieval method affects final RAG answer accuracy. The text-restricted setting isolates evidence selection within the gold source text, while the collection-wide setting also requires source-text selection. Dense BGE-M3 performs best in the text-restricted setting, suggesting that dense retrieval can be effective once the correct source text is known. In contrast, weighted hybrid retrieval performs best in the collection-wide setting, supporting the main finding that full Luxembourgish RAG benefits from combining subword lexical matching with dense retrieval.

Why character-level retrieval helps. Figure 3 illustrates the spelling and word-form variation between question terms and their corresponding evidence spans in the source texts. This analysis helps explain why character-level BM25 performs strongly in our retrieval experiments. In many cases, the question and the gold evidence express related lexical content but do not use exactly the same surface form. Word-level matching can miss such cases because it depends on full-token

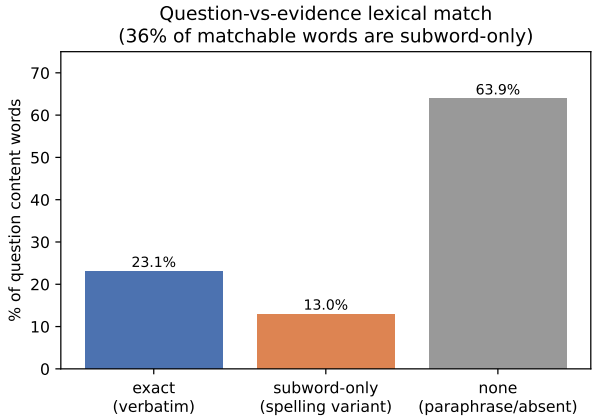


Figure 3: Spelling and word-form variation between question terms and gold evidence spans.

Retrieved Evidence	N	Distr. rate	Accuracy
Distractor only	124	0.161	0.661
Critical + distractor	1848	0.076	0.808
Critical only	256	**0.023**	**0.809**
Neither	280	0.175	0.500

Table 5: Retrieval-trap results for the best collection-wide retriever at $k = 5$, pooled over the four answering models. "N" gives the number of model-question examples in each group; "Distr. rate" is the proportion where the model selected the annotated distractor option.

overlap, whereas character n-gram retrieval can still match shared subword fragments. This makes character-level retrieval more robust to Luxembourgish spelling variation, inflectional differences, and informal question wording.

4.4 Retrieval Traps and Distractor Evidence

Beyond measuring whether the critical evidence is retrieved, we also examine whether the retrieved context contains annotated distractor evidence. Table 5 groups collection-wide RAG examples according to whether the top-k retrieved context contains the critical span, the distractor span, both, or neither. This analysis is pooled over the four answering models for the best collection-wide retriever at $k = 5$. The results show that the retrieved context can be incomplete or misleading. Accuracy is highest when the critical span is retrieved without the distractor span, and the distractor-choice rate is lowest in this condition. In contrast, accuracy is lowest when neither span is retrieved. The distractor-only condition is especially diagnostic: the model receives context associated with an incorrect option while missing the critical evidence needed for the correct answer. This shows why distractor-span retrieval is useful in addition to aggregate evidence recall.

Model	Closed	TR@10	CW@10	Oracle
Claude Opus 4.8	0.686	**0.831**	0.812	0.853
GPT-5.5	0.675	**0.827**	0.814	0.833
Claude Sonnet 4.6	0.595	**0.795**	0.787	0.830
DeepSeek v4-pro	0.581	**0.795**	0.755	0.805

Table 6: Closed-book, RAG@10, and oracle answering accuracy. TR denotes text-restricted retrieval and CW denotes collection-wide retrieval. Bold marks the stronger RAG setting for each model.

4.5 Answer Accuracy Across Settings and Context Size

We now evaluate whether retrieved context improves downstream answering. We use the main retrieval configuration from Section 4.3: weighted hybrid fusion over character BM25 and dense retrieval, with overlapping chunks and question+options queries. We compare text-restricted and collection-wide RAG to measure the effect of source-text selection on answer accuracy. Figure 4 shows that answer accuracy generally increases as more retrieved chunks are provided, with the largest gains after $k = 1$. Text-restricted RAG remains stronger than collection-wide RAG, indicating that source-text selection still affects downstream answering.

Table 6 shows that RAG substantially improves over closed-book answering for all models. For example, Claude Sonnet 4.6 improves from 0.595 in the closed-book setting to 0.795 with text-restricted RAG, while DeepSeek v4-pro improves from 0.581 to 0.795. These gains show that retrieved Luxembourgish context provides information not available from the question and answer options alone. Text-restricted RAG consistently outperforms collection-wide RAG at $k = 10$, indicating that source-text selection remains an important bottleneck. The remaining gap between RAG and the full-text oracle also shows that retrieved chunks do not fully replace access to the complete gold passage.

4.6 Qualitative Human Analysis

To complement the quantitative results, we conduct a targeted human analysis of selected collection-wide RAG failures. The goal is to inspect whether the retrieved chunks provide sufficient evidence for the target question and whether irrelevant or

Query	Retrieved evidence summary	Answer	Observation	Diagnostic interpretation
Text 13 (*Poker*), Interpretive. "Beschreiw d'Relatioun tëscht Marc an Karin." *Describe the relationship between Marc and Karin.*	Two Text 13 chunks are retrieved, but not the critical span; one Text 6 chunk is also retrieved, containing the longing passage "Ech denken nach vill un de Mike."	All models select distractor B; gold is C.	Retrieval mixes romance-themed chunks from **Text 6** with chunks from Text 13. The overlap is thematic and stylistic, but not evidential.	The Text 6 chunk is retrieved due to stylistic and thematic similarity rather than evidence relevance. It crowds out the critical span, and every model selects the distractor.
Text 5 (*Blues*), Interpretive. "Résumeier waat den Haptcharakter seu gaeren un senger Fra huet?" *Summarize what the main character loves about his wife.*	Chunks from **Text 6** and **Text 13** are retrieved; the critical Text 5 span is absent from the retrieved context.	All models select distractor D; gold is B.	Text 5 is the most abstract and philosophical text in the corpus. For this question, none of the top-5 chunks comes from Text 5.	With no source-text evidence retrieved, all the models lack the necessary basis for the answer and fail. This illustrates Text 5's high rate of critical-span absence and cross-text retrieval.

Table 7: Representative examples supporting the human observations in the collection-wide setting with the best retriever at $k = 5$. Queries are shown in Luxembourgish with an English gloss. Across the corpus, Text 5 has the most questions with the critical span absent from the top-3 retrieved chunks (20 of 40) and the most cross-text retrievals (38 of 40); Text 13's cross-text retrievals are almost entirely from Text 6 (16 of 18).

misleading chunks help explain incorrect answer selection. This analysis is not intended as a full-scale human evaluation but as a qualitative check on the diagnostic patterns identified by the automatic metrics. This analysis further supports the main quantitative finding that text-restricted RAG is easier than collection-wide RAG. The evaluation was carried out by two annotators. The primary annotator conducted the initial analysis, while a second annotator independently reviewed the findings to verify their consistency and interpretation. Both annotators are native Luxembourgish speakers, ensuring familiarity with the language's orthographic variation, regional usage, and cultural context. In the context of the manual inspection, a code-based pipeline was developed to randomly sample 100 items from the resulting answer paradigms. These sampled items were then subjected to a detailed manual analysis. During this process, recurring patterns within the generated texts were identified, prompting a more in-depth examination of text-specific responses. For these selected cases, all 40 corresponding questions associated with each targeted text were analyzed. This multi-stage evaluation procedure was designed to maximize the thoroughness and reliability of the overall analysis presented in this study.

Table 7 presents representative examples from the collection-wide setting, showing how retrieved evidence, generated answers, and human assessment help explain downstream errors. In the collection-wide setting, the chunk containing the critical span is often ranked among the top retrieved chunks in the $k = 5$ retrieval runs. However, a recurring pattern emerges for interpretive and inferential questions: the retriever sometimes ranks chunks from multiple source texts among the top results rather than prioritizing chunks from the target text. This introduces confounding context and can negatively affect answer selection. A noteworthy observation concerns the nature of the unrelated retrieved texts. For example, in Text 13, the retrieval system returns relevant chunks from the target text alongside chunks from Text 6. Although both texts involve romance-related content, the thematic overlap is superficial rather than evidential. This suggests that factors beyond direct semantic relevance may contribute to retrieval errors. In particular, similarities in writing style may increase the perceived relatedness between texts, thereby introducing confusion into the retrieval process. This pattern may be related to the pedagogical nature of the corpus. Many texts are written from the perspective of children or adolescents, resulting in overlap in themes, narrative structures, and writing styles. Text 5 provides a contrasting case: it differs from the rest of the corpus through its more abstract and philosophical style. The RAG system shows particularly weak retrieval performance on this text, with many questions lacking the critical span in the top retrieved chunks and with frequent retrievals from texts other than the

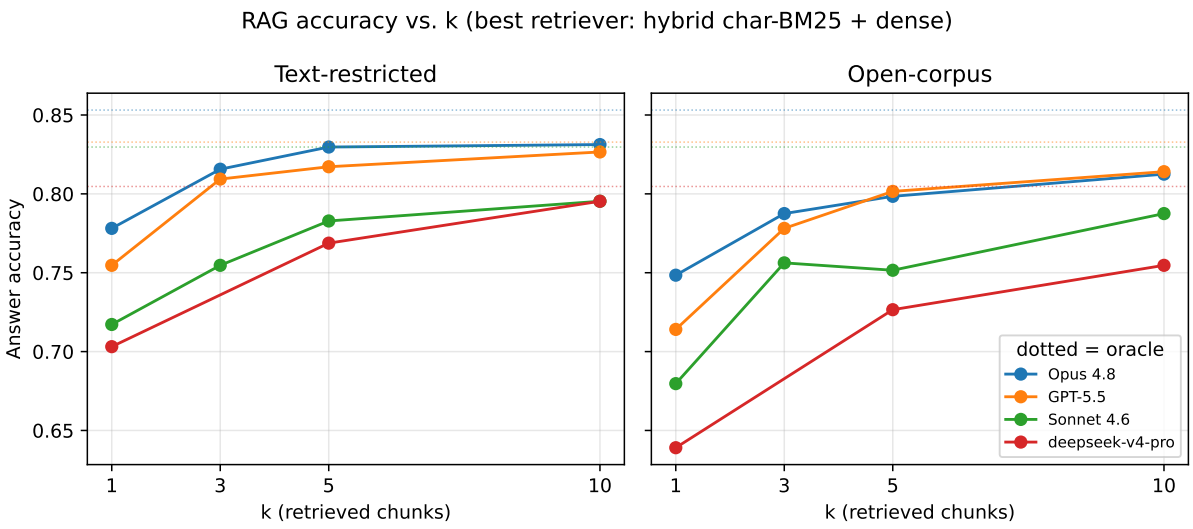


Figure 4: RAG answer accuracy as the number of retrieved chunks increases. Dotted horizontal lines indicate the corresponding full-text oracle accuracy for each model.

source. These observations suggest that stylistic and thematic characteristics may influence retrieval accuracy.

Overall, the qualitative analysis helps explain why text-restricted RAG is more reliable in this benchmark. By restricting retrieval to the gold source text, the system avoids cross-text confusions and focuses on evidence selection within a concrete passage.

5 Conclusion

This paper presented LuxDiag-RAG, a diagnostic evaluation framework for Luxembourgish retrieval-augmented reading comprehension. Rather than relying only on end-to-end answer accuracy, the framework separates several sources of RAG behavior: closed-book answerability, comprehension with the full gold text, evidence selection within the correct source text, and collection-wide document-plus-evidence retrieval. This makes it possible to analyze whether errors are associated with missing evidence, misleading distractor evidence, or answer selection despite useful context. The results suggest that current LLMs can make effective use of Luxembourgish context when the full passage is available, with oracle accuracy around 0.81–0.85. At the same time, the gap between text-restricted and collection-wide RAG indicates that retrieval remains an important bottleneck in this benchmark. The retrieval experiments further show that character-level lexical retrieval and weighted hybrid retrieval are strong baselines for this setting, outperforming the evaluated off-the-shelf multilingual dense retrievers in collection-wide evidence selection. This suggests that spelling variation and word-form differences between questions and source texts deserve more attention in Luxembourgish retrieval. The retrieval-trap analysis also shows the value of measuring distractor-span retrieval in addition to evidence recall, since retrieved context can be incomplete or misleading. Overall, LuxDiag-RAG offers a controlled and reproducible protocol for examining retrieval and comprehension failures in Luxembourgish RAG, and may provide a useful template for evaluating RAG in other low-resource languages.

6 Limitations

This study is intentionally controlled: LuxDiagRC contains 640 questions over 16 literary-style Luxembourgish texts, and retrieval is limited to this source collection rather than open-web documents. The results, therefore, should be interpreted as a diagnostic evaluation of current readily available RAG components, off-the-shelf multilingual retrievers and four instruction-tuned LLMs, rather than as an upper bound for Luxembourgish RAG. Future work should apply the protocol to broader Luxembourgish corpora, develop Luxembourgish-aware retrievers and rerankers, and use the diagnostic signals to build adaptive RAG systems that detect missing or misleading evidence.

References

Jumana Alsubhi, Mohammad D. Alahmadi, Ahmed Alhusayni, Ibrahim Aldailami, Israa Hamdine, Ahmad Shabana, Yazeed Iskandar, and Suhayb Khayyat. 2025. Optimizing rag pipelines for arabic: A systematic analysis of core components. *ArXiv*, abs/2506.06339.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 758–759, New York, NY, USA. Association for Computing Machinery.

Samhaa R El-Beltagy and Mohamed A Abdallah. 2024. Exploring retrieval augmented generation in arabic. *Procedia Computer Science*, 244:296–307.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Christophe Friezas Gonçalves, Salima Lamsiyah, and Christoph Schommer. 2026. LuxDiagRC: A diagnostic reading comprehension corpus for Luxembourgish with linguistic and cognitive annotation layers. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 532–541, Rabat, Morocco. Association for Computational Linguistics.

Sara Bourbour Hosseinbeigi, Sina Asghari, Mohammad Ali Seif Kashani, Mohammad Hossein Shalchian, and Mohammad Amin Abbasi. 2025. Advancing retrieval-augmented generation for persian: Development of language models, comprehensive benchmarks, and best practices for optimization. *arXiv preprint arXiv:2501.04858*.

Hossein Hosseini, Mohammad Sobhan Zare, Amir Hossein Mohammadi, Arefeh Kazemi, Zahra Zojaji, and Mohammad Ali Nematbakhsh. 2024. Persianrag: A retrieval-augmented generation system for persian language. In *2024 15th International Conference on Information and Knowledge Technology (IKT)*, pages 272–278. IEEE.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020a. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Cedric Lothritz, Jordi Cabot, and Laura Bernardy. 2026. Testing low-resource language support in LLMs using language proficiency exams: the case of Luxembourgish. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 2453–2476, Rabat, Morocco. Association for Computational Linguistics.

Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Lisa Veiber, Tegawende Bissyande, Jacques Klein, Andrey Boytsov, Clément Lefebvre, and Anne Goujon. 2022. LuxemBERT: Simple and practical data augmentation in language model pre-training for Luxembourgish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5080–5089, Marseille, France. European Language Resources Association.

Keerthana Murugaraj, Salima Lamsiyah, and Martin Theobald. 2026. RAGVUE: A diagnostic view for explainable and automated evaluation of retrieval-augmented generation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 512–526, Rabat, Marocco. Association for Computational Linguistics.

Fred Philippy, Siwen Guo, Jacques Klein, and Tegawende Bissyande. 2025. LuxEmbedder: A cross-lingual approach to enhanced Luxembourgish sentence embeddings. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11369–11379, Abu Dhabi, UAE. Association for Computational Linguistics.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. Ares: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354.

Suhang Wu, Jialong Tang, Baosong Yang, Ante Wang, Kaidi Jia, Jiawei Yu, Junfeng Yao, and Jinsong Su. 2024. Not all languages are equal: Insights into multilingual retrieval-augmented generation. *ArXiv*, abs/2410.21970.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A multi-lingual benchmark for dense retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 127–137, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

Assessment of Human-in-the-Loop Multi-LLMs for Low-Resource Educational Data Expansion

Christophe Friezas Gonçalves[*†] **, Hedi Tebourbi**[†], **Christoph Schommer**, **Salima Lamsiyah**

Faculty of Science, Technology and Medicine (FSTM),
University of Luxembourg, Luxembourg

christophe.friezas@uni.lu, hedi.tebourbi.001@student.uni.lu,
christoph.schommer@uni.lu, salima.lamsiyah@uni.lu

[†] Equal contribution [*] Corresponding author

Abstract

Low-resource languages often lack the pedagogically grounded datasets needed for language learning and assessment. To address this gap, we assess the potential of large language models (LLMs) to synthetically expand educational datasets in a low-resource language setting. We evaluate LLM performance within a complex, multi-layer annotation framework. Our approach builds upon LuxDiagRC, a Luxembourgish reading comprehension dataset fully annotated by human experts, which serves both as a methodological foundation and as a gold-standard benchmark. The evaluation follows the two-layer design of LuxDiagRC. In the first layer, three LLMs annotate linguistic features in Luxembourgish source texts. In the second layer, the models generate multiple-choice reading comprehension questions grounded in these annotations. Our results indicate that a multi-LLM approach combined with simple human curation outperforms individual and combined model approaches. The framework produces diagnostically grounded reading comprehension questions that closely align with the structure and pedagogical objectives of the original dataset. Overall, the findings suggest that combining multiple LLMs with expert human oversight form a promising strategy for the scalable expansion of educational corpora in a low-resource language setting, while highlighting the continued necessity of human expertise in dataset quality and pedagogical validity.[1]

1 Introduction

Second language acquisition is a complex process shaped by multiple interacting factors, including learners initial abilities, instructional conditions, and the availability of appropriate learning and assessment resources (Duke and Cartwright, 2021; Sabatini et al., 2024). Reading research and diagnostic assessment studies have shown that effective comprehension support requires not only accurate evaluation of answers, but also pedagogically grounded materials that capture the linguistic and cognitive sources of learner difficulty (Duke and Cartwright, 2021; Hemati and Baghaei, 2020; Liao et al., 2024). When it comes to low-resource languages, the availability of such materials is particularly constrained.

For Luxembourgish, a language spoken by approximately 400,000 people, the label of low-resource language applies (Gilles, 2019). Despite this classification, the digital era has led to a gradual increase in available data, largely driven by the expansion of social media platforms. In addition, the scientific community, specifically Natural Language Processing (NLP), found an increased interest in the language due to its inherently multilingual nature and culture (Plum et al., 2025). Recent work has contributed important Luxembourgish NLP resources and models, including LuxemBERT, LuxBank, and new benchmarks for generation and proficiency evaluation (Lothritz et al., 2022; PLUM et al., 2024; Plum et al., 2025; Lothritz et al., 2026). Nevertheless, within the educational sector, resource scarcity remains especially pronounced, as language teaching requires highly specialized and pedagogically tailored materials.

Reading comprehension (RC) datasets have played a central role in both NLP and educational assessment, with benchmarks such as SQuAD and RACE supporting large-scale evaluation, and frameworks such as STARC highlighting the importance of structured, diagnostically meaningful question design (Rajpurkar et al., 2016; Lai et al., 2017; Berzak et al., 2020). For Luxembourgish, LuxDiagRC addressed this gap by introducing a diagnostically rich reading comprehension corpus annotated by human experts, combining linguistic

[1] GitHub Repository

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 172–180
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_019

annotations with cognitively grounded multiple-choice questions (Gonçalves et al., 2026). However, because such resources require substantial expert effort, their expansion remains difficult to scale.

Recent studies show that LLMs can support annotation, synthetic data generation, and educational content creation, although human oversight remains important for quality control, especially in low-resource settings (Nasution and Onan, 2024; Pangakis and Wolken, 2024; Yu et al., 2024; Wang et al., 2024; Dutulescu et al., 2025; Hu et al., 2025; Säuberli and Clematide, 2024). In this paper, we assess a human-in-the-loop pipeline for synthetical expansion of a RC dataset for second language acquisition. Our assessment builds on LuxDiagRC, which serves both as the methodological foundation and as the gold-standard benchmark for evaluation (Gonçalves et al., 2026). Following its two-layer design, three LLMs first generate linguistic annotations for Luxembourgish short stories and then generate multiple-choice questions based on these annotations. We evaluate both individual and merged model outputs to assess the value of a multi-LLM strategy under human supervision.

The results show that the pipeline performs best when combining the outputs of multiple models in the annotation step, while also generating meaningful and textually grounded multiple-choice questions. These findings support the potential of using LLMs to expand diagnostically rich datasets, while also highlighting limitations in low-resource and instruction-sensitive settings.

2 Related Work

This section reviews the literature relevant to our study, focusing on three areas: Luxembourgish NLP, reading comprehension datasets, and LLM-based annotation and data synthesis.

2.1 Luxembourgish in NLP

Luxembourgish presents a complex multilingual context for NLP research, shaped by strong influences from German and French (Gilles, 2019). The language is further characterized by substantial orthographic and linguistic variation, partly due to limited standardization and formal teaching (Lutgen et al., 2025). These characteristics, combined with data scarcity, make Luxembourgish a challenging setting for both model development and evaluation. Recent work has led to the development of Luxembourgish-specific resources and models, including LuxemBERT (Lothritz et al., 2022) and LuxBank (PLUM et al., 2024). Despite these advances, such models remain constrained by limited training data and often struggle to capture the full variability of the language. In contrast, recent studies show that large commercial LLMs can achieve competitive performance on Luxembourgish tasks, leveraging cross-lingual transfer from high-resource languages (Lothritz et al., 2026). This makes them a promising tool for supporting resource development in low-resource settings.

2.2 Reading Comprehension Datasets

Reading comprehension datasets play a central role in both NLP and educational assessment. Large-scale benchmarks such as SQuAD (Rajpurkar et al., 2016) and RACE (Lai et al., 2017) have driven progress in extractive and multiple-choice question answering. More recent datasets, such as SciDQA (Singh et al., 2024) and MedMCQA (Pal et al., 2022), extend this paradigm to specialized domains. However, most existing datasets are designed for model evaluation rather than for pedagogical or diagnostic purposes, and often lack explicit links between textual evidence and answer choices. The STARC framework addresses this limitation by structuring answer options around text spans and supporting fine-grained diagnostic analysis of comprehension (Berzak et al., 2020). Building on this idea, LuxDiagRC (Gonçalves et al., 2026) introduces a diagnostically rich Luxembourgish RC dataset with a two-layer annotation scheme linking linguistic features to cognitively grounded questions. While this design enables detailed analysis of learner understanding, the reliance on expert human annotation limits scalability, motivating the need for automated expansion methods.

2.3 Automatic Annotation and Data Synthesis

Automatic annotation has been widely studied in NLP, particularly for tasks such as classification and linguistic tagging (Nasution and Onan, 2024; Pangakis and Wolken, 2024; Yu et al., 2024). Recent work has explored the use of LLMs as annotators, showing that they can achieve competitive performance while significantly reducing annotation cost (Yu et al., 2024; Nasution and Onan, 2024). However, their outputs can be inconsistent or biased, especially for complex or low-resource tasks, requiring human verification (Rønningstad et al., 2024). Hybrid human–LLM approaches have

therefore been proposed, where LLM-generated labels are filtered or corrected by human annotators to ensure quality (Wang et al., 2024). In parallel, synthetic data generation using LLMs has shown promise for improving downstream models and reducing reliance on manually labeled data (Pangakis and Wolken, 2024). In educational contexts, LLMs have also been used for generating assessment items, including multiple-choice questions, although ensuring pedagogical validity and high-quality distractors remains a challenge (Dutulescu et al., 2025).

2.4 Positioning of Our Work.

This work builds on these directions by combining LLM-based annotation, synthetic data generation, and human-in-the-loop validation in a unified assessment paradigm. We adopt a heterogeneous multi-LLM approach, employing three LLMs as annotators alongside a human overseer for quality control. Unlike prior work that focuses primarily on high-resource languages or single-model settings, we investigate a multi-LLM strategy for low-resource educational data expansion. Furthermore, our work explicitly targets the educational sector, leveraging the LuxDiagRC dataset as both a foundation and a benchmark. This allows us to evaluate not only annotation quality, but also the ability of LLMs to generate pedagogically grounded assessment items in a challenging low-resource setting.

3 Methodology

This section presents the human-in-the-loop multi-LLM pipeline for the assessment of the synthetical expansion of LuxDiagRC. We retain its two-layer design and extend it through LLM-based generation under human supervision. As shown in Figure 1, each model independently processes both layers. In Layer 1, models generate linguistic annotations reviewed at both individual and merged levels. In Layer 2, models generate diagnostically grounded questions, which are validated by the human annotator.

3.1 LuxDiagRC Overview

LuxDiagRC (Gonçalves et al., 2026) consists of 640 multiple-choice questions across 16 Luxembourgish short stories, fully annotated by human experts. The dataset follows a two-layer design. In the first layer, source texts are annotated with linguistic features that capture phenomena relevant to Luxembourgish reading difficulty. In the second layer, multiple-choice reading comprehension questions are constructed and linked to critical and distractor spans following the STARC framework (Berzak et al., 2020). Together, these layers support fine-grained analysis of learner comprehension by connecting answer choices to underlying linguistic phenomena.

In this work, we use LuxDiagRC as the reference dataset for evaluating LLM-generated annotations and questions. The linguistic feature set used in Layer 1 is summarized in Table 1. To align with token-level annotation constraints, we exclude tags requiring sentence-level analysis (e.g., LEX-IDIOM-ZLS, SYN-EMBEDDED). Experiments are conducted on 7 of the 16 texts in LuxDiagRC. We chose to use only texts with a word count higher than 1000 to provide sufficient context for LLM processing and evaluative analysis.

3.2 Automatic Annotation Pipeline

Rather than recreating the original dataset, we use LLMs to reproduce its two-layer annotation process and assess how well synthetic and human supervised outputs align with the fully human-created benchmark. The human-in-the-loop evaluation entails a shallow revision of the resulting two layer annotations.

3.2.1 Models

The pipeline employs three LLMs in a heterogeneous multi-LLM setup: GPT (5.4)[2], Claude Opus (4.6)[3], and DeepSeek (3.2)[4]. These models are selected based on prior evidence of complementary strengths in Luxembourgish tasks (Lothritz et al., 2026). The multi-LLM design aims to mitigate individual model weaknesses and improve coverage through aggregation.

For Layer 1, models are evaluated under zero-shot, few-shot, and chain-of-thought prompting. Zero-shot and few-shot serve as the primary evaluation settings, while chain-of-thought is used for qualitative analysis of model reasoning. In contrast, Layer 2 relies on zero-shot prompting, as few-shot settings led to imitation of the few-shot examples provided to the models.

3.2.2 Linguistic Annotation (Layer 1)

Layer 1 replaces human annotation with LLM-

[2]https://openai.com/index/introducing-gpt-5-4/
[3]https://www.anthropic.com/news/claude-opus-4-6
[4]https://api-docs.deepseek.com/news/news25120

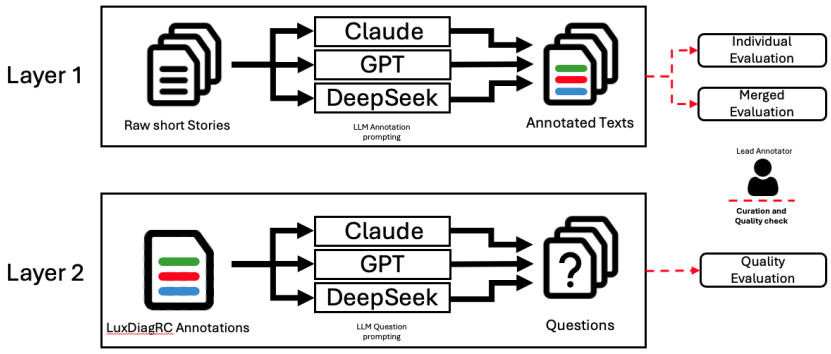


Figure 1: Overview of the automatic annotation pipeline and evaluation methodology. Layer 1 is evaluated in two steps: model outputs are assessed individually and after merging. All evaluations are supervised by a lead annotator for quality control.

based tagging of linguistic features. All three models receive identical prompts applied to the same seven raw Luxembourgish texts, including detailed instructions and linguistic feature definitions (Table 1). Each text is processed independently by each model, resulting in 21 prompt executions per prompting paradigm. To ensure comparability across models, strict formatting constraints are enforced, including token-level tagging, preservation of the original text, and exhaustive annotation of relevant features.

The prompting strategies differ only in the level of guidance: zero-shot uses base instructions, few-shot includes annotated examples, and chain-of-thought adds brief reasoning for ambiguous cases. This design allows us to assess both performance and the interpretability of model outputs.

3.2.3 Question Generation (Layer 2)

Layer 2 focuses on generating diagnostically grounded multiple-choice reading comprehension questions using the three selected LLMs. The generation process is guided by the STARC framework (Berzak et al., 2020) and the PIRLS cognitive taxonomy (Sabatini et al., 2024), ensuring both structural alignment and pedagogical validity. Each model is prompted to generate ten questions per short story, following the predefined distribution of four cognitive types: retrieval, interpretation, inferential, and evaluative (Gonçalves et al., 2026).

Prompts enforce strict constraints on both question structure and span selection. Each question must include four answer options: a correct answer, a misunderstanding-based option, a distractor-based option, and a textually unsupported option. All options are derived from critical and distractor spans explicitly grounded in the source text, in accordance with the STARC paradigm.

To isolate the evaluation of question generation, models are provided with gold-standard annotations from LuxDiagRC as input, ensuring a stable and controlled basis for generation independent of Layer 1 variability.

Table 1: Linguistic feature annotation tags with short descriptors used in layer 1 of the LuxDiagRC dataset creation (Gonçalves et al., 2026).

Tag ID	Description
LEX-LOAN-FR	Word of clear French origin.
LEX-LOAN-DE	Word of clear German origin.
LEX-FALSE-FRIEND	Word prone to cross-lingual confusion.
LEX-IDIOM-ZLS	Idiomatic expression.
LEX-RARE-ZLS	Rare or low-frequency word.
SYN-VERB-FINAL	Verb-final word order.
SYN-VERB-SEP	Split verbal construction.
SYN-EMBEDDED	Embedded or nested relative clause.
MORPH-N-RULE	Instance of the *Eifeler Regel*.
DISC-COREF-AMBIG	Pronoun with ambiguous referent.
ORTHO-PHONO-DIVERGE	Word exhibiting divergence between written and spoken forms.

3.3 Multi-LLM Framework with Human-in-the-Loop Verification

In Layer 1, the LLMs generate individual annotation outputs in parallel, which are subsequently combined using two distinct merging strategies. The first is a simple aggregation approach, in which all annotations are merged via union over token-level tags, resulting in a comprehensive but poten-

tially noisy annotation set. The second approach incorporates human-in-the-loop curation. A human annotator reviews the aggregated output and removes incorrect tags to produce a shallow yet refined annotation. This strategy mitigates common annotation errors while substantially reducing the required annotation time, requiring approximately one day compared with the considerably more time-intensive process used to construct the human-annotated dataset (Gonçalves et al., 2026). In Layer 2, the human annotator assumes an evaluative role, assessing both the difficulty level of the generated questions and the coherence between selected spans and corresponding answer options.

4 Results & Discussion

This section presents and discusses the results of the two annotation layers, with the goal of evaluating both model performance and the effectiveness. Beyond reporting quantitative scores, we analyze how different prompting strategies and aggregation mechanisms influence the quality of linguistic annotations and question generation. The results highlight both the limitations of current LLMs in low-resource language settings and the potential of multi-LLM and human-in-the-loop approaches.

4.1 Evaluation Methods

The evaluation protocol is structured along two complementary dimensions for each layer: quantitative and qualitative analysis. Quantitative evaluation provides performance measures based on predefined metrics, while qualitative analysis enables a deeper examination of model behavior, focusing on error patterns, reasoning strategies, and alignment with task requirements. The evaluation covers 7 Luxembourgish short stories, each annotated by 3 models under 3 prompting paradigms, yielding 56 valid annotation runs per experimental condition and 504 annotation-category pairs scored against the gold standard (GPT zero-shot excluded). For Layer 2, 208 questions are generated across 3 models and 7 texts.

For Layer 1, we employ the Jaccard similarity (Jaccard, 1901) coefficient to measure the overlap between predicted and gold annotations. The metric is computed over sets of unique word forms, ensuring that repeated occurrences do not bias the evaluation. This formulation emphasizes correct identification of linguistic phenomena rather than frequency. Qualitative analysis examines representative annotation cases per category, chain-of-thought reasoning traces, and systematic tagging error patterns to reinforce the quantitative findings.

For Layer 2, the quantitative evaluation assesses three properties: (i) distinctiveness of answer options, (ii) grounding of critical and distractor spans in the source text via exact string matching, and (iii) adherence to the predefined PIRLS distribution of cognitive question types. In addition, inter-model Jaccard similarity is computed to assess whether different models focus on similar textual regions. Qualitative evaluation examines the semantic validity and pedagogical coherence of each item, including alignment between cognitive type, question formulation, and supporting spans.

4.2 Linguistic Annotation Results

The results demonstrate that all models show limited capability in capturing fine-grained linguistic phenomena in Luxembourgish, as reflected by consistently low Jaccard scores across most categories, as seen in Table 2. Notably, the zero-shot configuration of GPT fails entirely to follow the annotation format, producing structured JSON outputs instead of inline annotations. This indicates a lack of controllability under constrained output specifications. Thus achieving no comparable results.

Overall performance remains relatively similar across models, with Claude achieving slightly higher scores and GPT consistently underperforming. The highest performance is observed for LEX-LOAN-FR (0.43), while several categories, such as DISC-COREF-AMBIG, show near-zero overlap with the gold standard. These results suggest that current LLMs struggle to reliably capture complex linguistic structures in low-resource settings.

Lexical categories achieve comparatively higher performance than syntactic and morphological ones. This can be attributed to the presence of loanwords from high-resource languages such as German and French, which are well represented in LLM training data. In contrast, categories requiring deeper structural or contextual reasoning, such as false friends or morphological rules, remain challenging. This pattern indicates that model performance is strongly influenced by training data distribution and lexical familiarity.

The LEX-FALSE-FRIEND category yields particularly low scores, reflecting the difficulty of capturing subtle cross-linguistic semantic and phonetic distinctions. These phenomena require contextual

and multilingual reasoning beyond surface-level lexical matching, which current models fail to consistently exhibit.

Among syntactic features, SYN-VERB-FINAL achieves relatively stronger performance, suggesting that models can capture local and consistent grammatical patterns. However, SYN-VERB-SEP remains challenging, as models often identify only one component of separable constructions. This indicates partial structural understanding without full adherence to the annotation scheme.

Qualitative analysis further shows systematic overgeneralization in the MORPH-N-RULE category. Models tend to associate the rule with specific lexical items rather than correctly identifying its contextual application. This behavior extends to the ORTHO-PHONO-DIVERGE category, suggesting that models rely primarily on surface-level cues rather than rule-based generalization. Taken together, these findings highlight a fundamental limitation of current LLMs in modeling linguistically grounded transformations, which is particularly critical for educational applications requiring precise and explainable annotations.

4.2.1 Attention and Context Window

To assess the impact of context length on annotation quality, we introduce a chunking strategy in which texts are split into segments of approximately 200 words. This modification eliminates output truncation observed in the full-text condition, reducing unscored outputs to zero. As a result, the average Jaccard score increases from 0.095 to 0.113. Performance gains are most noticeable for syntactic categories, particularly SYN-VERB-FINAL (+0.074), indicating that local grammatical patterns are easier to capture within shorter contexts. Improvements are also observed for word-level features such as MORPH-N-RULE (+0.025) and LEX-FALSE-FRIEND (+0.027). However, performance decreases for LEX-LOAN-FR (-0.061), suggesting that certain lexical judgments require a broader discourse context. These results demonstrate that context management plays a critical role in LLM-based annotation tasks. While reducing context improves local pattern detection and output reliability, it may remove essential global information. This trade-off highlights the importance of adaptive context strategies in multi-LLM systems.

4.2.2 Merged Annotation

The merged condition is based on a simple aggregation of the annotations produced by the individual models and yields mixed results. As shown in Tables 2, merging can improve weaker categories by increasing annotation coverage across models. However, it also reduces performance in categories where individual models already performed better. This decline is mainly caused by over-tagging, since each model introduces its own false positives and their combination expands the total number of errors, which is penalized by the Jaccard similarity metric. These findings indicate that naive ensemble strategies are insufficient for structured annotation tasks.

4.2.3 Human Guided Merging

To address this limitation, we evaluate a human-in-the-loop filtering step that removes over-tagged annotations from the aggregated outputs. This approach preserves the recall benefits of model aggregation while restoring precision through manual validation. As shown in Table 2, this hybrid strategy substantially outperforms both individual models and simple aggregation.

These findings provide strong evidence for the effectiveness of human-in-the-loop verification in multi-LLM settings. Rather than replacing human expertise, multiple imperfect models can be leveraged to maximize coverage, while human validation ensures accuracy. This hybrid paradigm is particularly well suited for complex linguistic annotation tasks in low-resource educational contexts.

Table 3: Quantitative results for the Layer 2 question generation task.

Metric	GPT 5.4	Claude 4.6	DeepSeek 3.2
PIRLS Distribution	70/70	68/70	70/70
Distinct Answers	70/70	68/68	70/70
Critical Span Found	61/70	60/68	58/70
Distractor Span Found	62/70	63/68	58/70

4.3 Question Generation Results

In contrast to the annotation task, models demonstrate good performance in question generation (Table 3). All models satisfy the distinctiveness constraint, and adherence to the PIRLS distribution is nearly perfect. A noteworthy observation regarding the Claude results is reported in the table. Specifically, the Claude output exhibited parsing errors for two questions (Text 2 and Text 7), which per-

Table 2: Jaccard similarity scores for Layer 1 annotations under the full-text condition and 200-word chunking. Each model was evaluated in three prompting settings: zero-shot (ZS), few-shot (FS), and chain-of-thought (CoT). The merged result is based on a simple aggregation of individual tags (Aggr) and a human-guided elimination of over-tagging (Hum).

Tag ID	Full Text										Chunking										
	GPT 5.4			Claude 4.6			DeepSeek 3.2			Merged	GPT 5.4			Claude 4.6			DeepSeek 3.2			Merged	
	ZS	FS	CoT	ZS	FS	CoT	ZS	FS	CoT	Aggr	ZS	FS	CoT	ZS	FS	CoT	ZS	FS	CoT	Aggr	Hum
LEX-LOAN-FR	/	0.39	0.42	0.27	0.27	0.28	0.29	0.32	**0.43**	0.18	/	0.35	0.36	0.24	0.20	0.23	0.24	0.25	0.31	0.12	**0.74**
LEX-LOAN-DE	/	0.16	0.15	0.26	**0.27**	0.23	0.17	0.25	0.25	0.22	/	0.17	0.23	0.21	0.21	0.23	0.18	0.22	0.21	0.17	**0.56**
LEX-FALSE-FRIEND	/	0.01	0.02	0.01	0.00	0.01	0.00	0.01	0.02	**0.06**	/	0.05	0.04	0.02	0.05	0.03	0.01	0.07	0.03	0.13	**0.17**
LEX-RARE-ZLS	/	0.13	0.12	**0.15**	0.13	0.02	0.13	0.13	0.14	0.14	/	0.12	0.11	0.12	0.14	0.13	0.13	0.18	0.28	0.12	**0.46**
SYN-VERB-FINAL	/	0.04	0.13	0.21	0.20	0.19	0.08	0.04	0.12	**0.22**	/	0.19	0.18	0.23	0.21	0.21	0.20	0.23	0.15	0.24	**0.50**
SYN-VERB-SEP	/	0.04	0.05	0.09	**0.11**	0.07	0.01	0.00	0.00	0.07	/	0.10	0.08	0.09	0.11	0.09	0.04	0.08	0.02	0.07	**0.22**
MORPH-N-RULE	/	0.01	0.01	0.06	0.06	0.06	0.00	0.01	0.00	**0.08**	/	0.02	0.02	0.09	0.12	0.09	0.01	0.03	0.02	0.14	**0.27**
DISC-COREF-AMBIG	/	0.00	0.01	**0.02**	**0.02**	0.00	0.00	0.00	0.00	0.01	/	0.01	0.01	0.02	0.01	0.01	0.01	0.08	0.05	0.09	**0.19**
ORTHO-PHONO-DIVERGE	/	0.01	0.01	0.01	0.02	0.02	0.00	0.00	0.01	**0.04**	/	0.02	0.02	0.02	0.03	0.03	0.00	0.03	0.02	0.08	**0.11**

sisted despite multiple prompt adaptations. Thus, reducing the total from 70 to 68.

Over 80% of the generated questions correctly ground both critical and distractor spans in the source text. A key finding is the high diversity of span selection across models, with only 8 overlapping cases out of 208 questions. This indicates that independently operating models capture complementary aspects of the input, reinforcing the value of a multi-LLM approach for generating diverse and pedagogically rich question sets. Qualitative analysis shows that models successfully construct structurally valid questions aligned with their intended cognitive type. However, limitations emerge for evaluative and inferential questions, where complexity often exceeds the intended learner level. This suggests that controlling cognitive difficulty remains an open challenge. Overall, these results demonstrate that LLMs are effective for structured question generation. Their outputs benefit from multi-LLM diversity. They do require additional control mechanisms to ensure pedagogical appropriateness.

5 Conclusion

This paper examined a the use of LLMs for linguistic annotation and question generation in Luxembourgish reading materials. The results show that the models performed better on question generation than on fine-grained linguistic annotation. For Layer 1, performance was generally low, especially for categories requiring structural or contextual understanding. Chunking improved annotation stability, while simple merging increased false positives. Human-guided filtering improved the merged results. For Layer 2, the models generated well-formed questions with strong span grounding and high adherence to the required distribution. The low overlap in selected spans also suggests complementary behavior across models. However, some evaluative and inferential questions remained too difficult or ambiguous. Overall, the findings suggest that current LLMs are more reliable for question generation than for detailed linguistic annotation in this setting, and that combining multiple LLMs is most useful when supported by human verification.

Limitations & Future Work Several limitations should be noted. First, the study uses general-purpose LLMs that are not specifically optimized for Luxembourgish. Low performance was therefore expected, especially for fine-grained linguistic phenomena. The focus of the study is not model specialization, but the effect of a unified curation strategy within a multi-LLM framework. Second, the human-in-the-loop component is limited to a simplified elimination-based curation process. Although this improves scores with limited additional effort, it does not allow full correction or refinement of the aggregated outputs. In addition, using a single annotator may introduce individual bias. Thirdly, the study evaluates Layer 2 under an optimal setting rather than in an end-to-end configuration, as the performance observed in Layer 1 was insufficient to support a reliable transition to Layer 2. Future work will evaluate Luxembourgish-specific models and involve multiple annotators to compare alternative human-in-the-loop strategies and better balance annotation quality with time efficiency.

References

Yevgeni Berzak, Jonathan Malmaud, and Roger Levy. 2020. STARC: Structured annotations for reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Lin-*

guistics, pages 5726–5735, Online. Association for Computational Linguistics.

Nell K. Duke and Kelly B. Cartwright. 2021. The science of reading progresses: Communicating advances beyond the simple view of reading. *Reading Research Quarterly*, 56(S1):S25–S44.

Andreea Dutulescu, Stefan Ruseti, Denis Iorga, Mihai Dascalu, and Danielle S. McNamara. 2025. Ymcq: Reasoning-enhanced mcq generation. In *Artificial Intelligence in Education*, pages 308–315, Cham. Springer Nature Switzerland.

Peter Gilles. 2019. *39. Komplexe Überdachung II: Luxemburg. Die Genese Einer Neuen Nationalsprache*, pages 1039–1060. De Gruyter Mouton, Berlin, Boston.

Christophe Friezas Gonçalves, Salima Lamsiyah, and Christoph Schommer. 2026. LuxDiagRC: A diagnostic reading comprehension corpus for Luxembourgish with linguistic and cognitive annotation layers. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 532–541, Rabat, Morocco. Association for Computational Linguistics.

Seyed Jamal Hemati and Purya Baghaei. 2020. A cognitive diagnostic modeling analysis of the english reading comprehension section of the iranian national university entrance examination. *International Journal of Language Testing*, 10(1):11–32.

Bihao Hu, Jiayi Zhu, Yiying Pei, and Xiaoqing Gu. 2025. Exploring the potential of llm to enhance teaching plans through teaching simulation. *npj Science of Learning*, 10(1):7.

Paul Jaccard. 1901. Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bull Soc Vaudoise Sci Nat*, 37:241–272.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Manqian Liao, Hong Jiao, and Qiwei He. 2024. Explanatory cognitive diagnosis models incorporating item features. *Journal of Intelligence*, 12(3):32.

Cedric Lothritz, Jordi Cabot, and Laura Bernardy. 2026. Testing low-resource language support in LLMs using language proficiency exams: the case of Luxembourgish. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 2453–2476, Rabat, Morocco. Association for Computational Linguistics.

Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Lisa Veiber, Tegawende Bissyande, Jacques Klein, Andrey Boytsov, Clément Lefebvre, and Anne Goujon. 2022. LuxemBERT: Simple and practical data augmentation in language model pre-training for Luxembourgish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5080–5089, Marseille, France. European Language Resources Association.

Anne-Marie Lutgen, Alistair Plum, Christoph Purschke, and Barbara Plank. 2025. Neural text normalization for Luxembourgish using real-life variation data. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 115–127, Abu Dhabi, UAE. Association for Computational Linguistics.

Arbi Haza Nasution and Aytuğ Onan. 2024. Chatgpt label: Comparing the quality of human-generated and llm-generated annotations in low-resource language nlp tasks. *IEEE Access*, 12:71876–71900.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Nicholas Pangakis and Sam Wolken. 2024. Knowledge distillation in automated annotation: Supervised text classification with LLM-generated training labels. In *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS 2024)*, pages 113–131, Mexico City, Mexico. Association for Computational Linguistics.

Alistair PLUM, Caroline DÖHMER, Emilia MILANO, Anne-Marie LUTGEN, and Christoph PURSCHKE. 2024. Luxbank: The first universal dependency treebank for luxembourgish. Association for Computational Linguistics. Accepted at 22nd Workshop on Treebanks and Linguistic Theories (TLT 2024).

Alistair Plum, Tharindu Ranasinghe, and Christoph Purschke. 2025. Text generation models for Luxembourgish with limited data: A balanced multilingual strategy. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 93–104, Abu Dhabi, UAE. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Egil Rønningstad, Erik Velldal, and Lilja Øvrelid. 2024. A GPT among annotators: LLM-based entity-level sentiment annotation. In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 133–139, St. Julians, Malta. Association for Computational Linguistics.

John Sabatini, Ann Kennedy, Erin Wry, and Matthias von Davier. 2024. PIRLS 2026 reading assessment framework. Technical report.

Andreas Säuberli and Simon Clematide. 2024. Automatic generation and evaluation of reading comprehension test items with large language models. In *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI) @ LREC-COLING 2024*, pages 22–37, Torino, Italia. ELRA and ICCL.

Shruti Singh, Nandan Sarkar, and Arman Cohan. 2024. SciDQA: A deep reading comprehension dataset over scientific papers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20908–20923, Miami, Florida, USA. Association for Computational Linguistics.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Danni Yu, Luyang Li, Hang Su, and Matteo Fuoli. 2024. Assessing the potential of llm-assisted annotation for corpus-based pragmatics and discourse analysis. *International Journal of Corpus Linguistics*, 29(4):534–561.

Human-Centred Approaches in Low-Resource Languages for Educational Applications and Language Learning: A Systematic Literature Review

Noura El Moussa
IRC for Intelligent Manufacturing and Robotics,
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia
research@elmoussa.ch

Abstract

Natural language processing increasingly targets education and human-centred design, yet these advances rarely reach low-resource languages, whose speakers remain underserved on both fronts. Little is known about how often low-resource NLP, human-centred AI, and education genuinely converge rather than being pursued separately. This systematic review asks whether educational NLP tasks extend to low-resource languages, whether progress is matched by fairness, inclusion, accessibility, and human-centred design, and how both converge in practice. We screened 24,795 main-conference papers from five core NLP venues, using keyword matching and a weighted scoring scheme to identify 196 papers with relevant signals. Grammatical error correction has progressed furthest for low-resource languages, while fairness- and accessibility-focused work is mature but rarely paired with broad low-resource coverage. A combined case study confirms that reused multilingual infrastructure can extend support to under-resourced languages.

1 Introduction

Educational technology increasingly draws on natural language processing for error correction, essay scoring, readability assessment, and second-language acquisition, applications with direct bearing on how learners worldwide access language education. Alongside this, human-centred concerns, fairness, accessibility, inclusion, participatory design, have gained prominence as a distinct thread in NLP research. Yet these threads, and the further question of which languages benefit, have largely developed as separate literatures: existing surveys map general low-resource NLP or NLP-for-education in isolation, leaving unexamined how often low-resource support, human-centred design, and educational application genuinely co-occur. This matters, since technology that is capable but not human-centred, or human-centred yet confined to well-resourced languages, cannot deliver equitable education. Motivated by this blind spot, we position this review at the intersection of these concerns define, asking whether educational NLP has reached low-resource languages, whether progress is matched by human-centred design, and where the two already converge. Our contribution is a systematic mapping of where this intersection stands across the field's major venues.

2 Methodology

This methodology investigates approaches in low-resource languages that are either human-centred in their design or focused on education and language learning. Papers were triaged across a large multi-venue corpus to identify where these themes appear individually and, more rarely, in combination.

The paper corpus was scoped through a set of explicit inclusion criteria: five core NLP venues (ACL, EMNLP, NAACL, EACL, COLING), the last ten editions actually convened per venue within the target window of the past ten years (2016–2025), and the main conference track only (long papers, short papers, and main proceedings), with Findings, workshops, demos, tutorials, industry tracks, student research workshops, and satellite events excluded to preserve a homogeneous, peer-reviewed sampling frame.

Given the irregular publication frequency across venues, the selection criterion was shifted from a calendar-year to an edition-based basis over the selected window. NAACL does not convene in years when ACL meets in North America, yielding ten editions spanning 2012–2025; EACL, being biennial, contributes only four editions (2017–2024) within the target window; and COLING, likewise biennial, contributes six editions (2016–2025), one of which (2024) constituted a joint proceeding with

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 181–190
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_020

LREC.

The ACL Anthology was selected as the bulk data source for all five NLP venues, as it publishes complete bibliographic metadata in structured XML format through the public GitHub repository `acl-org/acl-anthology`. This repository enables a single, consistent extraction procedure across venues, in contrast to the heterogeneous formatting of individual per-conference proceedings pages.

Volumes were classified for inclusion based on a case-insensitive match of their `<booktitle>` field against a predefined set of terms: the strings "Long Papers", "Short Papers", "Main Conference", or "Technical Papers" designated inclusion, whereas "Findings", "Demo", "Tutorial", "SRW", "Industry", "Workshop", or the name of a satellite workshop designated exclusion. This classification procedure isolated the main conference track from findings volumes, workshops, demonstrations, tutorials, industry submissions, and student research contributions.

Application of this pipeline yielded a total of 24,795 main-conference papers across the five venues, as summarized in Table 1.

Venue	Editions	Papers
ACL	10 (2016–2025)	7,573
EMNLP	10 (2016–2025)	8,376
NAACL	10 (2012–2025)	3,556
EACL	4 (2017, 2021, 2023, 2024)	1,068
COLING	6 (2016–2025)	4,222
Total		**24,795**

Table 1: Number of accepted papers per venue over the last 10 years.

Three keyword lists were constructed to operationalize the three thematic dimensions of the review. Matching was applied to paper titles only, as abstracts were not consistently available across all data sources. Term selection and matching followed three design principles: keywords were defined as exact phrases or single words, without stemming or fuzzy matching; matching employed regular-expression word boundaries to prevent partial-word matches; and matching was case-insensitive for all terms. For each paper title in the corpus, a compiled regular expression corresponding to each keyword in Table 3 was evaluated against the title string. Matches were recorded per list, with the specific keyword(s) triggering a match stored in three corresponding fields: `LRN` (Low-Resource NLP), `HC` (Human-Centred), and `ELL` (Education & Language Learning), enabling downstream aggregation of match counts as well as post-hoc inspection of which individual terms drove each classification. The scoring formula was designed to reflect the differential specificity of the three thematic dimensions, following a gradual power-of-two weighting scheme applied across dimensions. `LRN` (Low-Resource NLP) was assigned the lowest weight on the grounds that its associated keyword vocabulary captures a comparatively broader range of contributions, encompassing general multilingual and cross-lingual NLP work not necessarily aligned with the review's human-centred or educational focus. Two weighting variants were subsequently constructed to bias the ranked selection toward either an education-centred or a human-centred emphasis:

$$\begin{aligned} \text{Score}_{\texttt{ELL}} &= 2^2 \times (\texttt{ELL}) + 2^1 \times (\texttt{HC}) + 2^0 \times (\texttt{LRN}) \\ \text{Score}_{\texttt{HC}} &= 2^1 \times (\texttt{ELL}) + 2^2 \times (\texttt{HC}) + 2^0 \times (\texttt{LRN}) \end{aligned} \tag{1}$$

Under both variants, a paper's simultaneous alignment with two or three dimensions was rewarded additively, thereby surfacing contributions situated at the broader intersection of the theme's dimensions. Application of the title-matching procedure across the three keyword lists (`LRN`, `HC`, `ELL`) yielded 1,901 papers matching exactly one list and 34 papers matching two lists, for a combined total of 1,935 papers exhibiting at least one thematic signal, hereafter treated as the pool of promising candidates for the review. From this pool, the scoring formulas introduced above were applied as a ranking filter: $\text{Score}_{\texttt{ELL}}$ (education-focused) and $\text{Score}_{\texttt{HC}}$ (human-centred-focused) were each used to independently rank the pool and retain its top 100 highest-scoring papers, yielding two parallel shortlists of 100 papers each. Combining the two shortlists (200 papers total) revealed substantial overlap: 4 papers ranked in the top 100 of both lists, reducing the combined unique pool to 196 papers. The large majority of papers in each shortlist are single-focus matches (91 education-only, 88 human-centred-only), with a smaller number additionally carrying a low-resource signal (5 and 8, respectively), and only 4 papers matching all three dimensions simultaneously. It should be noted that the matching/scoring procedure relies on directed term-matching against the low-resource languages list rather than abstract semantic match-

ing, and deliberately excludes specific language terms that are too broad to serve as unambiguous low-resource indicators (e.g., "Hindi," which spans both low-resource dialects). As a result, the reported low-resource overlap counts likely underestimate the true overlap with the other two dimensions.

Category	Papers
ELL + HC + LRN (overlap)	4
ELL only	91
ELL + LRN	5
HC only	88
HC + LRN	8
Total (unique)	**196**

Table 2: Repartition of the combined top-100 shortlists by matched dimension.

Table 3 has been moved to our GitHub repository [1]

3 Two Fronts of Progress: Task Coverage and Human-Centred Design in Low-Resource Educational NLP

Taken together, the evidence assembled here addresses the Global Research Question by examining whether NLP's growing focus on education and human-centred design has genuinely reached low-resource languages, or whether their speakers remain underserved on both fronts. It does so from two angles: whether core educational NLP tasks have been extended to low-resource languages, and whether this task coverage matches the level of attention such tasks already receive for high-resource languages.

3.1 Extending Core Educational NLP Tasks to Low-Resource Languages

Across the seven educational domains reviewed, NLP research has begun building a genuine foundation for serving low-resource languages in education, even as progress remains unevenly distributed. **First**, task coverage itself is uneven. Grammatical Error Correction (GEC) leads the way, with dedicated, successful work on Hindi, Icelandic, Arabic, Russian, Czech, Swedish, and code-switched learner text showing low-resource GEC is a solvable problem (Rothe et al., 2021). The other domains (Essay Scoring, Language Learning, Speech Disorders, Second Language Acquisition, Readability Assessment, and Vocabulary Learning) remain concentrated mainly on English and a few well-resourced languages (Chinese, Spanish, Russian), reflecting an opportunity for expansion: GEC's trajectory shows what is achievable once a domain turns its attention here.

Second, encouragingly, a shared trend already cuts across domains: methods that reduce dependence on scarce annotated data recur wherever researchers look. GEC combines cross-lingual transfer (Yamashita et al., 2020), multilingual pretraining (Li et al., 2025) (Luhtaru et al., 2024), and language-agnostic architectures replacing hand-built tagging schemes; essay scoring shows unsupervised heuristic-signal aggregation and zero-shot LLM pairwise comparison reaching supervised-level performance (He and Li, 2024); readability assessment pairs adaptive pretraining with feature-informed prompt tuning, with gains growing larger as data grows scarcer; and SLA and vocabulary learning contribute reinforcement-learning simplification needing no parallel corpus and cross-lingually aligned embeddings needing no retraining, together forming a transferable toolkit for scarcity (Naous et al., 2024).

Third, this convergence is complemented by a finding on large language models: specialized, linguistically informed architectures consistently match or beat generic LLMs across GEC, essay scoring, readability, and vocabulary learning, and reframing a task often matters more than raw model scale. LLM performance on lower-resource languages does vary, reaching near-human quality on Swedish despite minimal training exposure while drifting semantically on others, and reference-based metrics can understate true quality for both high-performing and low-resource output, so measured gaps may be narrower than metrics suggest.

Fourth, a complementary insight follows: "low-resource" is not one fixed condition. Chinese readability corpora and Spanish learner-difficulty data are scarce even though both languages are otherwise well-resourced, showing targeted data collection, not wholesale language-technology development, can close many gaps. Code-switched GEC and multilingual pronunciation work extend this insight, addressing the register learners actually produce, while speech disorders mark the clearest next opportunity, since disordered-speech corpora remain rare beyond English. This points to a

[1] https://anonymous.4open.science/r/Human-Centred-SLR-E2E5/README.md.

natural next step: uniting the "low-resource" and "education" halves of the field's efforts. GEC's low-resource studies and its pedagogically rich contributions have progressed on parallel tracks, and vocabulary learning illustrates the same opportunity: a validated translation method and a validated usefulness predictor already exist separately, so combining them is a concrete step.

Taken together, NLP already addresses low-resource languages in education through a rich, growing set of solutions: cross-lingual transfer (Junczys-Dowmunt et al., 2018) (Kochmar and Shutova, 2016), multilingual pretraining, synthetic data generation, language-agnostic architectures, and label-efficient or zero-shot techniques all measurably narrow the data-scarcity gap wherever applied, most convincingly within GEC and, task-specifically, readability assessment. The clearest opportunity ahead is integration: pairing these proven techniques with authentic learner data and classroom validation across the remaining domains. Yet technical capability alone cannot guarantee equitable service; a parallel line of research asks whether the ecosystem surrounding these tools, its data, evaluation, and design, treats the communities it serves fairly, inclusively, and accessibly.

3.2 Designing for the Underserved: Human-Centred Designs in Low-Resource NLP

In this context, NLP research increasingly treats fairness, accessibility, and human-centered design as instruments for extending language technology to communities historically excluded from mainstream infrastructure. Low-resource status need not describe an entire national language: it can also describe a marginalized dialect, a sign language, or an underrepresented register within an otherwise well-resourced language. Sign language recognition, dialectal voice-assistant debiasing, and socially representative language identification each center an underserved variety while foregrounding the communities served.

First, a consistent set of technical strategies recurs. Cross-lingual and cross-modal transfer extends benefits from better-resourced varieties to underserved ones: pretraining on Indian Sign Language measurably improves recognition of American, Chinese, and Argentinian Sign Language despite no typological relation between them. Targeted, socially aware augmentation narrows performance gaps without sacrificing quality elsewhere: dialect-specific augmentation reduces error gaps for Arabic and German dialect speakers while preserving standard-language accuracy, and translation-based augmentation with deliberate annotation-budget allocation reduces the largest cross-language gap by about 60% while improving average accuracy with far less data. Infrastructure-building complements these advances: expanding a typological knowledge base to over 8,000 languages yields measurable gains for indigenous languages, while diversity audits identify which languages remain underserved (Khan et al., 2025).

Second, beyond these technical parallels, the strongest studies deliberately involve the affected human population in design and evaluation rather than relying on aggregate accuracy alone. Image-description accessibility research validates automatic metrics directly against blind and low vision users, revealing a widely used metric fails once real context is introduced, and Easy-to-Read text generation for cognitive disabilities is evaluated against official guidelines and native-speaker raters rather than generic fluency scores. A bias-evaluation resource co-authored by teachers and students through a structured educational process strengthened participants' critical AI literacy and achieved higher course completion than regional averages. Consequently, these strategies already yield measurable societal impact: equity-oriented language identification provided 30% more usable observations for the lowest-Human-Development-Index countries in a real public-health-monitoring application (Jurgens et al., 2017), with substantial recall gains for Nigeria and India.

Third, the literature also points toward clear opportunities ahead: importing human-centered design methods, participatory design, stakeholder theory, needs-finding, into machine translation research offers a roadmap for mitigating quality disparities, and the success of cross-lingual transfer in sign language (Selvaraj et al., 2022) recognition suggests comparable approaches remain available for other underserved modalities. Collectively, NLP addresses these challenges by combining technical methods that transfer knowledge and reallocate resources toward underserved languages and dialects (Abboud and Oz, 2024), infrastructure that expands linguistic coverage, and participatory methods that involve affected communities directly in design and evaluation, though most studies still

address technical capability or human-centered fairness in isolation, with genuine convergence remaining comparatively rare (Zhou et al., 2025) (T.y.s.s. et al., 2024) (Chalkidis et al., 2022).

Consistent with this convergence, across GEC and CEFR-based readability/proficiency assessment, this case study traces the same pattern from two angles (Imperial et al., 2025).

Fourth, NLP extends low-resource-language support by repurposing existing multilingual infrastructure rather than building bespoke systems: a 202-language MT backbone and explanation-based retrieval both improve Estonian GEC, while a consolidated CEFR resource brings together 13 typologically diverse languages, including Estonian, Hindi, and Welsh.

Fifth, fine-tuned multilingual encoders consistently outperform few-shot LLM prompting, indicating fine-tuning, not prompting, is the more promising route for scarce-data languages. Thus, NLP addresses low-resource language-education challenges by adapting large multilingual models and consolidating fragmented resources, extending measurable benefit to underserved learners: a maturing technical toolkit, an emerging commitment to fairness and human-centred design (Cunha et al., 2024), and this case study combining both trace the field's trajectory toward NLP that serves low-resource-language learners not only effectively, but equitably.

4 Analysis

Read together, the evidence points to a single interpretation: NLP is not starting from zero in serving low-resource-language learners, yet this early progress remains unevenly distributed, leaving clear gaps still to be addressed. These gaps surface along three fronts: whether core educational NLP tasks themselves have been extended to low-resource languages, whether that progress has been matched by fairness and human-centred design, and how far the two actually converge in practice. It is these gaps, rather than the progress itself, that the remainder of this discussion turns to.

4.1 Extending Core Educational NLP Tasks to Low-Resource Languages: Progress, Gaps, and Unrealized Convergence

First, across Grammatical Error Correction (GEC), Essay Scoring, Language Learning, Speech Disorders, Second Language Acquisition (SLA), Readability Assessment, and Vocabulary Learning, the literature reveals a field moving unevenly, but discernibly, from English-centric methodology toward an emerging capacity to serve low-resource languages in education. GEC shows the clearest trajectory: dedicated work spans Hindi, Icelandic, Arabic and its dialects, Russian, Czech, Swedish, and code-switched learner text, indicating low-resource GEC has moved from an unsolved problem toward an active, if unevenly applied, toolkit of transferable solutions. The remaining domains stay confined mainly to English and a few well-resourced languages (Chinese, Spanish, Russian), a pattern so consistent across independently reviewed corpora that it constitutes one of this synthesis's most robust findings. This asymmetry is instructive: low-resource applicability is not an inherent property of a technique but a research priority some domains, GEC foremost, have pursued while others have not yet turned their attention there.

Second, a shared technological trend nonetheless threads through every domain: mechanisms designed to reduce dependence on scarce annotated data recur regardless of application. GEC combines cross-lingual transfer, multilingual pretraining, synthetic data calibrated to pivot-language richness, and language-agnostic architectures replacing hand-built tagging schemes. Essay Scoring parallels this with unsupervised heuristic-signal aggregation and zero-shot LLM pairwise comparison approaching supervised performance, source-free model merging across prompts, and meta-learning acquiring prompt-generalized knowledge without target-prompt data. Readability Assessment pairs adaptive pretraining and corpus filtering with feature-informed prompt tuning, whose gains grow larger as data becomes scarcer. SLA and Vocabulary Learning contribute reinforcement-learning simplification requiring no parallel corpus and cross-lingually aligned embeddings extending translation to unseen languages without retraining. Speech Disorders converges on frozen self-supervised speech representations and zero-shot LLM scoring, both needing little labeled disorder-specific data. This convergence indicates the field already holds a coherent, reusable repertoire for scarcity; what varies is whether a domain has deployed it against a genuinely low-resource language or has so far validated it only on English.

Third, another cross-domain consensus concerns large language model behavior. Across GEC,

Essay Scoring, Readability Assessment, and Vocabulary Learning, generic LLMs consistently underperform specialized, linguistically informed architectures: general-purpose models struggle with long Chinese documents that lightweight feature-tuned models handle easily, zero-shot absolute essay scoring lags far behind fine-tuned encoders while zero-shot pairwise comparison closes most of that gap, and a prompted commercial LLM performs near baseline on educational context-supportiveness prediction. This indicates that how a task is framed often matters more than raw model scale. LLM reliability across languages is similarly inconsistent: GPT-3.5 reaches near-human grammaticality on Swedish despite minimal training exposure, yet drifts semantically on several other tested languages in the same corpus, implying apparent multilingual competence reflects incidental pretraining exposure rather than systematic capability. This connects to a recurring evaluation caution spanning GEC and Essay Scoring: reference-based metrics repeatedly underestimate quality for both high-performing and low-resource output, so measured gaps may partly reflect evaluation artifacts rather than genuine model weaknesses.

Fourth, a further comparative insight concerns the meaning of "low-resource" itself, which Readability Assessment, Essay Scoring, and Vocabulary Learning independently reframe as task-specific rather than language-general: Chinese readability corpora and Spanish learner-difficulty datasets are described as scarce specifically for the educational task, even though both languages are otherwise NLP-mature. Code-switched GEC and multilingual pronunciation work extend this, showing scarcity can originate in the register learners produce rather than a language's overall status. Speech Disorders sits at the far end: annotated disordered-speech corpora remain rare beyond English, making corpus construction itself, rather than modeling technique, the domain's principal bottleneck.

Fifth, the field's remaining limitations are best read not as fixed deficiencies but as propositions it is well positioned to test next: (a) GEC's toolkit, cross-lingual transfer, multilingual pretraining, language-agnostic tagging, proposes itself as an extension to the pedagogically rich but still English-only work in its own literature, such as learner-validated explanation generation, since both capabilities exist separately but remain uncombined; (b) reliance on non-learner or narrowly sourced text within dedicated low-resource studies proposes pairing existing pipelines with authentic classroom data collection, following the template Project MOSLA offers for SLA; (c) specialized architectures outperforming generic LLMs proposes prioritizing compact, linguistically informed models over large closed systems in low-resource deployments, given the infrastructure studies report as a reproducibility barrier; (d) authors repeatedly name untested extensions rather than leaving gaps unacknowledged, from stated uncertainty about efficacy beyond English to citations of scarce non-English parsers, so the missing prerequisite work is already identified and could be prioritized next; (e) GEC and Essay Scoring findings propose that future assessment adopt multi-reference or ranking-aligned metrics by default, since reference-based scoring understates quality precisely where it matters most for low-resource evaluation; and (f) finally, Vocabulary Learning's separation between a validated translation method and a validated educational-usefulness predictor proposes the clearest unification opportunity here: combining aligned embeddings with educational-context prediction for one low-resource language.

Collectively, this literature demonstrates that NLP addresses the challenges of low-resource languages in education by supplying a transferable methodological repertoire, cross-lingual transfer, multilingual pretraining, synthetic data generation, language-agnostic architectures, and label-efficient or zero-shot techniques, that measurably narrows the data-scarcity gap wherever applied, most convincingly within GEC and, task-specifically, Readability Assessment. This repertoire remains only partially deployed elsewhere: it is rarely combined with authentic learner data or classroom validation, and in Essay Scoring, Language Learning, SLA, Speech Disorders, and parts of Vocabulary Learning it has not been tested on a low-resource language at all. NLP therefore does not yet fully resolve the challenge, but it has already produced the constituent tools, and the propositions above, to do so; the field's task ahead is integration rather than invention.

4.2 Human-Centred Design in Low-Resource NLP: Strengths, Gaps, and a Path Toward Convergence

First, across the literature's own critical appraisals, NLP research addressing fairness, accessibility, and

human-centered design for low-resource-language communities emerges as methodologically mature and increasingly self-aware. Studies reviewed here are repeatedly credited with rigorous experimental design, transparent reporting, and genuine engagement with real, often hard-to-reach populations: accessibility research directly recruits blind and low vision participants alongside sighted crowdworkers within ecologically valid Wikipedia contexts, and Easy-to-Read text generation is validated against official European accessibility guidelines and native-speaker raters rather than generic fluency scores. This consistent emphasis on grounding technical work in authentic human experience, rather than treating fairness or inclusion as an abstract metric, characterizes the strongest contributions across every sub-literature surveyed.

Second, several complementary technological strategies recur as genuine strengths. Cross-lingual and cross-modal transfer learning, typified by open-source sign-language recognition infrastructure spanning six languages and a 1,129-hour pretraining corpus, shows that techniques built for one underserved modality can be responsibly extended to others, complete with explicit ethical treatment of privacy and bias. Targeted, socially aware data strategies recur similarly: dialect-specific augmentation and deliberate annotation-budget reallocation pair realistic deployment framing with thorough ablations that rule out confounds, while equity is treated as a first-class, quantified evaluation criterion, yielding actionable resource-allocation guidance for future dataset builders. Large-scale, transparent resource construction is a further recurring strength, from an 830,000-occurrence gender-inclusive German corpus to a 205-language benchmark and a typological knowledge base expanded to over 8,000 languages, each combining rigorous data engineering with empirical validation.

Third, a shared, encouraging trend across the strongest studies is the pairing of technical contribution with authentic participatory methodology. A large-scale bias-evaluation resource co-governed by informed consent, data-governance training, and a low-connectivity alternative for under-resourced classrooms combines rigorous participatory design with genuine human-outcome measurement, while a fully human-curated, locally sourced Arabic dialect dataset, built through weekly quality-control meetings and cross-review by native annotators, demonstrates that community-driven construction and low-resource-language coverage can be pursued jointly. Where automatic mitigation techniques are tested against cross-lingual transfer, the literature finds genuine, if uneven, benefit, with gains reported for Hindi, Tamil, Marathi, and Gujarati, and multilingual adaptive fine-tuning yielding concrete accuracy gains for previously unseen African languages, indicating fairness-oriented techniques developed on high-resource languages are not confined to them.

Fourth, these strategies already translate into concrete societal benefit. Equity-oriented, socially representative dataset construction is credited with delivering measurably reduced disparity for developing-country populations in a real downstream case study, and community-based bias-benchmark construction is associated with higher completion and retention in an accompanying teacher-training course alongside significant gains in critical AI literacy. Even where full deployment or end-user validation remains outstanding, several studies are praised for candid acknowledgment of real-world stakes, cautioning against premature deployment rather than overstating readiness.

Fifth, the literature's own appraisals point toward clear, tractable next steps rather than open-ended uncertainty. Expanding recruitment of hard-to-reach populations beyond the small samples achieved so far, extending cross-lingual debiasing transfer to languages where it does not yet generalize evenly, and pairing large-scale low-resource-language coverage with native-annotator, community-governed construction all represent concrete, already-modeled paths forward. Closing residual typological data gaps and adding direct end-user validation to studies currently reliant on automatic metrics are improvements the field has already specified with precision, indicating a mature research agenda rather than a stalled one.

Taken together, the collective evidence shows that NLP addresses these challenges by combining methodologically rigorous technical innovation, cross-lingual transfer, targeted data augmentation, large-scale infrastructure, with an increasingly serious commitment to participatory, community-governed design, exemplified by co-created bias benchmarks, natively curated dialectal datasets, and accessibility tools validated directly with disabled users. While the deepest participatory engagement and the broadest low-resource-language

coverage have not yet been achieved by the same study every time, the field's own critical self-assessment shows each capability has been independently and rigorously proven, and the concrete exemplars already combining both point toward a clear trajectory: technology progressively more inclusive, equitable, accessible, and empowering for low-resource-language communities worldwide.

Despite these advances, technical capability and human-centred design have so far been discussed largely in parallel. A natural next step is examining a single, tightly scoped body of work for both dimensions at once.

Sixth, NLP shows strong potential for low-resource language education by reusing large multilingual infrastructure: an off-the-shelf multilingual MT backbone and a training-free explanation-retrieval method extend correction capability to under-resourced languages without bespoke engineering, validated through rigorous ablations and strong baselines. Large-scale, typologically diverse CEFR resources, spanning over 505,000 texts across many scripts and languages, further broaden coverage and explicitly confront data-sharing inequities. While direct human evaluation of outputs and consistent low-resource coverage remain limited, these foundations show NLP is steadily building scalable, reusable infrastructure that extends effective language-learning support toward underserved language communities.

Overall, the three discussions converge on one interpretation: a maturing technical toolkit already narrows the data-scarcity gap wherever applied, a parallel body of work extends this progress with fairness, inclusion, accessibility, and participatory design, and the case study combining both confirms these threads can be brought together in practice. What remains is not invention but integration: pairing proven technical solutions with authentic learner data, classroom validation, and community participation.

5 Limitations

The scope of this work leaves several areas open for discussion and improvement. Venue scope is currently centred on five major NLP venues, favoring depth and consistency over broad regional coverage; this likely underrepresents low-resource work published in regional venues, which often host research most directly tied to local low-resource-language initiatives, and extending screening regionally is a natural next step for widening this foundation. Relatedly, title-only matching relies on title-level keywords rather than automated semantic selection, leaving title-proximity and abstract-semantic matching as a promising extension worth exploring. Similarly, findings are reported as aggregated syntheses rather than per-paper analyses, given space constraints; this necessarily smooths over some finer distinctions, and disaggregating them further is a clear opportunity for follow-up work. More broadly, this remains an initial-stage evaluation of the low-resource, human-centred, and education intersection rather than an exhaustive account, and is intended to orient and motivate further study rather than close the question. On the classification side, the LRN keyword list is deliberately conservative, excluding broad terms too ambiguous as low-resource indicators, including those spanning both high- and low-resource dialects; this means reported overlap counts likely underestimate the true overlap, and dialect-specific indicators could help sharpen this further. Finally, the power-of-two scoring weights, including assigning LRN the lowest weight since its vocabulary spans broader contributions, represent just one plausible design; testing alternative weightings could be a useful direction for future work, potentially surfacing a different, complementary shortlist.

6 Conclusion

This review examined whether NLP for education has reached low-resource languages and whether progress is matched by fairness, inclusion, accessibility, and human-centred design. Screening a large multi-venue corpus showed these dimensions rarely co-occur: only four of 196 relevant papers combined all three, and title-based matching likely underestimates even this overlap. Grammatical error correction has advanced furthest for low-resource languages via a shared, transferable toolkit, while fairness- and accessibility-focused work is mature but seldom paired with broad low-resource coverage; a combined case study nonetheless shows reused multilingual infrastructure can serve under-resourced learners. This mapping offers the first systematic account of where technical and human-centred progress diverge, giving researchers and educators a basis for prioritizing integration over new development. Given reliance on title-level matching and limited human evaluation of outputs, future work should extend semantic screening and

pair validated tools with authentic classroom participation. Overall, equitable low-resource language education appears within reach, provided the field consolidates what it has already built.

References

Khadige Abboud and Gokmen Oz. 2024. Towards equitable natural language understanding systems for dialectal cohorts: Debiasing training data. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia. ELRA and ICCL.

Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Schwemer, and Anders Søgaard. 2022. FairLex: A multilingual benchmark for evaluating fairness in legal text processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4389–4406, Dublin, Ireland. Association for Computational Linguistics.

Rossana Cunha, Thiago Castro Ferreira, Adriana Pagano, and Fabio Alves. 2024. A persona-based corpus in the diabetes self-care domain - applying a human-centered approach to a low-resource context. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1353–1369, Torino, Italia. ELRA and ICCL.

Junyi He and Xia Li. 2024. Zero-shot cross-lingual automated essay scoring. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17819–17832, Torino, Italia. ELRA and ICCL.

Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Muñoz Sánchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R. Jablonkai, Ekaterina Kochmar, Robert Joshua Reynolds, Eugénio Ribeiro, Horacio Saggion, Elena Volodina, Sowmya Vajjala, Thomas François, Fernando Alva-Manchego, and Harish Tayyar Madabushi. 2025. UniversalCEFR: Enabling open multilingual research on language proficiency assessment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9703–9755, Suzhou, China. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Shubha Guha, and Kenneth Heafield. 2018. Approaching neural grammatical error correction as a low-resource machine translation task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 595–606, New Orleans, Louisiana. Association for Computational Linguistics.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 51–57, Vancouver, Canada. Association for Computational Linguistics.

Aditya Khan, Mason Shipton, David Anugraha, Kaiyao Duan, Phuong H. Hoang, Eric Khiu, A. Seza Doğruöz, and En-Shiun Annie Lee. 2025. URIEL+: Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base. In *Proceedings of the 31st International Conference on Computational Linguistics*, Abu Dhabi, UAE. Association for Computational Linguistics.

Ekaterina Kochmar and Ekaterina Shutova. 2016. Cross-lingual lexico-semantic transfer in language learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 974–983, Berlin, Germany. Association for Computational Linguistics.

Wei Li, Wen Luo, Guangyue Peng, and Houfeng Wang. 2025. Explanation based in-context demonstrations retrieval for multilingual grammatical error correction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4881–4897, Albuquerque, New Mexico. Association for Computational Linguistics.

Agnes Luhtaru, Elizaveta Korotkova, and Mark Fishel. 2024. No error left behind: Multilingual grammatical error correction with pre-trained translation models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1209–1222, St. Julian's, Malta. Association for Computational Linguistics.

Tarek Naous, Michael J Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. 2024. ReadMe++: Benchmarking multilingual language models for multi-domain readability assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12230–12266, Miami, Florida, USA. Association for Computational Linguistics.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A simple recipe for multilingual grammatical error correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 702–707, Online. Association for Computational Linguistics.

Prem Selvaraj, Gokul Nc, Pratyush Kumar, and Mitesh Khapra. 2022. OpenHands: Making sign language recognition accessible with pose-based pretrained models across languages. In *Proceedings of the 60th*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2114–2133, Dublin, Ireland. Association for Computational Linguistics.

Santosh T.y.s.s., Nina Baumgartner, Matthias Stürmer, Matthias Grabmair, and Joel Niklaus. 2024. Towards explainability and fairness in Swiss judgement prediction: Benchmarking on a multilingual dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16500–16513, Torino, Italia. ELRA and ICCL.

Ikumi Yamashita, Satoru Katsumata, Masahiro Kaneko, Aizhan Imankulova, and Mamoru Komachi. 2020. Cross-lingual transfer learning for grammatical error correction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4704–4715, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yue Zhou, Barbara Di Eugenio, and Lu Cheng. 2025. Unveiling performance challenges of large language models in low-resource healthcare: A demographic fairness perspective. In *Proceedings of the 31st International Conference on Computational Linguistics*, Abu Dhabi, UAE. Association for Computational Linguistics.

SimuLe-Lux: A Neuro-Symbolic Teaching System for Diagnosing L2 Reading Comprehension

Christophe Friezas Gonçalves and **Christoph Schommer** and **Salima Lamsiyah**
Department of Computer Science, Faculty of Science, Technology and Medicine
University of Luxembourg, Esch-sur-Alzette, Luxembourg
{christophe.friezas,christoph.schommer,salima.lamsiyah}@uni.lu

Abstract

Understanding why second-language learners make reading comprehension errors is essential for effective instruction and teaching. However, providing fine-grained diagnostic feedback requires teacher expertise and time. While recent work has explored large language models as student simulators, existing approaches lack grounding in linguistic and educational evidence. We introduce *SimuLe-Lux*, a neuro-symbolic framework for simulating learner profiles and supporting teacher-facing diagnostic training in Luxembourgish reading comprehension. The system is built on *LuxDiagRC* and utilizes a large language model to create and simulate learner responses across six CEFR proficiency levels, including weaker-state variants. These responses are aggregated into interpretable diagnostic profiles. A retrieval-augmented teacher-facing agent then uses this structured knowledge base to generate grounded explanations of learner behaviour and instructional guidance. Evaluation results show that the simulated learners exhibit CEFR-consistent error patterns and that the teacher-facing agent produces relevant and informative diagnostic feedback for structured pedagogical queries.

1 Introduction

Learning to read in a second language inevitably involves making mistakes. Such errors are not merely failures of performance but indicators of underlying linguistic and cognitive difficulties. When teachers can identify and interpret these errors, they can better target instruction and support language development (Mullis and Martin, 2019; Suzuki et al., 2025). However, diagnosing the sources of reading comprehension errors requires substantial expertise and individual attention, which is difficult to provide at scale in classroom settings and for low-resource languages.

At the same time, large language models (LLMs) are increasingly used by students as learning aids (Ghimire and Edwards, 2024; Liu et al., 2025), and their widespread interaction with learner-generated text has opened new possibilities for modeling student behaviour. Recent studies have shown that LLMs can be prompted to simulate student responses in test and exam scenarios (Martynova et al., 2025; Benedetto et al., 2024; Xu et al., 2024), suggesting their potential as tools for generating synthetic learner data. Teachers also report positive attitudes toward AI-based educational support systems (Pokrivcakova, 2019). However, prior work has identified important limitations, including unrealistic response patterns in open-ended settings (Martynova et al., 2025) and poor calibration of learner proficiency levels (Benedetto et al., 2024; Xu et al., 2024). These shortcomings make it difficult to use existing LLM-based simulations for fine-grained diagnostic training.

In this paper, we propose *SimuLe-Lux*, a neuro-symbolic system designed to support teacher-facing diagnostic reasoning in reading comprehension. The system combines a diagnostically annotated multiple-choice reading comprehension dataset in Luxembourgishœ with large language model–based learner simulation across six CEFR proficiency levels. Rather than producing arbitrary errors, *SimuLe-Lux* generates learner responses that are grounded in annotated linguistic features and cognitively challenging question types, enabling structured analysis of why learners commit mistakes.

SimuLe-Lux is intended as a training partner for educators. It allows teachers to explore simulated learner profiles, examine patterns of misunderstanding, and receive explanations that link learner behaviour to specific linguistic and cognitive triggers. This design shifts the role of AI in education from task automation toward support for teacher diagnos-

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 191–198
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_021

tic expertise and reflective practice (Pruski et al., 2025).

Through a series of targeted evaluations, we show that a carefully constructed diagnostically rich dataset can serve as the symbolic core of a neuro-symbolic system, enabling both realistic learner simulation and grounded teacher-facing explanations. Our work demonstrates how dataset design, learner modeling, and retrieval-augmented reasoning can be combined to support human-centered educational AI in a low-resource language setting.

Contributions. This work makes two main contributions. First, we present *SimuLe-Lux*, a teacher-facing system that simulates second-language learner behaviour using linguistically and cognitively grounded annotations. Second, we propose a neuro-symbolic diagnostic control mechanism that conditions simulated learner errors on annotated linguistic features and question types, enabling interpretable and controllable learner profiles.

2 Related Work

This section places our work in relation to three research areas: reading comprehension datasets, diagnostic modeling of learner weaknesses, and teacher-facing educational systems.

Reading Comprehension Datasets. Datasets are a cornerstone for the development and evaluation of NLP systems. Large-scale benchmarks such as SQuAD (Rajpurkar et al., 2016) and RACE (Lai et al., 2017) have demonstrated strong performance in multiple-choice question answering and generation (Dutulescu et al., 2025). Domain-specific datasets such as MedMCQA (Pal et al., 2022) and SciDQA (Singh et al., 2024) extend this paradigm to medical and scientific texts. However, despite their scale and diversity, most of these datasets frame reading comprehension as a binary task, scoring responses simply as correct or incorrect. As a result, they provide little insight into the underlying causes of learner errors, which is a critical limitation in educational settings where diagnosing misconceptions is central to language acquisition and instruction (Berzak et al., 2020).

Student Weaknesses Detection. The detection of weaknesses in reading comprehension is a multifaceted problem that traditionally requires substantial teacher expertise and extensive one-on-one instructional time (Mullis and Martin, 2019; Suzuki et al., 2025). Diagnostic models analyse patterns in student responses to identify learner profiles and categories of misunderstanding (Hemati and Baghaei, 2020; Suzuki et al., 2025). While these approaches demonstrate the feasibility of AI-based weakness detection, they typically rely on specialized models, handcrafted diagnostic rules, and fixed category sets, which limit their adaptability across learners and domains. In contrast, our framework introduces a data-driven diagnostic controller that models emerging learner patterns and profiles from question responses, enabling flexible and dynamically evolving representations using state-of-the-art large language models.

Teacher-Facing Tools. LLM-based educational tools span a wide range of applications, from simple translation systems to intelligent tutoring systems (ITS) (Pokrivcakova, 2019). Recent research on teacher-facing AI has largely focused on automating instructional tasks such as curriculum design, content generation, and grading (Zhang et al., 2025; Hu et al., 2025; Gonnermann-Müller et al., 2025). While this increases efficiency, it risks encouraging forms of work substitution similar to those observed among students who over-rely on generative AI (Liu et al., 2025). In contrast, emerging work advocates for AI systems that support teacher skill development and reflective practice, enabling lifelong learning and pedagogical growth (Pruski et al., 2025).

Positioning of Our Work. *SimuLe-Lux* combines a diagnostically annotated reading comprehension dataset, learner simulation, and a teacher-facing analysis agent within a single neuro-symbolic framework. Unlike rule-based diagnostic systems, it derives learner profiles from simulated responses grounded in linguistic annotations. Compared to standard reading comprehension datasets that support only correct–incorrect scoring, the use of *LuxDiagRC* (Gonçalves et al., 2026) enables analysis of the sources of learner errors. The system is designed to support teacher diagnostic reasoning rather than automate instructional tasks.

3 Methodology

This section presents the methodological design of *SimuLe-Lux* , a neuro-symbolic framework for simulating second-language learners and supporting teacher-facing diagnostic training in reading comprehension. As illustrated in Figure 1, the

Table 1: Example *Retrieve* question from LuxDiagRC showing how answer options are grounded in annotated text spans and linguistic feature tags.

Question	Wéi schwätzt den Marc wann en rosen ass? *How does Marc talk when he is angry?*
Span Type	**Span**
Critical span	. . . seet de <ORTHO-PHONO-DIVERGE> <MORPH-N-RULE> Marc zu mir, wann hien <ORTHO-PHONO-DIVERGE> <MORPH-N-RULE> besonnesch rosen <LEX-FALSE-FRIEND> ass <LEX-FALSE-FRIEND>, schwätzt hien <ORTHO-PHONO-DIVERGE> <MORPH-N-RULE> franséisch, fir seng Roserei besser an <LEX-FALSE-FRIEND> der <LEX-FALSE-FRIEND> Gidd ze halen, seet en <DISC-COREF-AMBIG>.
Distractor span	wéi <ORTHO-PHONO-DIVERGE> wier et, wa <ORTHO-PHONO-DIVERGE> <MORPH-N-RULE> mir géifen erwuesse <ORTHO-PHONO-DIVERGE> ginn <SYN-VERB-FINAL>?!
Answer Type	**Answer**
Correct	En schwätzt Franséisch *He speaks in French*
Misunderstanding	En schwätzt séier fir sech an der Gidd ze halen *He speaks fast to keep his temper in check*
Distractor	En schwätzt wéi en Erwuessenen *He speaks like an adult*
No support	En schwätzt Däitsch *He speaks in German*

framework integrates three tightly coupled components: (i) *LuxDiagRC*, a diagnostically annotated reading comprehension dataset (Gonçalves et al., 2026), (ii) a learner simulation module with neuro-symbolic diagnostic control, and (iii) a retrieval-augmented teacher-facing rationalization agent. Together, these components form a pipeline that transforms annotated texts into simulated learner behaviours and grounded pedagogical explanations.

3.1 LuxDiagRC Dataset

The *LuxDiagRC* dataset is constructed using a two-layer human annotation paradigm designed to link reading comprehension errors to underlying linguistic and cognitive difficulties. Each answer option in the dataset is explicitly grounded in annotated spans of the source text, enabling diagnostic interpretation of learner choices (Gonçalves et al., 2026). The dataset applies their proposed annotation pipeline to 16 texts used in primary and secondary language teaching (Gonçalves et al., 2026).

The dataset contains 640 multiple-choice questions across 16 texts (40 per text, 10 per annotator), with each text comprising 8 *Retrieve*, 12 *Interpretation*, 12 *Inferential*, and 8 *Evaluative* items. Each answer option is explicitly grounded in annotated text spans, enabling direct mapping from learner choices to underlying linguistic and cognitive difficulties, as presented in Table 1.

3.2 Simulated Students and Neuro-symbolic Diagnostic Control

The neuro-symbolic diagnostic controller requires stable and interpretable probability vectors to represent learner profiles. To obtain these, we employ a large language model (GPT-5, August 2025) to simulate student responses across the six proficiency levels defined by the Common European Framework of Reference for Languages (CEFR): A1 to C2.

The model is prompted in a zero-shot setting to answer all questions in the dataset. It is instructed to respond as a student taking a language test and to produce realistic, level-appropriate errors. For each proficiency level, the model answers the questions twice: once at the nominal level and once in a deliberately weaker state. This weaker-state variant enables the diagnostic controller to model smooth transitions between adjacent proficiency levels, avoiding hard boundaries and allowing both weaker and stronger learner behaviours to be represented within each level.

Questions are grouped by text, with 40 questions per prompt, ensuring that the model's errors arise from reading comprehension difficulty within a single text rather than from cross-text interference or memory effects. This design emphasizes the emergence of systematic error patterns linked to linguistic and cognitive challenges rather than to recall alone.

The resulting simulated responses are aggregated to compute probability distributions over linguistic features, question types, and answer options for each proficiency level, yielding a diagnostic probability vector for each learner profile. In addition to these level-specific vectors, level-text-specific distributions are also computed, providing additional layers of diagnostic and contextual information in the knowledge base.

3.3 Teacher-Facing Diagnostic Agent

The teacher-facing component is implemented as a retrieval-augmented generation (RAG) system (Lewis et al., 2020) built on GPT-5 (August 2025), which integrates the *LuxDiagRC* dataset with the neuro-symbolic diagnostic controller to produce grounded explanations of learner behaviour. Through retrieval over the full knowledge base, the agent can link student responses to underlying linguistic features, cognitive question types, and proficiency-level vulnerability profiles. The model operates in a zero-shot setting and is instructed to assist teachers in analysing student errors, identifying the linguistic and cognitive sources of misunderstanding, and reflecting on instructional strategies for Luxembourgish second-language learners. By combining symbolic diagnostic representations with generative reasoning,

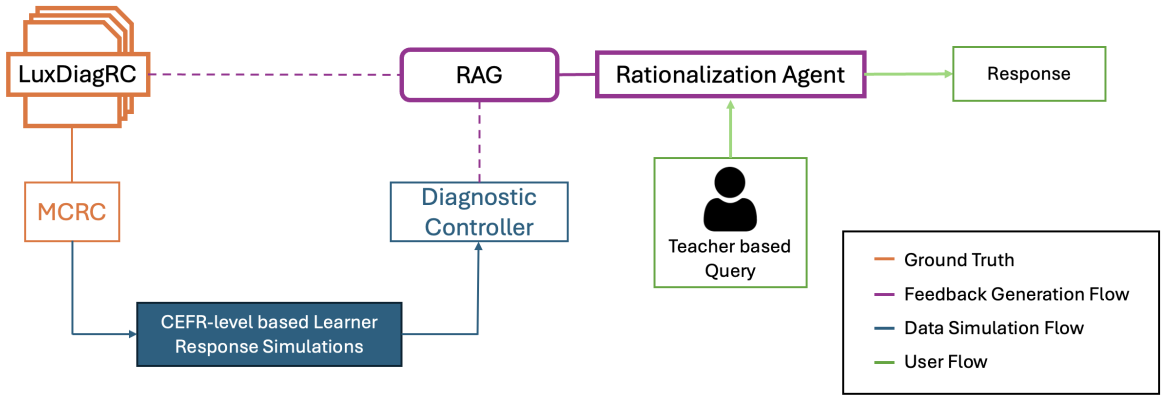


Figure 1: Overview of the *SimuLe-Lux* framework. *LuxDiagRC*) texts with the diagnostically rich multiple-choice reading comprehension questions (MCRC) are used to simulate CEFR-level learner responses. These responses are aggregated into neuro-symbolic diagnostic profiles and accessed by a retrieval-augmented teacher-facing agent to provide grounded explanations and pedagogical feedback.

the agent produces explanations that are both pedagogically meaningful and directly traceable to evidence in the dataset.

4 SimuLe-Lux Framework Evaluation

This section evaluates *SimuLe-Lux* as a multi-agent, teacher-facing framework for diagnostic reading comprehension training. In line with the methodology described in Section 3, the evaluation targets the two core computational components of the system: (i) the learner simulation and neuro-symbolic diagnostic controller, and (ii) the retrieval-augmented teacher-facing agent. Here we focus on whether the simulated learners exhibit CEFR-consistent behaviour and whether the teacher-facing agent produces grounded, pedagogically useful responses.

4.1 Evaluation of Learner Simulation

The learner simulation is evaluated with respect to its alignment with CEFR proficiency levels and the smoothness of transitions between adjacent learner profiles. After generating responses for all texts and proficiency levels, error patterns are analysed per question type and level to assess whether the neuro-symbolic diagnostic controller captures meaningful differences in reading comprehension ability.

4.1.1 Prompt Engineering and Simulation Protocol

The learner simulation underwent a systematic prompt engineering process. Initial prompts produced nearly identical response patterns across proficiency levels. This was resolved by including all CEFR levels (A1–C2) within a single prompt, enabling the model to differentiate between learner profiles within a shared context window. A minimum-correct-answer constraint was further introduced for the weaker-state simulations to avoid degenerate cases with unrealistically poor performance.

Table 2: Diagnostic controller probabilities derived from simulated learner responses. $_W$ denotes the weaker-state profiles. Each value represents the probability of committing an error for the corresponding question type.

Proficiency	Retrieve	Interpret	Infer	Evaluate
$A1_W$	0.77	0.81	0.79	0.76
A1	0.63	0.70	0.67	0.64
$A2_W$	0.68	0.79	0.73	0.78
A2	0.55	0.65	0.57	0.60
$B1_W$	0.66	0.77	0.66	0.72
B1	0.54	0.57	0.49	0.50
$B2_W$	0.64	0.64	0.62	0.65
B2	0.49	0.47	0.45	0.43
$C1_W$	0.59	0.61	0.58	0.55
C1	0.44	0.44	0.42	0.40
$C2_W$	0.52	0.51	0.51	0.50
C2	0.42	0.41	0.42	0.38

The final prompt was held fixed for all texts and question batches. It instructs the model to answer each question twice per proficiency level, once at the nominal level and once in a deliberately weaker state, with responses restricted to multiple-choice labels (A–D), ensuring consistency across runs:

- Beäntwer des Froen als géifs du als Schüler am Sproochentest setzen. Lëtzebuergesch

ass net deng Mammesprooch. Du mess realistesch an typesch Examensfeeler fir dei spezifesch Sproochen Niveauen (A1,A2,B1,B2,C1,C2). Äntwer simpel just als A, B, C oder D fir all Fro an fir all Sprochen Niveau. Pro Sproochen Niveau Beäntwer déi 40 Froën eng keier normal an nach eng kéier mä méi schwaach mat minimum 8 richtechen Äntwerten.

- Answer these questions as if you were a student in a language test. Luxembourgish is not your native language. You make realistic and typical exam mistakes for your specified language level (A1,A2,B1,B2,C1,C2). Answer only with A, B, C, or D for each question and for all language levels. For each language level, answer the 40 questions once normally and a second time in a weaker state with a minimum of 8 correct answers.

4.1.2 Neuro-Symbolic Diagnostic Controller

The aggregated error probabilities produced by the simulated learners are shown in Table 2. Each value represents the probability of committing an error for a given question type and proficiency level, with $_W$ denoting the weaker-state profiles. The results exhibit a CEFR-consistent trend: error rates decrease monotonically as proficiency increases across all question types.

Importantly, weaker-state profiles (e.g., $B2_W$) consistently fall between their nominal level and the adjacent lower level (e.g., B1), producing smooth transitions between learner profiles. This reflects a well-established property of second-language learning, where the strongest learners at a lower level may outperform the weakest learners at a higher level. Without weaker-state modeling, such transitional behaviour would be lost under hard level boundaries.

4.2 Evaluation of the Teacher-Facing Diagnostic Agent

The teacher-facing agent is evaluated on its ability to generate grounded, informative, and pedagogically appropriate responses to teacher queries. A total of 40 prompts were used, grouped into four categories: dataset-specific, proficiency-level–specific, teaching-skill–oriented, and vague. Representative prompts for each category are shown in Table 3.

Table 3: Sample prompt categories used to evaluate the teacher-facing agent. Prompts were originally in Luxembourgish and are translated here for clarity.

Prompt category	Example prompt
Dataset	A student performed very well on the "Blues" questions but struggled with the "D'Schnurreli" questions. Why might this be?
Proficiency level	Which difficulties does a B2 student encounter while reading?
Teaching skill	How can I improve the inferential skills of my weaker students?
Vague	How can I help strong students?

Table 4: Mean human evaluation scores for the teacher-facing agent across prompt categories. Scores are averaged over three judges on three categories with a five-point scale. Relevance (Rele.), Informativeness (Info.), Appropriateness (App.).

Prompt type	Rele.	Info.	Appr.
Dataset	4.43	3.92	3.97
Proficiency level	4.87	4.58	4.73
Teaching skill	4.27	4.37	4.77
Vague	3.10	2.60	3.83

4.2.1 Prompt Categories and Human Scoring

Dataset-specific prompts test whether the agent can retrieve and reason over LuxDiagRC content. Proficiency-level prompts assess its ability to connect learner profiles to expected weaknesses. Teaching-skill prompts evaluate whether the agent can support instructional reflection and diagnostic reasoning. Vague prompts test robustness to underspecified input.

All responses were scored by three independent human judges on a five-point scale along three rubrics: *Relevance* (use of dataset and diagnostic evidence), *Informativeness* (usefulness for teachers), and *Appropriateness* (pedagogical coherence). Mean scores for each prompt category are reported in Table 4.

4.2.2 Quantitative Results

As shown in Table 4, the agent achieves strong performance on dataset-specific, proficiency-level, and teaching-skill prompts, indicating that it effectively integrates the diagnostic controller with *LuxDiagRC* to generate grounded pedagogical explanations. Performance drops for vague prompts, particularly in relevance and informativeness, reflecting reduced access to dataset evidence and increased reliance on the model's internal knowledge.

4.2.3 Qualitative Analysis

Qualitative inspection of responses supports these findings. For teaching-skill prompts (Table 3), the agent explicitly references simulated learner performance and linguistic features when proposing instructional strategies, for example:

- *Start with texts where your learners already did better on interpret/retrieve and add one inferential item with strong evidence cues. Your own results show retrieve > inferential on Den Ersatzschoulmeeschter at lower levels; build up from there and explicitly neutralize false-friend traps in distractors during feedback.*

In contrast, for vague prompts the agent relies largely on generic pedagogical knowledge rather than diagnostic evidence, as illustrated by:

- *Compare across texts. Set quick synthesis tasks (e.g., compare how two texts set up a problem/solution or portray relationships) to push transfer of skills.*

- *Feedback that grows expertise. Use concise rubrics that reward evidence citation, metalinguistic precision, and quality of reasoning (not just speed or quantity). Ask students to annotate where their inference "leaps" occur using the text's key span labels.*

This contrast highlights the importance of *LuxDiagRC* and the neuro-symbolic diagnostic controller in enabling evidence-based teacher support.

5 Concrete Example: End-to-End SimuLe-Lux Workflow

To support understanding of *SimuLe-Lux's* operation, we present an end-to-end example following a LuxDiagRC input through to the diagnostic agent's final response, illustrating how the system's components interact throughout the process.

Input (LuxDiagRC) A Luxembourgish text annotated with linguistic features is used to generate a multiple-choice question for LuxDiagRC:

Question:

> *'Wisou hun d'Kanner forcéiert gelaacht?'* (*Why did the kids force a laughter?*)

Critical span (annotated):

> *Vun eis* `<LEX-FALSE-FRIEND>` *krut hien no all Witz* `<LEX-LOAN-DE>` *e* `<ORTHO-PHONO-DIVERGE>` `<MORPH-N-RULE>` *forcéiert* `<LEX-LOAN-FR>` *Laachen ze héieren* `<SYN-VERB-FINAL>`

Distractor span (annotated):

> *Seng Pedanterie* `<LEX-RARE-ZLS>` `<LEX-LOAN-FR>` *ass* `<LEX-FALSE-FRIEND>` *wait gang* `<SYN-VERB-FINAL>`

Answer options:

- A) *Vir dass se op dem Proff senger gudder seit stinn.* (Correct)
- B) *Well se wollten laachen.* (Misunderstanding)
- C) *Well den Prof pedant war.* (Distractor span)
- D) *Well d'Kanner den Witz gutt fannen.* (No textual support)

Simulated Learner Output (LLM Student Simulation). SimuLe-Lux generates CEFR-level learner responses:

Level	Answer	
	Normal	Weaker
A1	B	D
A2	B	D
B1	C	D
B2	A	B
C1	A	C
C2	A	A

The weaker-state simulation (e.g., $B2_W$) may still select either option B or C, allowing overlap between A2, B1 and B2. Thus, producing realistic transitional learner profiles.

Teacher Query.

> *"Why do my B2 learners often misunderstand this question?"*

Teacher-Facing Output (Rationalization Agent).

> *"B2 Learners often misread the LEX-LOAN-FR item forcéiert and map it to a volitional "wanted to laugh," so they pick the agentive answer "Well se wollten laachen" instead of the Inferential,*

social motive, which the critical span signals via "forcéiert Laachen". The Wisou prompt demands Inferential reasoning beyond surface form, but attention is drawn to clause-final structure (SYN-VERB-FINAL), so form-focused parsing overrides the causal reading and yields the misunderstanding."

This example demonstrates how *LuxDiagRC*, learner simulation, and the diagnostic controller are combined through retrieval-augmented generation to produce grounded, pedagogically meaningful teacher feedback.

6 Conclusion

This paper presents a novel paradigm for teacher-facing educational tools that addresses key gaps in the literature on reading comprehension datasets, learner simulation, and diagnostic AI systems. The proposed framework combines a dual-layer annotation pipeline with large language model–based learner simulation to construct a diagnostically grounded dataset and a retrieval-augmented teacher-facing agent capable of supporting pedagogical reasoning.

The results support three main conclusions. First, the evaluation of the teacher-facing agent shows that combining simulated learner data with a diagnostically rich dataset enables the generation of grounded and pedagogically meaningful feedback. Second, the learner simulation results demonstrate that large language models can be used to generate realistic and proficiency-sensitive student behaviour, strengthening their role as tools for learner modeling. Third, the inclusion of intermediate (weaker-state) proficiency profiles enhances the neuro-symbolic diagnostic controller by capturing smooth transitions and realistic overlaps between CEFR levels, allowing the system to model transitional learner weaknesses more faithfully.

Limitations and Future Work

This work stands as a proof of concept for the two main contributions of the paper, the *SimuLe-Lux* framework and the implementation of a neuro-symbolic diagnostic controller. However, several limitations remain. First, although *LuxDiagRC* is carefully designed to be diagnostically rich, it is relatively small in scale compared to large reading comprehension benchmarks. Its primary purpose is not broad coverage but the construction of a high-quality gold standard for diagnostic analysis. While this enables fine-grained grounding of learner errors, it limits the diversity of genres and topics.

Second, learner behaviour is generated using a single large language model. Although the CEFR-consistent patterns observed in Table 2 indicate that the simulation produces meaningful proficiency distinctions, the simulated learners may still reflect model-specific biases and stylistic regularities. In particular, some error patterns may be influenced by the internal representations of the underlying language model rather than by genuine second-language acquisition processes.

Third, the teacher-facing rationalization agent is evaluated using human judgments on synthetic prompts rather than through deployment with real teachers. While the results in Table 4 demonstrate strong grounding and pedagogical coherence for structured queries, the system's effectiveness in authentic classroom workflows and teacher training contexts remains to be validated.

Future work will address these limitations along three directions. First, *LuxDiagRC* will be expanded through controlled data synthesis, using the existing annotation framework to generate new diagnostically grounded texts and questions, thereby increasing coverage while preserving interpretability. Second, simulated learner profiles will be compared against real student response data to calibrate and validate the neuro-symbolic diagnostic controller, enabling more accurate modeling of second-language error distributions. Third, the teacher-facing agent will be evaluated in longitudinal user studies with educators, allowing the impact of *SimuLe-Lux* on diagnostic skill development and instructional decision-making to be assessed in realistic educational settings.

References

Luca Benedetto, Giovanni Aradelli, Antonia Donvito, Alberto Lucchetti, Andrea Cappelli, and Paula Buttery. 2024. Using LLMs to simulate students' responses to exam questions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11351–11368, Miami, Florida, USA. Association for Computational Linguistics.

Yevgeni Berzak, Jonathan Malmaud, and Roger Levy. 2020. STARC: Structured annotations for reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Lin-*

guistics, pages 5726–5735, Online. Association for Computational Linguistics.

Andreea Dutulescu, Stefan Ruseti, Denis Iorga, Mihai Dascalu, and Danielle S. McNamara. 2025. Ymcq: Reasoning-enhanced mcq generation. In *Artificial Intelligence in Education*, pages 308–315, Cham. Springer Nature Switzerland.

Aashish Ghimire and John Edwards. 2024. *Coding with AI: How Are Tools Like ChatGPT Being Used by Students in Foundational Programming Courses*, pages 259–267.

Christophe Friezas Gonçalves, Salima Lamsiyah, and Christoph Schommer. 2026. LuxDiagRC: A diagnostic reading comprehension corpus for Luxembourgish with linguistic and cognitive annotation layers. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 532–541, Rabat, Morocco. Association for Computational Linguistics.

Jana Gonnermann-Müller, Jennifer Haase, Konstantin Fackeldey, and Sebastian Pokutta. 2025. Facet: Teacher-centred llm-based multi-agent systems-towards personalized educational worksheets.

Seyed Jamal Hemati and Purya Baghaei. 2020. A cognitive diagnostic modeling analysis of the english reading comprehension section of the iranian national university entrance examination. *International Journal of Language Testing*, 10(1):11–32.

Bihao Hu, Jiayi Zhu, Yiying Pei, and Xiaoqing Gu. 2025. Exploring the potential of llm to enhance teaching plans through teaching simulation. *npj Science of Learning*, 10(1):7.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Chang Liu, Loc Hoang, Andrew Stolman, Rene F. Kizilcec, and Bo Wu. 2025. Understanding student engagement with large language model-powered course assistants. In *Artificial Intelligence in Education*, pages 3–10, Cham. Springer Nature Switzerland.

Daria Martynova, Jakub Macina, Nico Daheim, Nilay Yalcin, Xiaoyu Zhang, and Mrinmaya Sachan. 2025. Can llms effectively simulate human learners? teachers' insights from tutoring llm students. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 100–117.

Ina VS Mullis and Michael O Martin. 2019. *PIRLS 2021 Assessment Frameworks.* ERIC.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Silvia Pokrivcakova. 2019. Preparing teachers for the application of ai-powered technologies in foreign language education. *Journal of language and cultural education*, 7(3):135–153.

Cédric Pruski, Célia da Costa Pereira, Marcos Da Silveira, Gabriele Marconi, Marie Gallais, and Andrea Tettamanzi. 2025. Bridging skill gaps: Combining generative and symbolic ai for personalized lifelong learning pathways. In *Artificial Intelligence in Education*, pages 19–27, Cham. Springer Nature Switzerland.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Shruti Singh, Nandan Sarkar, and Arman Cohan. 2024. SciDQA: A deep reading comprehension dataset over scientific papers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20908–20923, Miami, Florida, USA. Association for Computational Linguistics.

Shungo Suzuki, Hiroaki Takatsu, Ryuki Matsuura, Miina Koyama, Mao Saeki, and Yoichi Matsuyama. 2025. Feedforwarding diagnostic language assessment: Artificial intelligence- (ai-) driven weakness identification and contextualised feedback for second language speaking. *Language Testing*, 42(4):476–507.

Songlin Xu, Xinyu Zhang, and Lianhui Qin. 2024. Eduagent: Generative student agents in learning.

Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, Lei Hou, and Juanzi Li. 2025. Simulating classroom education with LLM-empowered agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10364–10379, Albuquerque, New Mexico. Association for Computational Linguistics.

Towards Readability Assessment for Under-Resourced Arabic Dialects: A Study of Moroccan Darija

Houdaifa Atou
Mohammed VI Polytechnic University
houdaifa.atou@um6p.ma

Nouran Khallaf
University of Leeds; Alexandria University
n.khallaf@leeds.ac.uk

Salima Lamsiyah
University of Luxembourg
salima.lamsiyah@uni.lu

Ruslan Mitkov
University of Alicante
ruslan.mitkov@gcloud.ua.es

Abstract

Despite recent progress in readability assessment for Modern Standard Arabic (MSA), Moroccan Darija remains largely unexplored. This paper presents the first systematic study of sentence-level readability assessment for Moroccan Darija. To support this study, we introduce BAYAN, the first corpus and benchmark for Moroccan Darija readability assessment. It contains 2,030 Arabic-script sentences from seven sources covering journalistic, conversational, informal, spoken, lexical, figurative, and poetic language. Two native speakers independently annotated each sentence using four ordered levels: *Immediate* (DR1), *Accessible* (DR2), *Demanding* (DR3), and *Advanced* (DR4). We also introduce BAYANEX, a subset of 912 sentences annotated for orthographic, lexical, syntactic, and semantic difficulty. Using these resources, we evaluate standalone readability formulae, classical machine learning methods, pretrained encoders, transfer learning from MSA readability models, and large language models (LLMs). MARBERTv2 fine-tuned on BAYAN achieves the best overall performance, with 76.3% accuracy, 75.0% macro F1, and 0.838 quadratic weighted kappa (QWK). AARI is the strongest standalone formula at 0.740 QWK, Support Vector Regression using readability formula features reaches 0.756, and NileChat-3B achieves the best zero-shot performance, with 0.693. Our transfer learning experiments show that MSA readability pretraining benefits linear probing but does not improve full fine-tuning over a strong dialectal encoder.

1 Introduction

Readability assessment estimates the effort required by an intended reader to understand a text and supports educational applications, accessible communication, and text simplification (Collins-Thompson, 2014; Aluisio et al., 2010; Vajjala, 2022). Readability depends on interacting factors, including vocabulary familiarity, sentence structure, information density, abstraction, discourse relations, and the background knowledge assumed by the text (Collins-Thompson, 2014; Vajjala, 2022). These factors are particularly difficult to model for language varieties with limited annotated data and few reliable linguistic tools (Imperial and Kochmar, 2023).

Moroccan Darija presents a challenging setting for readability assessment. It is widely used in everyday conversation, social media, entertainment, advertising, and public communication, but it has no standardised written form. Written Darija shows substantial spelling and register variation, which affects word recognition and limits the direct application of readability methods developed for standardised languages (Outchakoucht and Es-Samaali, 2024; Bounhar, 2025; Shang et al., 2025). The benchmark developed in this study is restricted to Arabic-script sentences.

Existing Darija resources cover translation, morphology, summarisation, social media analysis, named entity recognition, and language model adaptation (Bouamor et al., 2018; Al-Shargi et al., 2016; Gaanoun et al., 2022; Aftiss et al., 2025; Moussa and Mourhir, 2023; Skiredj et al., 2025b; Shang et al., 2025), but they do not provide human judgements of reading difficulty. MSA readability research offers formulae, lexical resources, educational corpora, sentence-level datasets, simplification corpora, and annotation guidelines (Al-Khalifa and Al-Ajlan, 2010; Al Khalil et al., 2020; El-Haj et al., 2024; Alhafni et al., 2024; Khallaf and Sharoff, 2021; Naous et al., 2024; Elmadani et al., 2025; Habash et al., 2025). However, these resources reflect MSA vocabulary, educational grades, or language proficiency and may not transfer directly to Darija. Readability depends on both the intended reader and the language variety: a sentence that is accessible in MSA may not be equally accessible to a Darija-dominant reader, while fa-

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 199–225
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_022

miliar colloquial Darija may remain difficult for an MSA-based model.

We introduce BAYAN, which, to the best of our knowledge, is the first corpus and benchmark for sentence-level readability assessment in Moroccan Darija. The corpus includes journalistic, conversational, informal, spoken, lexical, figurative, and poetic language. Each sentence is assigned one of four ordered readability levels: *Immediate* (DR1), *Accessible* (DR2), *Demanding* (DR3), and *Advanced* (DR4). These levels reflect the reading effort required of a native or near-native Darija reader.

We also introduce BAYANEX, an enriched subset annotated for orthographic, lexical, syntactic, and semantic difficulty to support more fine-grained analyses of readability. Using these resources, we compare standalone readability formulae, classical machine learning methods, pretrained language models, transfer learning from MSA readability models, and large language models.

Our contributions are as follows:

- We present the first systematic study of sentence-level readability assessment for Moroccan Darija.
- We introduce BAYAN, the first human-annotated readability benchmark for Moroccan Darija, and BAYANEX, an enriched subset annotated for orthographic, lexical, syntactic, and semantic difficulty.
- We provide a comprehensive empirical evaluation spanning standalone readability formulae, classical machine learning, pretrained encoders, transfer learning from MSA readability models, and large language models.
- We investigate the transfer of readability knowledge from MSA to Moroccan Darija under different adaptation strategies.

Our experiments show that MARBERTv2 (Abdul-Mageed et al., 2021), when fine-tuned on BAYAN, achieves the best overall performance. Among traditional approaches, AARI (Al-Tamimi et al., 2014) provides the strongest standalone formula baseline, while Support Vector Regression with readability formula features performs best among feature-based models. Among large language models, NileChat-3B (El Mekki et al., 2025) achieves the best zero-shot performance. We also find that MSA readability knowledge transfers under linear probing, although it does not outperform a fully fine-tuned dialectal encoder.

2 Related Work

Research relevant to Moroccan Darija readability can be considered from four complementary perspectives: resources describing written Darija, datasets developed for other Darija NLP tasks, formula-based readability assessment, and supervised readability resources for Modern Standard Arabic.

Darija resources include parallel and multi-variety corpora such as MADAR, together with morphological and lexical resources such as MOR, DODa, the Moroccan Darija WordNet, and MDED (Bouamor et al., 2018; Al-Shargi et al., 2016; Outchakoucht and Es-Samaali, 2024; Mrini and Bond, 2017; Tachicart et al., 2014). More recent collections and adaptation resources include Goudsum, Atlaset, and Atlas-Chat (Issam and Mrini, 2022; Bounhar, 2025; Shang et al., 2025). These resources represent important characteristics of written Darija, including spelling variation and differences between formal and informal language. They provide source material for corpus construction, model pretraining, and lexical analysis, but they do not contain human judgements of reading difficulty. Appendix Table 19 summarises the resources most relevant to Darija readability research.

Task-specific Darija datasets cover sentiment and social-media analysis (Boujou et al., 2021; Oussous et al., 2020; Elouardighi et al., 2017; Jbel et al., 2023; Garouani and Kharroubi, 2021; Mihi et al., 2020), as well as offensive-language detection (Essefar et al., 2023; Ibrahimi and Mourhir, 2024). Other resources address text-to-SQL generation, culturally grounded translation, instruction following, and multimodal sentiment analysis (Chafik et al., 2025; Momayiz et al., 2024; Shang et al., 2025; Ben Cheikhi and Nfaoui, 2026). These datasets extend the topical and stylistic coverage of Darija NLP, but their labels describe objectives other than reading difficulty. They may support pretraining, augmentation, or robustness analysis, but they cannot serve as gold readability data. Additional resources of this type are listed in Appendix Table 20.

Formula-based methods provide lightweight readability estimates using surface properties of text. ARI, LIX, and RIX rely mainly on sentence

length, word length, and the proportion of long words (Smith and Senter, 1967; Björnsson, 1968; Anderson, 1983). Arabic-specific work includes the Al-Heeti formula and AARI (Cavalli-Sforza et al., 2018; Al-Tamimi et al., 2014), Arability (Al-Khalifa and Al-Ajlan, 2010), and OSMAN (El-Haj and Rayson, 2016). These methods are attractive in a low-resource setting because they can be computed without a parser, morphological analyser, or frequency lexicon. However, their original interpretations assume educational scales, standardised spelling, or properties of MSA that may not apply to Darija. We therefore evaluate ARI, LIX, RIX, Al-Heeti, and AARI as standalone continuous indicators using thresholds learned from the training partition. We also use the five formula scores as numerical features in the classical machine learning experiments. Their definitions and original references are provided in Appendix Section D.1.

Supervised Arabic readability research has focused primarily on MSA. The Leveled MSA Reading Corpus provides levelled textual supervision (Al Khalil et al., 2018), the SAMER Lexicon provides word-level difficulty labels (Al Khalil et al., 2020), and the SAMER corpus provides original and simplified text (Alhafni et al., 2024). DARES and DARES 2.0 use educational stages and curriculum difficulty (El-Haj et al., 2024; Almujaiwel et al., 2025), while BAREC provides a fine-grained ordinal scale covering a broad range of readerships (Elmadani et al., 2025; Habash et al., 2025). The Khallaf–Sharoff dataset and ReadMe++ organise sentences using CEFR-related levels (Khallaf and Sharoff, 2021; Naous et al., 2024). AraEyebility complements corpus annotation with subjective judgements and eye-tracking evidence (Baazeem et al., 2025).

These resources represent related but distinct forms of supervision. School grade reflects curriculum placement, CEFR primarily describes language proficiency, lexical resources estimate word difficulty, and simplification corpora represent relationships between original and simplified texts. None is directly equivalent to the reading effort required by a native Darija reader at sentence-level. CEFR is also difficult to apply directly because Darija lacks an established CEFR-aligned curriculum and standard reading benchmarks. We therefore use MSA resources for transfer and representation learning while evaluating all systems against independently annotated Darija labels. Appendix Table 21 compares the main MSA resources and their potential value for transfer.

To our knowledge, previous work has not provided a multisource corpus with human-annotated sentence-level readability labels for Moroccan Darija.

3 Task and Dataset

3.1 Task Definition

We formulate Moroccan Darija readability assessment as an ordinal sentence classification task. Given a written Darija sentence x_i, the objective is to predict

$$y_i \in \{\mathrm{DR1}, \mathrm{DR2}, \mathrm{DR3}, \mathrm{DR4}\},$$
$$\mathrm{DR1} < \mathrm{DR2} < \mathrm{DR3} < \mathrm{DR4}.$$

DR1 represents an immediate and explicit message. DR2 covers accessible sentences with familiar vocabulary and straightforward relations between ideas. DR3 represents demanding sentences containing several propositions, developed explanations, less familiar vocabulary, or moderate abstraction. DR4 represents advanced sentences requiring substantial effort because of specialised vocabulary, dense structure, implicit meaning, figurative language, cultural knowledge, or difficult decoding.

The labels represent increasing reading effort for a defined Darija reader rather than educational grade or language proficiency.

3.2 Bayan Corpus

BAYAN contains 2,030 Moroccan Darija sentences selected from six publicly available resources and one poetry collection compiled for this study. The sources were chosen to represent variation in register, provenance, and communicative setting, including journalistic, conversational, lexical, figurative, informal, spoken, and poetic language.

Of the final 2,030 sentences, 1,683 come from native Darija material, including 1,244 written sentences and 439 spoken transcripts, while the remaining 347 are human translations from English. Detailed source counts, registers, provenance, and descriptions are provided in Appendix Table 10.

3.3 Sources and Preprocessing

GOUD contains Moroccan news articles paired with their headlines and was released as the Goud-sum dataset for Moroccan Darija summarisation (Issam and Mrini, 2022). In our subset, longer articles were divided using common sentence delimiters.

Because the source contains substantial MSA, predominantly MSA articles were removed through an initial language screening, followed by sentence-level review by both annotators.

ALEXANDRIA is a machine translation dataset containing 107,000 conversational turns from 13 Arab countries and 11 domains (EL Mekki et al., 2026). We retained only the human Moroccan Darija translations and excluded the source side and all other Arabic varieties. Turns containing several sentences were divided using common sentence delimiters.

DODA is drawn from the Darija Open Dataset, a collaborative resource containing lexical entries, verb conjugations, and example sentences (Outchakoucht and Es-Samaali, 2021). We used the version distributed with Atlas Chat (Shang et al., 2025). This subset contributes short sentences and isolated examples that frequently represent the easier end of the readability range.

PALM contains instruction and response pairs in MSA and dialectal Arabic across 20 topics and 22 Arab countries (Alwajih et al., 2025). We retained Moroccan Darija items from the Language and Proverbs categories. Instructional framing was removed while preserving the expression being assessed so that each item could be read as standalone text.

MSM-MG contains native Moroccan Darija social media posts paired with synthetic variants developed for several NLP tasks (Shang et al., 2025). We retained only the original user-authored posts and excluded all generated content. Longer posts were divided into sentences using common delimiters.

WAK3I is a collection of Moroccan Darija audio and transcript pairs developed for speech processing (oddadmix, 2026). We retained the transcripts and excluded the audio. The resulting material consists mainly of spontaneous narratives about everyday events.

ZAJAL is a collection of Moroccan Darija poetry that we compiled for this study from publicly available poetry forums. It contributes figurative language, traditional vocabulary, and poetic structures that are uncommon in the other sources.

For every source, we retained only the Darija text required for readability annotation. Source names and original identifiers were preserved where available. We refer to the collections as source subsets rather than domains because several contain material from multiple topics and communicative settings.

3.4 Annotation Framework

3.4.1 Readability Levels

Existing Arabic readability formulae were developed primarily for MSA and often assume standardised spelling, formal vocabulary, or educational grade levels. CEFR primarily describes language learner proficiency and cannot be applied directly to Darija, which lacks an established CEFR-aligned curriculum and standard reading benchmarks.

We therefore defined four ordered readability levels specifically for Moroccan Darija, ranging from *Immediate* (DR1) for simple, explicit sentences to *Advanced* (DR4) for sentences requiring substantial reading effort because of lexical, syntactic, semantic, or orthographic complexity. *Accessible* (DR2) and *Demanding* (DR3) represent intermediate levels. Annotators began with DR1 and selected a higher level only when identifiable linguistic or interpretative evidence justified the increase. Sentence length alone did not determine the label. Full definitions of the four levels are provided in Appendix A.1.

3.4.2 Diagnostic Dimensions and BAYANEX

To identify the factors underlying the overall readability judgement, a subset of the corpus was also annotated for four diagnostic dimensions: **orthographic difficulty**, covering word recognition, segmentation, spelling, and written-form variation; **lexical difficulty**, covering word familiarity, register, and specialised expressions; **syntactic difficulty**, covering relations between clauses, propositions, and references; and **semantic difficulty**, covering abstraction, inference, figurative meaning, and cultural knowledge.

Each dimension was scored from 0, indicating no meaningful difficulty, to 3, indicating a major difficulty affecting understanding. The four scores were assigned independently and were not added together to determine the overall DR1 to DR4 label.

The resulting enriched subset contains 912 sentences and is referred to as BAYANEX. Full dimension definitions are provided in Appendix Section A.2.

3.4.3 Annotation Procedure

Following preprocessing, we sampled 4,000 candidate sentences for annotation. Two native speakers of Moroccan Darija independently annotated the

dataset. Both are fluent readers and writers of Darija and are familiar with Arabic-script, Arabizi, and code-switched forms, as well as with the domains represented in the corpus. Before annotation, both annotators reviewed the guidelines, which defined the four readability levels, provided representative examples, and described how to handle ambiguous cases. The guidelines remained available throughout the annotation process as a reference to support consistent annotation.

Annotators independently assessed each sentence. They first determined whether the item was suitable for readability assessment, then assigned an overall DR1–DR4 label. For items included in BAYANEX, they also assigned the four diagnostic scores. The overall readability label was assigned independently of the diagnostic scores.

3.4.4 Quality Control

Sentences were excluded if they were corrupted, incomplete, unintelligible, predominantly MSA, dependent on unavailable context, inappropriate for annotation, or otherwise unsuitable for reliable readability assessment. Difficult but interpretable sentences were retained. A complete breakdown of exclusion decisions by annotator and source is provided in Appendix Tables 8 and 9.

3.5 Annotation Agreement and Gold Labels

Of the 4,000 sampled sentences, 2,962 were retained as suitable by both annotators. Exact agreement on the readability level was 68.5%, with an unweighted Cohen's κ of 0.550.

Because the labels are ordered, we also calculated quadratic weighted Cohen's κ and ordinal Krippendorff's α. These measures reached 0.793 and 0.798, respectively. Table 1 reports the per-level annotation counts and specific agreement, which ranges from 62.7% for *Advanced* to 73.1% for *Demanding*. The annotators differed in how often they used some levels, particularly *Immediate*. Directional agreement matrices in Appendix B provide a more detailed view of these differences and show that most disagreements occurred between adjacent levels.

We assigned a gold label only when both annotators agreed on the readability level. This resulted in 2,030 gold-labelled sentences, while the remaining 932 were excluded rather than resolved through adjudication. We chose this strategy to prioritise label reliability, although it may also remove borderline or ambiguous cases that are difficult even for human annotators. Of the retained sentences, 912 belong to BAYANEX subset. Appendix Figure 4 shows the distribution of annotation distances.

Table 1: Per-level annotation counts and agreement. Agreed is the number of sentences assigned that level by both annotators. Specific agreement is calculated as $2n_{agreed}/(n_A + n_B)$.

Level	Annot. A	Annot. B	Agreed	Spec. agr.
Immediate	624	353	328	67.1
Accessible	939	1,087	668	65.9
Demanding	1,124	1,159	834	73.1
Advanced	275	363	200	62.7
Total	**2,962**	**2,962**	**2,030**	**68.5**

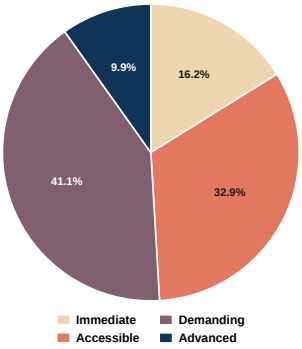


Figure 1: Distribution of the four readability levels in BAYAN.

The complete flow of items through the annotation process is shown in Appendix Figure 9, and representative level examples are provided in Appendix Figure 10.

3.6 Corpus Statistics

Following annotation and quality control, the final corpus comprises 2,030 sentences. As shown in Figure 1, the readability labels are distributed as follows: *Demanding* (41.1%), *Accessible* (32.9%), *Immediate* (16.2%), and *Advanced* (9.9%).

Sentence length shows a positive but imperfect association with annotated readability. As Figure 2 shows, longer sentences generally receive higher readability labels. The pattern therefore suggests that sentence length captures only one aspect of readability: longer sentences tend to be rated as more difficult, but the substantial overlap across levels indicates that other lexical and structural features also contribute to readability. Appendix Figure 6 shows sentence length distributions by readability level and corpus source. Appendix Figures 8 and 11 provide further statistics for BAYANEX, including the label distribution and diagnostic score

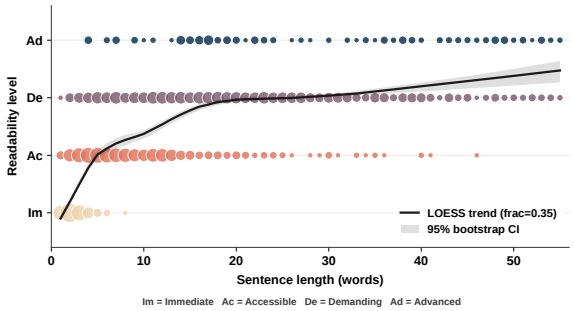


Figure 2: Association between sentence length and annotated readability level. The line shows the LOESS trend, and the shaded area represents the 95% bootstrap confidence interval.

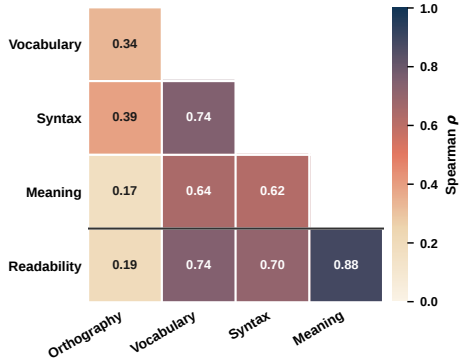


Figure 3: Spearman correlations between the diagnostic dimensions and the overall readability label in BAYANEX.

profiles.

The diagnostic annotations in BAYANEX provide further insight into the readability construct. Meaning shows the strongest correlation with the overall readability label ($\rho = 0.88$), followed by vocabulary ($\rho = 0.74$) and syntax ($\rho = 0.70$). Orthography has a much weaker correlation ($\rho = 0.19$). These results suggest that readability depends primarily on understanding the meaning of a sentence, with vocabulary and syntax also playing important roles.

4 Methodology

We conduct four groups of experiments: formula-based and classical baselines built directly from sentence text, general-purpose pretrained encoders, a transfer setup that uses MSA readability resources, and prompted large language models. All trainable models use the official training set BAYAN. For continuous baselines, we set the decision thresholds on the training set only. We choose models on the validation set and report final results on the held-out test set.

4.1 Formula-Based and Classical Baselines

We first evaluate five readability formulae as standalone baselines: ARI (Smith and Senter, 1967), LIX (Björnsson, 1968), RIX (Anderson, 1983), Al-Heeti (Cavalli-Sforza et al., 2018; Al-Tamimi et al., 2014), and AARI (Al-Tamimi et al., 2014). Each formula produces one continuous score for a sentence. Because these formulae were not developed for Moroccan Darija, we do not use their original educational or grade-based interpretations. Instead, three ordered thresholds are learned separately for each formula using the training set and are then applied unchanged to the validation and test sets.

We also evaluate four classical machine learning models: Logistic Regression (Cox, 1958), Support Vector Regression (Drucker et al., 1996), and Random Forest classification and regression (Breiman, 2001). Each model is evaluated with six sentence representations: (i) the five readability formula scores, (ii) explicit surface features, (iii) surface features combined with the formula scores, (iv) word TF–IDF, (v) character TF–IDF, and (vi) combined word and character TF–IDF.

The formula representation consists of the ARI, LIX, RIX, Al-Heeti, and AARI scores calculated for each sentence. The surface representation contains 17 features covering word and character counts, word-length statistics, long-word counts and proportions, lexical variation, punctuation, digits, symbols, and character elongation. Character TF–IDF captures subword and orthographic patterns, while word TF–IDF represents lexical items and short expressions. Formula definitions are provided in Appendix Section D.1, and the complete surface feature inventory is provided in Appendix Table 12.

Logistic Regression and Random Forest directly predict one of the four readability levels. Support Vector Regression and Random Forest Regression instead predict a continuous difficulty value. For the regression models, the ordered labels are represented numerically as

$$\mathrm{DR1} = 1, \quad \mathrm{DR2} = 2,$$
$$\mathrm{DR3} = 3, \quad \mathrm{DR4} = 4.$$

Continuous scores from the standalone formulae and regression models are converted to readability labels using three ordered thresholds, $t_1 < t_2 < t_3$. For a continuous score s_i, the predicted level is

$$\hat{y}_i = \begin{cases} \text{DR1}, & s_i \leq t_1, \\ \text{DR2}, & t_1 < s_i \leq t_2, \\ \text{DR3}, & t_2 < s_i \leq t_3, \\ \text{DR4}, & s_i > t_3. \end{cases} \quad (1)$$

For each continuous system, the thresholds are selected on the training partition to maximise quadratic weighted kappa. The selected thresholds are then fixed and applied unchanged to the validation and test predictions.

4.2 General-Purpose Encoders

We evaluate pretrained transformer encoders representing four pretraining settings: Moroccan Darija (Gaanoun et al., 2025), dialectal Arabic (Abdul-Mageed et al., 2021; Inoue et al., 2021), MSA (Abdul-Mageed et al., 2021; Antoun et al., 2020; Inoue et al., 2021), and multilingual text (Conneau et al., 2020; Devlin et al., 2019).

Each encoder is fine-tuned on BAYAN using a four-class classification head. This comparison examines whether pretraining on Moroccan Darija or dialectal Arabic provides an advantage over MSA and multilingual pretraining.

4.3 Transfer from MSA Readability Models

We examine whether representations learned from MSA readability supervision transfer to Moroccan Darija. We evaluate encoders previously trained on Arabic readability datasets (Naous et al., 2024; Elmadani et al., 2025) using two adaptation strategies.

Under sequential fine-tuning, all parameters of the readability-pretrained encoder are further updated using BAYAN. This setting tests whether previous MSA readability supervision provides a stronger starting point for Darija adaptation. In contrast, under linear probing, the encoder parameters remain frozen and only a new classification layer is trained using BAYAN. This setting assesses whether the frozen representation already contains transferable readability information.

MARBERTv2 is evaluated under the same adaptation conditions as a reference encoder pretrained on dialectal Arabic but not previously trained for readability assessment.

4.4 Large Language Models

We evaluate instruction-tuned large language models under zero-shot and few-shot prompting. The comparison includes models developed for Moroccan Darija (Shang et al., 2025; Skiredj et al., 2025a; El Mekki et al., 2025), Arabic-centred models (Fanar Team, 2025; Anwar et al., 2025; Bari et al., 2025), and general-purpose multilingual models (Gemma Team, 2025; Qwen Team, 2026; Grattafiori et al., 2024).

The prompt provides the four readability-level definitions and instructs the model to return one label from DR1 to DR4. The shared prompt structure and exact model checkpoints are provided in Appendix Section E.3 and Appendix Table 14.

4.5 Evaluation Measures

We report accuracy, macro F1, and quadratic weighted kappa (QWK).

Accuracy measures the proportion of sentences assigned the correct readability level. Macro F1 calculates F1 independently for each level and gives equal weight to all four classes, including the less frequent *Advanced* class.

QWK is the primary ordinal evaluation measure. It compares the predicted and gold labels while penalising errors according to their distance on the readability scale. QWK ranges from agreement below chance to perfect agreement, with a value of 1 indicating complete agreement.

5 Experimental Results

5.1 Implementation Details

All encoder models are fine-tuned using the training split, with hyperparameters selected on the validation set. The checkpoint achieving the highest validation performance is evaluated on the held-out test set. Unless otherwise stated, each trainable experiment is repeated using five seeds, and we report the mean and standard deviation across runs. LLMs are evaluated using one fixed run. Additional implementation details and hyperparameters are provided in Appendix Section E.1.

5.2 Standalone Readability Formulae

Table 2 reports the performance of individual readability formulae. AARI achieves the strongest performance across all evaluation measures. LIX ranks second, whereas ARI and RIX provide weaker baselines. Al-Heeti performs poorly, showing that Arabic-oriented formulae do not transfer equally well to Moroccan Darija.

5.3 Classical Machine Learning

Table 3 reports the best feature representation for each classical model. Hand-crafted features consis-

Table 2: Readability classification with individual formula scores and training-derived thresholds.

Formula	Accuracy	Macro-F1	QWK
AARI	**65.4**	**61.8**	**0.740**
LIX	56.3	54.3	0.596
ARI	51.4	49.0	0.522
RIX	42.8	41.8	0.441
Al-Heeti	45.5	29.1	0.107

Table 3: Best feature representation for each classical model, evaluated using quadratic weighted kappa (QWK). ∘ denotes thresholded regression.

Method	Features	QWK
SVR∘	formulas	**0.756**
Logistic Regression	surface+formulas	0.754
Random Forest	formulas	0.725
RF regression∘	surface+formulas	0.725

tently outperform word and character TF–IDF representations. The five formula scores perform best for Support Vector Regression and Random Forest. Surface features combined with formula scores perform best for Logistic Regression and Random Forest Regression. The full feature-by-model comparison is provided in Appendix Table 15.

5.4 General-Purpose Transformer Encoders

Table 4 compares encoders across Darija-specific, dialectal Arabic, MSA, and multilingual pretraining. MARBERTv2 (Abdul-Mageed et al., 2021), which was pretrained on large-scale Arabic data with broad dialectal coverage, achieves the strongest overall performance. Its exposure to dialectal Arabic may better capture the lexical and orthographic variation found in Darija. Overall, models pretrained on Darija or dialectal Arabic consistently outperform MSA and multilingual encoders, although ARBERTv2 (Abdul-Mageed et al., 2021) remains competitive.

5.5 Transfer from MSA Readability Models

Tables 5 and 6 compare sequential fine-tuning and linear probing. Sequential fine-tuning of the readability-pretrained encoders, readabert-ar (Naous et al., 2024) and arabertv02-word-CE (Elmadani et al., 2025), does not surpass MARBERTv2. This shows that MSA readability pretraining provides no advantage under full fine-tuning. Under linear probing, however, readabert-ar reaches 0.767 QWK and outperforms the frozen MARBERTv2 reference. The result indicates that MSA readability pretraining leaves transferable information in the representation, while also showing that its value depends on the adaptation strategy.

Table 4: General-purpose pretrained encoder results, reported as mean (standard deviation) across five seeds.

Model	Acc.	Macro-F1	QWK
Darija-specific			
DarijaBERT	75.6(6)	74.4(7)	0.833(3)
Dialectal Arabic			
MARBERTv2	**76.3(17)**	**75.0(13)**	**0.838(13)**
CAMeLBERT-DA	75.0(10)	73.9(10)	0.824(10)
Modern Standard Arabic			
ARBERTv2	75.4(8)	74.4(10)	0.823(8)
AraBERTv02	74.6(8)	73.7(8)	0.814(6)
CAMeLBERT-MSA	73.4(17)	72.3(20)	0.811(17)
AraBERTv2	70.5(10)	68.9(10)	0.782(2)
Multilingual			
XLM-R_{base}	72.8(3)	72.1(6)	0.806(6)
mBERT	71.0(7)	69.4(12)	0.786(7)

Table 5: Sequential fine-tuning results, reported as mean (standard deviation) across five seeds.

Model	Acc.	Macro-F1	QWK
Readability-pretrained			
readabert-ar	75.8(19)	74.8(22)	0.823(10)
arabertv02-word-CE	73.5(12)	72.8(17)	0.806(8)
Reference			
MARBERTv2	**76.3(17)**	**75.0(13)**	**0.838(13)**

Table 6: Linear probing results, reported as mean (standard deviation) across five seeds. MARBERTv2 is the reference encoder.

Model	Acc.	Macro-F1	QWK
Readability-pretrained			
readabert-ar	**70.0(28)**	**68.4(28)**	**0.767(15)**
arabertv02-word-CE	67.9(17)	66.1(17)	0.740(14)
Reference			
MARBERTv2	68.2(20)	63.0(52)	0.754(22)

5.6 Large Language Models

Table 7 presents zero-shot performance. All prompted models remain below the supervised encoders. NileChat-3B is the strongest evaluated zero-shot model at 0.693 QWK, while Gemma-3-12B-it (Gemma Team, 2025) is the strongest general-purpose model at 0.666. This advantage suggests that NileChat-3B's pretraining on dialect-rich Moroccan Arabic data may better align the LLM with the lexical and orthographic variability of Darija. Additional four-shot and eight-shot results are re-

Table 7: Zero-shot readability classification with instruction-tuned LLMs.

Model	Acc.	Macro-F1	QWK
Darija-specific			
Atlas-Chat-9B	59.7	48.4	0.592
GemMaroc-Qwen2.5-7B-darija	50.6	40.9	0.511
NileChat-3B	**63.4**	**51.9**	**0.693**
Arabic			
Fanar-1-9B-Instruct	52.8	43.3	0.559
Jais-2-8B-Chat	35.9	29.2	0.434
ALLaM-7B-Instruct	45.5	35.6	0.440
General-purpose			
Gemma-3-12B-it	62.4	50.7	0.666
Qwen3.5-9B	53.1	45.1	0.604
Llama-3.1-8B-Instruct	52.8	44.2	0.590

ported in Appendix Tables 16 and 17.

6 Conclusion

We presented, to our knowledge, the first systematic study of sentence-level readability assessment for Moroccan Darija. We introduced BAYAN, a 2,030-sentence Arabic-script benchmark with four ordered readability levels, and BAYANEX, a 912-sentence subset with diagnostic annotations. Across the evaluated systems, MARBERTv2 provides the strongest results, while formula-based and classical models remain competitive without requiring Darija-specific linguistic tools. Readability-pretrained MSA encoders retain useful information under linear probing, but sequential transfer does not surpass the strongest dialectal encoder. Large language models evaluated in the zero-shot setting remain below supervised encoders. These findings provide the first comprehensive benchmark for Moroccan Darija readability assessment and establish a foundation for future work in the area.

Limitations

Corpus coverage and distribution. BAYAN covers seven source subsets and four readability levels, but it is restricted to Arabic-script sentence-level material and therefore does not represent every form of written Darija or passage-level phenomena such as cohesion and discourse organisation. In addition, the label distribution is concentrated in the *Accessible* and *Demanding* classes, while the *Advanced* class is comparatively small.

Number of Annotators. The corpus was annotated by two native speakers. While the observed agreement scores indicate good annotation consistency, they should be interpreted in the context of a two-annotator study. A larger and more diverse pool of annotators would provide a more reliable estimate of variation in perceived readability across readers and help distinguish ambiguous sentences from annotator-specific differences.

Exact-agreement gold construction. The current gold corpus retains only sentences receiving the same overall label from both annotators. This provides reliable labels but favours clearer cases and may underrepresent ambiguous boundary instances. The excluded disagreement set should therefore be preserved for future adjudication or uncertainty-aware modelling.

Diagnostic annotation dependence. The diagnostic dimensions and overall labels were collected within the same annotation campaign. Although the dimensions were defined independently, complete cognitive independence between the two types of judgement cannot be guaranteed.

Transfer and feature availability. Our transfer learning and classical feature experiments should be interpreted with caution. The available Arabic readability datasets differ from BAYAN in both language variety and domain, making it difficult to isolate the effect of cross-variety transfer. In addition, the classical systems are limited to features that can be computed reliably without a Darija parser, morphological analyser, or validated frequency lexicon. Richer linguistic resources may improve performance, but their contribution remains to be evaluated.

Acknowledgment

This research has been partially supported by the CIDEGENT (CIDEXG/2023/13) project with funding provided by the Generalitat Valenciana.

References

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2021. QADI: Arabic dialect identification in the wild. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 1–10, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT and MARBERT: Deep bidirectional transformers for arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and*

the 11th International Joint Conference on Natural Language Processing, pages 7088–7105, Online. Association for Computational Linguistics.

Azzedine Aftiss, Salima Lamsiyah, Christoph Schommer, and Said Ouatik El Alaoui. 2025. Empirical evaluation of pre-trained language models for summarizing Moroccan Darija news articles. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 77–85, Abu Dhabi, UAE. Association for Computational Linguistics.

Hend S. Al-Khalifa and Amani A. Al-Ajlan. 2010. Automatic readability measurements of the arabic text: An exploratory study. *Arabian Journal for Science and Engineering*, 35(2C):103–124.

Muhamed Al Khalil, Nizar Habash, and Zhengyang Jiang. 2020. A large-scale leveled readability lexicon for standard Arabic. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3053–3062, Marseille, France. European Language Resources Association.

Muhamed Al Khalil, Hind Saddiki, Nizar Habash, and Latifa Alfalasi. 2018. A leveled reading corpus of modern standard Arabic. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association.

Faisal Al-Shargi, Ayal Kaplan, Ramy Eskander, Nizar Habash, and Owen Rambow. 2016. Morphologically annotated corpora and morphological analyzers for Moroccan and Sanaani Yemeni Arabic. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1300–1306, Portorož, Slovenia. European Language Resources Association.

Abdel-Karim Al-Tamimi, Manar Jaradat, Nuha Aljarrah, and Sahar Ghanim. 2014. AARI: Automatic arabic readability index. *The International Arab Journal of Information Technology*, 11(4):370–378.

Bashar Alhafni, Reem Hazim, Juan David Pineros Liberato, Muhamed Al Khalil, and Nizar Habash. 2024. The SAMER Arabic text simplification corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16079–16093, Torino, Italia. ELRA and ICCL.

Sultan Almujaiwel, Damith Premasiri, Tharindu Ranasinghe, Mo El-Haj, and Ruslan Mitkov. 2025. Complex concept-based readability estimation from arabic curriculum. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(11):1–21.

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9, Los Angeles, California. Association for Computational Linguistics.

Fakhraddin Alwajih, Abdellah El Mekki, Samar Mohamed Magdy, AbdelRahim A. Elmadany, Omer Nacar, El Moatez Billah Nagoudi, Reem Abdel-Salam, Hanin Atwany, Youssef Nafea, Abdulfattah Mohammed Yahya, Rahaf Alhamouri, Hamzah A. Alsayadi, Hiba Zayed, Sara Shatnawi, Serry Sibaee, Yasir Ech-chammakhy, Walid Al-Dhabyani, Marwa Mohamed Ali, Imen Jarraya, Ahmed Oumar El-Shangiti, Aisha Alraeesi, Mohammed Anwar AL-Ghrawi, Abdulrahman S. Al-Batati, Elgizouli Mohamed, Noha Taha Elgindi, Muhammed Saeed, Houdaifa Atou, Issam Ait Yahia, Abdelhak Bouayad, Mohammed Machrouh, Amal Makouar, Dania Alkawi, Mukhtar Mohamed, Safaa Taher Abdelfadil, Amine Ziad Ounnoughene, Anfel Rouabhia, Rwaa Assi, Ahmed Sorkatti, Mohamedou Cheikh Tourad, Anis Koubaa, Ismail Berrada, Mustafa Jarrar, Shady Shehata, and Muhammad Abdul-Mageed. 2025. Palm: A culturally inclusive and linguistically diverse dataset for Arabic LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32871–32894, Vienna, Austria. Association for Computational Linguistics.

Jonathan Anderson. 1983. Lix and rix: Variations on a little-known readability index. *Journal of Reading*, 26(6):490–496.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools*, pages 9–15, Marseille, France. European Language Resources Association.

Mohamed Anwar, Abdelhakim Freihat, George Ibrahim, Mostafa Awad, Abdelrahman Atef Mohamed Ali Sadallah, Gurpreet Gosal, Gokul Ramakrishnan, Sarath Chandran, Biswajit Mishra, Rituraj Joshi, Ahmed Frikha, Etienne Goffinet, Abhishek Maiti, Ali El Filali, Sarah Al Barri, Samujjwal Ghosh, Rahul Pal, Parvez Mullah, Awantika Shukla, Sajid Siddiki, Samta Kamboj, Onkar Pandit, Sunil Sahu, Abelrahman El Badawy, Amr Mohamed, Ahmad Chamma, Evan Dufraisse, Abdelaziz Bounhar, Dani Bouch, Hadi Abdine, Guokan Shang, Fajri Koto, Yuxia Wang, Zhuohan Xie, Ali Mekky, Rania Hossam Elmohamady Elbadry, Sarfraz Ahmad, Momina Ahsan, Omar Emad Mohamed El-Herraoui, Daniil Orel, Hasan Iqbal, Kareem Mohamed Naguib Abdelmohsen Fahmy Elzeky, Mervat Abassy, Kareem Ali, Saadeldine Eletter, Farah Atif, Nurdaulet Mukhituly, Haonan Li, Xudong Han, Aaryamonvikram Singh, Zain Quraishi, Neha Sengupta, Larry Murray, Avraham Sheinin, Joel Hestness, Natalia Vassilieva, Hector Ren, Zhengzhong Liu, Michalis Vazirgiannis, and Preslav Nakov. 2025. Jais 2: A family of Arabic-centric open large language models. Technical report, IFM.

Ibtehal Baazeem, Hend Al-Khalifa, and Abdulmalik Al-Salman. 2025. AraEyebility: Eye-tracking data for arabic text readability. *Computation*, 13(5):108.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham Abdullah Alyahya, Sultan AlRashed, Faisal Abdulrahman Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabdul-jabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Saad Amin Hassan, Dr. Majed Alrubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, Ahmed Abdelali, Jeril Kuriakose, Abdalghani Abujabal, Nora Al-Twairesh, Areeb Alowisheq, and Haidar Khan. 2025. ALLam: Large language models for arabic and english. In *The Thirteenth International Conference on Learning Representations*.

Ayoub Ben Cheikhi and El Habib Nfaoui. 2026. MSA-Moroccan dialect: A multimodal sentiment analysis dataset for Moroccan Arabic (Darija).

Carl-Hugo Björnsson. 1968. *Läsbarhet*. Liber, Stockholm.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association.

ElMehdi Boujou, Hamza Chataoui, Abdellah El Mekki, Saad Benjelloun, Ikram Chairi, and Ismail Berrada. 2021. An open access NLP dataset for arabic dialects: Data collection, labeling, and model construction. *arXiv preprint arXiv:2102.11000*.

Abdelaziz Bounhar. 2025. Atlaset: A moroccan darija text corpus. Hugging Face dataset.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Violetta Cavalli-Sforza, Hind Saddiki, and Naoual Nassiri. 2018. Arabic readability research: Current state and future directions. *Procedia Computer Science*, 142:38–49.

Salmane Chafik, Saad Ezzini, and Ismail Berrada. 2025. Dialect2SQL: A novel text-to-SQL dataset for Arabic dialects with a focus on Moroccan Darija. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 86–92, Abu Dhabi, UAE. Association for Computational Linguistics.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

David R Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 20(2):215–232.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Harris Drucker, Christopher J Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. 1996. Support vector regression machines. *Advances in neural information processing systems*, 9.

Mahmoud El-Haj and Paul Rayson. 2016. OSMAN: A novel arabic readability metric. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 250–255, Portorož, Slovenia. European Language Resources Association.

Mo El-Haj, Sultan Almujaiwel, Damith Premasiri, Tharindu Ranasinghe, and Ruslan Mitkov. 2024. DARES: Dataset for Arabic readability estimation of school materials. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context at LREC-COLING 2024*, pages 103–113, Torino, Italia. ELRA and ICCL.

Abdellah El Mekki, Houdaifa Atou, Omer Nacar, Shady Shehata, and Muhammad Abdul-Mageed. 2025. NileChat: Towards linguistically diverse and culturally aware LLMs for local communities. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10967–10991, Suzhou, China. Association for Computational Linguistics.

Abdellah EL Mekki, Samar M. Magdy, Houdaifa Atou, Ruwa AbuHweidi, Baraah Qawasmeh, Omer Nacar, Thikra Al-hibiri, Razan Saadie, Hamzah A. Alsayadi, Nadia Ghezaiel Hammouda, Alshima Mohammed Alkhazimi, Aya Hamod, Al-Yas Yaqoob Al-Ghafri, Wesam El-Sayed, Asila Ismail al Sharji, Mohamad Ballout, Anas Belfathi, Karim Ghaddar, Serry Sibaee, Alaa Aoun, Aeej Mohammed Aseri, Lina Abureesh, Ahlam Bashiti, Majdal Yousef, Abdulaziz Hafiz, Yehdih Mohamed, Emira Hamedtou, Brakehe Emehah, Rahaf Alhamouri, Youssef Nafea, Aya El Aatar, Walid Al-Dhabyani, Emhemed S. Hamed, Sara Shatnawi, Fakhraddin Alwajih, Khalid Elkhidir, Ashwag

Alasmari, Abdurrahman Gerrio, Omar Said Alshahri, AbdelRahim A. Elmadany, Ismail Berrada, Amir Azad Adli Al-kathiri, Fadi Zaraket, Mustafa Jarrar, Yahya Mohamed EL Hadj, Hassan Alhuzali, and Muhammad Abdul-Mageed. 2026. Alexandria: A multi-domain dialectal Arabic machine translation dataset for culturally inclusive and linguistically diverse LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32567–32592, San Diego, California, United States. Association for Computational Linguistics.

Khalid N. Elmadani, Nizar Habash, and Hanada Taha-Thomure. 2025. A large and balanced corpus for fine-grained Arabic readability assessment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16376–16400, Vienna, Austria. Association for Computational Linguistics.

Abdeljalil Elouardighi et al. 2017. ElecMorocco2016: A sentiment dataset of moroccan election facebook comments. Public dataset release. Associated with work on sentiment analysis of MSA and Moroccan Arabic Facebook comments.

Kabil Essefar, Hassan Ait Baha, Abdelkader El Mahdaouy, Abdellah El Mekki, and Ismail Berrada. 2023. OMCD: Offensive moroccan comments dataset. *Language Resources and Evaluation*, 57(4):1745–1765.

Fanar Team. 2025. Fanar: An arabic-centric multimodal generative AI platform. *arXiv preprint arXiv:2501.13944*.

Kamel Gaanoun, Abdou Mohamed Naira, Anass Allak, and Imade Benelallam. 2022. Automatic text summarization for Moroccan Arabic dialect using an artificial intelligence approach. In *Business Intelligence*, Lecture Notes in Business Information Processing. Springer.

Kamel Gaanoun, Abdou Mohamed Naira, Anass Allak, and Imade Benelallam. 2025. DarijaBERT: A step forward in NLP for the written moroccan dialect. *International Journal of Data Science and Analytics*, 20(2):917–929.

Moncef Garouani and Jamal Kharroubi. 2021. MAC: An open and free Moroccan Arabic corpus for sentiment analysis. In *Innovations in Smart Cities Applications Volume 5*. Springer. Corpus of approximately 18,000 manually labelled tweets.

Gemma Team. 2025. Gemma 3 technical report. Technical report, Google DeepMind.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Nizar Habash, Hanada Taha-Thomure, Khalid N. Elmadani, Zeina Zeino, and Abdallah Abushmaes. 2025. Guidelines for fine-grained sentence-level Arabic readability annotation. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 359–376, Vienna, Austria. Association for Computational Linguistics.

Anass Ibrahimi and Asmaa Mourhir. 2024. Moroccan darija dataset for offensive language and sentiment analysis.

Joseph Marvin Imperial and Ekaterina Kochmar. 2023. Basahacorpus: An expanded linguistic resource for readability assessment in central philippine languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6302–6309.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Abderrahmane Issam and Khalil Mrini. 2022. Goud.ma: A news article dataset for summarization in moroccan darija. In *Proceedings of the Third Workshop on African Natural Language Processing*.

Mouad Jbel, Imad Hafidi, and Abdelmoutalib Metrane. 2023. MYC: A Moroccan corpus for sentiment analysis. In *Advances in Machine Intelligence and Computer Science Applications*, pages 59–68. Springer.

Nouran Khallaf and Serge Sharoff. 2021. Automatic difficulty classification of Arabic sentences. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 105–114, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Soukaina Mihi, Brahim Ait Ben Ali, Ismail El Bazi, Sara Arezki, and Nabil Laachfoubi. 2020. MSTD: Moroccan sentiment twitter dataset. *International Journal of Advanced Computer Science and Applications*, 11(10):363–372.

Imane Momayiz, Aissam Outchakoucht, Omar Choukrani, and Ali Nirheche. 2024. TerjamaBench: A culturally specific dataset for evaluating translation models for moroccan darija. Hugging Face dataset.

Hanane Nour Moussa and Asmaa Mourhir. 2023. DarNERcorp: An annotated named entity recognition dataset in the Moroccan dialect. *Data in Brief*, 48:109234.

Khalil Mrini and Francis Bond. 2017. Building the Moroccan Darija Wordnet (MDW) using bilingual resources. In *Proceedings of the International Conference on Natural Language, Signal and Speech Processing*, pages 171–182, Casablanca, Morocco.

Tarek Naous, Michael J. Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. 2024. ReadMe++: Benchmarking multilingual language models for multi-domain readability assessment. In *Proceedings of*

the 2024 Conference on Empirical Methods in Natural Language Processing, pages 12230–12266, Miami, Florida, USA. Association for Computational Linguistics.

Naoual Nassiri, Violetta Cavalli-Sforza, and Abdelhak Lakhouaja. 2019. MoSAR: Modern standard arabic readability corpus for L1 learners. In *Proceedings of the Fourth International Conference on Big Data and Internet of Things*. Association for Computing Machinery.

oddadmix. 2026. WAK3I: Moroccan arabic audio and transcript collection. Hugging Face dataset. Accessed 17 July 2026.

Ahmed Oussous, Fatima-Zahra Benjelloun, Ayoub Ait Lahcen, and Samir Belfkih. 2020. ASA: A framework for arabic sentiment analysis. *Journal of Information Science*, 46(4):544–559.

Aissam Outchakoucht and Hamza Es-Samaali. 2021. Moroccan dialect-darija-open dataset. *arXiv preprint arXiv:2103.09687*.

Aissam Outchakoucht and Hamza Es-Samaali. 2024. The evolution of Darija open dataset: Introducing version 2. *arXiv preprint arXiv:2405.13016*.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Guokan Shang, Hadi Abdine, Yousef Khoubrane, Amr Mohamed, Yassine Abbahaddou, Sofiane Ennadir, Imane Momayiz, Xuguang Ren, Eric Moulines, Preslav Nakov, Michalis Vazirgiannis, and Eric Xing. 2025. Atlas-chat: Adapting large language models for low-resource Moroccan Arabic dialect. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 9–30, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abderrahman Skiredj, Ferdaous Azhari, Houdaifa Atou, Nouamane Tazi, and Ismail Berrada. 2025a. GemMaroc: Unlocking darija proficiency in LLMs with minimal data. *arXiv preprint arXiv:2505.17082*.

Abderrahman Skiredj, Ferdaous Azhari, Ismail Berrada, and Saad Ezzini. 2025b. DarijaBanking: A new resource for overcoming language barriers in banking intent detection for Moroccan Arabic speakers. *Natural Language Processing*.

Edgar A. Smith and R. J. Senter. 1967. Automated readability index. Technical report, Aerospace Medical Research Laboratories, Wright-Patterson Air Force Base, Ohio.

Rainer Storn and Kenneth Price. 1997. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341–359.

Rachid Tachicart, Karim Bouzoubaa, and Hamid Jaafar. 2014. Building a Moroccan dialect electronic dictionary (MDED). In *Proceedings of the Fifth International Conference on Arabic Language Processing (CITALA)*, Oujda, Morocco. Electronic lexical resource for Moroccan Arabic.

Sowmya Vajjala. 2022. Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 5366–5377.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Annotation Guidelines

A.1 Readability Level Definitions

Immediate (DR1). The sentence expresses one concrete and explicit message using familiar vocabulary and a straightforward structure. It can normally be understood immediately without rereading or additional context.

Accessible (DR2). The sentence remains easy to understand but contains modest complexity, such as two closely related ideas or a simple temporal, causal, conditional, or sequential relation. Its vocabulary and structure remain largely familiar.

Demanding (DR3). The sentence requires careful reading because it contains several propositions, a developed explanation or viewpoint, less familiar vocabulary, contrast, reference tracking, or moderate abstraction. The intended meaning remains recoverable from the sentence.

Advanced (DR4). The sentence requires substantial effort because of one or more major sources of difficulty, such as specialised or rare vocabulary, dense or embedded structure, figurative language, implicit meaning, cultural knowledge, or an unfamiliar written form.

A.2 Diagnostic Dimension Definitions

Each diagnostic dimension was scored from 0 to 3. A score of 0 indicates no meaningful difficulty, 1 indicates a minor difficulty that does not obstruct understanding, 2 indicates a moderate difficulty that requires noticeable effort, and 3 indicates a major difficulty that substantially affects understanding.

Orthography. Orthographic difficulty concerns recognition and decoding of the written form. It includes unclear segmentation, severe or unfamiliar spelling, abbreviations, and other written forms that impede recognition. Ordinary variation in Darija spelling was not treated as difficult unless it genuinely increased decoding effort.

Vocabulary. Vocabulary difficulty concerns word and expression familiarity. Annotators considered everyday frequency, register, specialisation, literary usage, and whether an unfamiliar expression could be inferred from context.

Syntax. Syntactic difficulty concerns the effort required to connect words, clauses, propositions, and references. Annotators considered coordination, subordination, embedding, long-distance dependencies, and reference resolution rather than length alone.

Meaning. Semantic difficulty concerns the effort required to construct the intended interpretation. It includes abstraction, implicit relations, inference, figurative meaning, cultural references, and background knowledge.

B Annotation Statistics

Tables 8 and 9 report the independent suitability decisions made before readability labelling. Both annotators reviewed the same 4,000-item candidate pool. Figure 4 complements the agreement matrix in the main paper by showing the ordinal distance between paired readability labels for the 2,962 items retained by both annotators.

Annotator 1 retained 3,452 candidates and Annotator 2 retained 3,074. Language issues, particularly MSA-heavy material in GOUD, were the most frequent exclusions. The intersection of their suitability decisions contained 2,962 sentences. Within this common set, 68.5% received the same readability label, 30.9% differed by one level, and approximately 0.5% differed by two or more levels.

Figure 5 shows the directional agreement between the two annotators. From Annotator A's perspective (i.e., with percentages normalised over the labels assigned by Annotator A), 46% of sentences labelled *Immediate* were labelled *Accessible* by Annotator B, whereas from Annotator B's perspective (i.e., with percentages normalised over the labels assigned by Annotator B), 93% of *Immediate* labels were also assigned as *Immediate* by Annotator A. A similar asymmetry appears at the upper end of the scale, where 44% of sentences labelled *Advanced* by Annotator B were labelled *Demanding* by Annotator A. These patterns reflect differences in how the annotators applied the boundaries between adjacent readability levels.

Table 10 provides a detailed breakdown of the seven source subsets that comprise BAYAN, including the number of retained sentences, register, provenance, and source description.

C Additional Corpus Analysis

C.1 Sentence Length

Figure 6a shows how the readability distribution changes as sentence length increases. Short sen-

Table 8: Annotation outcomes for Annotator 1 by source subset. *Annotated* counts items retained as suitable; the remaining rows are exclusion flags.

Flag	Alex.	DODa	Goud	MSM-MG	Palm	WAK3I	Zajal	All
Annotated	510	467	642	406	398	835	194	3452
Corrupted	1	0	0	0	0	2	0	3
Incomplete	1	0	3	5	0	3	0	12
Insufficient context	0	1	3	1	1	4	0	10
Language issue	0	1	240	9	1	1	4	256
Inappropriate	1	1	49	35	1	27	0	114
Spelling issues	2	25	1	7	3	19	1	58
Other	8	8	19	20	5	33	2	95
Total	**523**	**503**	**957**	**483**	**409**	**924**	**201**	**4000**

Table 9: Annotation outcomes for Annotator 2 by source subset.

Flag	Alex.	DODa	Goud	MSM-MG	Palm	WAK3I	Zajal	All
Annotated	512	439	504	330	406	690	193	3074
Corrupted	0	1	7	8	0	2	0	18
Incomplete	0	2	10	12	0	20	0	44
Insufficient context	0	0	3	0	0	3	0	6
Language issue	4	6	341	25	0	7	3	386
Inappropriate	0	7	46	45	0	59	4	161
Spelling issues	5	46	16	43	1	125	0	236
Other	2	2	30	20	2	18	1	75
Total	**523**	**503**	**957**	**483**	**409**	**924**	**201**	**4000**

Table 10: Composition of BAYAN. N is the number of items retained after sampling and annotation, following agreement between the two annotators.

Subset	N	Register	Provenance	Description
GOUD[1]	354	Journalistic	Native written	News article sentences from Goud-sum (Issam and Mrini, 2022).
ALEXANDRIA[2]	347	Conversational	Translated	Human Moroccan Darija translations of multi-domain conversations (EL Mekki et al., 2026).
DODA[3]	295	Lexical	Native written	Example sentences from the Darija Open Dataset (Outchakoucht and Es-Samaali, 2021; Shang et al., 2025).
PALM[4]	246	Figurative	Native written	Moroccan items from the Language and Proverbs categories (Alwajih et al., 2025).
MSM-MG[5]	239	Informal	Native written	Original user-authored Moroccan social-media posts (Shang et al., 2025).
WAK3I[6]	439	Spoken narrative	Native spoken	Sentences from an audio–transcript collection (oddadmix, 2026).
ZAJAL	110	Poetic	Native written	Darija poetry collected from public Moroccan poetry forums.
Total	**2,030**			

[1] `Goud/Goud-sum` [2] `UBC-NLP/alexandria` [3] `MBZUAI-Paris/DODa-10K` [4] `UBC-NLP/palm`
[5] `MBZUAI-Paris/MoroccanSocialMedia-MultiGen`
[6] `oddadmix/arabic-audio-collection-moroccan-wak3i`

tences are predominantly labelled as *Immediate* or *Accessible*, whereas longer sentences are increasingly assigned to the *Demanding* and *Advanced* levels. Nevertheless, all four levels occur across a wide range of lengths.

Figure 6b shows the distribution of sentence lengths across the seven source subsets. The sources exhibit distinct length profiles. DODA and MSM-MG contain predominantly short sentences, whereas GOUD and WAK3I include a broader range of sentence lengths.

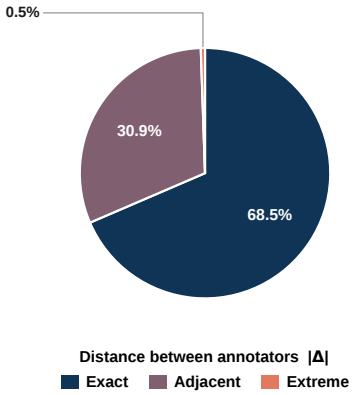


Figure 4: Distribution of ordinal distances between the two readability labels for sentences retained by both annotators.

C.2 BAYANEX

Figure 8 presents the readability distribution of BAYANEX. The distribution is close to that of the full corpus, indicating that the subset is broadly representative of the overall readability distribution.

C.3 Source–Readability Distribution

Figure 7 shows the relationship between corpus source and readability level from two complementary perspectives. Figure 7a shows the source composition within each readability level. *Immediate* and *Accessible* sentences come largely from DODA and ALEXANDRIA, whereas *Demanding* and especially *Advanced* sentences have greater representation from sources such as GOUD, WAK3I, PALM, and ZAJAL. Figure 7b shows the readability distribution within each source. DODA and ALEXANDRIA are dominated by the *Immediate* and *Accessible* levels, whereas PALM, MSM-MG, WAK3I, GOUD, and especially ZAJAL contain larger proportions of *Demanding* and *Advanced* sentences. Overall, the figure shows substantial differences in readability profiles across sources, while no source is associated with a single readability level.

C.4 Annotation Pipeline and Examples

Figure 9 summarises the flow of samples through the annotation pipeline, from the source corpora to the final readability labels. Filtering outcomes for each annotator and source subset are reported in Tables 8 and 9.

C.5 Diagnostic Dimensions

Figure 11 shows how the four diagnostic dimensions vary across the four readability levels. Orthographic difficulty remains generally low and does not follow a clear monotonic progression. Most sentences have an orthographic score of 0 at all levels. In contrast, vocabulary and syntactic difficulty show a clear progression with readability. Vocabulary shifts from predominantly score 0 for *Immediate* sentences to score 1 for *Accessible*, score 2 for *Demanding*, and scores 2–3 for *Advanced*. A similar pattern is observed for syntax.

Semantic difficulty provides the clearest separation between the four levels. A score of 0 accounts for 95% of *Immediate* sentences, while score 1 dominates *Accessible* sentences (78%), score 2 dominates *Demanding* sentences (72%), and score 3 accounts for 90% of *Advanced* sentences. These distributions suggest that semantic difficulty is the dimension most closely aligned with the overall readability progression, followed by syntactic and vocabulary difficulty.

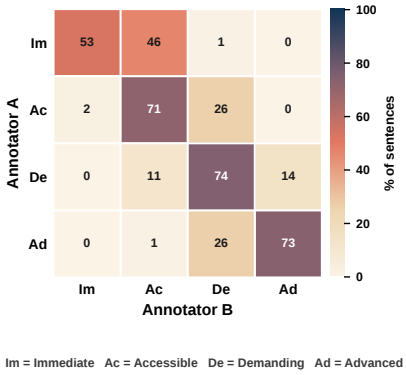
C.6 Sentence Length and Readability

Figure 12 shows the relationship between sentence length and readability within each source. The association is positive across all seven sources, with Spearman correlations ranging from $\rho = 0.30$ for ZAJAL to $\rho = 0.80$ for ALEXANDRIA. Relatively strong associations are also observed for WAK3I ($\rho = 0.75$), DODA ($\rho = 0.74$), and GOUD ($\rho = 0.72$). Although sentence lengths differ across sources, the within-source trends indicate that the overall association between length and readability cannot be explained solely by source or register.

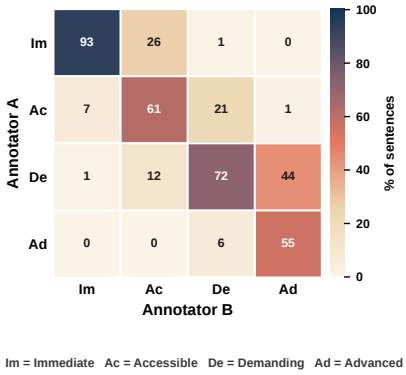
To further examine this relationship, we report the correlations between sentence length and the four diagnostic dimensions in BAYANEX. Sentence length is positively associated with all four dimensions, with the strongest association observed for syntactic difficulty ($\rho = 0.866$), followed by vocabulary ($\rho = 0.683$), meaning ($\rho = 0.494$), and orthography ($\rho = 0.432$). The tendency of longer sentences to receive higher readability levels may partly reflect the higher syntactic and lexical difficulty associated with longer sentences. Thus, although additional context may facilitate interpretation, longer sentences can simultaneously introduce more complex linguistic structure.

C.7 Data Splits

Table 11 reports the train, validation, and test splits used throughout the experiments. The splits preserve the overall readability distribution, with

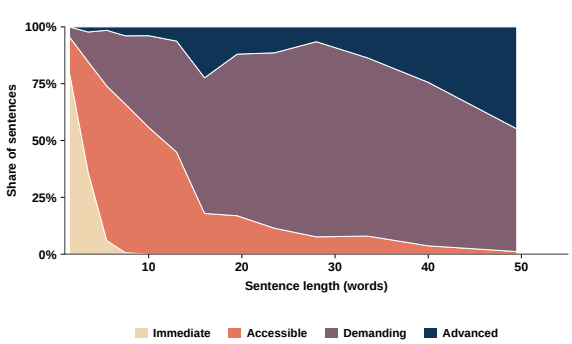


(a) Agreement normalised by Annotator A's labels.

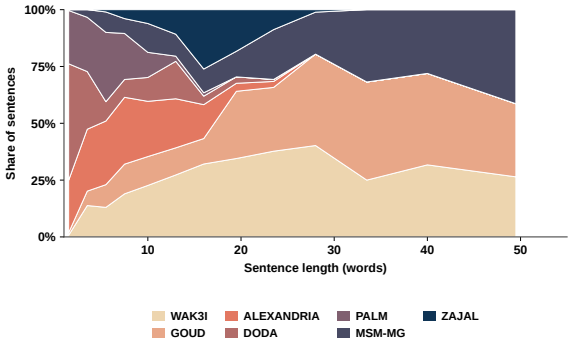


(b) Agreement normalised by Annotator B's labels.

Figure 5: Directional agreement between the two annotators. Each matrix shows the distribution of one annotator's labels relative to the labels assigned by the other annotator.

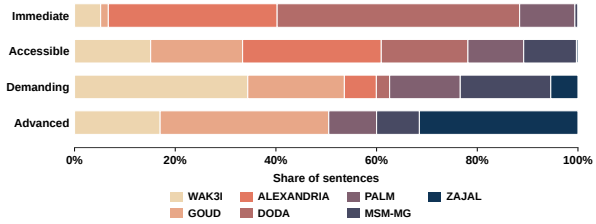


(a) By readability level.

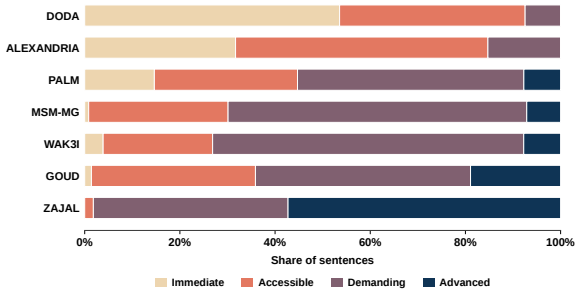


(b) By source subset.

Figure 6: Sentence-length distributions in BAYAN.



(a) Source distribution within each readability level.



(b) Readability-level distribution within each source.

Figure 7: Source composition across readability levels and distribution of readability levels across sources.

Demanding and *Accessible* remaining the largest classes and *Advanced* the smallest. This distribution provides a consistent evaluation setting across all experiments.

D Classical Readability Features

We evaluate two sets of hand-crafted features for the classical machine-learning models: *formulas* and *surface* features. All features are computed independently for each sentence after Unicode normalisation and whitespace tokenisation. Leading and trailing punctuation is removed for word-level calculations. Dense features are standardised using statistics computed on the training split for scale-sensitive models, whereas tree-based models operate on the original feature values.

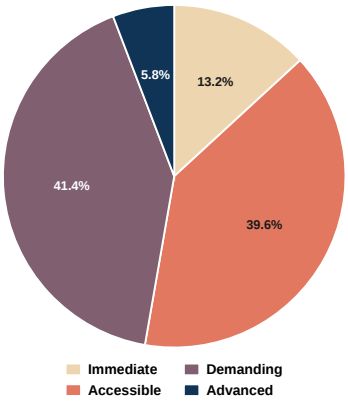


Figure 8: Readability distribution of BAYANEX.

Table 11: Distribution of readability levels across data splits.

Split	DR1	DR2	DR3	DR4	Total
Train	229	467	583	140	1,419
Val	33	67	84	20	204
Test	66	134	167	40	407
Total	**328**	**668**	**834**	**200**	**2,030**

D.1 Readability Formulae

Let C be the number of non-space characters, W the number of whitespace-delimited words, S the number of sentences, and LW the number of words containing more than six characters. Because every instance contains one sentence, $S = 1$.

Automated Readability Index.

$$\mathrm{ARI} = 4.71\left(\frac{C}{W}\right) + 0.5\left(\frac{W}{S}\right) - 21.43. \quad (2)$$

The formula follows Smith and Senter (1967).

LIX.

$$\mathrm{LIX} = \frac{W}{S} + 100\left(\frac{LW}{W}\right). \quad (3)$$

The formula follows Björnsson (1968).

RIX.

$$\mathrm{RIX} = \frac{LW}{S}. \quad (4)$$

The formula follows Anderson (1983).

Al-Heeti.

$$\mathrm{AlHeeti} = 4.414\left(\frac{C}{W}\right) - 13.468. \quad (5)$$

The formula is reported in Arabic readability reviews and comparative studies (Cavalli-Sforza et al., 2018; Al-Tamimi et al., 2014).

Automatic Arabic Readability Index.

$$\mathrm{AARI}_{\mathrm{Base}} = 3.28\,NOC + 1.43\,ACW + 1.24\,AWS, \quad (6)$$

where NOC is the number of characters, ACW is the average number of characters per word, and AWS is the average number of words per sentence (Al-Tamimi et al., 2014).

The formula outputs are treated as continuous numerical values. Their original grade interpretations are not used, and their three decision thresholds are learned from the training partition.

D.2 Surface Feature Inventory

The surface representation contains 17 language-independent features: two sentence-length features, three word-length features, two long-word features, two lexical-variation features, four punctuation features, two digit-and-symbol features, and two elongation features. The *surface + formulas* condition concatenates these features with the five formula scores.

E Experimental Details and Additional Results

E.1 Implementation Details

All transformer encoders were fine-tuned using PyTorch and the Hugging Face Transformers library (Wolf et al., 2020). We trained each model for a maximum of 50 epochs with early stopping using a patience of 5 epochs, AdamW, a learning rate of 2×10^{-5}, a batch size of 16, a maximum sequence length of 128 tokens, and a linear schedule with warmup. All experiments were conducted on a single NVIDIA A100 GPU. For linear probing, the encoder parameters were frozen and only the classification head was trained using a learning rate of 10^{-3}. Each trainable experiment was repeated using five random seeds: 0, 42, 2026, 123, and 1337. We report the mean and standard deviation across runs. Because of inference cost, each LLM prompting condition was evaluated in one fixed run.

E.2 Threshold Learning for Readability Formulae

For each readability formula, we first computed the Spearman correlation between the continuous score and the gold readability labels on the training partition. If the correlation was negative, the scores

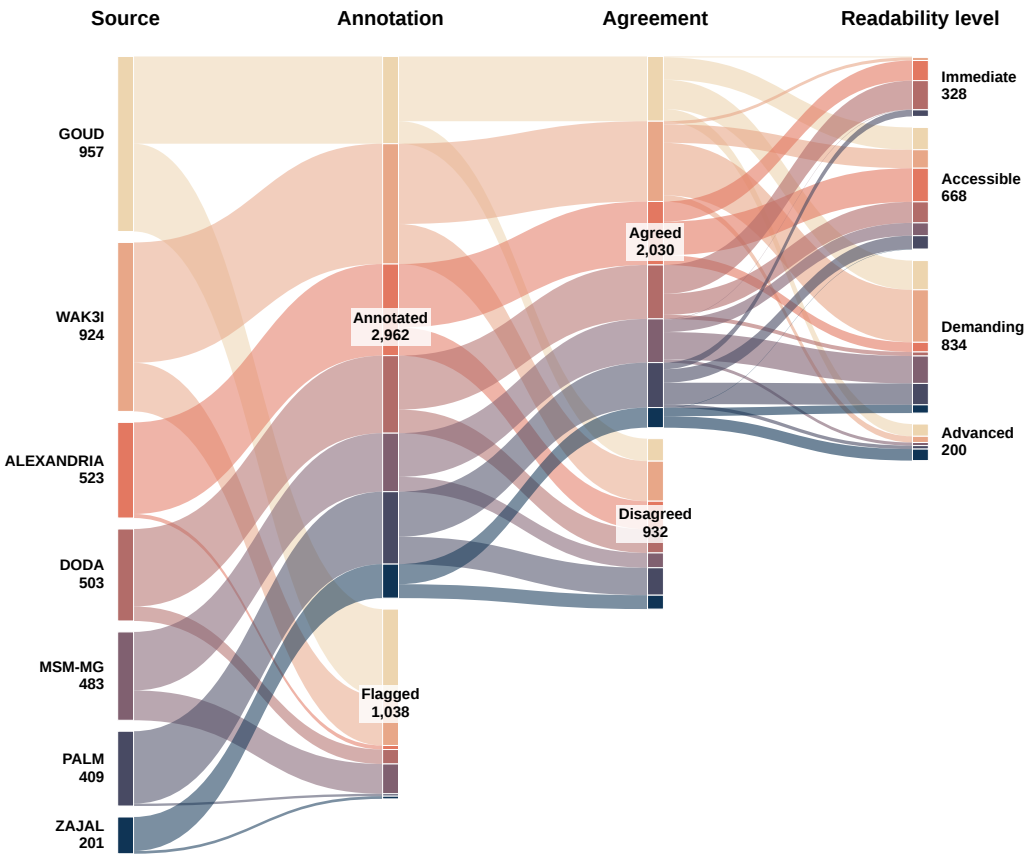
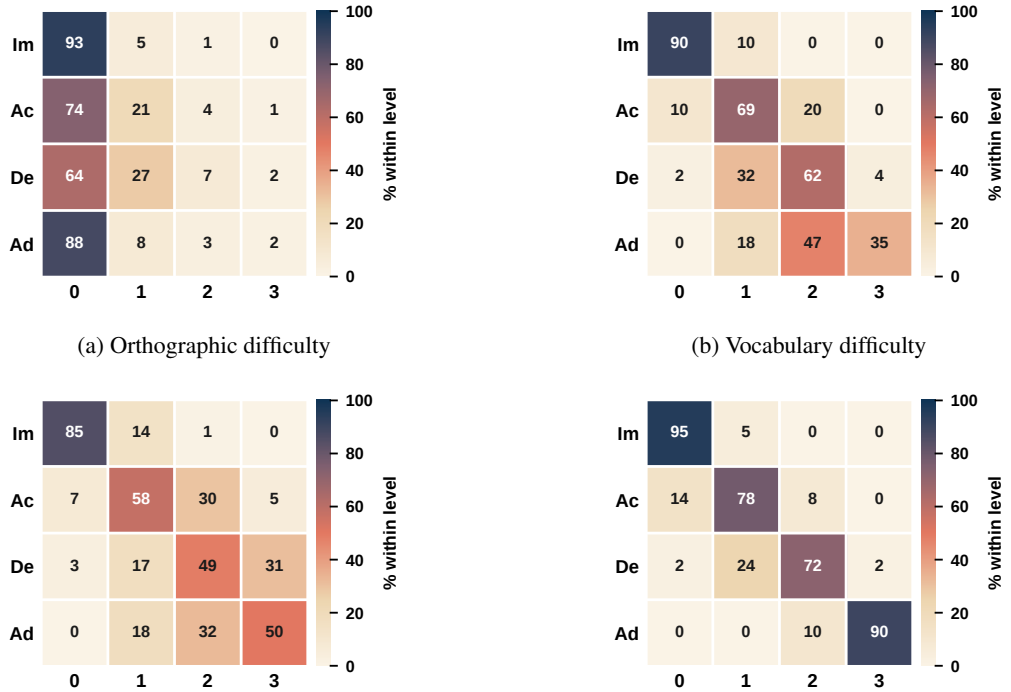


Figure 9: Flow from the 4,000 sampled candidates through suitability review and agreement filtering to BAYAN and BAYANEX.

LEVEL	SENTENCE · ENGLISH TRANSLATION	SOURCE
Immediate DR1	قرّيني *Teach me.*	DODA
Accessible DR2	كل وحدة سقيتها بصطل عامر ديال لما. *I watered each one with a bucket full of water.*	ALEXANDRIA
Demanding DR3	وبزااف السيد عمره مايحط عليه خبر بحال الناس ولكن فاش يكون شي فيديو دغيا يهزه ويحطه فقناته *And often the guy never posts any news about it like people do, but when there is a video he quickly grabs it and puts it up on his channel.*	MSM-MG
Advanced DR4	بقيت كنحسب ليام و كنتسنى شي نهار تبان، شي نقرة فين يغبر نحاسي، ولا شي ذهب فين يغبر جالوقي، حتى لواحد النهار من اليامات ديال سيدي ربي، كان حدانا واحد الكاراج ديال الميكانيسيان، عمرني مرديت ليه البال آش فيه ولا أش من طوموبيل واقفة ولا آش من معلم خدام ولا والو *I kept counting the days, waiting for one to finally come — some silver to bury my misfortune, or some gold to bury my rags — until, on one of the good Lord's days: right next to us stood a mechanic's garage I had never paid any mind — what was inside, which cars were parked there, which craftsman was at work, nothing at all.*	WAK3I

Figure 10: Representative examples of the four readability levels, accompanied by translations.



(a) Orthographic difficulty

(b) Vocabulary difficulty

(c) Syntactic difficulty

(d) Semantic difficulty

Figure 11: Distribution of difficulty-dimension scores across the four readability levels in BAYANEX. Each cell shows the percentage of pooled annotator ratings assigned a score from 0 to 3 within each readability level. Im, Ac, De, and Ad denote *Immediate*, *Accessible*, *Demanding*, and *Advanced*, respectively.

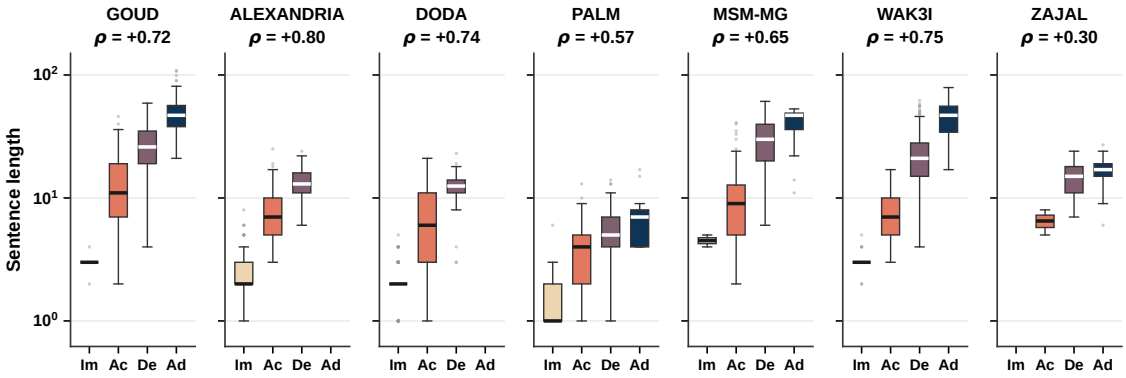


Figure 12: Sentence-length distributions across readability levels within each source.

were multiplied by -1 so that higher scores consistently indicated greater reading difficulty. We then learned three ordered thresholds on the training partition by maximising QWK using differential evolution (Storn and Price, 1997) with a population size of 20, a maximum of 1,000 iterations, and a random seed of 42. The resulting thresholds are presented in Table 13. For Al-Heeti, the score was negatively correlated with the Bayan readability labels and was therefore sign-reversed before threshold optimisation, which explains the descending order of the reported thresholds.

E.3 Prompt Structure and Checkpoints

All models used the same instruction template, shown in Figure 13. It contained the task description, definitions of the four readability levels, and an instruction to return exactly one label from `DR1`, `DR2`, `DR3`, or `DR4`. For four-shot prompting, we sampled one labelled demonstration from each readability level, while for eight-shot prompting, we sampled two demonstrations per level. Demonstrations were drawn from the validation set using a fixed random seed (42).

All LLM experiments used greedy decoding

Table 12: Surface features used for sentence-level Darija readability assessment. All features are extracted from the raw sentence and require no morphological analyser, syntactic parser, frequency lexicon, or syllable counter.

Feature	Definition or calculation	Readability relevance
Sentence length		
Number of words	Number of whitespace-separated tokens after punctuation trimming.	Longer sentences may contain more information and relations.
Number of characters	Number of non-space characters after punctuation trimming.	Measures length independently of token count.
Word length		
Average word length	Total token characters divided by the number of words.	Longer written forms may increase lexical or decoding effort.
Maximum word length	Character length of the longest token.	Identifies an unusually long form.
Standard deviation of word length	Standard deviation of token lengths.	Measures variation between short and long forms.
Long words		
Number of words longer than six characters	Count of tokens containing more than six characters.	Measures the absolute concentration of long words.
Proportion of words longer than six characters	Long-word count divided by word count.	Provides a length-normalised concentration of long words.
Lexical variation		
Number of unique words	Number of distinct lower-cased tokens.	Approximates the amount of different vocabulary.
Proportion of repeated words	Repeated token occurrences divided by word count.	Repetition may reduce lexical load.
Punctuation		
Total punctuation count	Number of punctuation marks.	May indicate segmentation, coordination, parenthetical information, or direct speech.
Comma count	Number of comma marks.	Multiple segments may reflect coordination or structural complexity.
Question mark indicator	1 when a question mark occurs, otherwise 0.	Marks interrogative structures and possible implied information.
Exclamation mark indicator	1 when an exclamation mark occurs, otherwise 0.	Captures emphatic sentence forms.
Digits and symbols		
Digit count	Number of digit characters.	Dates, quantities, and prices may affect processing.
Symbol count	Number of non-alphanumeric symbols excluding standard punctuation.	Captures currencies, notation, and abbreviations.
Orthographic elongation		
Repeated-character indicator	1 when any character occurs at least three times consecutively, otherwise 0.	Captures expressive or unusual spelling.
Number of elongated tokens	Number of tokens containing three or more identical consecutive characters.	Measures the extent of elongation.

Table 13: Score ranges and readability-level thresholds per formula.

Formula	Min	Max	t_1	t_2	t_3
AARI	15.4	1,662.4	56.9	154.4	670.3
LIX	1.0	117.2	3.7	11.9	56.0
ARI	-8.9	50.6	0.8	5.5	18.1
RIX	0.0	19.0	0.6	1.6	4.4
AlHeeti	-3.2	28.5	10.8	8.5	-2.5

(`do_sample=False`), a batch size of 8, and a maximum input length of 2,048 tokens, including any few-shot demonstrations. Table 14 lists the exact model identifiers.

E.4 Complete Classical Results

Table 15 presents the complete results for the classical machine-learning models. For each classifier, we report performance across all feature representations, including readability formulae, surface features, word and character TF–IDF, and their combinations. Formula and surface representations outperform the sparse TF–IDF conditions. Support Vector Regression with the five formula scores obtains the highest QWK, while Random Forest Regression with surface features obtains the highest macro F1.

```
Assign ONE overall readability level to the Darija sentence below.

## The four levels

- DR1 - Immediate: minimal effort. One simple, concrete, explicit message in highly familiar
  vocabulary and a simple structure. No rereading or background knowledge needed.
- DR2 - Accessible: low effort. Familiar everyday content that may connect two simple ideas
  through time, reason, purpose, condition or sequence. Mostly simple; the message stays clearly
  accessible.
- DR3 - Demanding: moderate effort. The reader must integrate several propositions, follow a
  developed explanation, viewpoint or contrast, or process moderately abstract content. The
  central meaning stays explicit and recoverable, though rereading may help.
- DR4 - Advanced: high effort from one major obstacle or several interacting difficulties:
  specialised or strongly formal vocabulary, dense embedding, strong abstraction, uncommon
  cultural knowledge, difficult figurative meaning, or serious but interpretable decoding
  ambiguity.

## Output format

Respond with a valid JSON object and nothing else:
{"level": "<DR1 | DR2 | DR3 | DR4>"}

## Example
Sentence: {{example sentence}}
Answer: {"level": "{{example level}}"}

## Sentence

{{sentence}}
```

Figure 13: Prompt template used for few-shot evaluation of LLMs.

E.5 Few-Shot LLM Results

Tables 16 and 17 report four-shot and eight-shot performance. Increasing the number of demonstrations does not consistently improve performance. Four-shot prompting generally performs worse than zero-shot prompting, with the exception of a small improvement for Gemma-3-12B-it. Eight-shot prompting improves Atlas-Chat-9B and Gemma-3-12B-it, but substantially reduces the performance of NileChat-3B. These results suggest that the effect of few-shot prompting is model dependent.

E.6 Confusion Matrix Analysis

Figure 14 compares representative confusion matrices for the strongest standalone readability formula (AARI), classical regression model (SVR), supervised encoder (MARBERTv2), and zero-shot LLM (NileChat-3B). MARBERTv2 produces the most balanced confusion matrix, while AARI and SVR more frequently classify *Advanced* sentences as *Demanding*. Across all systems, errors occur predominantly between adjacent readability levels. This behaviour is consistent with the ordinal structure of the task. The largest source of model error is the distinction between *Demanding* (DR3) and *Advanced* (DR4). This behaviour may be partly explained by the relatively small number of *Advanced* (DR4) examples in the training data, although the same pattern is also observed for the zero-shot LLM.

F Cross-Source Generalisation

Bayan combines material from heterogeneous sources, so a model trained on the full corpus could potentially exploit source-specific lexical or stylistic characteristics that correlate with readability levels. To study this effect, we evaluate the strongest supervised model, MARBERTv2, under a leave-one-source-out (LOSO) setting. In this setting, we remove all instances from each source from the training set, train the model on the remaining

Table 14: Model checkpoints. Short names are those used in the paper.

Short name	Hugging Face identifier	Reference
Encoders — general-purpose		
DarijaBERT	`SI2M-Lab/DarijaBERT`	(Gaanoun et al., 2025)
MARBERTv2	`UBC-NLP/MARBERTv2`	(Abdul-Mageed et al., 2021)
CAMeLBERT-DA	`CAMeL-Lab/bert-base-arabic-camelbert-da`	(Inoue et al., 2021)
ARBERTv2	`UBC-NLP/ARBERTv2`	(Abdul-Mageed et al., 2021)
AraBERTv02	`aubmindlab/bert-base-arabertv02`	(Antoun et al., 2020)
CAMeLBERT-MSA	`CAMeL-Lab/bert-base-arabic-camelbert-msa`	(Inoue et al., 2021)
AraBERTv2	`aubmindlab/bert-base-arabertv2`	(Antoun et al., 2020)
XLM-R_{base}	`FacebookAI/xlm-roberta-base`	(Conneau et al., 2020)
mBERT	`google-bert/bert-base-multilingual-cased`	(Devlin et al., 2019)
Encoders — readability-pretrained		
readabert-ar	`tareknaous/readabert-ar`	(Naous et al., 2024)
arabertv02-word-CE	`CAMeL-Lab/readability-arabertv02-word-CE`	(Elmadani et al., 2025)
LLMs — Darija-specific		
Atlas-Chat-9B	`MBZUAI-Paris/Atlas-Chat-9B`	(Shang et al., 2025)
NileChat-3B	`UBC-NLP/NileChat-3B`	(El Mekki et al., 2025)
GemMaroc-Qwen2.5-7B-darija	`GemMaroc/Qwen2.5-7B-Instruct-darija`	(Skiredj et al., 2025a)
LLMs — Arabic		
Fanar-1-9B-Instruct	`QCRI/Fanar-1-9B-Instruct`	(Fanar Team, 2025)
Jais-2-8B-Chat	`inceptionai/Jais-2-8B-Chat`	(Anwar et al., 2025)
ALLaM-7B-Instruct	`humain-ai/ALLaM-7B-Instruct-preview`	(Bari et al., 2025)
LLMs — general-purpose		
Gemma-3-12B-it	`google/gemma-3-12b-it`	(Gemma Team, 2025)
Qwen3.5-9B	`Qwen/Qwen3.5-9B`	(Qwen Team, 2026)
Llama-3.1-8B-Instruct	`meta-llama/Llama-3.1-8B-Instruct`	(Grattafiori et al., 2024)

sources, and evaluate it on the held-out source's original test set. The results in Table 18 compare the models on the same test instances. Thus, the test examples are identical between the two conditions.

The results show that performance decreases when the evaluated source is excluded from training, but the magnitude of the reduction varies across sources. For example, QWK decreases only slightly for MSM-MG (-0.027) and WAK3I (-0.049), while larger reductions occur for PALM (-0.282) and Zajal (-0.359). The drop for Zajal could be explained by its figurative and poetic material, which is less represented in the remaining sources and may make cross-source transfer more difficult. These results indicate some sensitivity to source-specific distributions, but the relatively strong LOSO performance on several sources shows that the model does not rely exclusively on source identity.

The magnitude of the LOSO reduction should not, however, be interpreted solely as the effect of source identity. Removing a source also reduces the amount of available training data and changes its distribution. In addition, source-specific estimates are based on test subsets of unequal size, ranging from 21 sentences for Zajal to 106 for WAK3I. The larger reductions for smaller subsets should therefore be interpreted cautiously.

G Resources Relevant to Darija Readability

Tables 21, 20, and 19 present the external resources used in this study. They cover MSA readability resources, Darija task-specific resources, and core Darija resources, respectively.

Table 15: Readability classification with classical models over handcrafted and TF-IDF feature sets. Each block is one classifier; rows are feature sets. **Bold** marks the best value per column.

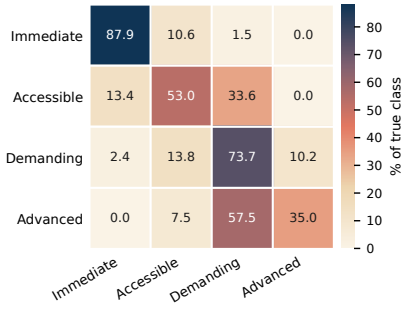
Features	Accuracy	Macro-F1	QWK
SVR → ordinal			
formulas	**68.6**	63.7	**0.756**
surface+formulas	64.4	58.3	0.714
surface	64.6	58.6	0.712
char TF-IDF	57.2	38.0	0.580
char+word TF-IDF	55.5	34.8	0.554
word TF-IDF	52.6	37.3	0.511
Logistic Regression			
surface+formulas	68.1	64.0	0.754
surface	68.1	63.9	0.753
formulas	68.3	62.1	0.747
char+word TF-IDF	56.8	41.8	0.547
char TF-IDF	56.8	38.9	0.542
word TF-IDF	50.4	31.4	0.391
Random Forest			
formulas	64.9	62.7	0.725
surface+formulas	67.1	63.8	0.710
surface	65.6	63.6	0.709
char+word TF-IDF	62.2	49.5	0.662
char TF-IDF	61.2	48.6	0.662
word TF-IDF	54.8	43.9	0.557
RF regression → ordinal			
surface+formulas	67.1	64.0	0.725
surface	66.6	**64.7**	0.717
formulas	63.4	60.5	0.713
char+word TF-IDF	62.4	52.2	0.675
char TF-IDF	61.9	51.8	0.672
word TF-IDF	54.5	50.0	0.571

Table 16: Four-shot readability classification with instruction-tuned LLMs. **Bold** marks the best value per column.

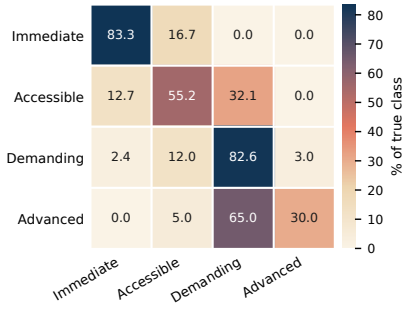
Model	Size	Accuracy ↑	Macro-F1 ↑	QWK ↑
Darija-specific				
Atlas-Chat-9B	9B	54.8	**51.5**	0.577
GemMaroc-Qwen2.5-7B-darija	7B	42.3	35.5	0.418
NileChat-3B	3B	53.6	45.2	0.560
Arabic				
Fanar-1-9B-Instruct	9B	41.0	32.6	0.480
Jais-2-8B-Chat	8B	37.3	32.9	0.406
ALLaM-7B-Instruct	7B	32.2	28.6	0.235
General-purpose				
Gemma-3-12B-it	12B	**59.2**	50.4	**0.674**
Llama-3.1-8B-Instruct	8B	51.1	36.4	0.523
Qwen3.5-9B	9B	50.9	42.7	0.585

Table 17: Eight-shot readability classification with instruction-tuned LLMs. **Bold** marks the best value per column.

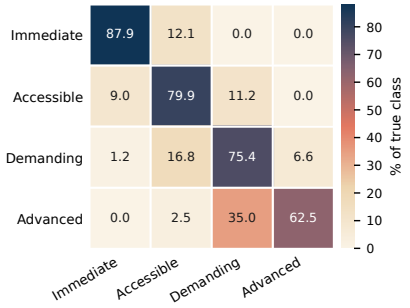
Model	Size	Accuracy ↑	Macro-F1 ↑	QWK ↑
Darija-specific				
Atlas-Chat-9B	9B	**62.4**	**61.9**	**0.723**
GemMaroc-Qwen2.5-7B-darija	7B	46.4	41.5	0.482
NileChat-3B	3B	51.4	42.0	0.522
Arabic				
Fanar-1-9B-Instruct	9B	56.5	49.3	0.600
Jais-2-8B-Chat	8B	45.9	44.7	0.581
ALLaM-7B-Instruct	7B	43.0	39.0	0.471
General-purpose				
Gemma-3-12B-it	12B	61.4	55.6	0.703
Qwen3.5-9B	9B	57.2	55.6	0.644
Llama-3.1-8B-Instruct	8B	56.3	41.8	0.583



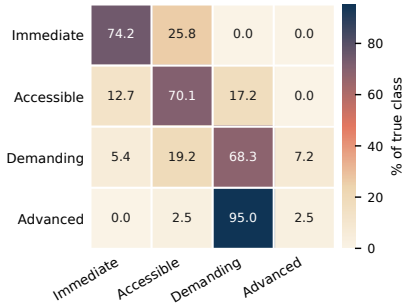
(a) AARI.



(b) Support Vector Regression.



(c) MARBERTv2.



(d) NileChat-3B, zero-shot.

Figure 14: Normalised confusion matrices for the strongest standalone formula, classical regression system, supervised encoder, and zero-shot LLM. Values are percentages within each gold class.

Table 18: In-domain versus leave-one-source-out (LOSO) MARBERTv2 performance, per source. Δ is LOSO minus in-domain.

		Accuracy			Macro-F1			QWK		
Source	n	In-dom.	LOSO	Δ	In-dom.	LOSO	Δ	In-dom.	LOSO	Δ
Overall	**407**	**76.3**	**66.6**	**−9.6**	**75.0**	**59.2**	**−15.8**	**0.838**	**0.729**	**−0.109**
WAK3I	106	75.3	71.7	−3.6	56.3	51.7	−4.6	0.707	0.658	−0.049
ALEXANDRIA	72	80.6	73.3	−7.2	76.4	73.0	−3.4	0.789	0.730	−0.059
GOUD	69	73.6	61.4	−12.2	56.7	44.2	−12.4	0.781	0.621	−0.160
DODA	51	83.5	70.6	−12.9	75.7	63.2	−12.5	0.788	0.679	−0.109
MSM-MG	46	83.9	81.7	−2.2	44.1	42.9	−1.2	0.764	0.737	−0.027
PALM	42	60.5	43.3	−17.1	49.3	34.5	−14.8	0.667	0.385	−0.282
ZAJAL	21	72.4	39.0	−33.3	70.7	32.9	−37.8	0.423	0.064	−0.359

Table 19: Core Darija resources relevant to readability corpus construction and analysis. Sizes follow the cited release.

Resource	Size	Data or domain	Supervision	Potential readability use
DODa v2 (Outchakoucht and Es-Samaali, 2024)	∼100K entries	Lexicon, variants, conjugations, sentences	Lexical and translation data	Candidate sampling, spelling variation, frequency estimates.
DODa-10K (Shang et al., 2025)	10K sets	Multilingual parallel data	Translation and native review	Controlled transfer and aligned comparisons.
MADAR–Rabat (Bouamor et al., 2018)	12K pairs	Travel-domain Rabat Darija and MSA	Professional translation	Cross-variety transfer with a narrow-domain caution.
MOR (Al-Shargi et al., 2016)	64,170 tokens	Forums, interviews, tales, textbooks, blogs	Manual morphology	Morphological and genre-sensitive analysis.
MArSum (Gaanoun et al., 2022)	∼19K pairs	Moroccan news articles and titles	Summarisation	Journalistic candidate pool; titles are not simplifications.
MA-v2 (Aftiss et al., 2025)	>50K pairs	News in 11 categories	Summarisation	Large journalistic pool and domain-held-out analysis.
QADI (Abdelali et al., 2021)	540K tweets	Tweets from 18 Arab countries	Country and dialect labels	Moroccan sampling and dialect filtering.
Atlaset (Bounhar, 2025)	∼155M tokens	Curated web and social text	Unlabelled pretraining data	Frequency, surprisal, retrieval, continued pretraining.

Table 20: Additional Darija datasets that may broaden coverage but do not provide readability labels.

Resource	Size	Content	Original task	Use and caution
MSDA (Boujou et al., 2021)	>50K tweets	Five Arabic dialects, including Moroccan	Dialect, topic, sentiment	Informal sampling after validation of the Moroccan subset.
MSAC (Oussous et al., 2020)	2K texts	Moroccan reviews and social media	Sentiment	Small informal sample; sentiment is unrelated to difficulty.
ElecMorocco2016 (Elouardighi et al., 2017)	∼10K comments	Public election discussions	Sentiment	Public discourse after Darija–MSA filtering.
MYC (Jbel et al., 2023)	20K comments	Moroccan video-platform comments	Sentiment	Informal challenge data with platform and topic controls.
OMCD (Essefar et al., 2023)	8,024 comments	Moroccan online comments	Offensive-language labels	Sensitive challenge data requiring ethical filtering.
Atlas-Chat data (Shang et al., 2025)	102,582 instruc-tions	Translated and generated instructions	Instruction following	Adaptation only; exclude from independent testing.
Dialect2SQL (Chafik et al., 2025)	9,428 pairs	Darija-focused questions and SQL	Text to SQL	Technical challenge set, too specialised for general training.
TerjamaBench (Momayiz et al., 2024)	Release dependent	Cultural translation examples	Translation	Held-out cultural analysis after overlap checks.

Table 21: Arabic readability resources that can support transfer to Moroccan Darija.

Resource	Size	Supervision	Population or domain	Transfer value and limitation
BAREC (Elmadani et al., 2025; Habash et al., 2025)	69,441 sentences	19 ordered levels and coarser mappings	Broad genres and readerships	Main ordinal-transfer source; MSA-specific levels require mapping.
DARES (El-Haj et al., 2024)	13,335 instances	Educational stages and grades	Saudi textbooks, Grades 1–12	Grade-based transfer; curriculum placement is not general readability.
MoSAR (Nassiri et al., 2019)	644 texts	Seven school levels	Moroccan primary textbooks	Culturally close but small and school-specific.
SAMER Corpus (Alhafni et al., 2024)	159K words	Original and two simplification levels	Arabic fiction for school readers	Simplification-aware transfer with literary-domain bias.
Khallaf–Sharoff (Khallaf and Sharoff, 2021)	22,740 sentences	Three CEFR bands and binary difficulty	Instructional and web data	CEFR transfer; some labels are source-derived.
ReadMe++ Arabic (Naous et al., 2024)	1,945 sentences	Six CEFR levels	Multi-domain learner material	Multilingual transfer; relatively small Arabic subset.
SAMER Lexicon (Al Khalil et al., 2020)	36K–40K entries	Five lexical levels	Expert-annotated MSA lemmas	Difficult-word features but no sentence complexity.
AraEyebility (Baazeem et al., 2025)	57,617 observations	Difficulty, gaze, ratings	MSA and Classical Arabic passages	Cognitive evidence with a different experimental unit.
Arability corpus (Al-Khalifa and Al-Ajlan, 2010)	Small corpus	Three textbook levels	Saudi textbooks	Historical formula baseline with limited scale.

Why Is Current XAI Not Enough for Arabic NLP? A Critical Survey of the Explainability Gap

Salima Lamsiyah[1], **Ruslan Mitkov**[2,3]
[1] University of Luxembourg, Luxembourg
[2] Lancaster University, United Kingdom
[3] University of Alicante, Spain
salima.lamsiyah@uni.lu

Abstract

Explainable AI (XAI) is now a major theme in NLP; however, Arabic NLP remains underexplained in three connected senses. First, there is a *method gap*: Arabic XAI relies heavily on a small set of post-hoc techniques such as LIME, SHAP, attention visualization, and saliency, while broader NLP XAI offers richer diagnostic, counterfactual, probing, rationale-based, and human-centered methods. Second, there is a *task gap*: existing Arabic XAI work is concentrated in classification tasks, especially sentiment analysis, hate/offensive language detection, fake news, and spam, with weaker coverage of generation, retrieval, translation, summarization, structured prediction, and dialogue. Third, there is a *linguistic gap*: many explanations identify influential tokens, but rarely explain Arabic-specific phenomena such as morphology, clitics, dialectal variation, diglossia, orthographic ambiguity, diacritics, code-switching, named entities, cultural references, or Classical and religious registers. This critical structured survey synthesizes the reviewed literature on Arabic XAI across text, speech, and multimodal settings. We argue that Arabic NLP does not only need explanations of model decisions; it needs explanations that are faithful to Arabic as a linguistic, cultural, and sociotechnical object. We introduce a taxonomy of tasks, methods, linguistic units, varieties, goals, and evaluation practices, and propose a research agenda for linguistically grounded Arabic XAI.

1 Introduction

Arabic NLP has advanced rapidly through Arabic-specific transformer models, pretrained multilingual and monolingual encoders, and Arabic-centric large language models (LLMs) (Abdul-Mageed et al., 2021; Inoue et al., 2021; Khader et al., 2025; Sengupta et al., 2023; Huang et al., 2024). However, the same developments that improve performance also make system behavior harder to inspect, a concern widely discussed in explainable NLP and LLM explainability research (El Zini and Awad, 2022; Luo et al., 2024; Zhao et al., 2024). This opacity is not merely a generic concern inherited from NLP. Arabic is morphologically rich, diglossic, dialectally diverse, and orthographically variable, with challenges involving clitics, segmentation, undiacritized text, dialectal variation, and register-specific usage (Habash, 2010; Darwish et al., 2021; Inoue et al., 2021). A model may rely on a clitic, subword fragment, dialect marker, missing diacritic, named entity boundary, Qur'anic expression, or culturally loaded phrase in ways that are difficult to interpret from surface tokens alone (Abdelali et al., 2022; Mustafa et al., 2024; Sahyoun and Shehata, 2023; Younes, 2026). Explanations that are adequate for one language, register, or task may therefore be misleading when applied unchanged to Arabic.

General NLP surveys describe a broad XAI landscape, including local interpretation, rationales, probing, counterfactuals, saliency, faithfulness evaluation, and LLM explanation (El Zini and Awad, 2022; Luo et al., 2024; Hassan et al., 2025; Zhao et al., 2024). Arabic-specific survey work has reviewed Arabic sentiment analysis (Shi and Agrawal, 2025) and explainability for Arabic sentiment analysis (Alsehaimi et al., 2025). However, the reviewed Arabic XAI literature reveals a broader *explainability gap*: Arabic systems are increasingly explained, but the explanations often remain methodologically narrow, task-concentrated, and weakly connected to Arabic linguistic and socio-cultural structure.

We unpack this gap along three dimensions. The **method gap** is that many Arabic studies use LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017), attention visualization, or saliency-style evidence as the main explanation device (Abdelwahab et al., 2022; Awadallah et al., 2022; Mustafa et al.,

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 226–236
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_023

2024; Aljrees, 2024; Sweidan et al., 2025; Bouke et al., 2026). These methods are valuable, but they represent only a subset of XAI for NLP. The **task gap** is that Arabic XAI is concentrated in classification: sentiment analysis, hate/offensive language, fake news, spam, reviews, news, authorship, and related social-media tasks (Atabuzzaman et al., 2023; Alwateer et al., 2025; Mahouachi, 2026; Aboulola and Umer, 2024; Alansari et al., 2026; Alshammari et al., 2025). The **linguistic gap** is that many explanations highlight tokens or features without explaining Arabic-specific evidence: morphology, clitics, dialectal variation, diglossia, diacritics, code-switching, orthographic ambiguity, named entities, cultural references, or Classical and religious registers. A smaller set of studies points toward a richer agenda by probing Arabic transformers for morphology and dialectal information (Abdelali et al., 2022), designing explainable metrics for dialectal automatic speech recognition (ASR) (Sahyoun and Shehata, 2023), developing diagnostic visual analytics for Arabic named entity recognition (NER) (Younes, 2026), and using knowledge-graph diagnostics for Arabic machine reading comprehension (MRC) hallucinations (AlGhamdi et al., 2026).

The central argument of this survey is that Arabic NLP does not only need more XAI, but explanations that are faithful to Arabic as a linguistic, cultural, and sociotechnical object. Word-level heatmaps may be useful, but they do not by themselves show whether a model relies on dialect cues, morphology, named entities, religious references, moderation norms, or dataset artifacts. Accordingly, this paper frames Arabic XAI through method, task, and linguistic gaps; proposes four explanation levels: prediction, model, linguistic, and socio-cultural; organizes the literature through a critical taxonomy and task-method coverage analysis; and outlines an agenda for linguistically grounded, faithful, useful, and reproducible Arabic XAI.

2 Survey Methodology and Scope

This paper is a *critical structured survey*, not a fully systematic review. We searched ACL Anthology, ACM Digital Library, IEEE Xplore, ScienceDirect, SpringerLink, MDPI, arXiv/preprint venues, and Google Scholar. The search covered work available up to August 2026 and used combinations of English keywords such as Arabic NLP, explainable AI, interpretability, transparency, XAI, LIME, SHAP, attention, saliency, LLMs, sentiment analysis, hate speech, fake news, spam, ASR, NER, MRC, hallucination, readability, and multimodal Arabic NLP. To reduce the risk of missing locally described work, we also used Arabic-script query variants for Arabic NLP, explainable AI, and model interpretation. General XAI-for-NLP and XAI-for-LLMs surveys are used for background and terminology (El Zini and Awad, 2022; Luo et al., 2024; Hassan et al., 2025; Zhao et al., 2024).

Dimension	Aggregate pattern
Study selection	58 records retrieved; 46 screened; 31 included.
Task focus	18 classification-oriented; 7 diagnostic or structured; 6 multimodal, LLM, or generation-adjacent.
Method family	17 post-hoc attribution or feature-importance; 8 attention, saliency, or Grad-CAM; 9 probing, visual analytics, knowledge graphs, rules, or rationales.
Arabic variety/register	18 MSA/general or social-media Arabic; 7 dialectal; 2 Classical/Qur'anic; 4 multimodal Arabic-language; limited Arabizi/code-switching.
Evaluation practice	20 predictive metrics plus examples; 7 method-comparison or diagnostic validation; 2 human/user studies; 2 explicit faithfulness or stability-oriented studies.

Table 1: Aggregate annotation summary for the included Arabic or Arabic-adjacent XAI studies. Counts are non-exclusive where method, variety, or evaluation labels overlap.

The search retrieved 58 candidate records. After removing duplicates, non-research items, and papers without an Arabic or XAI connection, 46 records were screened at title/abstract or full-text level. Thirty-one Arabic or Arabic-adjacent XAI studies met the inclusion criteria; 15 were excluded because they reported Arabic NLP performance without an explicit explanation, interpretability, transparency, diagnostic, or human-facing explanation component. We retained additional general XAI and Arabic NLP references only for background and positioning. We treat attention-based models cautiously because attention is not necessarily an explanation (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019). Each included work was annotated by task, model family, XAI method, Arabic variety or register, explanation goal, evaluation practice, and degree of linguistic grounding. Table 1 reports the aggregate annotation patterns used to support the taxonomy and coverage analysis in Table 2.

Positioning with Respect to Existing Surveys. The present survey differs from general XAI-for-NLP surveys, which provide broad taxonomies of

explanation methods but do not examine Arabic as a linguistically and sociotechnically specific setting (El Zini and Awad, 2022; Luo et al., 2024; Hassan et al., 2025). It also differs from LLM explainability surveys, which focus on large-model explanation challenges across languages and tasks (Zhao et al., 2024). These works are essential for terminology and evaluation criteria, but they do not address how Arabic morphology, dialectal variation, diglossia, diacritics, orthographic ambiguity, and cultural register affect what an explanation should explain.

Arabic-focused survey work is closer to our scope, but remains narrower. Existing surveys review Arabic NLP broadly (Darwish et al., 2021), contemporary Arabic sentiment analysis (Shi and Agrawal, 2025), or explainability for Arabic sentiment analysis (Alsehaimi et al., 2025). In contrast, this paper synthesizes XAI across Arabic NLP tasks and asks a different question: not only which explanation methods have been applied, but whether they explain Arabic-specific linguistic and socio-cultural evidence. This positioning motivates the three-gap framing developed in the rest of the paper.

3 What Should an Explanation Explain in Arabic NLP?

The phrase "explainable Arabic NLP" does not refer to a single explanatory target but encompasses several distinct ones. We propose a four-level distinction, synthesized from the reviewed literature, with the aim of rendering evaluation criteria more precise and tractable.

Prediction-level explanation. At the most fundamental level, an explanation accounts for why a model assigns a specific label, returns a particular score, retrieves a given result, or generates a specific output. Most Arabic XAI classification papers operate at this level, using local feature attribution to identify influential words or features for sentiment, fake news, spam, or hate-speech decisions (Abdelwahab et al., 2022; Awadallah et al., 2022; Aljrees, 2024; Aboulola and Umer, 2024; Bouke et al., 2026; Alwateer et al., 2025). Prediction-level explanations can support debugging and user trust, but they are not automatically faithful or linguistically meaningful.

Model-level explanation. A model-level explanation addresses the internal representations, layers, features, or mechanisms that the model employs. This level is especially relevant for Arabic transformers and multilingual encoders, where behavior may be distributed across layers and subword representations. Abdelali et al. (2022), for example, probe Arabic transformer models for morphology, syntax, and dialect identification. Attention-based sentiment architectures (Berrimi et al., 2023; Ghasemi and Momtazi, 2023; Sweidan et al., 2025) and LLM adaptation studies (Khader et al., 2025) are also relevant, although attention or benchmark performance alone does not settle whether the explanation is faithful.

Linguistic explanation. A linguistic explanation identifies which Arabic linguistic features – morphological, syntactic, semantic, dialectal, or orthographic – are operative in the model's predictions. This level is where the Arabic explainability gap is most visible. For Arabic, the relevant unit may be a stem, clitic, lemma, diacritic, named entity, multiword expression, dialect phrase, or orthographic variant rather than a whitespace-delimited token. Work on dialectal ASR evaluation (Sahyoun and Shehata, 2023), Arabic NER diagnostics (Younes, 2026), Arabic readability (Rabih, 2025), and transformer probing (Abdelali et al., 2022) begins to connect explanations to linguistic structure, but this remains less common than token attribution.

A simple worked example clarifies the problem. Suppose a classifier highlights the Arabic surface token وبالمدرسة (transliterated as *wabilmadrasa*, "and in/at the school") as evidence for a prediction. A whitespace-level explanation treats this as one unit, but a linguistically meaningful analysis may need to distinguish the conjunction *w+* وـ "and", the preposition *bi+* بـ "in/with", the definite article *al+* الـ "the", and the stem *madrasa* مدرسة "school".

Socio-cultural explanation. A socio-cultural explanation addresses how a model's interpretation is shaped by cultural context, moderation norms, religious or classical references, identity terms, user expectations, and the broader social meaning of language use. This level is essential for harmful-content moderation, propaganda, Qur'anic semantic search, and high-stakes applications. User-centered Arabic hate-speech explanation work (Al-Ansari et al., 2026), neuro-symbolic offensive-

language detection (Mahouachi, 2026), propagandistic meme explanation (Kmainasi et al., 2025), Qur'anic semantic search (Mustafa et al., 2024), and knowledge-graph diagnostics for Qur'anic MRC (AlGhamdi et al., 2026) illustrate why Arabic XAI cannot be reduced to visually plausible word highlights.

4 Task-Method Coverage and Critical Taxonomy

Table 2 combines the taxonomy and coverage analysis, showing how method, task, linguistic, variety, and evaluation gaps reinforce one another. It highlights the dominance of post-hoc explanations for classification tasks and the weaker coverage of generation, retrieval, structured linguistic analysis, cross-variety evaluation, and Arabic-aware faithfulness testing. The following subsections discuss these patterns by task cluster, moving from well-studied classification settings to more linguistically grounded, multimodal, and LLM-adjacent work.

Table 2 summarizes the main dimensions of the Arabic explainability gap by separating current practice, representative work, missing explanatory capacity, and Arabic-specific motivation.

The taxonomy shows that the gap is not only a lack of Arabic XAI papers, but a recurring mismatch between explanation methods, tasks, and linguistic targets. Most work applies post-hoc methods such as LIME, SHAP, attention visualization, or feature importance to classification tasks (Abdelwahab et al., 2022; Awadallah et al., 2022; Aljrees, 2024; Bouke et al., 2026). These methods can identify influential surface units, but they rarely show whether a model used Arabic-specific evidence such as clitics, stems, dialectal markers, diacritics, named-entity boundaries, or cultural references. Thus, explanations often remain at the prediction level even when the relevant failure is linguistic or socio-cultural.

The table also highlights the need to align explanation goals with evaluation practice. Some studies move beyond generic attribution through probing, visual analytics, ASR diagnostics, knowledge graphs, neuro-symbolic modeling, or user-centered evaluation (Abdelali et al., 2022; Sahyoun and Shehata, 2023; Younes, 2026; AlGhamdi et al., 2026; Mahouachi, 2026; Al-Ansari et al., 2026). However, these efforts remain scattered. The main challenge is therefore to embed existing and new XAI methods in Arabic-aware protocols that test faithfulness, stability, usefulness, and fairness across the linguistic variation Arabic NLP systems actually face.

4.1 Sentiment and Opinion Mining: Useful but Over-Represented

Arabic sentiment analysis is the clearest default testbed for XAI. Studies combine sentiment classifiers with LIME-style local explanations for domain-specific medical opinions (Abdelwahab et al., 2022), multi-dialect sentiment classification (Awadallah et al., 2022), and noisy deep explainable sentiment modeling (Atabuzzaman et al., 2023). Attention-based sentiment architectures and cross-lingual sentiment models use attention or contextual representations as interpretive evidence (Berrimi et al., 2023; Ghasemi and Momtazi, 2023), while newer work adds SHAP, multi-self-attention, empirical comparison of explanation techniques, or platform-level explanation interfaces (Hadj Azzem et al., 2024; Chafiqui et al., 2024; Sweidan et al., 2025).

These works contribute an important baseline: they make Arabic classifier decisions inspectable and show that XAI can be integrated into practical sentiment pipelines. However, much of this work remains prediction-level. It often explains which words influenced a label, not whether the model captured negation, dialectal intensification, sarcasm, stance, topic confounds, or morphology. This distinction matters because a visually plausible explanation can be useful for demonstration while still failing faithfulness, stability, or Arabic-awareness. The sentiment scoping review by Alsehaimi et al. (2025) is therefore important as a consolidation of the subfield, but it also underscores the limits of sentiment-centered Arabic XAI as a proxy for the whole language.

4.2 Harmful Content: Explanations Need Social Context

Hate speech, offensive language, and propaganda expose the socio-cultural level of explanation. Alwateer et al. (2025) study interpretable Arabic hate-speech detection with LLMs; Al-Ansari et al. (2026) evaluate perceived informativeness of XAI representations for Arabic hate-speech detection with Arabic-speaking users; Mahouachi (2026) uses a hybrid multi-view and neuro-symbolic approach for Arabic offensive-language detection; and Kmainasi et al. (2025)

study explanation-enhanced detection of propagandistic Arabic memes.

Compared with sentiment work, this cluster more clearly motivates user-centered and safety-oriented explanations. It also shows why token salience is insufficient. A word can be harmful, quoted, reclaimed, dialect-specific, target-dependent, or embedded in a meme. The contribution of this line of work is to connect explainability with moderation and social meaning. Its limitation is that the field still lacks shared protocols for testing whether explanations remain useful and fair across dialects, regions, identity terms, targets, and cultural references. In our synthesis, harmful-content XAI is one of the strongest arguments for treating Arabic as a sociotechnical object rather than a string of tokens.

4.3 Information Integrity: Attribution Is Not Evidence

Arabic fake-news and spam studies use explainability to inspect high-performing classifiers. Aljrees (2024) combines LIME with an ELMo-based tri-ensemble fake-news model; Aboulola and Umer (2024) use large-language-feature embeddings, CNN-LSTM ensembles, and explainable AI for Arabic fake-news classification; and Bouke et al. (2026) integrates LIME and SHAP into an Arabic spam-detection pipeline based on LightGBM.

These studies are useful for feature inspection and model debugging. The critical issue is that attribution is not the same as evidence. A highlighted token may reveal what the classifier used, but not whether the model identified deception, propaganda framing, source credibility, temporal inconsistency, or a culturally grounded claim. For Arabic misinformation, explanations should ideally connect entities, claims, sources, and context. The reviewed work moves toward transparency, but it rarely explains the information-integrity phenomenon itself.

4.4 Diagnostics Beyond Classification

Work that most substantively engages with Arabic-specific phenomena frequently falls outside the conventions of standard post-hoc explanation. Mustafa et al. (2024) use LIME and SHAP for Arabic transformer models in Qur'anic semantic search, where explanation must confront high-register and religious text. AlGhamdi et al. (2026) use knowledge-graph triple diagnostics for Arabic MRC hallucinations, shifting from word salience to relational grounding. Abdelali et al. (2022) probes Arabic transformers for morphology, syntax, and dialect information. Sahyoun and Shehata (2023) propose an explainable ASR metric for dialectal Arabic, and Younes (2026) develops visual analytics for Arabic NER evaluation. Rabih (2025) studies interpretable measures for Arabic readability.

This group shows what linguistically grounded Arabic XAI can look like. Rather than asking only which token affected a label, these works ask where linguistic information is encoded, what kind of recognition error occurred, which entity boundaries or annotations are problematic, which readability features matter, or whether a generated answer is semantically grounded. Their limitation is scale and standardization: they are promising case studies and frameworks, not yet a shared paradigm for Arabic XAI evaluation.

4.5 Multimodal and LLM-Adjacent Work

Arabic XAI is also expanding beyond text-only classification. Arabic sign language recognition uses LIME or Grad-CAM-style visual explanations (Baghdadi et al., 2024; Balat et al., 2025); Arabic captioning uses interpretable visual concept integration (Elchafei and Fashwan, 2025); and meme detection connects multimodal classification with rationale-oriented resources (Kmainasi et al., 2025). LLM-oriented Arabic work discusses adaptation, evaluation, and ethical deployment (Khader et al., 2025), while AraLingBench evaluates Arabic linguistic capabilities of LLMs (Zbib et al., 2026).

These works widen the scope of Arabic XAI, but they also make evaluation harder. For multimodal tasks, explanation quality must be separated from visual grounding and Arabic generation quality. For LLMs, explanation must be connected to factual grounding, linguistic competence, hallucination, and user expectations. Within the reviewed literature, Arabic LLM explainability remains more of an emerging requirement than a mature methodology.

5 Critical Discussion: Three Gaps and Their Evaluation Consequences

The four-level framework introduced earlier provides the lens for interpreting the three gaps. The method gap is largely a movement problem: many studies stop at prediction-level attribution even when model-level or linguistic explanations are needed. The task gap is a coverage problem: clas-

Axis	Aggregate pattern in the reviewed studies	Current practice and representative works	Main gap and Arabic-specific implication
Methods	Post-hoc attribution and feature-importance methods are the largest group; attention, saliency, and Grad-CAM also recur, while probing, visual analytics, knowledge graphs, rules, and rationales appear less often.	LIME, SHAP, attention visualization, saliency, and feature importance dominate (Abdelwahab et al., 2022; Awadallah et al., 2022; Hadj Azzem et al., 2024; Mustafa et al., 2024). Some studies use probing, ASR diagnostics, visual analytics, KG diagnostics, or neuro-symbolic rules (Abdelali et al., 2022; Sahyoun and Shehata, 2023; Younes, 2026; AlGhamdi et al., 2026; Mahouachi, 2026).	Counterfactuals, causal tests, Arabic-preserving perturbations, faithful rationale evaluation, mechanistic analysis, and systematic human evaluation remain rare. This matters because Arabic tokenization, morphology, spelling, and dialect variation can make word-level attributions unstable or misleading.
Tasks	Classification-oriented studies form the largest cluster, while diagnostic, structured, multimodal, LLM, and generation-adjacent tasks are less represented.	Most work targets sentiment, hate/offensive language, fake news, spam, app reviews, news classification, authorship, and mental health detection (Aljrees, 2024; Aboulola and Umer, 2024; Bouke et al., 2026; Alansari et al., 2026; Kumar et al., 2023).	Arabic QA, RAG, summarization, translation, dialogue, parsing, grammatical error correction, and information extraction remain weakly covered, although these tasks require explanations of grounding, factuality, alignment, and structure rather than only label support.
Linguistic grounding	Most explanations operate at token, word, feature, or example level; fewer studies target representations, ASR errors, NER components, readability features, semantic triples, or linguistic benchmarks.	More linguistically grounded directions appear in transformer probing, dialectal ASR evaluation, NER diagnostics, readability, Qur'anic semantic search, and MRC hallucination analysis (Abdelali et al., 2022; Sahyoun and Shehata, 2023; Younes, 2026; Rabih, 2025; Mustafa et al., 2024; AlGhamdi et al., 2026).	Explanations rarely target morphemes, clitics, stems, lemmas, diacritics, multiword expressions, named-entity boundaries, dialectal phrases, or cultural/religious references. Arabic evidence often lies below or above whitespace tokens.
Varieties and registers	MSA, social-media Arabic, dialectal Arabic, Classical/Qur'anic Arabic, and multimodal Arabic-language settings appear, but variety is rarely evaluated as an explanation variable.	Relevant work includes dialectal tweets, dialectal ASR, Arabic transformer probing, Qur'anic search and MRC, Arabic memes, and Arabic sign language/captioning (Awadallah et al., 2022; Sahyoun and Shehata, 2023; Abdelali et al., 2022; Mustafa et al., 2024; Kmainasi et al., 2025).	Cross-variety explanation evaluation remains limited for MSA, dialects, Classical Arabic, Arabizi, code-switching, and social-media spelling variation. This creates robustness and fairness risks when models rely on variety or register cues.
Evaluation	Most studies report predictive metrics and selected explanation examples; fewer compare explanation methods, perform diagnostic validation, or conduct human/user evaluation.	Some work moves toward user-centered evaluation, method comparison, visual diagnostics, and hallucination analysis (Al-Ansari et al., 2026; Hadj Azzem et al., 2024; Younes, 2026; AlGhamdi et al., 2026).	Faithfulness, plausibility, stability, usefulness, fairness, and reproducibility are not standardized. Without Arabic-aware evaluation, visually plausible explanations may hide brittle behavior under normalization, segmentation, diacritic, or dialect changes.

Table 2: Critical taxonomy and coverage summary of the Arabic XAI explainability gap across methods, tasks, linguistic grounding, Arabic varieties, and evaluation practices.

sification tasks encourage label-level explanations, while generative, retrieval, and structured tasks require grounding, alignment, and error decomposition. The linguistic gap is the deepest explanatory problem: Arabic-specific evidence often lives at a different level from the unit displayed by the explanation. Evaluation is therefore not a fourth independent gap, but a cross-cutting consequence of the three gaps.

Several reviewed studies already provide empirical evidence for this argument. Abdelali et al. (2022) show through layer and neuron analysis that Arabic transformer representations encode morphology, syntax, and dialectal information differently across layers, which supports the need for model-level rather than only prediction-level explanation. Sahyoun and Shehata (2023) show that standard WER is insufficient for dialectal Arabic ASR because it does not reveal morphological, lexical, and semantic sources of errors. Younes (2026) report that component-level visual analytics can reveal annotation mistakes, calibration issues, and entity-level error patterns that aggregate F1 scores obscure. AlGhamdi et al. (2026) similarly move from output correctness to triple-level grounding in Arabic MRC hallucination analysis. These findings demonstrate that the gaps identified here affect model analysis and evaluation, not only survey categorization.

5.1 The Method Gap

The reviewed literature shows a clear preference for methods that are easy to attach to existing classifiers: LIME, SHAP, attention visualization, saliency, or feature importance. This is understandable because model-agnostic methods are accessible and produce intuitive visual artifacts. The problem is that these artifacts are often treated as explanations without enough evidence that they are faithful, stable, or aligned with Arabic linguistic units. General NLP XAI surveys emphasize that local explanations, rationales, probing, counterfac-

tuals, and faithfulness tests answer different questions (El Zini and Awad, 2022; Luo et al., 2024; Hassan et al., 2025). In the Arabic corpus, richer methods appear in isolated forms, including probing (Abdelali et al., 2022), visual analytics (Younes, 2026), knowledge-graph diagnostics (AlGhamdi et al., 2026), user-centered evaluation (Al-Ansari et al., 2026), and neuro-symbolic modeling (Mahouachi, 2026), but they have not yet become common practice.

The added primary-method citations also clarify the scope of this claim. LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017), and Grad-CAM (Selvaraju et al., 2017) are influential explanation methods, but they were not designed specifically around Arabic morphology, dialectal spelling, or clitic segmentation. Attention-based explanations must also be handled carefully because the literature debates when attention can support interpretation (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019). Probing (Belinkov, 2022), rationale benchmarks (DeYoung et al., 2020), faithfulness evaluation (Jacovi and Goldberg, 2020), and counterfactual data methods (Kaushik et al., 2020) illustrate methodological options that Arabic XAI has not yet used systematically.

5.2 The Task Gap

Arabic XAI is disproportionately shaped by classification. Sentiment analysis, hate/offensive language, fake news, spam, app reviews, news classification, authorship, and mental health detection supply many of the examples and methods (Abdelwahab et al., 2022; Atabuzzaman et al., 2023; Alwateer et al., 2025; Aljrees, 2024; Bouke et al., 2026; Alansari et al., 2026; Kumar et al., 2023; Alshammari et al., 2025). Classification is a useful starting point, but it encourages explanations that identify label-supporting tokens. Modern Arabic NLP increasingly includes retrieval, generation, machine reading comprehension, captioning, LLM adaptation, and linguistic benchmarking (Mustafa et al., 2024; AlGhamdi et al., 2026; Elchafei and Fashwan, 2025; Khader et al., 2025; Zbib et al., 2026). These tasks require explanations of grounding, hallucination, alignment, fluency, factuality, entity tracking, and discourse, not only explanations of class labels.

5.3 The Linguistic Gap

We argue that the most important gap is linguistic. Many Arabic XAI papers explain which input words influenced a prediction, but not what Arabic evidence the model used. This is not a minor technical issue. Arabic preprocessing choices, including normalization, segmentation, stop-word removal, dialect filtering, subword tokenization, and diacritic handling, can change the unit being explained. If an explanation highlights a whitespace token while the model operates over subwords, or if preprocessing removes clitics or diacritics, the explanation may be difficult to interpret linguistically. Work on transformer probing, dialectal ASR, NER diagnostics, readability, and Qur'anic MRC demonstrates that explanations can be tied to morphology, dialect, recognition errors, entity boundaries, readability features, and semantic relations (Abdelali et al., 2022; Sahyoun and Shehata, 2023; Younes, 2026; Rabih, 2025; AlGhamdi et al., 2026). This is the direction in which Arabic XAI should move.

5.4 Evaluation Consequences

Because the three gaps affect what explanations can show, evaluation must separate several criteria rather than cite "interpretability" generically. Faithfulness asks whether an explanation reflects the model's actual decision process (Jacovi and Goldberg, 2020). Plausibility asks whether the explanation resembles human-recognizable evidence, a criterion operationalized in rationale benchmarks such as ERASER (DeYoung et al., 2020). Usefulness asks whether an explanation helps a target stakeholder, which is directly relevant to user-centered Arabic hate-speech explanation (Al-Ansari et al., 2026). Stability asks whether explanations remain consistent under meaning-preserving changes; for Arabic, this should include spelling, normalization, segmentation, dialectal paraphrase, and diacritic changes. Fairness asks whether explanation quality differs across varieties, registers, or identity terms. Many Arabic XAI papers still rely on predictive accuracy plus selected explanation examples, so these criteria should be reported separately in future work.

6 Toward Linguistically Grounded Arabic XAI

A linguistically grounded agenda for Arabic XAI should move beyond adding generic explanation methods to Arabic models. It should instead redesign explanation targets, evaluation protocols, and reporting standards around the linguistic and sociotechnical properties of Arabic itself. The

following directions outline how future work can make explanations more faithful to model behavior, more stable under Arabic-specific variation, more useful to Arabic-speaking users and domain experts, and more reproducible across tasks, datasets, and varieties.

Arabic-aware units of explanation. Explanations should be evaluated at levels that are meaningful for Arabic: morphemes, clitics, stems, lemmas, diacritics, named entities, multiword expressions, dialectal phrases, and discourse cues. Token-level heatmaps are insufficient when model decisions depend on morphology or orthography. Future papers should report the relationship between the model's internal tokenization and the units shown to users.

Arabic-preserving perturbation tests. Arabic XAI needs perturbation tests that preserve meaning while varying surface form: spelling variation, normalization, clitic segmentation, dialectal paraphrases, and diacritic restoration or removal. Such tests would help distinguish robust explanations from artifacts of preprocessing or tokenization. This direction connects faithfulness evaluation (Jacovi and Goldberg, 2020) and counterfactual data analysis (Kaushik et al., 2020) to Arabic-specific surface and morphological variation.

Cross-variety explanation evaluation. Benchmarks should evaluate explanations across MSA, regional dialects, Classical Arabic, social-media Arabic, Arabizi, and code-switched Arabic. Work on dialectal ASR and Arabic transformer probing shows that variety can be studied systematically (Sahyoun and Shehata, 2023; Abdelali et al., 2022); Arabic XAI should make variety an explanation variable, not merely a dataset description.

Human rationale benchmarks with Arabic expertise. Arabic XAI needs shared datasets with human rationales, expert annotations, and task-specific explanation quality labels. User-centered work on Arabic hate-speech explanation (Al-Ansari et al., 2026) and explanation-enhanced multimodal resources (Kmainasi et al., 2025) point to this direction. Annotators should include Arabic speakers with dialectal competence and, where necessary, domain expertise in moderation, health, education, religious text, or accessibility.

Explanation evaluation for Arabic LLMs, RAG, and generation. Future work should target Arabic LLMs, retrieval-augmented generation, QA, summarization, machine translation, dialogue, and hallucination detection. Arabic LLM adaptation and linguistic benchmarking are already emerging (Khader et al., 2025; Zbib et al., 2026), and knowledge-graph diagnostics show one path for hallucination analysis (AlGhamdi et al., 2026). Explanations for these systems should address factual grounding, retrieval evidence, morphology, syntax, discourse, and answer-source alignment.

Faithfulness, plausibility, stability, usefulness, and fairness protocols. Arabic XAI needs evaluation protocols that separate what the model actually used from what users find plausible. Faithfulness tests (Jacovi and Goldberg, 2020), rationale plausibility evaluation (DeYoung et al., 2020), stability under Arabic-preserving perturbations, usefulness for stakeholders (Al-Ansari et al., 2026), and fairness across varieties and identity terms should be reported separately. This distinction is especially important in moderation, health, education, religious-domain applications, and information integrity.

Reproducibility standards. Arabic XAI papers should report dataset variety, preprocessing, normalization, tokenization, segmentation, diacritic handling, model checkpoints, explanation parameters, evaluation metrics, and known validity limits. Where possible, studies should release code, trained models or inference scripts, explanation outputs, rationale annotations, and diagnostic error analyses. Without these details, it is difficult to know whether a method is better for Arabic or simply easier to visualize.

7 Conclusion

Arabic XAI is emerging, but it remains shaped by a clear explainability gap. Existing work shows the value of LIME, SHAP, attention visualization, probing, visual analytics, knowledge graphs, neuro-symbolic methods, and user-centered explanations. Yet the dominant pattern is still output-level explanation for classification tasks, especially sentiment, harmful-content detection, fake news, and spam. As a result, generation, retrieval, QA, summarization, translation, structured prediction, and LLM-based Arabic applications remain under-explained.

The main message of this survey is that Arabic NLP needs more than explanations of model decisions; it needs explanations that are faithful to Arabic as a linguistic, cultural, and sociotech-

nical object. Future work should therefore move from explaining predictions to explaining Arabic itself, through Arabic-aware units, cross-variety evaluation, perturbation testing, human rationale benchmarks, LLM/RAG explanation protocols, and stronger reproducibility standards. This agenda is essential for building Arabic NLP systems that are transparent, trustworthy, fair, and useful across the diverse contexts in which Arabic is used.

Limitations

This survey is a critical structured synthesis rather than an exhaustive systematic review. It focuses on studies that explicitly frame their contribution in terms of explainability, interpretability, transparency, or diagnostic analysis, so relevant work on Arabic model analysis or error diagnosis may be missed if it does not use XAI terminology.

Coverage is also uneven across tasks and publication stages. Sentiment and harmful-content detection are better represented than Arabic LLM explainability, RAG, hallucination analysis, generation, and multimodal XAI. Although the search used international databases, Google Scholar, and Arabic-script query variants, locally published, non-indexed, non-English, or differently framed work may still be absent. The taxonomy should therefore be read as a structured snapshot of a developing field, not a fixed or exhaustive classification.

Use of AI Assistance

In preparing this work, we used AI-assisted tools for language editing, such as spell checking and stylistic revisions with Grammarly and ChatGPT. The authors take full responsibility for the scientific content, analyses, and conclusions presented in this work.

References

Ahmed Abdelali, Nadir Durrani, Fahim Dalvi, and Hassan Sajjad. 2022. Post-hoc analysis of Arabic transformer models. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 91–103.

Youmna Abdelwahab, Mohamed Kholief, and Ahmed Ahmed Hesham Sedky. 2022. Justifying Arabic text sentiment analysis using explainable AI (XAI): LASIK surgeries case study. *Information*, 13(11):536.

Muhammad Abdul-Mageed, Abdel Rahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Omar Ibrahim Aboulola and Muhammad Umer. 2024. Novel approach for Arabic fake news classification using embedding from large language features with CNN-LSTM ensemble model and explainable AI. *Scientific Reports*, 14(1):30463.

Noor Al-Ansari, Dena Al-Thani, and Mariam Bahameish. 2026. Explaining in context: Perceived informativeness of explainable artificial intelligence (XAI) in Arabic hate speech detection. *Computers in Human Behavior Reports*, page 101083.

Aisha Alansari, Dorieh Alomari, Sajjad Mahmood, and Irfan Ahmad. 2026. Multi-label classification of Arabic app reviews with data augmentation and explainable AI. *Arabian Journal for Science and Engineering*, 51(9):12487–12506.

Najwa Abdullah AlGhamdi, Sadam Al-Azani, Kwabena Nuamah, and Alan Bundy. 2026. A knowledge graph based diagnostic framework for analyzing hallucinations in Arabic machine reading comprehension. In *Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script*, pages 413–421.

Turki Aljrees. 2024. Improving prediction of Arabic fake news using ELMo's features-based tri-ensemble model and LIME XAI. *IEEE Access*, 12:63066–63076.

Afnan Alsehaimi, Amal Babour, and Dimah Alahmadi. 2025. Toward transparent modeling: A scoping review of explainability for Arabic sentiment analysis. *Applied Sciences*, 15(19):10659.

Bunder Sebail Alshammari, Khaled Nasser Alfraidi, Mohammad Omar Aljudaiey, and Walid Abdelhalim. 2025. Robustness and explainability in Arabic authorship models: Tackling the challenge of AI-generated texts. Preprint.

Majed Alwateer, Ibrahim Gad, Mahmoud Elmarhomy, Ghada Elmarhomy, Hasan Hashim, Malik Almaliki, and El-Sayed Atlam. 2025. Interpretable Arabic hate speech detection using large language model. In *2025 2nd International Conference on Advanced Innovations in Smart Cities*, pages 1–8. IEEE.

Md Atabuzzaman, Md Shajalal, Maksuda Bilkis Baby, and Alexander Boden. 2023. Arabic sentiment analysis with noisy deep explainable model. In *Proceedings of the 2023 7th International Conference on Natural Language Processing and Information Retrieval*, pages 185–189.

Mohamed Salah Awadallah, Francisco de Arriba-Perez, Enrique Costa-Montenegro, Mohamed Kholief, and Nashwa El-Bendary. 2022. Investigation of local

interpretable model-agnostic explanations (LIME) framework with multi-dialect Arabic text sentiment classification. In *2022 32nd International Conference on Computer Theory and Applications*, pages 116–121. IEEE.

Nadiah A. Baghdadi, Yousry M. Abdulazeem, Hanaa ZainEldin, Tamer Ahmed Farrag, Mansourah K. Aljohani, Amer Malki, Mahmoud Badawy, and Mostafa A. Elhosseini. 2024. Toward robust Arabic sign language recognition via vision transformers and local interpretable model-agnostic explanations integration. *Journal of Disability Research*, 3(8):20240092.

Mazen Balat, Rewaa Awaad, Ahmed B. Zaky, and Salah A. Aly. 2025. Revolutionizing communication with deep learning and XAI for enhanced Arabic sign language recognition. *arXiv preprint arXiv:2501.08169*.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Mohamed Berrimi, Mourad Oussalah, Abdelouahab Moussaoui, and Mohamed Saidi. 2023. Attention mechanism architecture for Arabic sentiment analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–26.

Mohamed Aly Bouke, Omar Imhemed Alramli, Alsadg Ahmed Albadwi Abdalhafid, and Azizol Abdullah. 2026. A novel LightGBM model for Arabic spam detection integrated with XAI for enhanced explainability. *Computers and Electrical Engineering*, 133:111032.

Youssef Chafiqui, Houda Anoun, Latifa Aloiradi, and Fatima Ezzahraa Nazih. 2024. XS2A: A platform for explainable sentiment analysis in Arabic comments. In *International Conference on Advances in Communication Technology and Computer Engineering*, pages 309–320. Springer.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R El-Beltagy, Wassim El-Hajj, et al. 2021. A panoramic survey of natural language processing in the arab world. *Communications of the ACM*, 64(4):72–81.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. Eraser: A benchmark to evaluate rationalized nlp models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458. Association for Computational Linguistics.

Julia El Zini and Mariette Awad. 2022. On the explainability of natural language processing deep models. *ACM Computing Surveys*, 55(5):1–31.

Passant Elchafei and Amany Fashwan. 2025. Multimodal Arabic captioning with interpretable visual concept integration. *arXiv preprint arXiv:2510.03295*.

Rouzbeh Ghasemi and Saeedeh Momtazi. 2023. How a deep contextualized representation and attention mechanism justifies explainable cross-lingual sentiment analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(11):1–15.

Nizar Y Habash. 2010. *Introduction to Arabic natural language processing*. Morgan & Claypool Publishers.

Yousra Chahinez Hadj Azzem, Fouzi Harrag, and Ladjel Bellatreche. 2024. Exploring explainability in Arabic language models: An empirical analysis of techniques. *Procedia Computer Science*, 244:212–219.

Md Mehedi Hassan, Anindya Nag, Riya Biswas, Md Shahin Ali, Sadika Zaman, Anupam Kumar Bairagi, and Chetna Kaushal. 2025. Explainable artificial intelligence for natural language processing: A survey. *Data & Knowledge Engineering*, 160:102470.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Song Dingjie, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, et al. 2024. Acegpt, localizing large language models in arabic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8139–8163.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in arabic pre-trained language models. In *Proceedings of the sixth Arabic natural language processing workshop*, pages 92–104.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 3543–3556. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Khalid A. Khader, Mohammed S. Hussein, and Abdallatif S. Abu-Issa. 2025. Adapting large language models for Arabic: Comparative evaluation, fine-tuning, and ethical deployment. *IEEE Access*.

Mohamed Bayan Kmainasi, Abul Hasnat, Md Arid Hasan, Ali Ezzat Shahroor, and Firoj Alam. 2025. MemeIntel: Explainable detection of propagandistic and hateful memes. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30263–30279.

Abhinav Kumar, Jyoti Kumari, and Jiesth Pradhan. 2023. Explainable deep learning for mental health detection from English and Arabic social media posts. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30.

Siwen Luo, Hamish Ivison, Soyeon Caren Han, and Josiah Poon. 2024. Local interpretations for explainable natural language processing: A survey. *ACM Computing Surveys*, 56(9):1–36.

Rim Mahouachi. 2026. A hybrid multi-view and neuro-symbolic approach for Arabic offensive language detection: A step towards safer online spaces. *Expert Systems with Applications*, page 131854.

Ahmad M. Mustafa, Saja Nakhleh, Rama Irsheidat, and Raneem Alruosan. 2024. Interpreting Arabic transformer models: A study on XAI interpretability for Qur'anic semantic-search models. *Jordanian Journal of Computers and Information Technology*, 10.

Nour Rabih. 2025. Interpretable measures for Arabic readability. Master's thesis, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM.

Abdulwahab Sahyoun and Shady Shehata. 2023. AraDiaWER: An explainable metric for dialectical Arabic ASR. In *Proceedings of the Second Workshop on NLP Applications to Field Linguistics*, pages 64–73.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, et al. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Zhiqiang Shi and Ruchit Agrawal. 2025. A comprehensive survey of contemporary Arabic sentiment analysis: Methods, challenges, and future directions. *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3760–3772.

Asmaa Hashem Sweidan, Nashwa El-Bendary, Shereen A. Taie, Amira M. Idrees, and Esraa Elhariri. 2025. Explainable deep learning for COVID-19 vaccine sentiment in Arabic tweets using multi-self-attention BiLSTM with XLNet. *Big Data and Cognitive Computing*, 9(2):37.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 11–20. Association for Computational Linguistics.

Ahmed Mustafa Younes. 2026. DeformAR: A visual analytics framework for evaluation of Arabic named entity recognition. In *Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script*, pages 253–275.

Mohamad Bilal Zbib, Hasan Abed Al Kader Hammoud, Ammar Mohanna, Nadine Rizk, Fatima Karnib, Sina Moukaled, and Bernard Ghanem. 2026. AraLingBench: A human-annotated benchmark for evaluating Arabic linguistic capabilities of large language models. In *Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script*, pages 385–393.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.

Benchmarking Speech Foundation Models and LLMs for Sentiment Analysis in Low-Resource Najdi Arabic of Saudi Arabia

Nadia Ghezaiel Hammouda
University of Hail, Saudi Arabia
ghezaielnadia.ing@gmail.com

Maram I. Alharbi
Lancaster University, UK
Jazan University, Saudi Arabia
m.i.alharbi@lancaster.ac.uk

Ruslan Mitkov
Lancaster University, UK
University of Alicante, Spain
r.mitkov@lancaster.ac.uk

Abstract

This paper addresses a gap on two fronts simultaneously in the Najdi dialect: unsupervised speech-based pseudo-labeling and instruction-tuned LLM classification. On the speech side, fourteen audio foundation models–spanning self-supervised learning, audio-language alignment, spectrogram encoding, and ASR–are evaluated through a K-Means pseudo-labeling pipeline applied to 181 recordings (15,377 segments). Cluster quality is assessed via Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index, and Shannon entropy; inter-encoder agreement is quantified through pairwise Cohen's κ and Fleiss' κ_F, treating each encoder as an independent annotator. XLSR-53 leads all geometric metrics by a wide margin (SS = 0.320, DBI = 1.309, CHI = 8660.7), attributable to its Arabic-inclusive pre-training. On the text side, three LLMs; SILMA-9B, LLaMA3-8B, and Qwen2.5-7B–are evaluated across zero-shot, few-shot, LoRA zero-shot, and LoRA few-shot conditions on 181 conversational transcripts (289,888 words). Analysis of prediction distributions across all twelve model–condition pairs reveals that label skew, output-parsing failures, and LoRA adaptation interact differently across architectures: Qwen2.5-7B zero-shot most closely mirrors the ground-truth label distribution, while SILMA-9B suffers from severe Positive bias and low parse rates under all conditions. Together, the results yield a composite encoder ranking–an annotation-free, practical tool for selecting speech models in downstream multimodal fusion pipelines targeting Najdi Arabic and similarly under-resourced Saudi dialects.

Keywords: Najdi Arabic, Saudi Arabia, sentiment analysis, speech encoders, LLM, LoRA, cluster quality, inter-rater agreement

1 Introduction

The challenge of sentiment analysis for low-resource dialects remains substantially underexplored, it extends beyond the linguistic characteristics of Arabic. In fact, Arabic famous by a rich morphology, dialectal variation, and diglossia and a limited availability of high-quality annotated corpora. Najdi Arabic exemplifies this problem. Najdi is the principal dialect of the Najd region in central Saudi Arabia, spoken natively by approximately 13 million people within the Kingdom (Habberrih and Abuzaraida, 2024; Alhedayani, 2024). Despite being the dominant vernacular of the world's largest Arabic-speaking economy, Najdi remains one of the least-resourced major Arabic dialects for sentiment analysis. Several existing benchmarks and works, such as NADI (Abdul-Mageed et al., 2023) and AHaSIS (Alharbi et al., 2025), have accelerated research on several Arabic varieties, such as Egyptian and Levantine Arabic, but dedicated sentiment resources for Najdi Arabic are still lacking. As a result, the effectiveness of both speech-based and text-based sentiment analysis methods for this Saudi dialect remains largely unknown. This resource gap motivates the need for systematic evaluation frameworks capable of assessing sentiment representations without relying on extensive manual annotation.

One of the fundamental challenges is the modality gap between speech and text. In fact, transcripts lose everything prosodic such as the flat delivery, the rising pitch, the stretched vowel that signals irritation besides the speech encoders most likely to recover that information were benchmarked almost exclusively on English (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2021; Baevski et al., 2022). Whether any of it transfers to Najdi conversational recordings is an empirical question that cannot be answered by inspection. On the text side, instruction-tuned LLMs have proven surprisingly effective at Arabic sentiment classification even without task-specific data (Zouidine and Khalil, 2025; Abdelali et al., 2024), and LoRA Alharbi

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 237–245
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_024

et al. (2025) makes targeted domain adaptation feasible with only a few hundred labeled examples.

This paper addresses both tracks simultaneously on the same dialect. Fourteen speech encoders go through an unsupervised K-Means pipeline on 181 Saudi Najdi recordings; we treat the resulting model disagreements as inter-rater reliability data, computing pairwise kappa and Fleiss' κ_F treating each encoder as an annotator–to our knowledge the first time this framing has been applied to speech pseudo-labelers. Three LLMs are tested under four prompting conditions. Unlike high-resource languages, Najdi Arabic lacks both large annotated speech corpora and established benchmarks for speech-based sentiment analysis. Consequently, the proposed framework combines complementary unsupervised and prompt-based approaches to evaluate speech and text representations without relying on extensive manual annotation. Rather than proposing a new learning algorithm, the objective is to systematically benchmark existing speech foundation models and LLMs to identify those that transfer most effectively to this low-resource dialect and provide practical guidance for future multimodal sentiment analysis.

The contributions are:

- First benchmark for Najdi Saudi Arabic speech and transcript sentiment analysis.
- Novel evaluation framework for annotation-free speech sentiment assessment in a low-resource Saudi dialect.
- Comprehensive comparison of 14 speech encoders and 3 LLMs on Saudi Najdi data, including cross-modal analysis and recommendations for multimodal fusion.

2 Related Work

2.1 Arabic Dialect Sentiment Analysis

Sentiment analysis for Arabic dialects has always had a data problem. Work on MSA is comparatively rich; work on spoken regional varieties is not. Habberrih and Abuzaraida Habberrih and Abuzaraida (2024) document the drop in pipeline performance once you leave MSA, and dialect-specific tools are rarely available to offset it. Things have improved somewhat at the shared-task level: NADI (Abdul-Mageed et al., 2023) extended evaluation to province-level dialects; AHaSIS 2025 Alharbi et al. (2025) introduced Saudi Arabic and Moroccan benchmarks–covering Gulf and Saudi varieties broadly but not Najdi specifically–with a top ensemble (MARBERTv2, SaudiBERT, DarijaBERT) at 0.81 micro-F1. Classifier comparisons consistently favour transformers over classical methods (Alosaimi et al., 2024). None of it works on audio.

2.2 Speech Emotion and Sentiment Recognition

SSL changed SER in roughly the same way it changed ASR, though the timeline is a few years behind. Wang et al. (2022) fine-tuned Wav2Vec2 and HuBERT to 79.58% weighted accuracy on IEMOCAP, beating hand-crafted feature baselines by a clear margin. Morais et al. (2022) showed that fine-tuning is not always necessary, as frozen SSL representations already outperform earlier baselines. Jafarzadeh et al. (2024) replicated this across RAVDESS, SHEMO, SAVEE, and Emo-DB. More recent work has broadened the space: emotion2vec (Ma et al., 2024) incorporates affect into pre-training itself; CLAP (Wu et al., 2024) grounds audio in natural language at scale; AST rethinks the input entirely using spectrogram patches. The English training data dominates throughout.

2.3 LLM-Based Sentiment Analysis for Arabic

LLaMA-family models currently lead the Arabic sentiment benchmarks among open-source options (Zouidine and Khalil, 2025). AraReasoner (Hasanaath et al., 2025) attributes much of that advantage to example selection–13+ F1 point swings from prompting strategy alone. LAraBench (Abdelali et al., 2024) benchmarked GPT-3.5, GPT-4, and Jais across 33 Arabic tasks and confirmed few-shot gains are robust. LoRA adaptation has been shown to improve performance when labeled data are available (Alkaoud, 2025); however, Wu et al. (2025) noted that morphological complexity can make these gains uneven or even negative.

2.4 Multimodal Sentiment Analysis

Gandhi et al. (2023) survey CMU-MOSI, CMU-MOSEI, and CH-SIMS and find a consistent pattern: text usually dominates, but audio fills in what text misses. Wu et al. Wu et al. (2024) combined data2vec audio with text in a multi-loss setup and topped three benchmarks; Abdulhalim et al. (2025) reached similar conclusions with frozen encoder fusion. Both rely on audio representations that

actually carry affect–exactly what our encoder evaluation tries to identify.

For ease of comparison, Table 1 provides a concise summary of the representative studies discussed in this section, highlighting their research focus, language, modality, main contribution, and the remaining limitations addressed by the present work.

3 Dataset

One hundred and eighty one recordings, native Saudi Najdi speakers, no sentiment labels. The corpus is drawn from the Najdi Arabic Corpus introduced by Alhedayani (2024), a resource built specifically for this under-represented Saudi dialect. Recordings were collected from speakers in the Najd region of Saudi Arabia, ranging from under a minute to around 18 minutes; each has speaker metadata and a human transcription. The absence of audio-level labels is not unusual for low-resource dialects.

Segmentation used 10-second non-overlapping windows with a 3-second floor. The mean duration of the segment of 9.97 s (std 0.37 s) reflects that the short segments come almost exclusively from the endpoints of the recording, not the cuts in the middle of the speech. The full set of 181 transcripts was used for LLM training; the 53-transcript evaluation split was reviewed label-by-label before anything was run on it. The transcripts are substantive in length (mean 1,601.6 words, median 1,628 words), ranging from 86 to 3,072 words, providing rich conversational context for both speech and text-based analysis.

4 Methodology

Figure 1 illustrates the end-to-end pipeline. The speech track (left, teal) segments 181 recordings into 10-second windows, extracts representations from 14 encoders spanning SSL, ASR, CLAP, and spectrogram families, then applies K-Means (K=3) to obtain pseudo-labels. The choice of $K = 3$ reflects the three target sentiment categories (Positive, Negative, and Neutral). Since no single metric can fully characterize clustering performance, multiple complementary evaluation measures are employed. Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Index assess different aspects of cluster compactness and separation, Shannon entropy measures the balance of the resulting cluster assignments, while pairwise Cohen's κ and Fleiss' κ_F quantify the agreement between pseudo-labels produced by different speech encoders. The transcript track (right, purple) passes the same 181 Arabic transcripts through three LLMs, SILMA-9B, LLaMA3-8B, and Qwen2.5-7B (silma-ai, 2024; Meta AI, 2024; Qwen Team, 2025), under four prompting conditions: zero-shot, few-shot, LoRA (Hu et al., 2021) zero-shot, and LoRA few-shot. Both tracks converge in a cross-modal comparison whose output is a composite encoder ranking, combining consensus rate, average κ, Silhouette Score, and Shannon entropy to provide an annotation-free assessment of speech representation quality for downstream multimodal fusion.

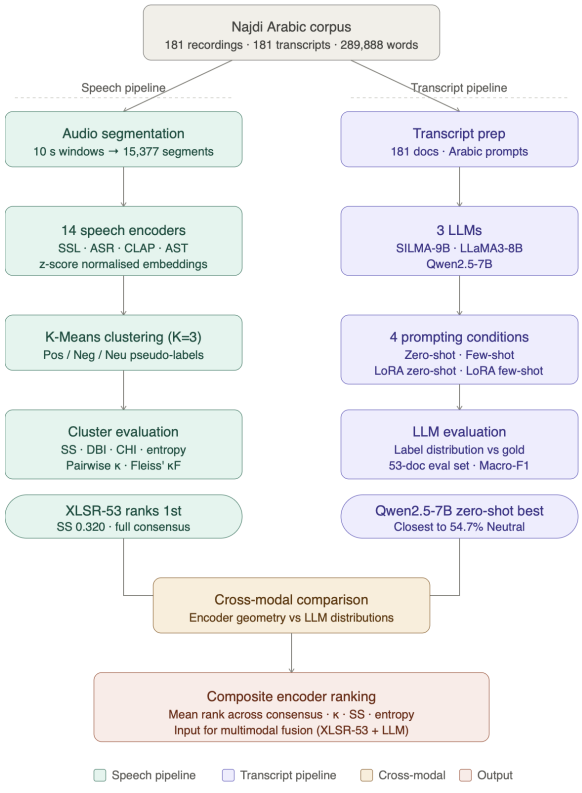


Figure 1: Overview of the dual-pipeline methodology. The speech track (teal) runs 14 encoders through an unsupervised K-Means pipeline on 181 Najdi recordings and evaluates cluster quality and inter-encoder agreement. The transcript track (purple) tests three LLMs under four prompting conditions on the same 181 transcripts. Both tracks converge into a cross-modal comparison that yields the composite encoder ranking used for multimodal fusion.

Table 1: Summary of the related work discussed in Section 2.

Reference	Task	Language	Modality	Main contribution	Remaining gap
Habberrih and Abuzaraida (2024)	Survey	Arabic dialects	Text	Review of Arabic dialect sentiment analysis	No speech evaluation
Abdul-Mageed et al. (2023)	Dialect benchmark	Arabic	Text	NADI shared task	Not sentiment; no Najdi speech
Alharbi et al. (2025)	Sentiment benchmark	Saudi Arabic	Text	AHaSIS shared task	No Najdi speech
Wang et al. (2022); Morais et al. (2022); Jafarzadeh et al. (2024); Ma et al. (2024)	Speech representation	Mainly English	Speech	SSL models outperform conventional SER methods	No Arabic evaluation
Zouidine and Khalil (2025); Abdelali et al. (2024); Alkaoud (2025); Hasanaath et al. (2025)	LLM sentiment analysis	Arabic	Text	LLMs achieve strong Arabic sentiment performance	No speech modality
Gandhi et al. (2023); Abdulhalim et al. (2025)	Multimodal sentiment	English/Chinese	Speech+Text	Fusion improves sentiment prediction	No Arabic multimodal benchmark
This work	**Benchmark**	**Najdi Arabic**	**Speech + Text**	**14 speech foundation models + 3 LLMs on the same corpus**	**Addresses the above gaps**

Statistic	Value
Total recordings	181
Segments (most models)	15,377
Segments (Whisper Medium)	30,663
Total predictions	>200,000
Seg. duration: mean	9.97 s
Seg. duration: std	0.37 s
Seg. duration: min	3.31 s
Total transcripts	181
Transcripts (train)	181
Transcripts (eval)	53
Total words	289,888
Mean words / transcript	1,601.6
Median words / transcript	1,628
Min words / transcript	86
Max words / transcript	3,072
Total characters	1,451,880
Mean characters / transcript	8,021.4
Median characters / transcript	8,281
Min characters / transcript	474
Max characters / transcript	14,227

Table 2: Najdi Saudi Arabic Speech and Transcript Dataset statistics.

5 Speech-Based Results and Analysis

5.1 Sentiment Distribution and Cluster Balance

Table 3 shows predicted sentiment distributions and entropy for all 14 encoders. HuBERT's values reflect the corrected full pipeline run (cluster 0: Neutral 58.7%, cluster 1: Positive 23.8%, cluster 2: Negative 17.5%, entropy 1.384); an earlier partial run had produced an artefactual 99.11% Positive result due to incomplete embedding extraction and is no longer used.

The two genuinely odd models are WavLM Base and Whisper Small. Both exceed 65% Positive; nothing about a Najdi conversational corpus justifies that, and it reflects embedding geometry rather than audio content. At the other end, w2v-BERT 2.0 nearly reaches the theoretical entropy ceiling (1.565 versus 1.585 maximum). HuBERT lands in the high-balance tier with entropy 1.384, similar to XLSR-53 (1.362) and Whisper Tiny (1.350)–a balanced, Neutral-dominant distribution consistent with the conversational nature of the Saudi Najdi corpus. The Whisper family shows a size-related pattern: Neutral bias grows with capacity, from 25% at Small up to 61% at Medium, consistent with larger models encoding more lexical content.

5.2 Cluster Quality Evaluation

Table 4 reports SS, DBI, and CHI for all 14 encoders.

XLSR-53 leads by a margin large enough to be unambiguous: SS 0.320 against 0.182 for the next model; CHI 8660.7 against 3558.2 for the runner-up. Pre-training data that includes Arabic is almost certainly the reason–every other model in the set was trained purely on English and must generalise across a phonetic gap it never saw. XLSR-53 at least shares part of the Saudi Arabic acoustic inventory, giving it a meaningful head start on Najdi speech.

The w2v-BERT 2.0 case is worth noting. It has the best entropy in the study–nearly uniform cluster sizes–yet its DBI of 3.151 is the worst value in the table. Uniform cluster sizes and geometrically separated clusters are two different things. Any workflow that used entropy alone to select an encoder would be led straight to the wrong model here.

Model	Positive (%)	Negative (%)	Neutral (%)	Entropy H (bits)
AST	19.4	53.6	27.1	1.451
CLAP	29.3	22.0	48.7	1.505
UniSpeech-SAT	43.0	42.2	14.9	1.457
Wav2Vec2 Base	41.8	16.3	41.9	1.478
Wav2Vec2 Large	19.5	34.5	46.0	1.505
WavLM Base	**65.5**	14.7	19.8	1.269
Whisper Base	23.4	46.1	30.5	1.528
Whisper Medium	29.3	9.3	61.4	1.270
Whisper Small	**67.5**	7.5	25.0	**1.162**
Whisper Tiny	34.2	10.8	55.1	1.350
XLSR-53	17.9	21.6	60.4	1.362
data2vec	17.5	41.6	40.9	1.494
w2v-BERT 2.0	26.4	33.8	39.7	**1.565**
HuBERT	23.8	17.5	58.7	1.384
Maximum possible				1.585

Table 3: Sentiment label distribution (%) and Shannon entropy for all 14 encoders. **Green**: w2v-BERT 2.0 (entropy 1.565)–most balanced distribution. **Red**: WavLM Base and Whisper Small–Positive bias >65%, least reliable for pseudo-labeling.

Model	Silhouette ↑	Davies-Bouldin ↓	Calinski-Harabasz ↑
XLSR-53	**0.320**	**1.309**	**8660.7**
UniSpeech-SAT	0.137	2.062	3558.2
Wav2Vec2 Base	0.122	2.202	3045.4
Whisper Medium	0.097	2.548	3775.2
WavLM Base	0.182	2.544	1737.2
Whisper Small	0.173	2.241	1778.6
Whisper Tiny	0.125	2.217	3136.3
CLAP	0.130	2.324	1855.0
Whisper Base	0.111	2.742	2116.8
AST	0.110	2.771	1390.2
Wav2Vec2 Large	0.087	2.925	1350.6
data2vec	0.130	2.097	92.9
w2v-BERT 2.0	0.059	3.151	844.8
HuBERT	0.110	2.752	1217.3

Table 4: Cluster quality metrics. ↑ = higher is better; ↓ = lower. **Green**: XLSR-53 leads all three metrics. **Light green**: UniSpeech-SAT and Wav2Vec2 Base–strong CHI (>3000) and DBI (<2.3), recommended secondary encoders. **Red**: w2v-BERT 2.0–best entropy (Table 3) but worst separability, illustrating the entropy–quality dissociation. Underlined: high SS but moderate CHI.

5.3 Inter-Model Agreement Analysis

Table 5 shows pairwise kappa across 10 models.

The nominally highest value is UniSpeech-SAT with Wav2Vec2 Base at 0.176– agreement driven by shared class bias, not shared signal. Both lean Positive/Neutral; they assign the same label for the same reason. The informative numbers are the negative ones. AST with w2v-BERT 2.0 lands at -0.402; the two Wav2Vec2 variants hit -0.325 with each other. Systematic opposition like that is potentially useful: in an ensemble, models that disagree coherently can cancel each other's errors. WavLM's average pairwise kappa of -0.193 (the most negative in the updated results) is the clearest signal to exclude it from any pseudo-labeling ensemble.

Fleiss' $\kappa_F = -0.029$ (Table 6) means observed agreement is lower than random chance–a consequence of strong Positive biases inflating the expected Positive term $\bar{P}_e$ well above what actually occurs. The calculation covers 5 recordings (the intersection including data2vec, which ran on a small representative sample). That is a conservative scope; more recordings would likely show more divergence.

5.4 Consensus-Based Evaluation and Encoder Ranking

Table 7 combines four criteria–consensus percentage, average pairwise kappa, Silhouette Score, Shannon entropy–ranked independently and combined by mean rank. It is a practical tool for deciding which encoders feed downstream pseudo-

	AST	CLAP	HuBERT	USAT	W2V-B	W2V-L	WavLM	W-Base	W-Med	w2vB2
AST	1.00	−.14	.00	−.14	.08	−.14	.01	−.02	.00	−.40
CLAP	−.14	1.00	.00	.05	−.07	.08	.09	.07	−.02	.26
HuBERT	.00	.00	1.00	.00	.00	.00	.00	.00	.00	.00
USAT	−.14	.05	.00	1.00	.18	.05	−.09	.15	.00	−.01
W2V-B	.08	−.07	.00	.18	1.00	−.33	−.12	−.07	.12	−.18
W2V-L	−.14	.08	.00	.05	−.33	1.00	.01	.18	−.02	.14
WavLM	.01	.09	.00	−.09	−.12	.01	1.00	−.09	−.06	−.07
W-Base	−.02	.07	.00	.15	−.07	.18	−.09	1.00	−.06	−.01
W-Med	.00	−.02	.00	.00	.12	−.02	−.06	−.06	1.00	.02
w2vB2	−.40	.26	.00	−.01	−.18	.14	−.07	−.01	.02	1.00

Table 5: Pairwise Cohen's κ (10 models). USAT = UniSpeech-SAT; W2V-B/L = Wav2Vec2 Base/Large; W-Base/Med = Whisper Base/Medium; w2vB2 = w2v-BERT 2.0. HuBERT kappa values of 0.00 across all pairs reflect its Neutral-dominant distribution (58.73%), which aligns with the majority-vote consensus label and produces zero disagreement beyond chance on the 176-recording file-level intersection.

Metric	Value
Annotators (encoders)	14
Items (recordings)	5
Sentiment classes	3
Fleiss' κ_F	−0.029
Interpretation	Poor (below chance)

Table 6: Global Fleiss' kappa (Landis & Koch, 1977). Negative value: observed agreement below chance, driven by divergent class priors.

labeling.

XLSR-53 at rank 1 is the only encoder that leads cluster geometry and reaches full consensus simultaneously. The other two models at 100% consensus (Whisper Tiny, Whisper Medium) share the same average kappa (0.172), but neither leads on geometry–they reach full consensus because their Neutral-heavy distributions dominate the majority vote.

HuBERT rises to rank 4, its consensus reach 60% and its average pairwise kappa is 0.087–the highest among the 60%-consensus group. Its label distribution across the 5 shared recordings shows 3 unique labels (Neutral: 2, Positive: 2, Negative: 1), confirming genuine three-way discrimination rather than a single-class collapse.

WavLM is the worst case. Rank 14, SS of 0.182. Tight clusters, but average kappa of −0.193 is now even more negative than the previous estimate (−0.142). A model that consistently carves out compact groups that no other encoder replicates is producing wrong labels, not precise ones.

Whisper Base hits zero consensus. Its 46% Negative distribution sits far enough outside the set's centre of gravity that the majority vote never lands on Negative.

HuBERT ranks 4th on all four criteria (consensus 60%, average κ 0.087, SS 0.110, entropy 1.384). Its cluster geometry is mid-table: SS of 0.110 matches AST and sits below the SSL leaders, but DBI of 2.752 and CHI of 1217.3 are consistent with meaningful three-way separation. Given its positive kappa and balanced distribution, it remains a viable secondary encoder for pseudo-labeling pipelines.

6 Transcript-Based Sentiment Analysis

6.1 Prediction Distributions

Table 8 shows the label distribution produced by each model–condition pair on the 53-document evaluation set (181 total transcripts used for training where applicable). Notable patterns emerge across configurations. SILMA-9B zero-shot assigns 82.1% Positive labels and produces zero Neutral predictions–extreme positive skew that is unlikely to reflect corpus content. LLaMA3-8B shifts markedly under few-shot conditions toward Neutral (41.5%), reflecting greater sensitivity to in-context demonstrations. Qwen2.5-7B is consistently the most Neutral-dominant model: its zero-shot and LoRA variants all exceed 55% Neutral, closely matching the 54.7% Neutral ground-truth distribution. After LoRA adaptation, SILMA-9B gains Neutral coverage (26.9%) but raises Positive predictions to 61.5%. The SILMA-9B LoRA zero-shot row reflects a partially parsed output (52 valid predictions out of 53).

6.2 Analysis by Configuration

Zero-shot prompting reveals clear architectural differences among the evaluated LLMs. SILMA-9B consistently exhibits a pronounced Positive bias

Rk	Model	Cons. (%)	Avg. κ	SS ↑	Entropy ↑	Mean Rank
1	XLSR-53	100	0.172	**0.320**	1.362	**3.25**
2	Whisper Tiny	100	0.172	0.125	1.350	5.00
3	data2vec	60	0.009	0.130	1.494	5.50
4	HuBERT	60	0.087	0.110	1.384	6.00
5	w2v-BERT 2.0	60	0.068	0.059	**1.565**	5.75
6	Whisper Medium	100	0.172	0.097	1.270	6.25
7	Wav2Vec2 Base	60	0.038	0.122	1.478	6.25
8	CLAP	60	−0.139	0.130	1.505	6.25
9	UniSpeech-SAT	20	0.061	0.137	1.457	6.75
10	Wav2Vec2 Large	40	−0.064	0.087	1.505	9.00
11	Whisper Small	20	0.000	0.173	1.162	9.00
12	Whisper Base	**0**	−0.034	0.111	1.528	9.00
13	AST	20	−0.012	0.110	1.451	9.50
14	WavLM Base	20	−0.193	0.182	1.269	9.75

Table 7: Composite encoder ranking (lower Mean Rank = better). **Dark green**: top tier–100% consensus and highest avg κ (0.172). **Light green**: Ranks 3–5–60% consensus; HuBERT at rank 4 with $\bar{\kappa}$=0.087 and SS = 0.110 from the completed full run. **Pale green**: usable middle tier (Ranks 6–9). **Pale red**: Ranks 10–13–negative or zero avg κ, low consensus. **Dark red**: WavLM Base–highest neg. $\bar{\kappa}$ (−0.193), anti-correlated with all others.

Model	Setting	Total Trans.	Valid Preds	Pos	Neg	Neu	Pos %	Neg % / Neu %
SILMA-9B	Zero-shot	181	28	23	5	0	82.1	17.9 / 0.0
	Few-shot	181	25	15	10	0	60.0	40.0 / 0.0
	LoRA Zero-shot	181	52	32	6	14	61.5	11.5 / 26.9
	LoRA Few-shot	181	53	40	10	3	75.5	18.9 / 5.7
LLaMA3-8B	Zero-shot	181	53	31	6	16	58.5	11.3 / 30.2
	Few-shot	181	53	19	12	22	35.8	22.6 / 41.5
	LoRA Zero-shot	181	53	26	11	16	49.1	20.8 / 30.2
	LoRA Few-shot	181	53	25	14	14	47.2	26.4 / 26.4
Qwen2.5-7B	Zero-shot	181	53	14	8	31	26.4	15.1 / 58.5
	Few-shot	181	53	11	14	28	20.8	26.4 / 52.8
	LoRA Zero-shot	181	52	20	2	30	38.5	3.8 / 57.7
	LoRA Few-shot	181	52	15	12	25	28.8	23.1 / 48.1

Table 8: LLM prediction distributions over the 53-document evaluation set. **Dark green**: LLaMA3-8B LoRA zero-shot and Qwen2.5-7B zero-shot– closest to the true 26.4 / 18.9 / 54.7% label distribution. **Dark red**: SILMA-9B zero-shot (82.1% Positive, 0% Neutral) and Qwen2.5-7B LoRA zero-shot (only 3.8% Negative). †One transcript produced an unparseable output; 52 valid predictions.

together with frequent malformed outputs, whereas Qwen2.5-7B most closely follows the ground-truth label distribution under zero-shot prompting. LoRA adaptation improves the performance of some models, particularly LLaMA3-8B, but does not consistently reduce prediction bias across all architectures. Similarly, combining LoRA with few-shot prompting produces mixed effects, indicating that the effectiveness of adaptation strategies remains model dependent for low-resource Najdi Arabic sentiment classification.

7 Cross-Modal Comparison

Comparing different types of evidence. Speech encoder results are characterised by geometric cluster scores and inter-model agreement; LLM results here are characterised by prediction distributions relative to the known gold-label distribution (26.4% Pos, 18.9% Neg, 54.7% Neu). These two evidence types cannot be stacked directly. What can be said is that the 14 speech encoders together land at κ_F=−0.029–below chance agreement–while Qwen2.5-7B zero-shot (26.4% Pos, 15.1% Neg, 58.5% Neu) and LLaMA3-8B LoRA zero-shot (49.1% Pos, 20.8% Neg, 30.2% Neu) are the LLM conditions whose distributions most closely shadow the true label proportions. That distributional alignment is a necessary (though not sufficient) condition for good classification. Formal Macro-F1 and Cohen's κ evaluation against gold labels remains as future work once the gold-labeling pipeline is complete.

Encoder selection. The ranking (Table 7) is the output designed for this step. XLSR-53 first; HuBERT (rank 4), data2vec, and w2v-BERT 2.0 as credible secondary choices; UniSpeech-SAT and Wav2Vec2 Base as further options. Whisper Base and WavLM Base are excluded entirely–WavLM's average κ of -0.193 is the most negative in the set.

Fusion direction. Pitch contours and lexical content carry different and sometimes contradictory sentiment signals. A fusion model can exploit both. XLSR-53 pseudo-labels combined with LLM predictions–particularly from the condition whose distribution best matches the corpus–can enter a Dynamic Attention Fusion framework (Abdulhalim et al., 2025) or a multi-loss architecture (Wu et al., 2024) without requiring audio annotations.

8 Conclusion

We set out to evaluate what speech encoders and LLMs can do for Najdi Arabic sentiment in Saudi Arabia without labeled audio. Results are unambiguous on the encoder side; the LLM side reveals important patterns in output behaviour that inform the path to formal evaluation.

XLSR-53 is the best encoder for this task–the only one combining strong cluster geometry with full consensus agreement. The remaining 13 encoders collectively disagree more than chance (κ_F=-0.029), making encoder selection mandatory before any pseudo-labeling application. HuBERT rises to rank 4 in the updated results (consensus 60%, average κ 0.087, SS 0.110, DBI 2.752, CHI 1217.3, entropy 1.384), confirming it as a reliable secondary encoder for pseudo-labeling pipelines.

On the LLM side, prediction distribution analysis shows that Qwen2.5-7B zero-shot and LLaMA3-8B LoRA zero-shot produce distributions most closely aligned with the 54.7% Neutral ground-truth. SILMA-9B's severe Positive bias (82.1% zero-shot) and low valid-parse rate point to prompt–dialect mismatch as a practical obstacle before formal scoring is possible on Saudi Najdi data. Formal Macro-F1 and Cohen's κ evaluation is the immediate next step, pending completion of gold-label annotation across the full evaluation set.

The longer-term goal is a multimodal fusion model combining XLSR-53 acoustic pseudo-labels with LLM predictions for Saudi Najdi Arabic (Abdulhalim et al., 2025). The encoder ranking this paper provides is the input that step needs.

References

Ahmed Abdelali, Hamdy Mubarak, Shammur Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Samir Abdaljalil, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, Nizi Nazar, Youssef Elshahawy, Ahmed Ali, Nadir Durrani, Natasa Milic-Frayling, and Firoj Alam. 2024. Larabench: Benchmarking arabic ai with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 487–520, St. Julian's, Malta. Association for Computational Linguistics.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, El Moatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. Nadi 2023: The fourth nuanced arabic dialect identification shared task. In *Proceedings of ArabicNLP 2023*, pages 600–613, Singapore (Hybrid). Association for Computational Linguistics.

Sadia Abdulhalim, Muaz Albaghdadi, and Moshiur Farazi. 2025. Multi-modal sentiment analysis with dynamic attention fusion. *arXiv preprint arXiv:2509.22729*.

Maram Alharbi, Salmane Chafik, Saad Ezzini, Ruslan Mitkov, Tharindu Ranasinghe, and Hansi Hettiarachchi. 2025. Ahasis: Shared task on sentiment analysis for arabic dialects. *arXiv:2511.13335*.

Rukayah Alhedayani. 2024. The Najdi Arabic corpus: A new corpus for an underrepresented Arabic dialect. *Language Resources and Evaluation*, 59(2):1593–1612.

Mohamed Alkaoud. 2025. On the effectiveness of limited-data large language model fine-tuning for arabic. *PLOS ONE*.

Wael Alosaimi, Hossam Saleh, A. A. Hamzah, Noha El-Rashidy, A. Alharb, A. Elaraby, and Samir Mostafa. 2024. Arabbert-lstm: Improving arabic sentiment analysis based on transformer model and long short-term memory. *Frontiers in Artificial Intelligence*, 7:1408845.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv:2006.11477*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, and Furu Wei. 2021. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *CoRR*, abs/2110.13900.

Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444.

Abdullah Habberrih and Mustafa Ali Abuzaraida. 2024. Sentiment analysis of arabic dialects: A review study. In *Computing and Informatics*, Communications in Computer and Information Science (CCIS), pages 137–153. Springer.

Ahmed Abul Hasanaath, Aisha Alansari, Ahmed Ashraf, Salmane Chafik, Hamzah Luqman, and Saad Ezzini. 2025. Arareasoner: Evaluating reasoning-based llms for arabic nlp. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18898–18914, Suzhou, China. Association for Computational Linguistics.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *arXiv preprint arXiv:2106.07447*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Pourya Jafarzadeh, Amir Mohammad Rostami, and Padideh Choobdar. 2024. Speaker emotion recognition: Leveraging self-supervised models for feature extraction using wav2vec2 and hubert. *arXiv:2411.02964*.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, ShiLiang Zhang, and Xie Chen. 2024. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15747–15760, Bangkok, Thailand. Association for Computational Linguistics.

Meta AI. 2024. LLaMA 3: Open foundation and chat models.

Edmilson Morais, Ron Hoory, Weizhong Zhu, Itai Gat, Matheus Damasceno, and Hagai Aronowitz. 2022. Speech emotion recognition using self-supervised features. *arXiv:2202.03896*.

Qwen Team. 2025. Qwen2.5: A party of foundation models.

silma-ai. 2024. SILMA 9B Instruct v1.0.

Yingzhi Wang, Abdelmoumene Boumadane, and Abdelwahab Heba. 2022. A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. *arXiv:2111.02735*.

Chengyan Wu, Bolei Ma, Zheyu Zhang, Ningyuan Deng, Yanqing He, and Yun Xue. 2025. Evaluating zero-shot multilingual aspect-based sentiment analysis with large language models. *International Journal of Machine Learning and Cybernetics*, 16(10):8079–8101.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2024. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. *arXiv:2211.06687*.

Mohamed Zouidine and Mohammed Khalil. 2025. Large language models for arabic sentiment analysis and machine translation. *Engineering, Technology and Applied Science Research*, 15(2):20737–20742.

Spatial Entity Extraction Methods from Arabic Texts in the Context of Epidemiological Surveillance: A Comparative Study

Fatima Ezzahra El Houbri[1,2,3], **Najlae Idrissi**[1], **Mathieu Roche**[2,3], **Sarah Valentin**[2,3]
[1]IDC Team, FST, Sultan Moulay Slimane University, Beni-Mellal, Morocco
[2]CIRAD, UMR TETIS, F-34398 Montpellier, France
[3]TETIS, Univ. Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France
elhoubri.fatimaezzahra@usms.ac.ma, n.idrissi@usms.ma
{mathieu.roche, sarah.valentin}@cirad.fr

Abstract

The location of events plays a key role in epidemiological surveillance, as it allows epidemic outbreaks to be identified and their spread to be tracked. Several event-based surveillance systems rely on English translation before extracting spatial entities using pretrained models adapted to English. However, the step of translating Arabic texts prior to extraction can influence the quality of spatial entity recognition. This study compares two strategies for extracting spatial entities from Arabic texts related to epidemiological surveillance evaluated in terms of precision, recall, and F-measure. The first approach relies on machine translation using DeepL, Microsoft Azure, and Reverso, followed by the extraction of entities using the spaCy (small, medium, and large) and GLiNER models. The second strategy consists of directly extracting entities from Arabic texts by fine-tuning the AraBERT, AraELECTRA, ARBERT, and MARBERT models. The results indicate that the second approach offers superior performance, particularly with the AraBERT and AraELECTRA models, demonstrating the effectiveness of direct extraction on Arabic data compared to the translation-based strategy.

1 Introduction

Epidemiological information systems monitor various channels to detect and report emerging and re-emerging infectious diseases. These systems are classified into two main categories of surveillance: indicator-based surveillance (IBS), which uses official sources such as medical records, and event-based surveillance (EBS), which relies on informal sources such as social media or news articles on the internet. Despite its complex implementation, EBS stands out for its ability to provide early and timely alerts about health threats. This method is particularly suitable for remote areas or areas with inadequate health coverage, where unofficial information channels are the main source of information (Menya et al., 2024; Delon et al., 2025).

Recently, a lot of studies (Joshi et al., 2019) have focused on the use of textual data from unofficial sources for epidemic surveillance, particularly in the context of animal health surveillance. While organizations such as the World Organization for Animal Health publish official notifications, informal sources such as online news sites offer data that may be more timely. In practice, event-based epidemiological surveillance systems often rely on the automatic translation of text sources collected from different languages into a pivot language, usually English, to enable the extraction of epidemiological information using artificial intelligence and machine learning methods (Valentin et al., 2021).

This information includes spatial data. Accurate location determination is essential for assessing epidemic risks, enabling epidemiologists to identify potential outbreaks or monitor the spread of disease. This spatialization is a prerequisite for implementing appropriate control and protection measures and therefore requires high-quality translation. However, to our knowledge, the effect of translation prior to spatial entity extraction has not been evaluated for epidemiological surveillance, particularly for certain languages due to their stylistic and lexical diversity. These issues are particularly important for the Arabic language, which encompasses a wide variety of dialects, linguistic registers, and cultural nuances (Akan et al., 2019; Guellil et al., 2021). This complexity highlights the need to evaluate two methods/approaches of spatial entity extraction for detecting epidemiological events from Arabic-

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 246–256
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_025

language data.

In this context, our work consists of comparing two Arabic data extraction strategies. The first strategy is based on a traditional translation-based approach, where Arabic data is translated into English before applying named entity recognition models trained in English (**Trans-NER-EN**). The second strategy consists of performing named entity recognition directly on Arabic data using models trained specifically for Arabic (**NER-AR**). Our objective is to evaluate the effectiveness of both strategies to identify the best approach to Arabic data extraction for accurately detecting relevant spatial entities for epidemiological surveillance.

The remainder of the article is organized as follows. Section 2 presents related work on text-based epidemiological surveillance systems and pre-trained language models for Arabic. Section 3 describes the corpus used and the process for annotating spatial entities. Section 4 details the first strategy evaluated, namely the machine translation approach followed by named entity extraction in English (Trans NER-EN). Section 5 presents the second strategy, which consists of directly extracting spatial entities from Arabic texts using pre-trained models for Arabic (NER-AR). Section 6 outlines the evaluation metrics used and the results obtained for both strategies. Section 7 discusses the results obtained. Finally, Section 8 concludes the article and outlines prospects for future work.

2 Related Work

Text-based health monitoring systems, particularly HealthMap (Freifeld et al., 2008) and PADI-web (Valentin et al., 2021), aim to detect and extract epidemiological information from textual sources, such as online news. These systems automate several stages of text data processing, from document collection (e.g., online news) to the extraction of epidemiological information that can be used for analysis through the incorporation of artificial intelligence and machine learning methods.

Historically, these systems have been designed primarily for English-language texts (Arsevska et al., 2018), which limits their ability to identify events reported in local languages, particularly Arabic. This limitation has led to the development of multilingual systems that increasingly target different languages to guaranty direct access to local information sources. These systems generally include a machine translation step into a target language, often English, using open-source or proprietary translation systems (e.g., Google Translate, Microsoft Translator), to enable the use of pre-trained models for downstream tasks such as classification, named entity recognition, and event localization. Although spatial entity recognition is essential for correctly identifying epidemics and tracking their spread, to our knowledge, the impact of translating Arabic texts before automatic extraction has not been evaluated for epidemiological surveillance.

Recently, pre-trained language models (PLM) for Arabic have demonstrated superior representation capabilities for capturing contextual information within sentences. In general, PLMs based on Transformer-type architectures allow for direct analysis of Arabic texts while taking into account their lexical, morphological, and syntactic specificities, which significantly improve various NLP tasks, including Arabic Named Entity Recognition, which involves detecting and classifying named entities in a given text, including spatial entities (El Moussaoui and Loqman, 2024; Qu et al., 2023).

AraBERT (Antoun et al., 2020a) is a pre-trained language model, based on the architecture of BERT (Devlin et al., 2019) in English and specially designed for the Arabic language. The model is based on bidirectional learning using masked language modeling (MLM) and next sentence prediction (NSP) tasks. In order to take into account the linguistic structure of Arabic, AraBERT uses external segmentation to break words down into roots, prefixes, and suffixes. AraBERT's performance was evaluated on three Arabic language comprehension tasks, specifically named entity recognition, question answering, and sentiment analysis.

AraELECTRA (Antoun et al., 2020b) is another pre-trained model for Arabic, based on the architecture of ELECTRA (Clark et al., 2020) in English. It contains a generator that predicts the original tokens from corrupted sentences and a discriminator that determines which tokens are masked and which ones are original. During fine-tuning, the input sentence is fed to the discriminator with an additional linear layer for downstream tasks. AraELECTRA was

evaluated on the ANERCorp dataset, and the results clearly show its superiority (Qu et al., 2023).

MARBERT (Abdul-Mageed et al., 2021), meanwhile, is a large-scale pre-trained masked language model for Arabic dialects and Modern Standard Arabic. Based on the BERT-base architecture, it uses a 128 GB corpus of Arabic tweets and removes the next-word prediction task. Pre-trained on a wide range of dialects, MARBERT achieves state-of-the-art performance on named entity recognition tasks, particularly on the TW-NER dataset (Darwish, 2013).

Finally, **ARBERT** (Abdul-Mageed et al., 2021) is a large-scale masked language model specifically focused on Modern Standard Arabic (MSA). ARBERT follows the same architecture as BERT base: 12 attention layers, each with 12 attention heads and 768 hidden dimensions, with a vocabulary of 100,000 WordPieces, representing approximately 163 million parameters. It was trained on a 61 GB corpus of Arabic text, equivalent to 6.2 B tokens.

Although several studies have used fine-tuning of these models to extract named entities, to our knowledge, none have focused on the extraction of spatial entities within epidemiological surveillance. In this context, we conducted a comparative study on the performance of these fine-tuned models, evaluated on Arabic data related to animal epidemiological surveillance. This approach was then compared to the strategy based on translation before extraction to determine the best method for retrieving these spatial entities.

3 Corpus and annotation

To evaluate and compare the performance of both strategies to extract spatial entities in English and Arabic for spatial entity recognition, we constructed two datasets from 96 Arabic news articles harvested by the PADI-web platform between February 2025 and February 2026. These datasets contain 328 manually annotated spatial entities, corresponding to 142 distinct spatial entity values. Spatial entities were manually annotated and extracted by the first author, a native Arabic speaker. Among these entities, there are 275 Absolute Spatial Entities (ASE). These entities can be found on a map or in a geographic database because they are direct references to specific, locatable geographic places, such as continents, countries, and cities. The remaining 53 entities are Relative Spatial Entities (RSE), which are specified in relation to other spatial entities utilizing topological relationships. Examples include " West of Mosul " and " northwestern Hungary ".

The first, PADI-Location-AR-EN[1], is a parallel Arabic-English dataset dedicated to spatial entities from epidemiological texts. It combines manually validated translations using geographic databases and machine translations performed by automatic translators, as well as normalization through the GeoNames database using a unique identifier. This corpus is developed to evaluate the impact of machine translation of spatial entities before extraction and the performance of English named entity recognition models for spatial entity recognition applied to translated texts. For more details on the corpus construction, annotation, extraction, and normalization process see our previous study (Houbri et al., 2026).

The second, PADI-Location-AR-IOB, is a dataset containing the same spatial entities as PADI-Location-AR-EN but annotated according to the IOB schema, in which the labels B-LOC and I-LOC mark the beginning and interior of spatial entities, respectively, while the label O refers to other tokens that do not correspond to any entity. This corpus is constructed to evaluate the performance of named entity recognition models in Arabic for spatial entity recognition.

4 Trans-NER-EN Strategy

This section aims to evaluate the preservation of spatial information across two classical and sequential steps in multilingual text processing: (1) the translation of Arabic locations into English and (2) the automated extraction of locations from the translated text.

4.1 Translation

A spatial entity extracted from texts in Arabic can undergo significant alterations when translated into another language, such as English, which can lead to a partial loss of meaning or distortion of

[1] https://dataverse.cirad.fr/dataset.xhtml?persistentId=doi:10.18167/DVN1/ZA1YUO

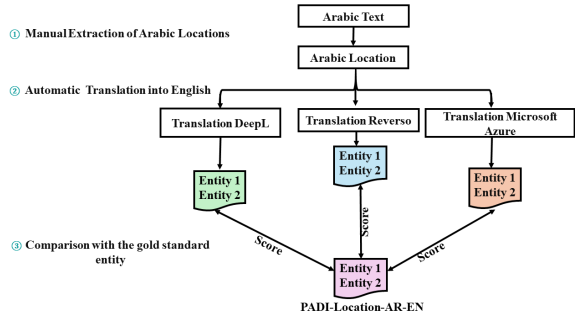


Figure 1: The Proposed Process for Translation Evaluation.

the spatial information (Almohish, 2024). In this context, computer-aided translation techniques play an essential role in identifying the most suitable machine translation system for converting spatial entities from Arabic to English. To this end, we are evaluating the performance of DeepL, Microsoft Azure, and Reverso by comparing their translations of spatial entities to the ground truth.

The three systems selected are briefly presented below, along with the reasons for their selection in our evaluation:

- **DeepL**[2]: One of the most renowned machine translation systems, based on convolutional neural networks (Laki and Yang, 2024). We chose DeepL because of its international reputation for excellent translations.
- **Microsoft Azure**[3]: The Microsoft Azure Translation API is a cloud-based machine translation service that you can use to translate texts and documents with a simple call to the REST API, which uses modern neural machine translation technology. This system is an operational service often integrated into text-mining tools, which makes its evaluation particularly relevant for epidemiological surveillance.
- **Reverso**[4]: Reverso is a suite of online applications and services using artificial intelligence to help translate, write, and improve language proficiency. It offers contextual translation, bilingual dictionaries, synonyms, conjugation, interactive grammar tools, as well as spelling and grammar correction based on neural machine translation [5]. We chose Reverso because it not only provides translations of individual words but also gives examples of paragraphs containing these translated words, enabling the translation to be checked in context.

Figure 1 shows the evaluation process for these three translation systems. The Arabic spatial entities, extracted and annotated manually from the 96 articles, served as a reference. We then automatically translated the sentences containing these entities using the three translators. The translated spatial entities were manually extracted from these texts, ensuring that they were translated in context and not in isolation. Finally, a comparison was made between the translated entities and those in our reference dataset. This comparison was done using two evaluation metrics: Exact Matching and Fuzzy Matching

4.2 Location extraction

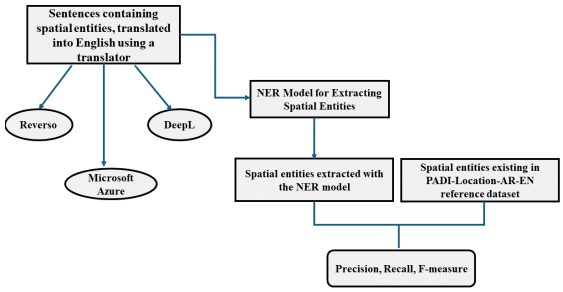


Figure 2: Spatial entities extraction pipeline based on extraction models, along with their evaluation.

As illustrated in Figure 2, to assess the quality of the named entity extraction models, we selected two named entity recognition models: spaCy[6] NER model, including its small[7], medium[8], and large[9] versions, and the GLiNER model (Zaratiana et al.,

[2] https://www.deepl.com/fr/translator
[3] https://learn.microsoft.com/en-us/azure/ai-services/translator/
[4] https://www.reverso.net/traduction-texte
[5] https://fr.wikipedia.org/wiki/Reverso_(logiciel)
[6] https://spacy.io/
[7] https://github.com/explosion/spacy-models/releases/tag/en_core_web_sm-3.8.0
[8] https://github.com/explosion/spacy-models/releases/tag/en_core_web_md-3.8.0
[9] https://github.com/explosion/spacy-models/releases/tag/en_core_web_lg-3.8.0

2024), capable of identifying any type of entity using a BERT-type bidirectional transformer encoder. We used the models to extract spatial entities from the English-translated sentences, resulting in a set of extracted entities. Finally, we compared these entities to those in our reference dataset using exact matching, based on the classic metrics of precision, recall, and F-measure. It should be noted that the objective here is not to exhaustively evaluate all existing NER tools, such as NuNER (Bogdanov et al., 2024), UniversalNER (Zhou et al., 2024), or fine-tuned transformer-based models like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), but rather to assess a small selection of representative and state-of-the-art models — some of which are often integrated into epidemiological surveillance platforms or text-mining tools — to demonstrate the usefulness of the PADI-Location-AR-EN dataset, which can serve as a resource for evaluating other NER tools in future work.

5 NER-AR Strategy

To evaluate the strategy of extracting spatial entities directly from Arabic data for epidemiological surveillance, as shown in Figure 3, we consider this approach as a named entity recognition (NER) task in Arabic. To do this, we fine-tuned several pre-trained language models for Arabic, including AraBERT, AraELECTRA, ARBERT, and MARBERT, to conduct a comparative study of their performance in spatial entity extraction.

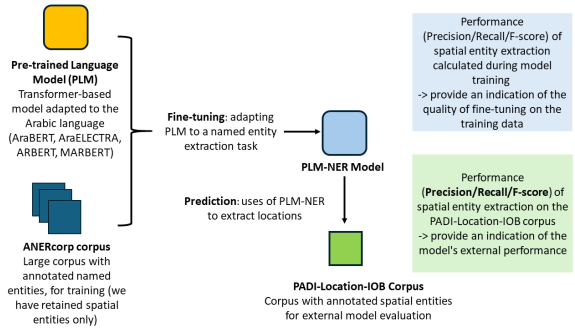


Figure 3: The proposed NER-AR Strategy on PADI-Location-AR-IOB

.

These models were trained on normalized training data containing only location entities from the ANERCorp corpus. The ANERCorp (Benajiba et al., 2007) corpus is a manually annotated Arabic corpus designed specifically for named entity recognition tasks. It consists of two subsets: train and test, and it contains over 150,000 tokens. Each token is associated with one of the following categories: person, place, organization, or other. It is crucial to note that in ANERCorp, location corresponds exclusively to absolute spatial entities, such as named toponyms (e.g., cities, countries, and regions). The corpus does not include or annotate spatial relative entities, which are not considered named entities within its annotation scheme.

During the normalization stage, we retained only annotations corresponding to spatial entities. More specifically, we retained the B-LOC and I-LOC labels, indicating the beginning and interior of spatial entities, respectively, as well as the O label for tokens that do not correspond to any entity. In total, the dataset contains **3,776 B-LOC**, **525 I-LOC**, and **120,801 O** labels. The Arabic PLMs were fine-tuned exclusively on the ANERCorp training split, using the following pre-trained checkpoints from the Hugging Face Hub: aubmindlab/bert-base-arabertv02 (AraBERT), aubmindlab/araelectra-base-discriminator (AraELECTRA), UBC-NLP/ARBERT (ARBERT), and UBC-NLP/MARBERT (MARBERT). Fine-tuning was performed using the Hugging Face Trainer with the AdamW optimizer, a learning rate of 5 × 10, a batch size of 16, gradient accumulation over two steps, a maximum sequence length of 256 tokens, weight decay of 0.01, six training epochs, a cosine learning rate scheduler with a 10% warm-up ratio, mixed-precision (FP16) training, and a fixed random seed of 42. Due to computational constraints, training was conducted with a single random seed; results may therefore be subject to some variance that was not quantified in this study. Precision, recall, and F-measure were computed using weighted averaging; since only one entity type (LOC) is annotated, this is numerically equivalent to a standard micro-average over entities.

We then evaluated the fine-tuned models on the PADI-Location-AR-IOB corpus, which is annotated with the same labels (B-LOC, I-LOC, and O) and contains exclusively spatial entities extracted from Arabic texts related to epidemiological surveillance. This dataset includes **328 B-LOC**, **115 I-LOC**, and **19,110 O** labels. No samples from the PADI-Location-AR-IOB corpus were used to update model weights or

to select the best checkpoint: no early stopping or best-model selection based on validation performance was applied, and training ran for a fixed number of epochs. Since our objective was to evaluate cross-domain generalization, no separate validation split was held out; test-set metrics were nonetheless logged periodically during training for monitoring purposes only, without influencing the training process or hyperparameter choices, and the fine-tuned models were evaluated directly on this independent test dataset at the end of training. Finally, each fine-tuned model was evaluated to conduct a comparative study of the models' performance. The results obtained were then compared with those of the Trans-NER-EN strategy to determine the best strategy for extracting spatial entities for the detection of epidemiological events from Arabic data. It should be noted that this comparison reflects two practical workflows rather than two strictly equivalent experimental settings, since the Arabic strategy relies on fine-tuned Arabic PLMs, whereas the Trans-NER-EN strategy uses off-the-shelf English NER systems after machine translation.

6 Evaluation

In this section, we introduce the evaluation metrics for translation and for the extraction of entities (from English-translated text and from Arabic text).

6.1 Evaluation Metrics for Translation

- **Exact Matching**: For each spatial entity translated by each of the translation systems, if the translation corresponds exactly to the character string present in our reference dataset, it is considered correct.
  The exact matching score corresponds to the percentage of spatial entities correctly translated by the translation system, calculated using Equation (1):

$$\text{Exact Matching Score} = \frac{\text{Number of correctly translated spatial entities}}{\text{Total number of spatial entities}} \quad (1)$$

- **Fuzzy Matching**: For each spatial entity translated by the translation systems, the translation is considered correct if it meets one of the following criteria. Fuzzy matching was performed manually according to predefined annotation guidelines. Each translated spatial entity was compared with the corresponding reference translation by the first author and subsequently reviewed by the co-authors to ensure consistency. Ambiguous cases were discussed collectively until a consensus was reached. No automatic semantic similarity metric was used. (1) exact match with the reference entity (e.g. 'Morocco' vs 'Morocco') (2) corresponds to an alternative name recognized in GeoNames or Google Maps (e.g. 'Fes' vs 'Fez') (3) differs from the reference entity by a single character operation according to the Levenshtein distance (e.g. 'Joumine' vs 'Jomine', by deleting the 'u' character), (4) has the same contextual meaning despite differences in wording while referring to the same geographic location (e.g. 'Northwestern China' vs. 'Northwest China'), or (5) is a reformulation or lexical rearrangement that retains the same meaning without modifying the referenced spatial entity (e.g. 'Oran Province' vs. 'Province of Oran').
  The fuzzy matching score corresponds to the percentage of spatial entities that satisfy at least one of these criteria, including exact matches, calculated according to Equation (2):

$$\text{Fuzzy Matching Score} = \frac{\text{Number of spatial entities meeting at least one of the criteria}}{\text{Total number of spatial entities}} \quad (2)$$

6.2 Evaluation Metrics for Extraction

To evaluate the performance of named entity extraction models, we used three metrics: precision, recall and F-measure. These metrics are calculated from true positives (TP), false positives (FP) and false negatives (FN), defined as follows:

- **True positives (TP)**: number of spatial entities correctly extracted by the model that appear in our reference dataset.
- **False positives (FP)**: number of spatial entities extracted by the model but not appearing in our reference dataset.
- **False negatives (FN)**: number of spatial entities not extracted by the model but appearing in our reference dataset.

Precision corresponds to the percentage of correct spatial entities extracted by the model, calculated

using the following formula:

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{\text{Number of spatial entities correctly extracted by the model}}{\text{Total number of spatial entities extracted by the model}} \quad (3)$$

Recall represents the percentage of spatial entities existing in our reference set that are found by the model and is calculated using formula (4):

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{\text{Number of spatial entities correctly extracted by the model}}{\text{Total number of spatial entities present in the reference dataset}} \quad (4)$$

A measure that incorporates both precision and recall is the harmonic mean of these two measures, known as the F-measure (formula (5)):

$$\text{F-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

6.3 Results

6.3.1 Results of the translation

Table 1 shows the overall scores obtained by the translation systems evaluated. DeepL and Microsoft achieve the highest scores for exact matching, successfully translating 226 and 222 out of 328 entities, respectively. On the other hand, Reverso lags behind, with a score of 59.75%, reflecting a significantly lower level of accuracy and translation quality than DeepL and Microsoft Azure.
In terms of fuzzy matching, Microsoft Azure came out on top with a score of 86.58%, demonstrating a strong ability to produce translations that closely align with the intended meaning, even when not literal. DeepL follows closely with a score of 85.06%, confirming its consistency in translation quality. In contrast, Reverso scored lower at 74.39%, underlining weaker performance in preserving accurate translations.

Translator	Exact Matching Score (%)	Fuzzy Matching Score (%)
Reverso	196 / 328 = 59.75	244 / 328 = 74.39
DeepL	226 / 328 = **68.90**	279 / 328= **85.06**
Microsoft Azure	222 / 328 = **67.68**	284 / 328 = **86.58**

Table 1: Performance of the three translators for the translation of spatial entities in terms of exact and fuzzy match scores.

6.3.2 Results of entity extraction (from English-translated text)

According to the results presented in Table 2, Table 3, and Table 4, the spaCy small model achieves the lowest performance of the medium and large spaCy models, particularly in terms of recall. For example, it achieves an F-measure of 51.46% when used with Microsoft Azure, and 48.79% with DeepL. These results reflect the model's limitations in terms of complexity and ability to recognize entities in a variety of contexts, making it a less suitable choice for tasks requiring a high degree of language understanding. As shown in Table 3 and Table 4,

Translator	Precision (%)	Recall (%)	F-measure (%)
Reverso	64.70	33.53	44.16
Microsoft Azure	**72.37**	**39.93**	**51.46**
DeepL	**70.93**	**37.19**	**48.79**

Table 2: Evaluation metrics of spatial entity extraction with spaCy's small model, according to the translations of the three translators.

the spaCy medium model delivers slightly lower performance than the large model, but is still superior to the small model. For example, when used with Microsoft Azure, it achieves an F-measure of 63.05%, while with DeepL, it achieves an F-measure of 51.04%.

Translator	Precision (%)	Recall (%)	F-measure (%)
Reverso	64.28	35.67	45.88
Microsoft Azure	**71.65**	**56.30**	**63.05**
DeepL	**68.91**	**40.54**	**51.04**

Table 3: Evaluation metrics of spatial entity extraction with spaCy's medium model, according to the translations of the three translators.

Table 4 shows that spaCy's large model outperforms the small and medium models in spatial feature extraction, irrespective of which of the three evaluated translators is used. Its F-measure scores are higher than those of the other models. For example, with Microsoft Azure, it achieves an F-measure of 58.03%, while with DeepL, it achieves an F-measure of 54.53%.

Table 5 demonstrates the competitiveness of

Translator	Precision (%)	Recall (%)	F-measure (%)
Reverso	62.28	38.10	47.27
Microsoft Azure	**73.70**	**47.86**	**58.03**
DeepL	**69.66**	**44.81**	**54.53**

Table 4: Evaluation metrics of spatial entity extraction with spaCy's large model, according to the translations of the three translators.

the GLiNER model, especially when paired with DeepL and Microsoft Azure translations. It surpasses the spaCy large model's scores with an F-measure of 67.96% with Microsoft Azure and

65.49% with DeepL. These results show that GLiNER offers improved spatial entity recognition capability, particularly in multilingual or translated contexts, and is a powerful alternative to traditional spaCy models. However, even with the best-performing approach (GLiNER), approximately 36.29% of locations are not extracted (63,71% recall), which is an important consideration for epidemiological surveillance applications.

Translator	Precision (%)	Recall (%)	F-measure (%)
Reverso	61.70	53.04	57.04
Microsoft Azure	**72.82**	**63.71**	**67.96**
DeepL	**68.79**	**62.50**	**65.49**

Table 5: Evaluation metrics of spatial entity extraction with the GLiNER model, according to the translations of the three translators.

6.3.3 Results of entity extraction (from Arabic text)

We evaluated the performance of the fine-tuned AraBERT, AraELECTRA, ARBERT, and MARBERT models for extracting annotated locations from the EpiDataset2 dataset (Table 6).

Arabic PLM	Precision (%)	Recall (%)	F-measure (%)
AraBERT	**79.73**	**81.35**	**80.53**
AraELECTRA	**79.86**	**80.67**	**80.26**
ARBERT	79.5	78.45	78.98
MARBERT	79.6	67.68	73.16

Table 6: Performance of Arabic pre-trained language models (PLMs) for spatial entity extraction in terms of precision, recall, and F-measure. The two best values for each metric are shown in bold.

The obtained results demonstrate that the AraELECTRA model offers the best accuracy (79.86%), while AraBERT presents the best balance between accuracy and recall, with an F-measure of 80.53%. ARBERT performs slightly worse but remains competitive, and MARBERT achieves decent accuracy (79.6%) but lower recall (67.68%), which reduces its F-score (73.16%). Overall, these performances confirm the effectiveness of fine-tuning pre-trained models on standard Arabic for direct extraction of spatial entities.

7 Discussion

The results obtained show that the performance of absolute and relative spatial entity detection depends heavily on the choice of translator and extraction model used. Indeed, the quality of the translation directly influences the ability of the NER model to correctly identify spatial entities. The aim is to determine which translator provides the most accurate results while faithfully preserving the original spatial meaning. It should be noted that the aim of this study is not to comprehensively evaluate all current translation tools, especially since many other AI-based systems, such as Google Translate and others, may offer competitive performance. Instead, we selected a representative panel of three translators to provide an initial direction and demonstrate the usefulness of the manually created set in supporting this type of evaluation. Future work can easily extend this approach to other translation systems. Furthermore, the inclusion of Microsoft Azure is particularly significant. This allows us to assess the reliability of the translation tool currently used for epidemiological surveillance. In addition, when analyzing errors, translators generally do not make mistakes with the same spatial entities.

In the Trans-NER-EN strategy based on translation, even when combining the best translation system with the most effective extraction model (GLiNER), around 36.29% of spatial entities are not extracted, representing a recall of 63.71%, which is important to take into account for epidemiological surveillance applications. Extrapolating, this means that 36.29% of locations, and therefore potentially outbreaks, are ”missed”. There is therefore considerable room for improvement. An important perspective would be to compare the performance obtained in this study with that reported for NER models when fine-tuned on English training data, to assess the extent to which translation from Arabic impacts their performance. Furthermore, the NER-AR strategy, which relies on the direct extraction of spatial entities from Arabic texts, appears to offer better performance for the identification of spatial entities, highlighting the relevance of an approach that is linguistically adapted to the Arabic context. Encoder-based PLMs such as AraBERT and AraELECTRA are specifically designed for sequence labeling tasks like NER and can be efficiently fine-tuned on small specialized corpora such as ours (328 spatial entities) with limited computational resources. In contrast, LLMs such as GPT-4 and LLaMA are generally used in few-shot or zero-shot settings without fine-tuning, which makes them less suitable for specialized and small-scale corpora

such as those encountered in epidemiological surveillance. However, a direct comparison with the translation-based strategy remains to be established, as the spatial entities extracted in Arabic have not yet been normalized or translated into English. In this context, relative spatial entities (RSEs) represent a major challenge in terms of normalization. This can be explained by the very nature of these entities: they do not correspond to officially recognized administrative units but rather to directional or contextual descriptions of known places, such as ”northwestern China” or ”west of Mosul.” The next step will therefore be to evaluate the geocoding of the spatial entities extracted using both strategies in order to assess the consistency between the two methods. It would also be interesting to examine the feasibility of direct geocoding (Mubarak and Hassan, 2021), which involves converting spatial entities into geographic coordinates, from Arabic entities, without going through translation, which would reduce dependence on machine translation tools and improve the overall accuracy of the process.

8 Conclusion

In this work, we conducted a comparative study comparing two spatial entity extraction strategies for detecting epidemiological events from Arabic data, using precision, recall, and F-measure as evaluation metrics. The first, known as the traditional strategy, involves translating Arabic texts into English before applying named entity recognition models trained in English (Trans-NER-EN) to extract spatial entities from the translated texts. The second strategy (NER-AR) consists of directly extracting spatial entities from Arabic data using pre-trained models adapted to this language. The objective of this study was to evaluate the effectiveness of these two approaches to identify the most suitable strategy for accurately extracting relevant spatial entities for epidemiological surveillance. The results show that the NER-AR strategy gives the best results in terms of precision, recall, and F-measure, particularly thanks to the application of fine-tuning on pre-trained models for Arabic. In contrast, the translation-based Trans-NER-EN strategy performs less well, highlighting the need for high-quality translation and high-performance pre-trained models for the target language to extract translated entities correctly. However, these results are based on a relatively small dataset drawn from a single surveillance source/domain, and should not be taken to imply that the same performance ranking will necessarily hold across other Arabic dialects, data sources, or types of epidemiological events. Future work could explore the extraction of other types of epidemiological entities, such as hosts and symptoms, and evaluate the performance of translation-based approaches using newer translation tools based on large language models (LLMs). It would also be valuable to extend the comparison to additional Arabic pre-trained language models, such as CAMeLBERT, as well as multilingual models, to provide a broader benchmark. Furthermore, future studies could complement the evaluation with translation quality metrics such as BLEU and ChrF++ (Kumar et al., 2026), perform more detailed error analyses according to spatial entity type (e.g., absolute versus relative spatial entities), and report uncertainty estimates or statistical significance analyses to better assess the robustness of the observed differences. Finally, qualitative examples comparing the outputs of the two extraction strategies would help illustrate their respective strengths, limitations, and practical relevance for epidemiological surveillance.

Acknowledgments

The authors thank Julien Rabatel for using the Microsoft Azure technology linked with PADI-web to translate phrases into English.
Under the ”PhD Associate Scholarship – PASS” initiative for doctorate student monitors in the digital field by 2027—National Strategy Morocco 2030—the CNRST is supporting this work.
The MOOD (Monitoring Outbreaks for Disease surveillance in a data science context) project, DGAL, and the French National Research Agency under the Investments for the Future Program: ANR-16-CONV-0004 (DigitAg) have provided funding for this work.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Md Faruquzzaman Akan, Md Rezaul Karim, and Abdullah Mohammad Kabir Chowdhury. 2019. An analysis of arabic-english translation: Problems and prospects. *Advances in Language and Literary Studies*.

N. R. I. Almohish. 2024. Spatial and diachronic challenges encountering literary arabic translators of the classical arabic literature into english. *Research Journal in Advanced Humanities*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020a. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020b. Araelectra: Pre-training text discriminators for arabic language understanding. *ArXiv*, abs/2012.15516.

Elena Arsevska, Sarah Valentin, Julien Rabatel, Jocelyn De Goër de Hervé, Sylvain Falala, Renaud Lancelot, and Mathieu Roche. 2018. Web monitoring of emerging animal infectious diseases integrated in the French animal health epidemic intelligence system. *PLoS ONE*, 13(8). Publisher: Public Library of Science.

Yassine Benajiba, Paolo Rosso, and José-Miguel Benedí. 2007. Anersys: An arabic named entity recognition system based on maximum entropy. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing. CICLing 2007*, volume 4394 of *Lecture Notes in Computer Science*, pages 148–159. Springer, Berlin, Heidelberg.

Sergei Bogdanov, Alexandre Constantin, Timothée Bernard, Benoit Crabbé, and Etienne Bernard. 2024. Nuner: Entity recognition encoder pre-training via llm-annotated data.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint.

Kareem Darwish. 2013. Named entity recognition using cross-lingual resources: Arabic as an example. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1558–1567, Sofia, Bulgaria. Association for Computational Linguistics.

François Delon, Gabriel Bédubourg, Léo Bouscarrat, Jean-Baptiste Meynard, Aude Valois, Benjamin Queyriaux, Carlos Ramisch, and Marc Tanti. 2025. Infectious risk events and their novelty in event-based surveillance: new definitions and annotated corpus. *Language Resources and Evaluation*, 59(1):277–295.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Taoufiq El Moussaoui and Chakir Loqman. 2024. Advancements in arabic named entity recognition: A comprehensive review. *IEEE Access*.

Clark C Freifeld, Kenneth D Mandl, Ben Y Reis, and John S Brownstein. 2008. Healthmap : Global infectious disease monitoring through automated classification and visualization of internet media reports. *Journal of the American Medical Informatics Association*, 15:150–157.

Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. 2021. Arabic natural language processing: An overview. *Journal of King Saud University - Computer and Information Sciences*, pages 497–507.

Fatima Ezzahra El Houbri, Najlae Idrissi, Mathieu Roche, and Sarah Valentin. 2026. Padi-location-aren: A normalized arabic–english spatial entity dataset for epidemiological surveillance. *Data in Brief*.

Aditya Joshi, Sarvnaz Karimi, Ross Sparks, Cecile Paris, and C Raina MacIntyre. 2019. Survey of text-based epidemic intelligence: A computational linguistic perspective. *ACM Computing Surveys*.

Sanjeev Kumar, Preethi Jyothi, and Pushpak Bhattacharyya. 2026. Evaluating extremely low-resource machine translation: A comparative study of chrf++ and bleu metrics.

László János Laki and Zijian Győző Yang. 2024. Improving machine translation capabilities by fine-tuning large language models and prompt engineering with domain-specific data. *2024 IEEE 3rd Conference on Information Technology and Data Science (CITDS)*, 2014:1-5.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Edmond Menya, Roberto Interdonato, Dickson Owuor, and Mathieu Roche. 2024. Explainable epidemiological thematic features for event-based disease surveillance. *Expert Systems with Applications*, 250:123894.

Hamdy Mubarak and Sabit Hassan. 2021. UL2C: Mapping user locations to countries on Arabic Twitter. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 145–153, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Xiaoye Qu, Yingjie Gu, Qingrong Xia, Zechang Li, Zhefeng Wang, and Baoxing Huai. 2023. A survey on arabic named entity recognition: Past, recent advances, and future trends. *IEEE Transactions on Knowledge and Data Engineering*, 36:943–959.

Sarah Valentin, Elena Arsevska, Julien Rabatel, Sylvain Falala, Alizé Mercier, Renaud Lancelot, and Mathieu Roche. 2021. Padi-web 3.0 : A new framework for extracting and disseminating fine-grained information from the news for animal disease surveillance. *One Health*, 13.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. Gliner: Generalist model for named entity recognition using bidirectional transformer. *Proc. of the 2024 Conf. of the North American Chapter of the Ass. for Comp. Ling.: Human Language Technologies, NAACL 2024*, pages 5364–5376.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2024. Universalner: Targeted distillation from large language models for open named entity recognition.

Building an Aligned Speech Corpus for Northern Russian

Kseniia Protonina
Higher School of Economics
kiprotonina@edu.hse.ru

Daniil Ignatev
Utrecht University
d.ignatev@uu.nl

Abstract

Russian dialects remain underrepresented in speech recognition resources despite phonological and lexical divergence from the standard language. The paper reports the results of a pilot study on constructing a corpus for dialect-aware speech processing. We first make an argument in favor of dialect-aware ASR – particularly, for Russian. We then describe a reproducible procedure for converting long archival recordings, coarse timestamp indices and manual transcripts into utterance-level aligned data, and introduce a pilot corpus of the Vologda dialect of Russian that is highly scored by expert evaluators. Testing zero-shot ASR models against the corpus reveals that current models remain sensitive to dialectal speech, motivating further expansion of our data collection.

1 Introduction and Background

In recent years, there have been growing efforts to extend speech recognition technologies to underrepresented languages and dialects (Faisal et al., 2024; Joshi et al., 2025), particularly in the corpus domain (Bentum et al., 2024; Oberkircher et al., 2026; Thillainathan et al., 2026). This development broadens the ASR coverage of natural speech and voice-controlled applications, while highlighting performance disparities between language varieties, even when the standard language is well supported (Faisal et al., 2024; Dipto et al., 2025).

Russian is well-represented in the existing speech corpora such as the Whisper dataset (Radford et al., 2022), Russian LibriSpeech (Pratap et al., 2020), Mozilla Common Voice (Ardila et al., 2020), and Golos (Karpov et al., 2021). However, dialectal Russian remains limitedly represented: regional variation is rarely distinguished in large ASR corpora, while dialect-focused resources are often designed for linguistic documentation rather than speech recognition and appear scant in size or accessibility. The dialectal part of the Russian National Corpus (Kachinskaya and Sitchinava, 2015; Savchuk et al., 2024) is an instance of this trend, although it includes a large amount of dialect recordings. Similarly, smaller corpora also provide valuable dialectological material (e.g., Daniel et al. (2013); Knyazev (2023)), but they are often only accessible via their search interfaces and thus are not readily available as ASR data collections. This data shortage limits the robust handling of Russian dialects by speech-to-text tools and may, thereby, reinforce their perceived marginal societal position.

In our pilot study, we motivate dialect-aware ASR and introduce a novel aligned speech corpus of the Vologda dialect, a subtype of Northern Russian. Specifically, our research questions are: (1) how accurately does the pipeline recover word-level temporal boundaries, evaluated against manual annotation? (2) how well do current ASR models perform on the resulting dialectal data in a zero-shot setting?

We describe the dialect and source materials, our audio-transcription alignment method and evaluation, and the resulting corpus. Finally, we benchmark ASR models on the corpus in a zero-shot setting. [1]

2 Source Data

2.1 Dialectal Properties

Vologda dialect is considered a subtype of Northern Russian, a group of dialects spoken primarily in the northern regions of European Russia, including the Arkhangelsk, Vologda, and Kostroma regions (Daniel et al., 2019). Northern Russian dialects are characterized by a set of features that distinguish them from the Moscow-based standard. The most

[1] Our data is available at https://huggingface.co/datasets/hse-prs-folklore/pez. The code is available at https://github.com/ruthenian8/alignment.

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 257–263
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_026

prominent phonological feature is *okanye*: the absence of vowel reduction in unstressed syllables, where unstressed /o/ is preserved as a distinct vowel rather than merging with /a/. Additional features include *tsokanye* (the merger of the affricates *ch* and *ts* sounds in some subdialects) and the retention of archaic morphological and lexical forms.

Our study does not cover Vologda Russian in its entirety, as we focused on a set of recordings collected in Pezhma, a village in the South of Arkhangelsk oblast that falls within the Vologda dialectal area. Still, although a limited number of younger speakers exhibit none or few dialectal features, most speakers in the set do represent Vologda Russian.

2.2 Provenance and Formatting

The Folklore Archive of the Russian State University for the Humanities (Moroz, 2015) incorporates the field data collected in expeditions to the Arkhangelsk oblast, taking place from 1993 on a yearly basis. These expeditions secure novel linguistic materials and focus thematically on folklore studies and dialectology. The collection procedure involves conducting in-depth recorded interviews with local participants, followed by partial transcription of the recordings.

Interviews are manually transcribed during expeditions by linguistically trained collectors who receive prior ear training in the local dialect. Transcriptions cover only thematically relevant fragments, leaving parts of the audio untranscribed. Manually curated index files for each interview record the transcribed and skipped segments with timestamps, along with metadata such as recording time, location, and speaker and collector names.

The archive currently comprises interviews from 12 localities in the Arkhangelsk region totaling to 808 audio recordings and approximately 1046 hours of recorded speech. For our pilot study, we focus on the Pezhma subcorpus, including 48 recordings (around 57 hours). Discarding 3 files with corrupt transcripts leaves us with a final set of 45 interviews. Each recording is accompanied by a manual transcription and a coarse timestamp index serving as the entry point for our pipeline.

The manual transcriptions pose a challenge for straightforward processing due to their specific transcription format. First, they follow a non-standard orthography reflecting dialectal pronunciation, including stress marks on vowels. This orthographic divergence can affect audio-text alignment, as the dialectal spellings differ from the forms represented in standard Russian acoustic models, and can consequently impact word-level ASR metrics. At the same time, they retain standard syntax as opposed to pause marks, standard in phonetic transcription. Secondly, utterances from two or more speakers, often overlapping, are transcribed in a linearized manner as opposed to multilevel transcription. Finally, collector's notes (e.g., [laughter] or [unclear]) are interleaved with the transcription and only marked by brackets. Cf. the excerpt below (backward slash denotes stress by convention):

[S002:] Ну бы\ли э\ти зна\хари... [нрзб.]
nu byli èti znaxari
"Well, there were these healers... [unclear]"

3 Alignment

3.1 Protocol

In our approach, we draw on the increasingly popular method of leveraging ASR outputs as a proxy for their respective audio sources, shifting the alignment task to text-to-text matching of gold captions and these intermediary targets. Some examples of this approach are the procedure adopted by Bentum et al. (2024) in the Spoken Dutch Corpus and the system for Russian speech presented in Borodin et al. (2025). These works leverage intermediary ASR transcriptions from Wav2vec 2.0 (Baevski et al., 2020) and a family of models including GigaAM (Kutsakov et al., 2025) respectively.

Inspired by the said studies, we designed a five-stage pipeline. The first step is audio preprocessing, which involves converting all audio to 16kHz WAV as well as additional loudness normalization and denoising, as described in Appendix A. It is followed by the joining of timestamped indices and fragment-level manual transcriptions, initially loosely matched, thereby ensuring a strict one-to-one mapping. These are joined by their positions; in case of mismatches, word overlap is used as a heuristic. Third, we transcribe each timestamped fragment with WhisperX (Bain et al., 2023), resulting in utterance-level segments. Our choice of WhisperX is dictated by its out-of-the-box capabilities in segmentation and diarization. Thus, even though the direct ASR performance of WhisperX appears inferior to newer, Russian-focused models (see Section 5), this does not impact our pipeline strongly.

The fourth step is transcript alignment and corpus export. The alignment stage finds the best match between each WhisperX-produced segment and a fragment of the manual transcript via dynamic programming, using the token-set F1 score ($\frac{2|A \cap B|}{|A|+|B|}$) as a similarity measure. The use of intermediary WhisperX output is restricted to the temporal scaffold. Manual transcripts remain the source of truth for the exported text. Speaker labels are inferred from the manual transcript when possible, including speaker initials at the end of transcript blocks.

At the fifth and final, export stage, we filter interviewer turns from the output, so as to keep them separate from dialectal speech. Our reason is that these utterances are intentionally left unattributed in some of the corpus files, unlike the speaker responses. Nevertheless, we retain this data in the collection for future use.

3.2 Evaluation

Our evaluation protocol focused on the utterance level. In **question 1**, we adopted the schema from Bentum et al. (2024) that covers 5 categories: "good" for full match, "bad" for full mismatch, as well as "start match", "end match", and "middle match". Meanwhile, we also noted that this taxonomy does not reflect the quality of partial matches. To estimate the latter, we introduced **question 2**: a clip and a caption are "well-matched" if 75% of the caption words match the audio and not otherwise. For example, for a 6-word clip, a mismatch of 2 words would trigger the answer "no", while a shorter mismatch still triggers "yes". Despite its coarseness, this metric appeared intuitive for the annotators, resulting in high inter-worker consistency.

We recruited 5 paid workers among trained students with linguistic background to score a sample of 1000 clips, stratified across the 45 source interviews, based on the above schema. Each annotator received from one to three batches of 100 items. We used a self-hosted Potato (Jurgens et al., 2026) instance for data collection.

500 of the clips were doubly annotated to estimate the inter-worker agreement; we chose Krippendorf's alpha as our IAA metric. The obtained scores equal 0.872 for question 1 and 0.934 for question 2, pointing at acceptable annotation quality. Likewise, we scored all workers using the MACE algorithm (Hovy et al., 2013), resulting in scores above 0.7, i.e., moderate to high reliability. Doubly-annotated labels were aggregated based on the MACE scores.

Label	Ratio
Q1: Good	57.90%
Q1: Start match	16.30%
Q1: End match	16.20%
Q1: Middle match	5.30%
Q1: Bad	4.30%
Q2: Well-matched	78.03%
Q2: No	21.97%

Table 1: Evaluation label distribution.

The results are reported in Table 1. We observe that completely mismatched files constitute a relatively small fraction of 4.3%, indicating that our solution is generally viable. At the same time, the proportion of partial matches is high, similarly to results by Bentum et al. (2024). In turn, Q2 shows that the ratio of files with none- or minor mismatches is high overall.

4 Corpus Design

4.1 Metadata Schema

The metadata is stored in a single "manifest" table, where each row corresponds to one audio segment and its associated transcription. Speaker information is included separately for every utterance.

The folder structure is multilevel. At the global level, files are organized per interview. Within each interview folder, utterances are stored in subfolders per thematic fragment, reflecting the original index structure of the archive. Each segment is identified by a filename including the interview ID, the fragment number, the pseudonymized speaker tag, and the timestamp. Each audio file is paired with a plain-text transcription file. The corpus manifest includes, for each segment: the audio path, the transcript path, the speaker label and quality flags indicating whether the segment was manually reviewed or flagged as possibly misaligned.

4.2 Corpus Statistics

Table 2 provides an overview of the pilot subcollection of the Pezhma corpus. To compute the overall duration and utterance number, we automatically diarized multiple speakers per interview using WhisperX and filtered out the interviewer turns. Considering the speaker groups, male and female speakers are comparably represented. Meanwhile,

the age groups are highly imbalanced, as more than half of the speakers fall into the 40+ category. However, this also entails a better representation of the dialectal features, as they are more prominent in the 40+ stratum.

Parameter	Value
Selected interviews	45
Total duration	$\approx$ 57 h.
Aligned utt.	9 641
Mean utt. length (words)	8.5
Median utt. length (words)	7
Max utt. length (words)	53
Total speakers	63
Female / male speakers	37 / 26
Age: 0-20 / 20-40 / 40+	13 / 2 / 48

Table 2: Final corpus overview.

4.3 Consent and Anonymization

All participants provided informed consent before recording, and their identities were concealed for privacy reasons; they are pseudonymized using hash-based speaker tags in both the transcripts and the corpus metadata.

5 Model Benchmarking

5.1 Models

In order to verify how well existing speech recognition solutions handle the dialect under scrutiny, we evaluate a number of pretrained ASR models, both multilingual and Russian-specific, on our data. This evaluation is performed zero-shot, without further fine-tuning or dialect-specific adaptation. The list of models is as follows. **Whisper Large-v3** (Radford et al., 2022), a state-of-the-art solution for multilingual ASR. Second, **MMS**, Meta's massively multilingual model. Finally, **GigaAM** (Kutsakov et al., 2025), a Russian-specific Conformer-based model yielding SOTA results on Russian; GigaAM offers **RNNT**- and **CTC**-decoding versions, both of which are covered in our evaluation.

5.2 Evaluation

As the evaluation is fully zero-shot, we report the metrics obtained on the entirety of corpus data. We use WER as our primary evaluation criterion and also report CER, which helps account for whether the phonetic spelling makes a difference. Punctuation marks and capitalization are ignored in the evaluation. Diacritics used in transcripts to indicate phone duration during evaluation are stripped away and ignored in the WER and CER computation.

5.3 Results

Table 3 reports zero-shot ASR performance on the corpus data.

Model	WER	CER
GigaAM-RNNT	36.02	24.57
GigaAM-CTC	38.00	24.10
Whisper Large-v3	42.33	29.27
MMS	67.94	34.94

Table 3: Zero-shot ASR results on the corpus data. WER and CER are percentages computed on normalized references and predictions.

5.4 Discussion

The strongest result is obtained by GigaAM RNNT, but even this model yields a WER above 36%. The two GigaAM variants show lower CER than Whisper, suggesting that many errors are short character-level deviations rather than complete word failures, but the word-level mismatch rate remains high for all systems. MMS performs substantially worse, as expected from a primarily multilingual model, not tailored to Russian dialectal speech.

At the same time, pretrained models appear to struggle to process dialectal data as opposed to regular, normative speech. This trend is further confirmed by the error analysis below.

Qualitative inspection of low-overlap segments revealed three systematic error types. (1) *Dialectal phonology*: northern Russian features *okanye*—the preservation of the $/o/–/a/$ distinction in unstressed syllables—that has been lost in standard Russian; standard-trained models often fail to recognize these words (Daniel et al., 2019). (2) *Non-standard lexicon*: dialect-specific words absent from standard training data are substituted or omitted. In addition, the systems may struggle with the general domain shift due to the focus of the interviews on folklore and traditional culture. (3) *Spontaneous speech phenomena*: filled pauses, false starts and ellipses create mismatches between manual transcripts and model output. Filtering interviewer turns and manually reviewing boundary segments mitigated these issues.

6 Conclusion

We presented a compact pipeline for turning partially transcribed archival field recordings into an aligned utterance-level corpus for dialect-aware ASR. The pilot on Pezhma recordings shows that coarse archival indices and manual transcripts can be converted into ASR-ready data while preserving the original dialectal orthography. The boundary evaluation indicates that the resulting segmentation is usable but still imperfect, and zero-shot ASR benchmarks reveal a clear performance gap on dialectal speech. Future work will expand the pipeline to the remaining localities in the archive, improve manual validation of boundary cases and use the resulting data for dialect-aware model adaptation.

7 Limitations

The authors admit that the proposed alignment procedure is not error-free, and that the corpus can occasionally contain erroneous alignments. However, thanks to the corpus metadata, files with low aligner confidence, which largely coincide with errors, can be filtered out in downstream applications.

As indicated earlier, our study focused on recordings from one village for pilot purposes; hence, our current release does not represent Northern Russian in full.

We did not test the produced alignment system on different languages and dialects, even though it should be compatible with more WhisperX-compatible languages.

7.1 Disclosure of AI assistance

In the current project, we used AI-based coding assistant to refactor and consolidate an early version of our code-base. The authors manually verified the code and the results.

8 Acknowledgments

We are grateful to Daria Savina and Eleonora Izmailova for their contributions to the project. We thank the anonymous reviewers for helping us improve the paper. Finally, we are deeply thankful to the makers and maintainers of the RSUH Folklore archive as well as the recorded speakers.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. ArXiv:1912.06670 [cs.CL].

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Martijn Bentum, Eric Sanders, Antal PJ van den Bosch, Douwe Zeldenrust, and Henk van den Heuvel. 2024. Corpus creation and automatic alignment of historical dutch dialect speech. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4021–4029.

Kirill Borodin, Nikita Vasiliev, Vasiliy Kudryavtsev, Maxim Maslov, Mikhail Gorodnichev, Oleg Rogov, and Grach Mkrtchian. 2025. A data-centric framework for addressing phonetic and prosodic challenges in russian speech generative models. *arXiv preprint arXiv:2507.13563*.

Michael Daniel, Nina Dobrushina, and Ruprecht von Waldenfels. 2013. The language of the ustja river basin. a corpus of north russian dialectal speech.

Michael Daniel, Ruprecht von Waldenfels, Aleksandra Ter-Avanesova, Polina Kazakova, Ilya Schurov, Ekaterina Gerasimenko, Daria Ignatenko, Ekaterina Makhlina, Maria Tsfasman, Samira Verhees, and et al. 2019. Dialect loss in the russian north: Modeling change across variables. *Language Variation and Change*, 31(3):353–376.

Tawsif Tashwar Dipto, Azmol Hossain, Rubayet Sabbir Faruque, Md Rezuwan Hassan, Kanij Fatema, Tanmoy Shome, Ruwad Naswan, Md Foriduzzaman Zihad, Mohaymen Ul Anam, Nazia Tasnim, et al. 2025. Are asr foundation models generalized enough to capture features of regional dialects for low-resource languages? In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 178–188.

European Broadcasting Union. 2011. Recommendation r 128: Loudness normalization and permitted maximum level of audio signals. EBU Technical Recommendation. First issued February 2010; revised 2011, 2014, 2020 and 2023.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECTBENCH: An NLP Benchmark for Dialects, Varieties, and Closely-Related Languages. In *Proceedings of the*

62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14412–14454, Bangkok, Thailand. Association for Computational Linguistics.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2025. Natural language processing for dialects of a language: A survey. *ACM Computing Surveys*, 57(6):1–37.

David Jurgens, Michael Chen, and Lina Iyer. 2026. Potato 2.0: A comprehensive annotation platform with support for ai-in-the-loop and agentic systems. In *Proceedings of the The 64th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Irina B. Kachinskaya and Dmitri V. Sitchinava. 2015. Today's state of the dialectal subcorpus. In *Proceedings of the V.V. Vinogradov Russian Language Institute. № 6.*, pages 142–163.

Nikolay Karpov, Alexander Denisenko, and Fedor Minkin. 2021. Golos: Russian Dataset for Speech Research. ArXiv:2106.10161 [eess.AS].

S. V. Knyazev. 2023. Corpus of the russian dialect spoken in the village of veegora. [Online resource]. Available at: https://lingconlab.ru/veegora (accessed: 2026-05-19).

Aleksandr Kutsakov, Alexandr Maximenko, Georgii Gospodinov, Pavel Bogomolov, and Fyodor Minkin. 2025. GigaAM: Efficient Self-Supervised Learner for Speech Recognition.

Andrei Moroz. 2015. The folklore archive of the russian state university for the humanities in moscow. *Studia mythologica Slavica*, 7:175–179.

Lena S Oberkircher, Jesujoba O Alabi, Dietrich Klakow, and Jürgen Trouvain. 2026. Saar-voice: A multi-speaker saarbr\" ucken dialect speech corpus. *arXiv preprint arXiv:2604.11803*.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A Large-Scale Multilingual Dataset for Speech Research. In *Interspeech 2020*, pages 2757–2761.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Resemble AI. 2023. Introducing resemble enhance: Ai powered speech denoising and enhancement. Blog post. Accessed 20 May 2026.

S. O. Savchuk, T. Arkhangelskiy, A. A. Bonch-Osmolovskaya, O. V. Donina, Yu. N. Kuznetsova, O. N. Lyashevskaya, B. V. Orekhov, and M. V. Podryadchikova. 2024. Russian national corpus 2.0: New opportunities and development prospects. *Voprosy Jazykoznanija*, 2:7–34.

Hendrik Schroter, Alberto N Escalante-B, Tobias Rosenkranz, and Andreas Maier. 2022. Deepfilternet: A low complexity speech enhancement framework for full-band audio based on deep filtering. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7407–7411. IEEE.

Adsajan Thillainathan, Nishanthini Kanthakumar, Nivethiga Rasan, and Kengatharaiyer Sarveswaran. 2026. Talk-corpus: A regionally diverse evaluation set for sri lankan tamil speech. In *Proceedings of Speech Language Models in Low-Resource Settings: Performance, Evaluation, and Bias Analysis (SPEAKABLE) @ LREC 2026*, pages 194–200, Palma de Mallorca, Spain. ELRA and ICCL.

A Normalization

All source audio was converted to 16kHz mono WAV prior to alignment for compatibility with ASR. Denoising was applied selectively based on signal-to-noise ratio (SNR) estimates. Table 4 summarizes our three-level denoising strategy. For severely noisy segments (SNR $<$ 9 dB) we employed the Resemble Enhance denoiser–enhancer pipeline (Resemble AI, 2023). For moderate noise (9 dB $\leq$ SNR $<$ 11 dB) we used DeepFilterNet, a low-complexity deep filtering framework for full-band speech enhancement (Schroter et al., 2022). Segments with SNR $\geq$ 11 dB were left unaltered. A high-pass filter at 80Hz and a low-pass filter at 7,500 Hz removed low-frequency rumble and high-frequency interference. We then performed loudness normalization using the `ffmpeg` double-pass `loudnorm` filter implementing the EBU R128 standard for loudness normalization and permitted maximum level (European Broadcasting Union, 2011), targeting an integrated loudness of -16 LUFS. Table 5 summarizes the effect of preprocessing on audio quality across the corpus.

SNR (dB)	Method
<9	Resemble AI (2023)
9–11	Schroter et al. (2022)
>11	-

Table 4: Denoising strategy by SNR level

After normalization the loudness standard deviation dropped from 3.97 to 2.39 LUFS, reflect-

Metric	Before	After
Mean LUFS	-12.4	-17.4
Std LUFS	3.97	2.39
Mean SNR (dB)	14.1	15.1
SNR range (dB)	6.0–27.9	6.7–32.2
Files with SNR < 10dB	7.6% (731)	< 1% (91)

Table 5: Audio preprocessing statistics (N = 9,617 segments)

ing greater consistency across segments. The final mean LUFS of -17.4 remains slightly below the -16 LUFS target, likely due to wind noise and fluctuating speech activity in field recordings, which affect the gating mechanism of the EBU R128 algorithm.

Probing Geographic Performance Gaps in Moroccan ASR with ALAMA: Annotated Local Audio of Moroccan Arabic

Avery C. Kanel[1,2] **Christian Schuler**[1,2]
Bouazza Laracha[3,*] Imrane Lbouhli[4,*] Yassine Chaouri[5,*]
Yusser Al Ghussin[1,2] **Timo Baumann**[6]

[1]Saarland University, [2]German Research Center for Artificial Intelligence (DFKI),
[3]Ibn Zohr University, [4]Mohammed V University, [5]Hassan II University of Casablanca
[6]OTH Regensburg

Abstract

Moroccan Arabic (Darija) is routinely treated as a single dialect in NLP research, yet it exhibits substantial regional variation across its geography. This pilot study questions the monolithic framing by collecting a regionally annotated speech corpus, 60 minutes of broadcast audio from Casablanca, Tangier, and Oujda, transcribed and verified by native speakers, and evaluating 17 state-of-the-art ASR systems against it. Our results show architecture-dependent performance gaps across regions: systems exhibit significant gaps in WER points between the best- and worst-served regions, and fine-tuned "Moroccan Arabic" checkpoints exhibit inconsistent regional profiles that are hard to reconcile with balanced training data. These findings suggest that national-level dialect labels obscure regionally uneven model coverage with real performance consequences. We release the dataset[1] and support the adoption of more fine-grained provenance labeling as standard metadata for Arabic dialect resources.

1 Introduction

Moroccan Arabic (Darija) is spoken by nearly all of Morocco's 37 million inhabitants as a first or second language (Heath, 2020). Many dialectical datasets are labeled simply as 'Moroccan Arabic', foregoing further geographic specification, glossing over substantial regional variation, phonological, lexical, and contact-driven, that characterizes the language (Heath, 2013; Aguadé, 2008). Darija also exists within a state of diglossia with Standard Arabic and features heavy code-switching into French and Tamazight languages as well as occasional use of English and Spanish, with patterns

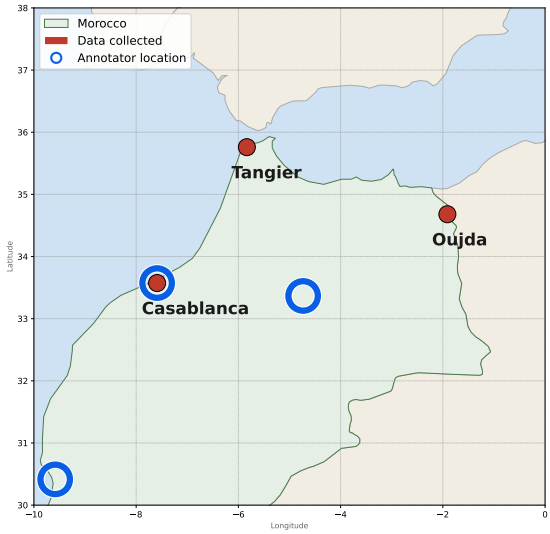


Figure 1: Data sources and human annotator locations.

that differ systematically by region and social context (Skiredj et al., 2024; Hermas, 2025).

Recent years have seen substantial progress in Darija ASR, with reported improvements of fine-tuned systems (Bounhar and Majjodi, 2025; Atlasia, 2026; Boumehdi and Yousfi, 2024). It is, however, an open question whether these advances equally benefit speakers across all of Morocco's regions. ASR models inherit the distributional features of their training data (Koenecke et al., 2020); if Moroccan Arabic corpora are, in fact, regionally concentrated as they are in other North African countries (Mohamed Eida and Habash, 2025), without disclosing this at the national-label level, speakers from less-represented areas face systematic performance penalties with no way to diagnose this from published benchmarks or model cards (Keleg and Magdy, 2023; Zaidani et al., 2024).

In this pilot study we present ALAMA, a regionally stratified Darija dataset spanning Casablanca, Tangier, and Oujda (Figure 1), with transcriptions verified by native speakers and metadata covering accent, register, and loan-word use, among others.

[1]https://github.com/Ackanel/alama_audio_dataset
*: Equal contribution (Human annotators).

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 264–271
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_027

Table 1: Salient samples to highlight phonological and morphological differences across the Northern, Central and Eastern regions of Morocco, illustrated through closely paired locations Tangier and Larache, Casablanca and Temara, and Oujda and Berkane, respectively (Boumans, 2005; Heath, 2013; Falchetta and Guerrero, 2023).

Linguistic Feature	**"Northern"**	**"Central"**	**"Eastern"**
Historical Dialect Group	Pre-Hilalian	Hilalian	Hilalian
Genitive Exponent	*d-* / *dyal*	*dyal*	*dyal* / *ntaʕ*
Affrication of /t/	Alveo-dental affricate [t͡s]	Alveo-palatal affricate [t͡ʃ]	Dental plosive [t̪]
Reflex of etymological *q	[q]	[g] and [q]	[g] and [q]

The remainder of this work presents our findings by evaluating state-of-the-art ASR-systems against ALAMA.

2 Related Work

Arabic NLP has long focused on distinguishing broad dialect groups from Modern Standard Arabic (MSA), while fine-grained intra-dialectal variation remains underexplored. The ADI17 dataset (Shon et al., 2020) advanced dialect identification to 17 country-level Arabic varieties, and work on micro-dialect identification (Abdul-Mageed et al., 2020) further narrowed the scope. The ALDi framework (Keleg et al., 2023, 2024) shows that utterances within a single dialect label vary along a dialectness continuum correlated with inter-annotator disagreement, a finding directly relevant to our annotation quality. Despite these advances, regional variation within country, including Moroccan Arabic, remains largely unaddressed.

Broadcast speech has anchored Arabic ASR research since MGB-2 (Ali et al., 2019a), extended by the QASR corpus (Mubarak et al., 2021) and MGB-5 (Ali et al., 2019b), which provided an early Moroccan regional reference point. Talafha et al. (2023) studied zero-, few-, and full-shot Whisper adaptation across Arabic dialects as a methodological predecessor to our evaluation. The most direct prior work is the Casablanca dataset (Talafha et al., 2024), which grounded the NADI 2025 shared task (Talafha et al., 2025) and its top system (Salhab et al., 2025); contemporaneous benchmarks Sullivan et al. (2026); Abdullah et al. (2025) similarly evaluate multi-system Arabic ASR but at national granularity. Our work complements these by making regional provenance within Moroccan Arabic a primary experimental variable. Hutiri and Ding (2022) operationalize fairness through worst-group and max-minus-min gap metrics, a framework that we apply at the regional level. Bootstrap methods for Word-Error Rate (WER) confidence intervals (Bisani and Ney, 2004; Liu et al., 2021) are discussed further in our evaluation strategy.

Moroccan Arabic exhibits substantial inter-regional variation as summarized in Table 1, shaped by urbanization, pre- and post-Hilalian mixing, and Berber contact influence (Heath, 2020; Ramos et al., 2024; Francisco, 2021; D'Anna, 2024), documented in linguistic studies of Casablanca (Aguadé, 2003; Hachimi, 2018, 2012), Tangier (Montes, 2015; Grigore and Bițună, 2016), Oujda (Fernández, 2017), and cross-regionally (Boumans, 2005). The ARCADE dataset (Nacar et al., 2026), which contributes to our corpus, illustrates how regional asymmetries arise even in carefully constructed resources, containing forty audio samples from Casablanca, fifteen from Tangier, and none from the Oriental region where Oujda is located. This motivates the balanced sampling pursued here.

3 Data

ALAMA consists of 60 minutes of regional broadcast speech drawn evenly from Casablanca, Tangier, and Oujda with approximately 20 minutes each. Audio was drawn from two sources: first, our manually gathered recordings from local radio channels, specifically call-in programs, in-studio interviews, and occasional sports broadcasts and radio dramas. Second, the ARCADE dataset (Nacar et al., 2026), which provided 9 minutes for Casablanca and 3 minutes for Tangier. To limit speaker-level bias, no single speaker contributes more than 3.4% of the to-

Table 2: Per-region word counts of human-curated transcriptions. **Radio region**: location of the broadcast; **Gold strict**: 3/3 agree; **Gold majority**: 2/3 agree.

Region	**Radio region**	**Gold strict**	**Gold majority**
Casablanca	2,402	1,912	4,318
Tangier	2,498	605	719
Oujda	2,483	0	1,208
Total	7,383	2,517	6,245

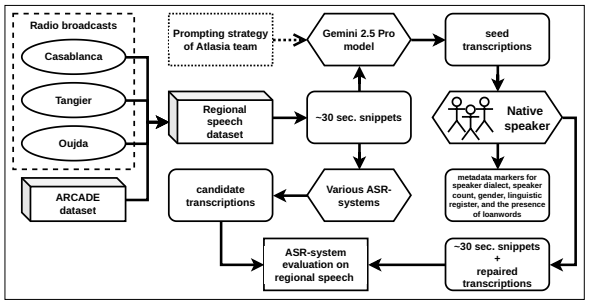


Figure 2: ASR experiment pipeline.

tal dataset. All segments were trimmed to between 27 and 30 seconds for Whisper-compatibility and no audio enhancement or denoising was applied, preserving raw broadcast conditions. Table 2 shows how human-annotated detected speaker accents differ from recording source labels.

Transcriptions were first produced by Gemini 2.5 Pro (Google, 2025) following the same prompting strategy used by Atlasia (2026), an open-source research group based in Morocco that developed one of the models examined in this paper, MoulSot. During this process we observed that identical interviews were occasionally broadcast across multiple regional channels, meaning the broadcast city did not always reflect the speaker's geographic origin. The geographic metadata was designed specifically to track these discrepancies. All transcriptions were subsequently verified and corrected by three native speakers of Moroccan Arabic (Table 3), producing the gold-label references used throughout our evaluation.

4 Methodology

Figure 2 provides an overview of the experimental pipeline and Table 5 in Appendix A shows information about the included ASR systems. We evaluate each of the 17 ASR systems (see Table 4 for details) against all three native-speaker reference transcriptions and report both Word-Error Rate (WER) and Character-Error Rate (CER) in two variants. **Raw** scores strip only chat-template artifacts present in generative model outputs; **Normalized** scores additionally apply the Open Universal Arabic ASR Leaderboard normalization (punctuation removal, diacritic stripping, hamza and madda normalization, Eastern-to-Western numeral conversion). We also report **Oracle WER**, the per-utterance minimum WER across the references, which provides a lower bound exposing the contribution of inter-annotator disagreement to observed error rates.

Scoring was performed using NIST SCTK `sclite` for word-level error decomposition (substitutions, deletions, insertions) and a `jiwer`-based port of the Leaderboard normalization pipeline, both run on all reference-hypothesis pairs. Per-region WER is computed separately for Casablanca, Tangier, and Oujda. To operationalize regional fairness we follow Hutiri and Ding (2022) and report the max-minus-min WER gap across regions as a disparity proxy.

We flag the `Gemini 2.5` system for gold-label circularity throughout: reference transcriptions were seeded by Gemini 2.5 Pro before native-speaker correction, and this system also relies on the Gemini model family for inference.

Table 3: Native-speaker annotator demographics.

Region	Sex	Age	Education
Souss-Massa	male	25	bachelor's student
Beni Méllal-Khenifra	male	25	master's student
Casablanca-Settat	male	20	bachelor's degree

5 Results and Analysis

Table 4 reports overall WER and CER and per-region breakdown for all 17 systems. Normalized WER spans a wide range: MoulSot v0.3 achieves the best non-circularity-flagged result at 35.5 %, while other systems exceed 90 %.

The per-region picture is the more suggestive one. 12 of 17 systems perform worst on the Oujda dialect which is a significant deviation from chance (χ^2 goodness-of-fit, df=2, p=.023), with gaps ranging from 0.87 to 17.8 WER points relative to the best-served region. Crucially, the direction of this disadvantage is not uniform: several systems fine-tuned on data labeled only as "Moroccan Arabic" perform *better* on Oujda than on Casablanca or Tangier (e.g., WSmall-D and WLarge-M fine-tunes), an inconsistency hard to reconcile with the assumption of homogeneous regional coverage in their training data.

The gold-circularity diagnostic makes this argument sharper. Gemini 2.5 Pro, which seeded our reference transcriptions, scores near-zero WER on Casablanca and Tangier (2.9 % and 4.4 %) but 23.1 % on Oujda (Table 4). Error decomposition (last column group in Table 4) shows this Oujda gap is driven almost entirely by insertions (7.0 %) rather than substitutions or deletions, indicating that the native-speaker post-editing effort required to correct the Gemini-seeded transcriptions differed

Table 4: ASR evaluation results on the 60-min ALAMA dataset (all values in %). **Normalized (LB)** shows leaderboard-normalized WER/CER and Oracle WER (per-utterance minimum across three references); **Per-Region WER** shows normalized WER per city, with **Gap** as the max-minus-min regional disparity; **sclite Decomposition** shows word-level Substitution/Deletion/Insertion breakdown. Green/yellow/red: column best/second-best/worst. Orange-tinted row is gold-circularity flagged because the reference transcripts were seeded by this system before native-speaker correction, making its WER an artifact rather than a comparable baseline.

	Normalized (LB)			Per-Region WER				sclite Decomposition		
Model	**WER**	**CER**	**Oracle**	**Casa.**	**Tang.**	**Oujda**	**Gap**	**Sub**	**Del**	**Ins**
DEDICATED ASR *(single-purpose end-to-end ASR)*										
WLarge-DL	45.5	24.9	41.0	43.2	41.2	51.6	10.5	32.9	12.8	3.5
WLarge	51.8	26.9	47.6	52.1	47.1	55.9	8.8	36.8	16.0	3.1
WLarge-M	56.9	21.8	53.2	57.3	57.2	55.6	1.7	42.9	8.1	9.3
WLargeCER-M	57.4	21.9	53.7	58.0	57.1	56.6	1.4	43.6	8.1	8.9
WMedium-D	53.4	26.9	49.9	45.5	52.6	61.6	16.1	30.2	8.9	18.3
WMedium	62.5	33.4	59.0	60.7	60.4	65.9	5.5	42.5	19.6	2.8
WSmall	72.9	36.3	70.0	71.0	72.3	74.9	3.9	52.6	18.7	3.4
WSmall-D	81.3	59.2	77.9	81.4	87.3	74.8	12.5	40.0	29.0	13.7
WTurbo	55.4	29.3	51.3	53.1	50.7	62.0	11.3	36.5	19.3	2.9
WTiny	91.4	56.3	90.3	91.6	90.5	92.1	1.6	58.2	31.6	2.0
W2V2-MD	55.6	21.4	51.7	58.4	56.5	51.5	6.9	45.2	10.1	3.2
W2V2-M	69.1	25.9	67.3	68.9	69.5	68.6	0.9	52.8	8.1	10.4
MoulSot	35.5	16.0	31.4	33.4	35.2	37.4	4.0	27.2	10.3	2.4
Qwen3ASR	54.2	23.1	49.6	51.6	51.4	59.0	7.6	43.1	8.9	4.9
MULTIMODAL LLM / MULTITASK SPEECH *(transcription as one of many tasks)*										
Gemini	9.7	5.7	0.2	2.9	4.4	23.1	20.2	2.7	0.1	7.0
Gemma4	42.1	20.1	38.1	38.8	38.6	48.4	9.8	32.2	11.2	2.2
Omni-7	50.1	30.1	45.8	47.1	47.3	55.3	8.2	28.2	23.6	1.2
Seamless	84.8	58.7	81.5	73.6	91.4	89.2	17.8	42.9	36.0	8.9

systematically by region. Taken together, these results suggest that regional origin is a measurable, architecture-dependent factor in ASR performance, and that training-data provenance shapes regional generalization in ways not visible from model labels alone.

6 Discussion

Our results provide evidence for ASR performance differences between regions of Morocco. A potential cause of which could be treating Moroccan Arabic as a single homogenous target in dataset construction, obscuring regional variation. The inconsistent, sometimes reversed regional skew across "Moroccan Arabic"-labeled checkpoints is hard to explain under the assumption of regionally balanced training data; a more parsimonious explanation is that these corpora are regionally concentrated, introducing undisclosed biases invisible at the national level.

This concern extends beyond the specific systems we evaluate. Labeling speech data as "Moroccan Arabic" while training predominantly on one region creates models whose blind spots are invisible to downstream users, and we argue that region-level provenance should become standard metadata for datasets and model cards not only for Arabic dialects.

Several caveats apply. Our three-region, broadcast-based setup cannot fully disentangle dialectal variation from acoustic recording conditions; isolating these effects would require larger per-region samples and more controlled environments. Because we lack provenance information for the fine-tuning corpora we evaluate, our account of their regional skew remains an inference rather than a verified claim. The gold-circularity in our annotation pipeline means WER figures are measured against a non-fully-independent standard, though the Gemini 2.5 Pro diagnostic corroborates region-dependent transcription difficulty. Additionally our gold-strict labels count for Oujda is n = 0. This could be due to annotator's unfamiliarity with Oujda shibboleths and subsuming the Oujda accent into the Casablanca category as none of the annotators come from Oujda. It is also worth considering that the general pattern of ASR models performing worse on Oujda data is a trickle-down effect of Gemini's worse performance as it was the basis for the annotator's transcriptions. This pilot motivates, rather than conclusively demonstrates, the need for regionally-stratified Darija ASR research.

7 Conclusion

ALAMA provides evidance that regional origin is a potential, architecture-dependent factor in Darija ASR, one that is invisible from national-level labels alone. The inconsistent profiles of "Moroccan Arabic"-labeled fine-tunes further suggest their training corpora are regionally skewed – a hypothesis that this pilot motivates but cannot conclusively verify at scale. We suggest the adoption of region-level provenance as metadata for Arabic dialect datasets and model cards when feasible, and to scale regionally-stratified Darija evaluation beyond this pilot.

For the future, we hope for dialect awareness *within* models rather than only in their description.

Acknowledgments

Christian Schuler gratefully acknowledges financial support from GraduSaar, University of Saarland.

References

Muhammad Abdul-Mageed, Chiyu Zhang, Abdel-Rahim Elmadany, and Lyle Ungar. 2020. Toward Micro-Dialect Identification in Diaglossic and Code-Switched Environments.

Badr M. Abdullah, Yusser Al Ghussin, Zena Al-Khalili, Ömer Tarik Özyilmaz, Matias Valdenegro-Toro, Simon Ostermann, and Dietrich Klakow. 2025. Saarland-Groningen at NADI 2025 Shared Task: Effective Dialectal Arabic Speech Processing under Data Constraints. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 745–751, Suzhou, China. Association for Computational Linguistics.

Jordi Aguadé. 2003. Notes on the Arabic dialect of Casablanca (Morocco). In *Proceedings of the Fifth Conference of AIDA (Association Internationale de Dialectologie Arabe). Ed. by I. Ferrando & JJ. Sánchez Sandoval, Cádiz, Servicio de Publicaciones de la Universidad de Cádiz*, pages 301–308.

Jordi Aguadé. 2008. Morocco [dialectological survey]. In Kees Versteegh, Mushira Eid, Alaa Elgibali, Manfred Woidich, and Andrzej Zaborski, editors, *Encyclopedia of Arabic Language and Linguistics*, volume 3, pages 287–297. Brill, Leiden–Boston.

Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. 2019a. The MGB-2 Challenge: Arabic Multi-Dialect Broadcast Media Recognition.

Ahmed Ali, Suwon Shon, Younes Samih, Hamdy Mubarak, Ahmed Abdelali, James Glass, Steve Renals, and Khalid Choukri. 2019b. The MGB-5 Challenge: Recognition and Dialect Identification of Dialectal Arabic Speech. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1026–1033.

Atlasia. 2026. MoulSot: A curated moroccan darija speech dataset and fine-tuned ASR model.

M. Bisani and H. Ney. 2004. Bootstrap estimates for confidence intervals in ASR performance evaluation. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–409.

Louis Boumans. 2005. Lexical variation in Moroccan Arabic: The genitive exponent in experimental data from four cities. *EDNA, Estudios de dialectología norteafricana y andalusí*, 9:125–137.

Ahmed Boumehdi and Abdellah Yousfi. 2024. MDVC corpus: Empowering Moroccan Darija speech recognition. *Indonesian Journal of Electrical Engineering and Computer Science*, 34(1):290–301.

Abdelaziz Bounhar and Abdeljalil El Majjodi. 2025. Atlaset dataset for moroccan darija: From data collection, analysis, to model trainings. *Hugging Face Blog*.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023. SeamlessM4T: Massively Multilingual & Multimodal Machine Translation.

Luca D'Anna. 2024. A dialect type of Eastern first-layer Maghrebi Arabic. *Kervan. International Journal of African and Asian Studies*, 28(2).

Jacopo Falchetta and Jairo Guerrero. 2023. Koineization trends in spoken arabic across three moroccan towns. *Kervan – International Journal of Afro-Asiatic Studies*, 27(1).

Montserrat Benítez Fernández. 2017. Variation linguistique dans le parler arabe d'Oujda: Nord-est marocain vs. communauté marocaine dans la diaspora. *Dialectologia*, pages 1–18.

Felipe Benjamin Francisco. 2021. The Southern Moroccan Dialects and the Hilāli Category. *Languages*, 6(4):192.

Google. 2025. Gemini 2.5 Pro.

George Grigore and Gabriel Bițună, editors. 2016. *Arabic Varieties: Far and Wide: Proceedings of the 11th International Conference of AIDA, Bucharest, 2015*. editura universității din București, București.

Atiqa Hachimi. 2012. The urban and the urbane: Identities, language ideologies, and arabic dialects in morocco. *Language in Society*, 41(3):321–341.

Atiqa Hachimi. 2018. Arabic Dialect Contact and Change in Casablanca: The Role of Simplification and Salience in the Adoption of a Morphosyntactic Variable. *Anthropological Linguistics*, 60(1):60–93.

Jeffrey Heath. 2013. *Jewish and Muslim Dialects of Moroccan Arabic*. Routledge, London.

Jeffrey Heath. 2020. Moroccan arabic. *Arabic and contact-induced change*, 1:213.

Abdelkader Hermas. 2025. Managing Morocco's linguistic market: A linguistic citizenship proposal. *European Journal of Language Policy*, 17(2):237–254.

Wiebke Toussaint Hutiri and Aaron Ding. 2022. Bias in Automated Speaker Recognition. In *2022 ACM Conference on Fairness Accountability and Transparency*, pages 230–247.

Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. ALDi: Quantifying the Arabic Level of Dialectness of Text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10597–10611, Singapore. Association for Computational Linguistics.

Amr Keleg and Walid Magdy. 2023. Arabic Dialect Identification under Scrutiny: Limitations of Single-label Classification. In *Proceedings of ArabicNLP 2023*, pages 385–398, Singapore (Hybrid). Association for Computational Linguistics.

Amr Keleg, Walid Magdy, and Sharon Goldwater. 2024. Estimating the Level of Dialectness Predicts Interannotator Agreement in Multi-dialect Arabic Datasets.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Chunxi Liu, Michael Picheny, Leda Sarı, Pooja Chitkara, Alex Xiao, Xiaohui Zhang, Mark Chou, Andres Alvarado, Caner Hazirbas, and Yatharth Saraf. 2021. Towards measuring fairness in speech recognition: Casual conversations dataset transcriptions.

Mai Mohamed Eida and Nizar Habash. 2025. Beyond Cairo: Sa'idi Egyptian Arabic Literary Corpus Construction and Analysis. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 292–304, Albuquerque, USA. Association for Computational Linguistics.

Valerie Susana Montes. 2015. *Urban Mediterranean Dialects of Arabic : Tangier and Tunis*. Master of Arts, The University of Texas at Austin.

Hamdy Mubarak, Amir Hussein, Shammur Absar Chowdhury, and Ahmed Ali. 2021. QASR: QCRI Aljazeera Speech Resource A Large Scale Annotated Arabic Speech Corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2274–2285, Online. Association for Computational Linguistics.

Omer Nacar, Serry Sibaee, Adel Ammar, Yasser Alhabashi, Nadia Samer Sibai, Yara Farouk Ahmed, Ahmed Saud Alqusaiyer, Sulieman Mahmoud AlMahmoud, Abdulrhman Mamdoh Mukhaniq, Lubaba Raed, Sulaiman Mohammed Alatwah, Waad Nasser Alqahtani, Yousif Abdulmajeed Alnasser, Mohamed Aziz Khadraoui, and Wadii Boulila. 2026. ARCADE: A City-Scale Corpus for Fine-Grained Arabic Dialect Tagging.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision.

Carmen Berlinches Ramos, Jairo Guerrero, and Montserrat Benítez Fernández. 2024. *AIDA Granada: A Pomegranate of Arabic Varieties*, volume 21. Prensas de la Universidad de Zaragoza.

Mahmoud Salhab, Shameed Sait, Mohammad Abusheikh, and Hasan Abusheikh. 2025. Munsit at NADI 2025 Shared Task 2: Pushing the Boundaries of Multidialectal Arabic ASR with Weakly Supervised Pretraining and Continual Supervised Fine-tuning.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. 2026. Qwen3-ASR Technical Report.

Suwon Shon, Ahmed Ali, Younes Samih, Hamdy Mubarak, and James Glass. 2020. ADI17: A Fine-Grained Arabic Dialect Identification Dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8244–8248.

Abderrahman Skiredj, Ferdaous Azhari, Ismail Berrada, and Saad Ezzini. 2024. DarijaBanking: A New Resource for Overcoming Language Barriers in Banking Intent Detection for Moroccan Arabic Speakers. ArXiv:2405.16482 [cs.CL].

Peter Sullivan, AbdelRahim Elmadany, Alcides Alcoba Inciarte, and Muhammad Abdul-Mageed. 2026. Arab Voices: Mapping Standard and Dialectal Arabic Speech Technology.

Bashar Talafha, Karima Kadaoui, Samar Mohamed Magdy, Mariem Habiboullah, Chafei Mohamed Chafei, Ahmed Oumar El-Shangiti, Hiba Zayed, Mohamedou Cheikh Tourad, Rahaf Alhamouri, Rwaa Assi, Aisha Alraeesi, Hour Mohamed, Fakhraddin Alwajih, Abdelrahman Mohamed, Abdellah El Mekki, El Moatez Billah Nagoudi, Benelhadj Djelloul Mama Saadia, Hamzah A. Alsayadi, Walid Al-Dhabyani, Sara Shatnawi, Yasir Ech-chammakhy, Amal Makouar, Yousra Berrachedi, Mustafa Jarrar, Shady Shehata, Ismail Berrada, and Muhammad Abdul-Mageed. 2024. Casablanca: Data and Models for Multidialectal Arabic Speech Recognition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21745–21758, Miami, Florida, USA. Association for Computational Linguistics.

Bashar Talafha, Hawau Olamide Toyin, Peter Sullivan, AbdelRahim Elmadany, Abdurrahman Juma, Amirbek Djanibekov, Chiyu Zhang, Hamad Alshehhi, Hanan Aldarmaki, Mustafa Jarrar, Nizar Habash, and Muhammad Abdul-Mageed. 2025. NADI 2025: The First Multidialectal Arabic Speech Processing Shared Task.

Bashar Talafha, Abdul Waheed, and Muhammad Abdul-Mageed. 2023. N-Shot Benchmarking of Whisper on Diverse Arabic Speech Recognition.

Omnilingual ASR team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, Kehan Lyu, Sagar Miglani, Vineel Pratap, Kaushik Ram Sadagopan, Safiyyah Saleem, Arina Turkatenko, Albert Ventayol-Boada, Zheng-Xin Yong, Yu-An Chung, Jean Maillard, Rashel Moritz, Alexandre Mourachko, Mary Williamson, and Shireen Yates. 2025. Omnilingual ASR: Open-source multilingual speech recognition for 1600+ languages.

Hajar Zaidani, Abderrahim Maizate, Mohammed Ouzzif, and Rim Koulali. 2024. Building a Corpus for the Underexplored Moroccan Dialect (CFMD) Through Audio Segmentations. *Revue d'Intelligence Artificielle*, 38(3):857–866.

A ASR Systems

Table 5 provides additional information about all ASR-systems currently included in the experiments of this pilot study. Our aim is to cover a broad spectrum of model sizes for base- as well as fine-tuned models, to allow comparisons across modalities.

Label	Name	Type	Size	Reference
DEDICATED ASR *(single-purpose end-to-end ASR)*				
MoulSot	MoulSot v0.3	*Fine-tune*	–	(Atlasia, 2026)
Omni7B	Omni ASR 7B	*Base model*	7 B	(team et al., 2025)
Qwen3ASR	Qwen3-ASR	*Base model*	1.7 B	(Shi et al., 2026)
W2V2-M	W2V2 Moroccan	*Fine-tune*	300 M	`hf:/othrif/wav2vec2-large-xlsr-moroccan`
W2V2-D	W2V2 Darija	*Fine-tune*	300 M	(Boumehdi and Yousfi, 2024)
WLarge	Whisper Large	*Base model*	1.55 B	(Radford et al., 2022)
WLarge-DL	Whisper Darija LoRA	*Fine-tune*	1.55 B	`hf:/Hafsa0/whisper-large-v3-darija-lora`
WLarge-M	Whisper Morocco FT	*Fine-tune*	1.55 B	(Abdullah et al., 2025)
WMedium	Whisper Medium	*Base model*	769 M	(Radford et al., 2022)
WMedium-D	Whisper Medium Darija	*Fine-tune*	769 M	`hf:/ychafiqui/whisper-medium-darija`
WSmall	Whisper Small	*Base model*	244 M	(Radford et al., 2022)
WSmall-D	Whisper Small Darija	*Fine-tune*	244 M	`hf:/KandirResearch/Whisper-Small-Darija`
WTurbo	Whisper Turbo	*Base model*	809 M	`hf:/openai/whisper-large-v3-turbo`
WTiny	Whisper Tiny	*Base model*	39 M	(Radford et al., 2022)
MULTIMODAL LLM / MULTITASK SPEECH *(transcription as one of many tasks)*				
Gemini	Gemini 2.5	*Proprietary*	–	(Google, 2025)
Gemma4	Gemma4 E4B Instruct	*Base model*	4 B (effective)	`hf:/google/gemma-4-E4B-it`
Qwen3Omni	Qwen3-Omni	*Base model*	30 B (3 B active)	`hf:/Qwen/Qwen3-Omni-30B-A3B-Instruct`
Seamless	Seamless-M4T	*Base model*	2.3 B	(Communication et al., 2023)

Table 5: ASR systems evaluated in this study, grouped by purpose. *Type*: whether the system is a general-purpose base model or a Darija-specific fine-tune. *Size*: approximate parameter count (– = undisclosed or proprietary). *Reference*: citation or path of the model on `huggingface.co`.

Generator-Guided Amount Recovery for Voice-Based Financial Record-Keeping in Mooré–French Code-Switched Speech

Maimouna Ouattara
University of Luxembourg

El-Hacen Diallo
University of Luxembourg

Fred Philippy
University of Luxembourg

Abdoul Kader Kaboré
University of Luxembourg

Jacques Klein
University of Luxembourg

Tegawendé F. Bissyandé
University of Luxembourg

{maimouna.ouattara, el-hacen.diallo, fred.philippy
abdoulkader.kabore, jacques.klein, tegawende.bissyande}@uni.lu

Abstract

Voice interfaces promise to bring digital financial services to the roughly 800 million adults worldwide who cannot read, yet informal markets demand more than broad speech-recognition coverage. In Burkina Faso, traders record credit in Mooré and in French–Mooré code-switched speech, where the value of a spoken amount depends on the active counting convention, a historical base-five *wakir* system in Mooré against base ten in French, so that a correctly transcribed numeral can still encode the wrong sum, and standard cascades and frontier audio Large Language Models (LLMs) fail even when the convention is supplied in context. We present a voice-to-ledger pipeline that recovers the customer, the transaction direction, and the monetary amount from naturalistic market speech. Its core is Generator-Guided Analysis-by-Synthesis (GG-Abs), a deterministic amount-recovery module that treats each spoken amount as a latent integer. GG-Abs proposes candidates with a bidirectional Mooré numeral generator, renders each phonetically, keeps only those that re-synthesise to the noisy transcript, and abstains otherwise, so every accepted amount carries an explicit reconstruction; a conservative LLM judge supplies the customer and direction. We also release a voice-ledger corpus of Mooré–French market recordings spanning four credit intents. On a sealed real-speech holdout, the pipeline attains **76.0%** amount accuracy on Mooré and **85.0%** on code-switched speech, with direction accuracy above 90% in both regimes, improving Mooré amount recovery over Automatic Speech Recognition (ASR)→LLM and audio-native LLM baselines by 16 to 47 points.[1]

1 Introduction

Recent advances in multilingual AI now reach hundreds of under-represented languages, spanning translation models such as No Language Left Behind (NLLB) (Costa-Jussà et al., 2022) and speech recognition systems such as Massively Multilingual Speech (MMS) (Pratap et al., 2024) and Omnilingual ASR (Keren et al., 2025). Yet broader coverage alone does not guarantee *inclusive* access for the millions of people who cannot read. UNESCO estimates that 739 million adults still lack basic literacy skills, including 225 million in sub-Saharan Africa (UNESCO, 2025). This challenge is particularly acute in countries like Burkina Faso, where the adult literacy rate stood at about 42% in 2023 (World Bank, 2023). Voice-first services promise to lift this barrier, replacing written input with speech and opening digital access in the languages people actually speak (Medhi Thies, 2015; da Silva et al., 2024).

However, these barriers are, in practice, contextual and cultural. Everyday communication is shaped by social and cultural conventions that demand particular attention. One of the most common is *code-switching*, the mixing of two languages within a single utterance. In Burkinabè markets, for instance, traders fluidly alternate between French and Mooré (Myers-Scotton, 1997; Amuzu and Singler, 2014). A historical artefact of the CFA franc [2] makes this code-switching especially treacherous. An amount spoken in French denotes its exact base-ten value in CFA. The same amount spoken in Mooré, however, follows a base-five system anchored on the *wakir*, a five-franc unit inherited from colonial currency: it is counted in units of five francs, not one. A trader who says "twenty" in Mooré therefore means one hundred CFA. This fivefold gap is invisible to anyone pars-

[1] Code is available at https://github.com/mouna80/latell2026-moore-french-cs.

[2] CFA refers to a currency used in several West and Central African countries. In this paper, it specifically denotes the West African CFA franc (ISO 4217: XOF), distinct from the Central African CFA franc (ISO 4217: XAF). See https://en.wikipedia.org/wiki/CFA_franc.

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 272–283
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_028

ing the words alone. In voice-first financial applications, such a discrepancy becomes a critical source of error. The system must accurately identify the language of each amount to determine its true value.

```
[TASK] Audio-to-Text Transcription &
    Financial Normalization

[INPUT_AUDIO]  "<Audio clip: a trader
    saying 'Moussa [first name] owes me
    2,000 [XOF]' in moore>"

[EXPECTED_VALUE] .....2,000 XOF
[SYSTEM_OUTPUT]  .......400 XOF
```

Listing 1: Example of an audio-native LLM transcription failure[3]. The system maps the acoustics literally, masking the fivefold currency gap.

Our contributions:

- **GG-Abs: Generator-Guided Analysis-by-Synthesis.** To recover Mooré amounts, this framework treats each value as a latent integer, proposes candidates using a bidirectional Mooré numeral generator, and scores them via phonetic re-synthesis against unfused transcripts. Requiring no LLM call, this deterministic method recovers 72.5% of Mooré amounts, improving recall by 10.0 points over parse-only (cf., 5). When used as a hint for `gemini-3.1-flash-lite` (approximately one-sixth the estimated judge-only API cost of `gemini-3.5-flash`), it improves Mooré amount recovery from 53.1% to 70.0%.
- **A Mooré–French voice-to-ledger pipeline for informal credit.** Built on GG-Abs, the pipeline recovers the customer, transaction direction, and amount from monolingual Mooré as well as French–Mooré code-switched speech, without requiring explicit numeral-language diarization within this evaluated language pair.
- **A Mooré–French voice-ledger corpus and rigorous evaluation.** We construct and release a task-specific corpus covering four credit intents, built in three phases, from seed recordings through template-based synthesis to a human-validated gold set, and evaluate on a sealed real-speech holdout. GG-Abs is the only system with strong performance across both regimes (76.0%/85.0% for Mooré/code-switched amounts, and about 90% for direction in each), improving Mooré amount recovery over the baselines by 16 to 47 points.

[3] https://gemini.google.com/share/2ba4289a0b13

2 Related Work

Code-switched speech and cascade pipelines. Production ASR still handles code-switching poorly. Multilingual training on English-Bantu pairs in South Africa beats monolingual baselines (Yılmaz et al., 2018; Biswas et al., 2019, 2020), and self-supervised representations help under-resourced African languages (Ògúnrèmí et al., 2023), but most results come from controlled, single-pair settings. Surveys of end-to-end code-switching ASR (Agro et al., 2025; Mustafa et al., 2022) show that Mandarin-English and Hindi-English dominate the literature, with little coverage of African pairs. Turning such speech into structured output typically relies on cascades. ASR followed by machine translation (ASR→Machine Translation (MT)) is the de-facto approach for low-resource speech translation, including African languages and French targets (Kponou et al., 2025), while ASR followed by a language model (ASR→LLM) feeds transcripts to an LLM for correction or understanding (Chen et al., 2023). Beyond standard error propagation, these cascades miss cultural context shifts like the base-five Mooré counting system (Naveen and Trojovský, 2024; Al Sharoufi and Al-Fadhli, 2025), which leaves the numeral correctly transcribed but wrongly valued.

Language identification and diarization. To handle code-switched amounts, a natural approach is to identify the language of each spoken numeral and apply the corresponding counting convention. In audio, this requires spoken language diarization to segment utterances by language. While recent diarization systems show progress (Chua et al., 2023; Dhanush et al., 2025; Baroudi et al., 2026), they primarily target high-resource pairs and require abundant training data. Text-side language identification is no easier at scale (Burchell et al., 2024), and French-African pairs remain barely studied, so this route is not yet viable in our context.

Task-oriented dialogue. While widely studied for high-resource languages and financial domains (Tur and De Mori, 2011; Altinok, 2018), dialogue systems for African languages are only just emerging through datasets like WAXAL (Diack et al.,

2026) and AfriWOZ (Adewumi et al., 2023). Although cross-lingual transfer helps bridge these gaps (Thangaraj et al., 2024), in-domain financial corpora remain scarce, and no existing system targets a structured financial ledger from code-switched market speech. The output is a structured ledger comprising the amount, direction, and customer. Recent African benchmarks such as INJONGO (Yu et al., 2025) include banking intents across multiple languages, but target text-based intent classification rather than speech-based, value-normalized ledger extraction.

Accessibility and financial inclusion. Financial inclusion depends on low-cost services reaching unbanked populations (Kunjal, 2026). Yet text-centric banking in dominant languages imposes severe literacy barriers on rural communities (Chatterjee and Das, 2021), limiting benefits even where mobile access exists (Ozili, 2025). FinTech providers respond with simplified interfaces (Koefer et al., 2024), and mobile connectivity expands inclusion in West Africa (Senou et al., 2019), but infrastructure alone is insufficient. This literature motivates voice as an accessible interface capable of relaxing literacy constraints (Kunjal, 2026), yet it stops at design and leaves open the technical challenge of recovering correct monetary values from code-switched speech. GG-Abs addresses this gap.

3 Methodology

Figure 1 illustrates the proposed pipeline. It operates in two phases. An offline phase derives correction rules by analyzing ASR drift (Sections 3.1 and 3.2), and an online phase applies these rules to reconstruct amounts from noisy transcripts (Section 3.3).

3.1 Offline: Synthetic Corpus

No existing corpus covers Mooré–French code-switched speech for informal credit ledgers. We built a task-specific corpus in two steps.

Step 1: Intent definition. We fix the task scope with 20 development intents spanning four types, six debt increases, six repayments, five combined multi-transaction cases, and three balance queries, each paired with a structured XOF gold ledger (`debt_up`, `debt_down`, `queries`). Speakers produced free Mooré and Mooré–French versions with French numerals, establishing the schema and linguistic regimes rather than a benchmark.

Step 2: Synthetic augmentation. Bilingual contributors mapped French ledger frames to Mooré templates with entity, product, and amount slots. A generator then fills these slots with regional names, market roles, products, and CFA amounts across the four intent types. Gold JSON ledgers are derived from the source slots and deterministic amount decoding, never from ASR output, yielding 4,570 utterances rendered with the fine-tuned XTTS-v2 Mooré voice (Ouedraogo et al., 2026), which we use as a compositional stress test.

3.2 Offline: ASR drift analysis

Mooré counts money in *wakir* (units of five CFA) rather than single francs. For example, *kobs-naase* means 100×4 *wakir* (2,000 CFA), and *tus pis-nu* means $1{,}000 \times 50$ *wakir* (250,000 CFA). To invert this convention, we implement a generator-parser pair $G(n) \leftrightarrow \mathrm{parse}(t)$, verified to round-trip ($\mathrm{parse}(G(n)) = n$) up to 10,000,000 CFA. The generator proposes canonical numeral candidates, which are matched against the transcript to recover the amount. This works perfectly on canonical Mooré numerals but fails on raw ASR output. Noise introduces a systematic drift between canonical forms and surface transcripts, corrupting numerals into unparseable fragments or near-homophones.

This drift is patterned rather than random. It exhibits recurring distortions such as compound-boundary loss (*pis-nu* split), final nasal drift (*pisnii* for *pis-nii*), vowel neutralization, and digit fallback (*tus 14*). Because the corruptions follow predictable patterns, a separate normalization phase reverses them with deterministic rules before matching, mapping the noisy surface form back to a canonical one. The generator itself remains untouched, proposing candidates only, while the rules clean the incoming transcript.

Rule mining occurs strictly offline. The deterministic GG-Abs module makes no LLM call at inference. We use three foundation models (GPT-5, Gemini Pro, Claude Opus 4.8) solely to analyze the drift and propose collision-safe rewrites (i.e., rewrites that never map two distinct amounts to the same form). They are supplied with the canonical numeral lexicon, clean utterances, and paired noisy transcripts. Model consensus ($\geq 2/3$) serves merely as a pre-filter. The definitive gate is a one-rule ablation on the development split, where a rule is kept only if it raises recall or reduces false hints

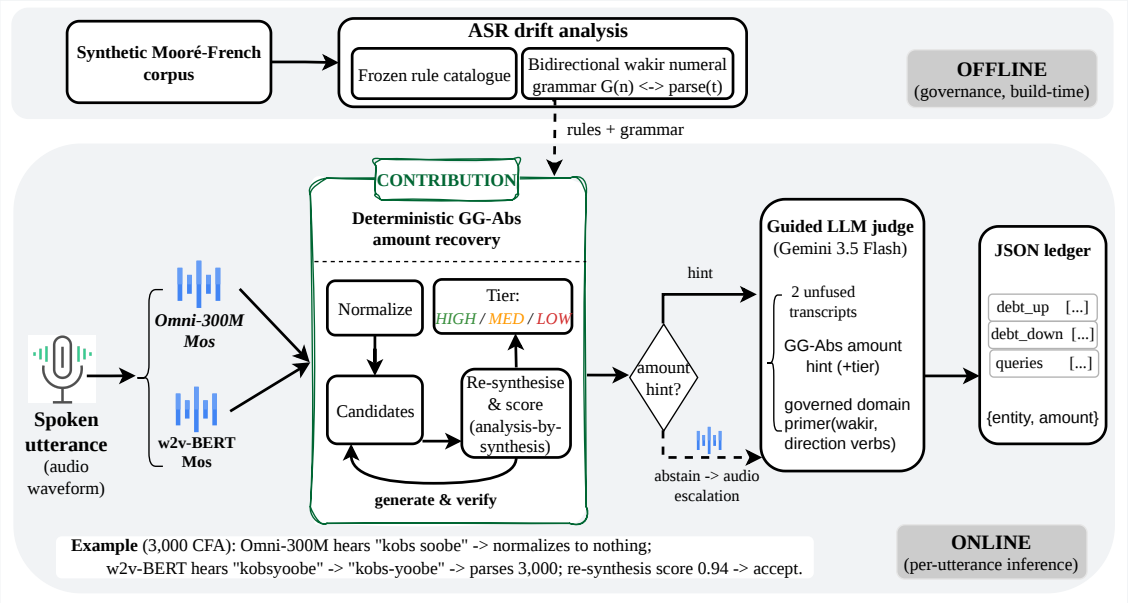


Figure 1: The GG-Abs voice-to-ledger pipeline. *Offline*, we study ASR drift on a synthetic Mooré–French corpus and freeze a rule catalogue and a bidirectional *wakir* numeral grammar $G(n) \leftrightarrow \mathrm{parse}(t)$. *Online*, two unfused ASR channels feed the deterministic GG-Abs module, which normalises each transcript, enumerates candidates, scores them by re-synthesis, and assigns a confidence tier. A candidate amount yields a tagged hint, and the absence of any round-tripping candidate yields abstention. The guided judge (Gemini 3.5 Flash) reads the two transcripts, the hint with its tier, and a governed domain primer, and emits the ledger. On abstention the raw audio is escalated to the judge. The deterministic module (our contribution) recovers the amount without an LLM.

without adding misses, and retired otherwise. All mining and ablation rely exclusively on the development split, never the sealed evaluation holdout. In parallel, an expert track adds rules grounded in descriptive Mooré literature (e.g., discourse-marker strips for *samde* and *wakir*), so every deployed rule carries a precise provenance tag (Table 1).

Rule (prov.)	Drift	→	Canonical
Expert			
samde strip (E)	*samde* / *samda*	→	∅
wakir strip (E)	*wakir*	→	∅ (keep *tus*)
Heuristic ($k/3$ *consensus*)			
final-*nii* (3/3)	*pisnii*	→	*pis-nii*
pis-X (3/3)	*pisnaase*	→	*pis-naase*
thousand (3/3)	*piir*	→	*piig*
yoobe (3/3)	*yogbe*	→	*yoobe*
yopoe (3/3)	*yokoe*	→	*yopoe*
final-*u* (2/3)	*nung*	→	*nu*
teen (3/3)	*tus 14*	→	*tus piig la a naase*

Table 1: Frozen rules that rewrite recurring ASR drift of Mooré numerals back to the canonical form the generator parses. E marks expert, linguist-validated rules and $k/3$ marks rules proposed by k of three LLMs.

System robustness is further reinforced by architectural diversity. We analyze drift through two distinct Mooré systems, Omnilingual ASR and a fine-tuned w2v-BERT ASR (Wav2Vec-BERT 2.0, a multilingual self-supervised model adapted to Mooré through supervised fine-tuning on a dedicated speech corpus (Ouedraogo et al., 2026)). The two systems exhibit complementary failure modes. For 3,000 CFA (*kobs-yoobe*), w2v-BERT elides the boundary as `kobsyoobe`, whereas Omnilingual splits it and drifts the vowel as `kobs soobe`. Both transcripts are normalized by the rules and matched against the generator's candidates without fusing the channels. Because their error distributions differ, a correct amount routinely survives even when one channel corrupts it.

3.3 Online Inference

neural ASR front ends. Their transcripts then enter the deterministic post-ASR amount-recovery path described in Algorithm 1, before the downstream LLM judge.

At inference, each utterance is processed by two frozen Mooré-targeted ASR front ends, W2V-mos and Omni-mos. Their unfused transcripts enter the deterministic GG-Abs amount-recovery path described in Algorithm 1 and are subsequently passed

Algorithm 1 GG-Abs inference for one utterance

Require: audio x of one utterance
Ensure: ledger ℓ, or abstain
$T \leftarrow \text{ASR}(x)$ ▷ two unfused channels
$T \leftarrow \text{NORMALISE}(T)$ ▷ frozen Mooré rules
$C \leftarrow \text{CANDIDATES}(T)$ ▷ runs and windowed spans
if C has no complete run **then**
  return abstain; attach audio x
end if
for all $c \in C$ **do**
  $s(c) \leftarrow \text{MATCH}(G(c), T)$ ▷ re-synthesis score
end for
$\hat{a} \leftarrow \arg\max_{c \in C} s(c)$
$\tau \leftarrow \text{TIER}(\hat{a}, T)$ ▷ hint confidence
$\ell \leftarrow \text{JUDGE}(T, \hat{a}, \tau)$
return ℓ ▷ customer, direction, amount

to the downstream judge together with the resulting hint. When GG-Abs abstains, raw audio is additionally attached. The system decodes the audio, normalizes the transcripts, enumerates candidate amounts, verifies them via phonetic re-synthesis, reconciles readings across channels, and finally passes a tagged hint to the judge. When GG-Abs emits a Mooré amount hint, that hint is constructed and round-tripped by the deterministic layer. When GG-Abs abstains, the downstream judge may instead infer an amount from the ASR transcripts and raw audio.

The normalizer applies the frozen rules from Section 3.2 in four steps, running character normalization, then skeleton-based fuzzy matching (τ=0.80), compound repair behind magnitude prefixes (τ=0.65), and sequence cleanup. Fuzzy matching relies on a normalized Levenshtein similarity over pronunciation-stable skeletons, identical to the re-synthesis scoring metric. The pipeline operates with zero learned parameters, and both match thresholds are deterministic constants fixed on the development set. Normalized transcripts are grouped into maximal contiguous runs. A run is valid only if it pairs a magnitude with a quantifier or explicit digits, and a bare unit triggers abstention. To bound the otherwise $O(n^2)$ search, candidates are parseable windows ($\leq$12 tokens) within each channel. Each candidate a is scored by re-synthesis,

$$s(a) = \max_{c} \text{SIM}\big(\text{skel}(G(a)), \text{skel}(t_c)\big),$$

where $G(a)$ is the generated spelling, $\text{skel}(\cdot)$ extracts a pronunciation-stable skeleton by folding diacritics and vowel length, and SIM is the normalized Levenshtein similarity (1 minus edit distance over maximum length). Phonetically identical spellings therefore score equally, so the candidate that best matches the acoustic transcript wins. On the 3,000 CFA example, the correct candidate (*kobs-yoobe*) scores 0.94 against the w2v-BERT skeleton, far above distractors such as 300 (0.63), 600 (0.47), and 60,000 (0.21). The same score resolves merge-versus-split ambiguity, since merged and split renderings compete directly and ties default to the split to preserve genuine compound transactions.

Per-channel readings are then reconciled slot by slot. An amount split across channels is merged only when the merge raises the overall re-synthesis score, which blocks spurious combinations, while single-channel recoveries are corroborated by the windowed enumeration. A stage-attribution audit shows that $\sim$90% of multi-transaction failures are upstream, where the second amount never forms a complete run, so reconciliation is not the bottleneck and these repairs close the rest. Each surviving hint is tiered HIGH when two channels agree and MED for a single channel. Fragile hints are downgraded rather than deleted, so the judge is never deprived of a valid Mooré amount.

4 Experiments

We evaluate GG-Abs on spontaneous market speech, focusing on structured ledger recovery. The full task requires extracting transaction direction and customer identity, but amount recovery remains the primary stress test, compounding ASR noise with non-decimal counting conventions. Systems output a constrained JSON ledger scored against gold records. Amount and direction use exact-set match, while customer extraction is scored by identity equivalence rather than exact string matching, using human labels for Mooré.

To prevent leakage, we enforce a strict three-split regime. All deterministic rules, match thresholds, and LLM prompts were developed and frozen exclusively on a *synthetic dev* pool (Section 3). The end-to-end system was then evaluated exactly once on a *sealed real holdout* of 312 utterances (194 Mooré, 118 code-switched) from seven disjoint speakers, never touched in development, tuning, or calibration. The holdout introduces 60 novel intents, covering debt increases, repayments, multi-transaction scenarios, and balance queries, all distinct from the 20 development intents, so no evaluation data influenced any rule, threshold, or prompt. Table 2 details the holdout subsets used per met-

ric. For joint evaluation we report *transaction-core* (correct amount and direction) and *transaction-exact* (correct amount, direction, and customer). To isolate judge variance, all final scores are the mean±std over three runs at temperature 0.1.

Mooré holdout subset	n	Used for
Amount-bearing	160	amount, core, exact
balance query	22	direction (CHECK)
missing-amount	12	no-amount control
No-amount control (22+12)	34	false-insertion rate
All Mooré	194	direction
Code-switch (CS)	118	CS amount and direction

Table 2: Real-speech holdout dataset ($n = 312$).

5 Results

Table 3 reports the sealed real-speech test. The two single-evidence baselines fail in opposite regimes, ASR→LLM on code-switching and the audio-native LLM on Mooré and direction. GG-Abs is the only balanced system, so it is not a one-regime amount recogniser but a voice-to-ledger pipeline that keeps the critical fields high in both regimes that break single-evidence systems. For fairness, the LLM baselines in Table 3 receive the same domain-context block as GG-Abs (the *wakir* convention, ledger semantics, and direction cues), so they are evaluated with task knowledge. Amount accuracy is computed only on the 160 of 194 Mooré utterances that carry an amount, the remaining 34 being balance queries or missing-amount cases (Table 2).

	Mooré (n=160)		**Code-switch** (n=118)	
System	Amount	Dir.	Amount	Dir.
ASR→MT	6.9	—	4.2	—
ASR→LLM	60.2	73.5	27.1	72.9
Audio→LLM	28.5	58.1	**90.1**	57.6
GG-Abs (full)	**76.0**	**90.2**	85.0	**92.7**

Table 3: Amount and direction exact-set accuracy (%), averaged over three runs. The audio-native baseline is higher on code-switched amount recovery (90.1 versus 85.0), whereas GG-Abs is higher on the other three reported fields.

Cross-regime performance. The 85.0 versus 90.1 code-switch amount gap is about six utterances on n=118, roughly three rows where a spurious Mooré hint blocks the rescuing audio escalation (fixable) and roughly three of genuine residual (Section B.1). It is a wrinkle in one field, not a hole in the main claim, since GG-Abs wins code-switch direction decisively (92.7 vs. 57.6) and is the better full-ledger system on code-switching. The fixable rows point to a concrete next step, which we leave for future work to keep the sealed holdout scored once. Escalating to audio even when the deterministic module emits a low-confidence (MED) Mooré hint alongside French numeral evidence would stop a spurious hint from suppressing the rescue, and all three blocking rows here carry a MED hint, so this rule would target exactly them.

Mooré amount recovery before the judge. The defining property of GG-Abs is that the critical balance-changing field is recovered *before* the LLM judge. On amount-bearing Mooré utterances it emits a hint in 93.8% of cases (77.3% precision, 72.5% recall). The deterministic post-ASR module requires no additional LLM call. Emitted amounts are structurally valid and traceable. Neither deterministic variant emits an amount on the 34 no-amount controls, while carrying an explicit *wakir* reconstruction and a re-synthesis score that traces the value back to the recognized transcript rather than taking it on trust. Recovery stays consistent across transaction types and does not degrade on compound utterances (Table 4), reaching 69.3% on debt increases, 71.1% on repayments, and 80.0% on combined multi-transaction scenarios, which require every spoken amount to match. The judge therefore receives evidence graded by confidence tier (87.5% accuracy for HIGH two-channel hints, 71.3% for MED single-channel hints) rather than blind guesses.

Intent type	Mooré amount (%)	n
Debt increase (credit)	69.3	75
Repayment	71.1	45
Combined (multi-transaction)	80.0	40
All	72.5	160

Table 4: Deterministic GG-Abs Mooré amount recovery by intent type (exact-set match, n=160 amount-bearing utterances, no LLM). Performance stays consistent across types and is strongest on compound multi-transaction utterances.

GG-Abs secures the balance-changing fields, amount and direction, in both regimes. Direction reaches 90.2% on Mooré and 92.7% on code-switched (CS) speech, and transaction-core accuracy (amount and direction jointly correct) is 73.8%

on Mooré and 82.2% on CS, with a single IN/OUT confusion across the 132 single-direction Mooré utterances. The free-text customer name, left to the LLM judge, is the one open field, scored by identity rather than string match (e.g. *moussa* and *musa*) through a human same-person judgment. Strict transaction-exact accuracy (all three fields correct) is therefore 49.4% on Mooré and 50.8% on CS.

Complementary Mooré ASR channels. On Mooré, the deterministic amount module reads two complementary ASR channels, a fine-tuned w2v-BERT Mooré specialist and Omnilingual decoded Mooré-target (Omni-Mooré). Table 5 gives each channel's effect on the deterministic Mooré amount (exact-set match on the 160 amount-bearing rows, no LLM). The two channels are strongly complementary. The w2v-BERT specialist is the workhorse, and dropping it collapses amount recovery from 72.5 to 38.8 (−33.7), while Omni-Mooré rescues amounts it mishears (−8.1 amount, 10 points of coverage). Keeping the two transcripts unfused lets either channel rescue an amount the other loses, rather than letting one channel's error contaminate the pool. GG-Abs is therefore not a single recogniser followed by a parser but a generator-guided recovery layer over complementary ASR views.

Channel configuration	Mooré amount (%)	Δ
Full	72.5	–
− w2v-BERT Mooré (`w2v`)	38.8	−33.7
− Omni-Mooré (`omni_mos`)	64.4	−8.1

Table 5: Per-channel effect on the deterministic GG-Abs Mooré amount (exact-set match, n=160 amount-bearing, no LLM). The two Mooré ASR channels are complementary.

Cheap-judge viability. In constrained deployments, pairing the amount hint with a lightweight judge lifts Mooré amount accuracy from 53.1% to 70.0%, bridging the gap to larger models (Table 6). With a strong judge the hint instead trades a slight accuracy dip (76.0% vs. 78.8%) for strict auditability, since the model already exploits the raw ASR transcripts effectively. Direction resolution stays judge-dependent across all configurations.

The deterministic hint incurs no API cost and negligible CPU latency, so end-to-end latency is dictated by the ASR front-end and the LLM call. On an RTX 6000 Ada, single-utterance ASR latencies (0.30 s for the Omnilingual pass and 0.08 s for w2v-BERT), and the LLM brings the median end-to-end latency to 1.9 s for the strong judge and 1.5 s for the lightweight one. Switching to the smaller judge therefore speeds up the pipeline and cuts per-call API cost sixfold, from $2.20 to $0.36 per thousand calls. The estimates multiply the observed input and output token counts by their respective published prices[4].

Judge	Hint	Amt (%)	Dir (%)	Latency (s)
Flash	yes	76.0	90.2	1.9
Flash	no	**78.8**	**93.1**	1.9
Flash-Lite	yes	**70.0**	76.2	1.5
Flash-Lite	no	53.1	76.9	1.5

Table 6: Deployment frontier on the Mooré holdout (n=160). Flash is `gemini-3.5-flash`, Flash-Lite is `gemini-3.1-flash-lite`. Latency reflects the end-to-end median including the ASR front-end.

6 Conclusion

We presented GG-Abs, a generator-guided analysis-by-synthesis method for voice-based financial record-keeping in Mooré and French–Mooré market speech. Its central idea is to turn a local monetary convention into an executable recovery mechanism. Rather than relying on ASR coverage or free-form LLM interpretation, GG-Abs treats each spoken amount as a latent integer, reconstructs it through the *wakir* counting system, verifies it against noisy ASR outputs, and abstains when no auditable match exists.

On a sealed real-speech holdout of 312 recordings from seven speakers, the full pipeline reaches 76.0% amount accuracy on Mooré and 85.0% on code-switched speech, with direction accuracy above 90% in both regimes. These results show that inclusive voice finance requires more than multilingual recognition. It requires systems that encode local knowledge, protect balance-changing fields with deterministic checks, and make each deterministically recovered amount traceable. We therefore present GG-Abs not as a deployment-ready product but as an assistive, confirmation-gated step toward low-resource financial AI that is accessible, culturally grounded, auditable, and trustworthy in real use.

[4] https://ai.google.dev/gemini-api/docs/pricing, accessed 1 Jun 2026.

7 Limitations

Evaluation scope. We evaluate an assistive recovery tool, not a deployment-ready product. The sealed holdout contains 312 recordings from seven speakers in one language pair, Mooré and French. This supports the core hypothesis but leaves demographic, acoustic, and cross-lingual generalization untested, and the single-voice synthetic corpus may miss natural market noise and prosody. Larger field studies with more speakers and sessions remain future work.

Language and rule dependence. GG-Abs relies on a hand-built *wakir* grammar and frozen, linguist-validated ASR-drift rules. This supports auditability but limits portability, since a new language or monetary convention requires a new grammar, parser, and rule catalogue. Our claim is a generator-guided design principle, not a universal rule set. We also note that the drift rules change the final output on only 7 of 388 Mooré-side transcripts on the sealed holdout, so the deployed accuracy rests on re-synthesis scoring and skeleton matching rather than the rules, which serve auditability and validity.

Compute and latency. The two parallel ASR channels improve robustness but raise compute, and a production system would need to learn when fewer suffice. Low-latency use may require GPU access, compression, or an on-device recognizer. We have not yet evaluated deployment on low-end devices or poor connectivity.

Dependencies and open problems. Only the amount is recovered deterministically, while the pipeline still relies on the LLM judge for ledger structure, direction, and customer identity, so a cheaper judge loses direction accuracy (Table 6). Free-text name extraction is the primary bottleneck for strict full-ledger accuracy, and the direction verb lexicon is corpus-induced and awaits full linguistic validation. Routing audio to a commercial API on abstention also introduces privacy and cost concerns. Real-world deployment will therefore require stronger identity resolution, local inference, and explicit user confirmation before any ledger entry is committed.

8 Ethical considerations

Speech data was collected voluntarily in two stages, crowdsourced seeds for template design and recordings from native Mooré-speaking colleagues for evaluation. Participants gave informed consent, received no compensation, and may withdraw their data at any time. All transactional content is fictional. The only personal data collected is biometric voice, stored securely under Burkina Faso's Loi n°010-2004 and GDPR principles, with participants specifically consenting to its collection, research use, and possible public release. To protect privacy, the corpus will be published under strict access control, withholding raw audio while sharing derived text and metadata more openly. Since the system's audio-escalation path sends raw waveforms to a commercial API, any real-world deployment must disclose this transmission and obtain separate consent. We used large language models only to polish the writing of this paper, for grammar and phrasing. They were not used to generate research ideas, design or run experiments, or produce results. All scientific content and claims are the authors' own.

Acknowledgments

This work is supported by the Luxembourg Ministry of Foreign and European Affairs through their Digital4Development (D4D) portfolio under the project LuxWAyS (Luxembourg/West-Atrica Lab for Higher Education Capacity Building in Cyber-Security and Emerging Topics in ICT4Dev).
This research was also supported by Google.org through the provision of Google Cloud Platform credits used to conduct our experiments. We are grateful to the Centre d'Excellence Interdisciplinaire en Intelligence Artificielle (CITADEL) for facilitating data collection through its fellows. We also thank all individuals who contributed to the translation and recording tasks.

References

Tosin Adewumi, Mofetoluwa Adeyemi, Aremu Anuoluwapo, Bukola Peters, Happy Buzaaba, Oyerinde Samuel, Amina Mardiyyah Rufai, Benjamin Ajibade, Tajudeen Gwadabe, Mory Moussou Koulibaly Traore, Tunde Oluwaseyi Ajayi, Shamsuddeen Muhammad, Ahmed Baruwa, Paul Owoicho, Tolulope Ogunremi, Phylis Ngigi, Orevaoghene Ahia, Ruqayya Nasir, Foteini Liwicki, and Marcus Liwicki. 2023. Afriwoz: Corpus for exploiting cross-lingual transfer for dialogue generation in low-resource, african languages. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Maha Tufail Agro, Atharva Kulkarni, Karima Kadaoui,

Zeerak Talat, and Hanan Aldarmaki. 2025. Code-switching in end-to-end automatic speech recognition: A systematic literature review.

Hussain Al Sharoufi and Waleed Al-Fadhli. 2025. Bridging the gap: Pragmatic and cultural challenges in machine translation. *International Journal of Society, Culture & Language*, 13(2):29–43.

Duygu Altinok. 2018. An ontology-based dialogue management system for banking and finance dialogue systems. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Evershed K Amuzu and John Victor Singler. 2014. Codeswitching in west africa.

Séverin Baroudi, Yanis Labrak, Shashi Kumar, Joonas Kalda, Sergio Burdisso, Pawel Cyrta, Juan Ignacio Alvarez-Trejos, Petr Motlicek, Hervé Bredin, and Ricard Marxer. 2026. Doctor or patient? synergizing diarization and asr for code-switched hinglish medical conditions extraction.

Astik Biswas, Febe De Wet, Ewald Van der westhuizen, and Thomas Niesler. 2020. Semi-supervised acoustic and language model training for English-isiZulu code-switched speech recognition. In *Proceedings of the 4th Workshop on Computational Approaches to Code Switching*, pages 52–56, Marseille, France. European Language Resources Association.

Astik Biswas, Emre Yılmaz, Febe de Wet, Ewald van der Westhuizen, and Thomas Niesler. 2019. Semi-Supervised Acoustic Model Training for Five-Lingual Code-Switched ASR. In *Interspeech 2019*, pages 3745–3749.

Laurie Burchell, Alexandra Birch, Robert Thompson, and Kenneth Heafield. 2024. Code-switched language identification is harder than you think. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 646–658, St. Julian's, Malta. Association for Computational Linguistics.

Amrita Chatterjee and Simontini Das. 2021. Information communication technology diffusion and financial inclusion: an inter-state analysis for india. *Innovation and Development*, 11(1):1–23.

Chen Chen, Yuchen Hu, Chao-Han Huck Yang, Sabato Marco Siniscalchi, Pin-Yu Chen, and Eng-Siong Chng. 2023. Hyporadise: An open baseline for generative speech recognition with large language models. *Advances in Neural Information Processing Systems*, 36:31665–31688.

Victoria YH Chua, Hexin Liu, Leibny Paola Garcia Perera, Fei Ting Woon, Jinyi Wong, Xiangyu Zhang, Sanjeev Khudanpur, Andy WH Khong, Justin Dauwels, and Suzy J Styles. 2023. Merlion ccs challenge: A english-mandarin code-switching child-directed speech corpus for language identification and diarization. *arXiv preprint arXiv:2305.18881*.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

G. Mohan Dhanush, J. Gopichand, and B. Balasaigayatri. 2025. Language diarization model for bilingual code-switched speech analysis. In *2025 7th International Conference on Intelligent Sustainable Systems (ICISS)*, pages 656–661.

Abdoulaye Diack, Perry Nelson, Kwaku Agbesi, Angela Nakalembe, MohamedElfatih MohamedKhair, Vusumuzi Dube, Tavonga Siyavora, Subhashini Venugopalan, Jason Hickey, Uche Okonkwo, Abhishek Bapna, Isaac Wiafe, Raynard Dodzi Helegah, Elikem Doe Atsakpo, Charles Nutrokpor, Fiifi Baffoe Payin Winful, Kafui Kwashie Solaga, Jamal-Deen Abdulai, Akon Obu Ekpezu, Audace Niyonkuru, Samuel Rutunda, Boris Ishimwe, Michael Melese, Engineer Bainomugisha, Joyce Nakatumba-Nabende, Andrew Katumba, Claire Babirye, Jonathan Mukiibi, Vincent Kimani, Samuel Kibacia, James Maina, Fridah Emmah, Ahmed Ibrahim Shekarau, Ibrahim Shehu Adamu, Yusuf Abdullahi, Howard Lakougna, Bob MacDonald, Hadar Shemtov, Aisha Walcott-Bryant, Moustapha Cisse, Avinatan Hassidim, Jeff Dean, and Yossi Matias. 2026. Waxal: A large-scale multilingual african language speech corpus.

Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, et al. 2025. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages. *arXiv e-prints*, pages arXiv–2511.

Franziska Koefer, Amber Bokkens, Massimo Preziuso, and Michel Ehrenhard. 2024. Addressing financial and digital literacy challenges for inclusive finance: Insights from microfinance institutions and fintech organisations. *EIF Working Paper*, (2024/97).

D Fortuné Kponou, Salima Mdhaffar, Fréjus AA Laleye, Eugène Cokou Ezin, and Yannick Estève. 2025. Ffstc 2: Extending the fongbe to french speech translation corpus. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 145–152.

Damien Kunjal. 2026. From access to ability: The impact of financial inclusion on financial knowledge, capabilities, and literacy among gen z university students in south africa. *Journal of Risk and Financial Management*, 19(3).

Indrani Medhi Thies. 2015. User interface design for low-literate and novice users: Past, present and future. *Found. Trends Hum.-Comput. Interact.*, 8(1):1–72.

Mumtaz Begum Mustafa, Mansoor Ali Yusoof, Hasan Kahtan Khalaf, Ahmad Abdel Rahman Mahmoud Abushariah, Miss Laiha Mat Kiah, Hua Nong Ting, and Saravanan Muthaiyah. 2022. Code-switching in automatic speech recognition: The issues and future directions. *Applied Sciences*, 12(19).

Carol Myers-Scotton. 1997. *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press.

Palanichamy Naveen and Pavel Trojovský. 2024. Overview and challenges of machine translation for contextually appropriate translations. *iScience*, 27(10):110878.

Tolúlọpẹ́ Ògúnrèmí, Christopher D. Manning, and Dan Jurafsky. 2023. Multilingual self-supervised speech representations improve the speech recognition of low-resource African languages with codeswitching. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 83–88, Singapore. Association for Computational Linguistics.

Fayçal S. A. Ouedraogo, Maimouna Ouattara, Rodrique Kafando, Abdoul Kader Kabore, Aminata Sabane, and Tegawendé F. Bissyandé. 2026. Contributing to speech-to-speech translation for African low-resource languages : Study of French-mooré pair. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 623–629, Rabat, Morocco. Association for Computational Linguistics.

Peterson K. Ozili. 2025. Financial literacy theory of financial inclusion. *SSRN Electronic Journal*.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Melain Modeste Senou, Wautabouna Ouattara, and Denis Acclassato Houensou. 2019. Financial inclusion dynamics in waemu: Was digital technology the missing piece? *Cogent Economics & Finance*, 7(1):1665432.

T. H. O. da Silva et al. 2024. How do illiterate people interact with an intelligent voice assistant? *International Journal of Human–Computer Interaction*. Voice assistants and access for low-literacy users (publisher page). Accessed 2026-02-24.

Harish Thangaraj, Ananya Chenat, Jaskaran Singh Walia, and Vukosi Marivate. 2024. Cross-lingual transfer of multilingual models on low resource african languages.

Gokhan Tur and Renato De Mori. 2011. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons.

UNESCO. 2025. Youth and adult literacy: International literacy day 2025. promoting literacy in the digital era. factsheet. UIS education database, September 2025.

World Bank. 2023. Literacy rate, adult total (% of people ages 15 and above) - burkina faso. `https://data.worldbank.org/indicator/SE.ADT.LITR.ZS?locations=BF`. Most recent value: 42% in 2023. Source: UNESCO Institute for Statistics. Accessed: 2026-06-15.

Hao Yu, Jesujoba Alabi, Andiswa Bukula, Jian Yun Zhuang, En-Shiun Annie Lee, Tadesse Kebede Guge, Israel Abebe Azime, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson Koffi Kalipe, et al. 2025. Injongo: A multicultural intent detection and slot-filling dataset for 16 african languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9429–9452.

Emre Yılmaz, Astik Biswas, Ewald van der Westhuizen, Febe de Wet, and Thomas Niesler. 2018. Building a Unified Code-Switching ASR System for South African Languages. In *Interspeech 2018*, pages 1923–1927.

A Judge Prompt and Output Schema

We reproduce the conservative-judge prompt and JSON schema from `gg_abs/judge.py` and `gg_abs/main.py`. The prompt is assembled by `_build_prompt` from a context block, the two raw ASR transcripts, a decision-dependent body, and the schema tail, and is identical across regimes. Every call uses `gemini-3.5-flash` at temperature 0.1 with JSON output.

A.1 Context block

The domain context carries the *wakir* numeral system, the log-odds-induced direction verbs, the phonetic-tolerance note, and the tense rule.

```
Merchant credit-ledger utterance in Moore /
Moore-French code-switch (Burkina Faso,
L2-French speakers).  Customers may take
goods/cash on credit (debt INCREASES) or
pay back (debt DECREASES) ; the speaker may
also ask a balance question.  A single
utterance may contain several transactions
(compound).  All amounts are in CFA.
Entity = customer's name or social role,
in lowercase snake_case, or null when the
speaker does not mention one.
Moore numeral system (only when speaker uses
Moore numerals; not needed for French):
* Speakers count in WAKIR units;
  CFA amount = wakir x 5.
* Magnitudes: tus = 1000 wakir,
  kobs = 100 wakir, piiga = 10 wakir.
* Units 1-9: yembre, yi, ta, naase, nu,
  yoobe, yopoe, nii, wae.
* Compound multiplies magnitude by unit:
  kobs-naase = 100 x 4 = 400 wakir = 2000 CFA.
* Connector "la a": tus la a kobs-yi
  = 1000 + 200 wakir = 6000 CFA.
French numerals are read directly in CFA.
Moore direction verbs (induced from synth via
log-odds; validated direction-discriminative):
* debt DECREASES: yaoo/yao/yaoome (pay back),
  koo (customer gives -> repays), reegd.
* debt INCREASES: tara/reega/ket (took credit).
* query: wana ("how much").
ASR Phonetic Tolerance: treat close phonetic
matches of these verbs as valid.
TENSE matters: count debt_down only if payment
ALREADY happened; a promised future payment
stays debt_up only.
```

A.2 Evidence, hint body, and tiering

The two channels follow under a fixed header (`W2V-mos`, `Omni-mos`), one transcript per line, empty channels as `(none)`. With a Mooré hint (`DECISION_MOORE_HINT`) the body adds `GG-Abs: amount(s) = {final_amounts} CFA` and the judge finalizes direction and entity, one entry per amount. On abstention (`DECISION_NO_HINT`) the judge self-extracts amounts and returns `null` if none, the arm that routes to audio. Each emitted hint carries a categorical tier from independent ASR support, not a tunable score: HIGH when at least two channels agree (`support_count` ≥ 2) and MED for a single channel. LOW is the abstention state (no amount, no hint, routes to audio) and never appears in the tier-precision table. The amount-bearing holdout splits into 56 HIGH and 94 MED (Appendix C).

A.3 Schema tail and escalation

The prompt closes by demanding strict JSON.

```
Return STRICT JSON matching this compound
ledger schema :
{"currency": "XOF",
"debt_up":   [{"entity": <"snake_case"|null>,
"amount": <int CFA|null>}, ...],
"debt_down": [{"entity": <"snake_case"|null>,
"amount": <int CFA|null>}, ...],
"queries":   [{"type": <"who_owes_me"
|"debt_of_entity"|"total_due_to_me"
|"remaining_total">,
"entity": <"snake_case"|null>}, ...],
"is_proposal_correct": <bool>}
```

Lists let compound utterances be expressed naturally, and output is parsed fence-tolerantly (`parse_ledger`). Escalation fires exactly when the module abstains (`hint.decision == "no_hint"`): the same prompt is sent with the raw audio as a second `inline_data` part (`audio/wav`), additional evidence rather than a replacement. This produces the headline CS amount 85.0 ± 0.5 and CS direction 92.7.

B Governed Rule Catalogue

The deterministic path is governed regex rewrites with no learned parameters: CPU-only, reproducible from cached transcripts without GPU or API. Table 7 is the deployed Mooré catalogue with a one-rule ablation on the male synthetic development split (`intents_mos_fr`, `pool_dev`, n=800, baseline amount-exact 77.0%, gold misses 4, false hints 180). The ablation is a deterministic replay with no LLM call: we disable one regex at a time and recompute the hint. `Net` is helps-minus-hurts, and every kept rule has `net` ≥ 0 and leaves gold misses unchanged.

The most load-bearing rule is *kobs-nii* (H-3/3, net +32). Expert rules carry literature anchors in the governance file: *samd[e/a/ã]* stripping is expert-validated as non-amount-bearing, and the truncated magnitudes and the phonetic-skeleton fold follow descriptive Mooré sources.

B.1 Code-switch amount gap

On code-switching the module should abstain and route audio to the judge, so the 85.0 versus 90.1 gap (5.1 points, about six utterances on n=118) splits in two. In three cases the module fails to abstain and emits a spurious MED hint on a French amount (gold 1000, 2500, 20000 read as 5, 40, 5). Because audio is not attached to a row that already

Rule	Prov.	Fires	Exact	Net	Δmiss
Expert (linguist / literature)					
samd[e/a/ã] strip	EXP	361	78.7	0	0
wakir+tus preserve	EXP	7	71.4	+3	0
wakir strip	EXP	307	73.6	0	0
Heuristic (cross-LLM clusters)					
kobs,nii→kobs-nii	H-3/3	63	63.5	+32	0
pis,nii→pis-nii	H-3/3	1	100.0	+1	0
nung/nuy/num→nu	H-2/3	22	86.4	+2	0
pis{X}→pis-X	H-3/3	120	88.3	0	0
piir(ã)→piig	H-3/3	5	40.0	0	0
tus piir→tus piig	H-3/3	2	100.0	0	0
yogbe→yoobe	H-3/3	0	—	0	0
yokoe→yopoe	H-3/3	0	—	0	0
*tus 1d→*teen	H-3/3	0	—	0	0

Table 7: Deployed Mooré catalogue, one-rule dev ablation (`pool_dev`, n=800, baseline 77.0%). EXP=expert, H-k/3=proposed by k of three LLM analyses. Fires=utterances matched; Exact=amount accuracy (%) on those rows; Net=helps minus hurts; Δmiss=change in gold misses if removed (all 0).

carries a hint, these lose the rescuing escalation and are all wrong. Attaching audio even when a low-confidence hint fires on a likely-French row would lift CS amount toward 87.5. The other three are genuine residual. The gap is a wrinkle in one field, and GG-Abs still wins CS direction (92.7 vs. 57.6).

C Threshold and Reliability Ablations

The module has two frozen knobs: `TAU_FUZZY`, the ungated per-token fuzzy threshold (default 0.80), and `TAU_GATE`, the looser threshold for compound repair behind a matched prefix (default 0.65). Both sweeps are deterministic replays over the n=800 development pool, with the sealed hold-out untouched.

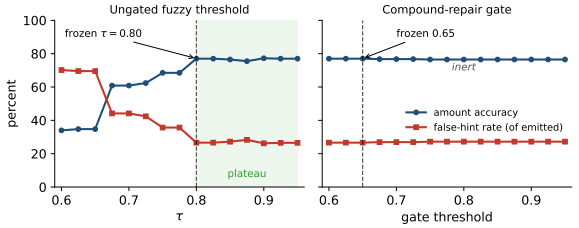


Figure 2: Threshold sweeps on the Mooré development pool (n=800, no LLM). Left: amount accuracy plateaus at and above τ=0.80 and collapses below 0.65, where the false-hint rate overtakes accuracy. Right: the compound-repair gate is inert across its range. Both were fixed before the sealed test.

At the deployed τ=0.80, accuracy is 77.0% with a 26.6% false-hint rate, and τ=0.85 is similar (76.5%, 27.2%). Loosening to τ=0.65 collapses the module to 34.8% accuracy and a 69.6% false-hint rate, since indiscriminate fuzzy matching maps noise onto magnitude markers. The frozen design makes magnitudes reachable only by exact token, exact skeleton, or an adjacency-gated rewrite. Sweeping `TAU_GATE` from 0.60 to 0.95 (with τ=0.80) leaves accuracy in $[76.5, 77.0]$%, so precision comes from the prefix, not the gate value.

LëtzCross: A Cross-Lingual Page-Level Benchmark for Multimodal Retrieval over Luxembourgish Documents

Omar El Bachyr[1,*] **Fred Philippy**[1] **Laura Maria Bernardy**[1]
Saad Ezzini[2] **Jacques Klein**[1] **Tegawendé F. Bissyandé**[1]

[1]University of Luxembourg, Luxembourg
[2]King Fahd University of Petroleum and Minerals, Saudi Arabia
[*]**Correspondence:** omar.elbachyr@uni.lu

Abstract

Recent page-image retrievers such as ColPali (Faysse et al., 2025) have improved retrieval over visually rich documents, yet little is known about how they behave in cross-lingual, low-resource settings. We introduce *LëtzCross*, a benchmark for cross-lingual page-level retrieval over Luxembourgish PDF documents, with document pages indexed as images and queries provided in English, French, German, and Luxembourgish. The benchmark combines text-focused QA pairs with visually grounded QA pairs, covering both textual and visual retrieval needs in PDF-based RAG. We use *LëtzCross* to compare OCR-based text-only retrievers with ColPali-style page-image retrievers and find that the latter perform better across query languages in this system-level comparison. We also examine single-language and multilingual fine-tuning. Fine-tuning transfers across query languages, with French yielding the highest mean performance on Luxembourgish queries among the single-language settings. In the multilingual setting, including Luxembourgish gives the strongest results and substantially improves retrieval for Luxembourgish queries.

1 Introduction

Retrieval-augmented generation (RAG) and document retrieval systems increasingly operate in multilingual environments, where user queries and source documents may be written in different languages. In this setting, cross-lingual information retrieval (CLIR) has advanced substantially with dense retrieval architectures and multilingual encoders (Karpukhin et al., 2020; Conneau et al., 2020). However, most current evaluations focus on text-centric, high-resource scenarios and do not reflect the challenges of visually rich PDF documents.

Recent vision–language retrieval models, such as DSE (Ma et al., 2024) and ColPali (Faysse et al., 2025), improve retrieval over document pages by jointly modeling textual and visual evidence. Yet their behavior in low-resource language settings remains underexplored. We study this challenge through the case of Luxembourgish, a low-resource language for which NLP resources are growing but still limited compared to high-resource languages (Lothritz et al., 2022; Bernardy, 2022; Plum et al., 2025).

To address this gap, we introduce *LëtzCross*, a benchmark for cross-lingual page-level retrieval over Luxembourgish PDF pages, where pages are represented as images. The benchmark combines text-focused and visually grounded QA pairs, and is designed around realistic page-level retrieval needs in RAG pipelines, where answers may depend on textual content, visual structure, or both, and where queries may be issued in multiple languages.

Our main contributions are:

- We introduce LëtzCross, a benchmark for cross-lingual page-level retrieval over Luxembourgish PDF pages, where pages are represented as images.
- We benchmark multilingual text-only retrievers and late-interaction page-image retrievers on this task.
- We study query-language-specific fine-tuning for cross-lingual retrieval over Luxembourgish document pages.
- We analyze multilingual fine-tuning settings for retrieval over Luxembourgish document pages as a low-resource language use case.

Overall, our findings show that late-interaction page-image retrievers are strong candidates for cross-lingual retrieval over Luxembourgish document pages, outperforming the evaluated OCR-

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 284–294
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_029

based text-only baselines in a system-level comparison. We further show that query-language-specific fine-tuning transfers across languages, while multilingual fine-tuning is most effective when it includes the target low-resource language. Together, these results position *LëtzCross* as a useful benchmark for studying cross-lingual page-level retrieval over Luxembourgish PDF documents as a low-resource language in RAG settings.

We release LëtzCross and all experimental code and resources through our GitHub repository.

2 Related Work

Luxembourgish NLP-Resources The Luxembourgish NLP landscape is still under development. Natural Language Processing techniques for Luxembourgish are primarily developed within a low-resource setting and rely heavily on manual efforts, such as data collection and annotation. In recent years, this approach has enabled the development of several foundational resources and applications, including sentiment analysis systems (Sirajzade et al., 2020; Gierschek, 2022), a Luxembourgish dependency parsing treebank (Plum et al., 2024), and zero-shot topic classification (Philippy et al., 2024). These initiatives have contributed to a better understanding of the linguistic properties of Luxembourgish through computational methods. Within this context, Lothritz et al. introduced LuxemBERT (Lothritz et al., 2022), a language model trained using artificial data augmentation strategies. In addition to the model itself, they expanded Luxembourgish resources for several downstream tasks, including part-of-speech tagging, named entity recognition, and news classification. Two further language models have since been established for Luxembourgish: LuXGPT (Bernardy, 2022), developed using transfer learning techniques, and, more recently, LUXT5 (Plum et al., 2025), which benefits from multilingual pretraining. Beyond language modeling, practical NLP applications have also emerged. These include LUX-ASR (Gilles et al., 2023) for automatic speech recognition, as well as systems for automatic comment moderation (Ranasinghe et al., 2023) and orthographic correction (Purschke, 2020).

Cross-lingual retrieval Transformer-based Cross-Lingual Information Retrieval (CLIR) builds on multilingual representation learning and dense retrieval architectures that enable semantic matching across languages. Early foundations include BERT (Devlin et al., 2019), which introduced contextual transformer representations for retrieval models, and XLM-R (Conneau et al., 2020), a multilingual transformer that provides strong cross-lingual embeddings. Dense Passage Retrieval (DPR) (Karpukhin et al., 2020) later established the dual-encoder architecture widely used in dense retrieval systems. Building on these foundations, more recent work explores unsupervised and generative approaches, such as Unsupervised Multilingual Dense Retrieval (UMR) proposed by (Huang et al., 2024), which uses generative pseudo-labeling to train dense retrievers without parallel supervision, and research on adapting LLMs for retrieval tasks (Guo et al., 2024). Recent multilingual embedding models such as BGE-M3 (Chen et al., 2024), Qwen3 Embedding (Zhang et al., 2025), and voyage-4-nano (AI, 2026) further improve cross-lingual semantic representations and are widely used as baselines in modern CLIR systems.

Vision Language Retrieval Text-only encoders have achieved strong performance in information retrieval (IR), with models such as (Chen et al., 2024; Zhang et al., 2025; AI, 2026) demonstrating high effectiveness across multilingual and dense retrieval tasks. However, recent studies in IR have advanced beyond conventional text-only systems by incorporating visual features. DSE (Ma et al., 2024) propose the use of Vision–Language Models (VLMs) to encode full document screenshots, successfully preserving both text and visual information without the need for optical character recognition (OCR). Similarly, ColPali (Faysse et al., 2025) addresses the shortcomings of text-focused retrieval in the context of visually rich documents by adapting ColBERT (Khattab and Zaharia, 2020) late-interaction similarity scoring mechanism to work with page images. This produces multi-vector visual embeddings that enable fine-grained matching between query tokens and specific visual areas on document pages.

3 LëtzCross Benchmark Construction

LëtzCross is a cross-lingual page-level retrieval benchmark for Question-Answering (QA), built from Luxembourgish documents, with queries in English (EN), French (FR), German (DE), and Luxembourgish (LB). Documents are represented as page images, and answers are grounded in the corresponding pages. Figure 1 summarizes the process,

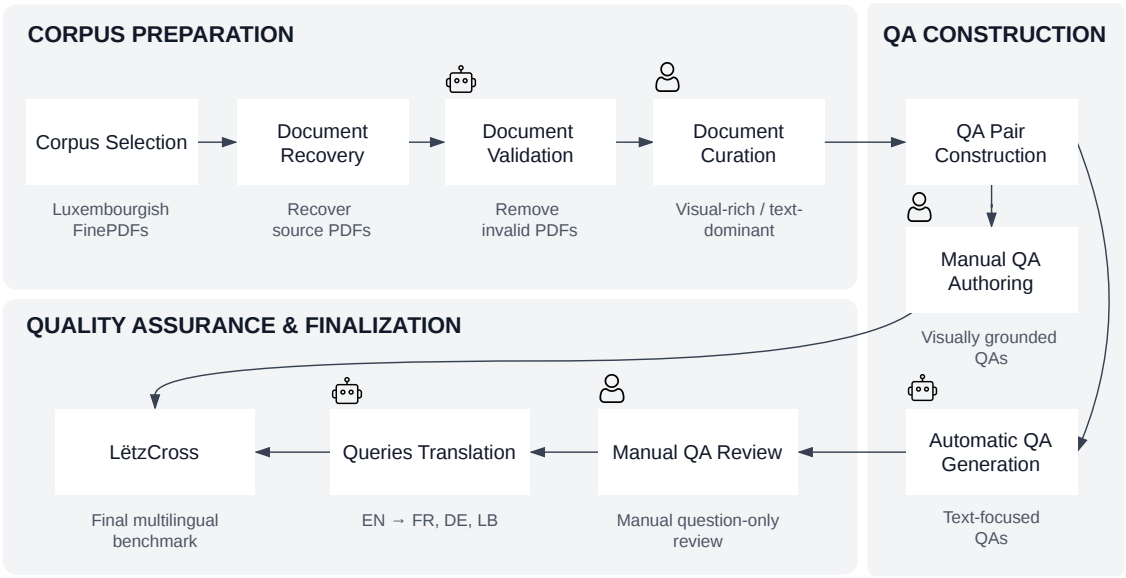


Figure 1: Overview of the LëtzCross benchmark construction pipeline.

with steps described below.

3.1 PDF Documents Acquisition

We obtain the original PDF documents by iterating over the Luxembourgish (LB) split of FinePDFs (Kydlíček et al., 2025) dataset and retrieving each document from its source. When documents are archived in Common Crawl, we extract the corresponding PDF directly from the web archive; otherwise, we download the PDF from the original URL provided in the dataset. This process allows us to recover the original PDF files associated with the dataset entries.

3.2 Filtering and Curation of PDF Documents

After downloading the PDFs, we perform an automatic validation step to remove corrupted or unreadable documents by verifying that each PDF can be successfully parsed and processed. We then manually classify the validated PDFs into visually rich and text-dominant subsets. A document is considered visually rich when it contains informative visual elements, such as charts, diagrams, or tables, that convey substantive information beyond the surrounding text. Non-informative visuals, such as portraits, landscapes, and logos, are disregarded. Documents whose information is conveyed primarily through continuous text, with few or no informative visual elements, are classified as text-dominant. The resulting corpus serves as the foundation for the subsequent Question–Answer generation step.

3.3 Question–Answer Generation

We then use the validated PDFs from the previous step to generate Question–Answer (QA) pairs. We generate queries in English to ensure high-quality and consistent question formulation, as this yields more reliable QA pairs on Luxembourgish pages. The benchmark combines two complementary QA sources: manually authored visually grounded QA pairs for visually rich pages, and automatically generated text-focused QA pairs for text-dominant pages.

Manual visually grounded QA authoring. For the subset of PDFs containing rich visual components, a native Luxembourgish speaker manually crafts page-level QA pairs that intentionally target visually grounded evidence, such as charts, plots, diagrams, figures, and other layout-dependent elements, rather than relying solely on textual content. Questions are phrased to be self-contained and answerable from a single page image, with concise answers grounded in the visual content.

Automatic text-focused QA generation. Using the text-dominant PDF subset, we automatically generate page-level QA pairs with GPT-5-mini. Each page is processed independently as an image and prompted to produce self-contained questions with concise answers grounded in its textual content. To validate the generated QA pairs, each query and its corresponding PDF are provided to Gemini-2.5-Flash, which extracts an answer from

the document. We then compare the extracted answer with the *gold answer* using both exact and semantic matching; the latter is performed by prompting GPT-5-mini to assess whether the two answers convey the same meaning. Overall, 90.28% of the QA pairs were judged semantically equivalent by GPT-5-mini, and 71.50% contained the exact literal gold answer. QA pairs that do not yield consistent answers are discarded, resulting in a consistency-filtered set for the subsequent manual validation step.

3.4 Quality Control and RAG Alignment

As a follow-up step, the manually authored QA pairs are included directly in the final benchmark. In contrast, the automatically generated QA pairs undergo a manual validation phase focused exclusively on question formulation: annotators are shown only the questions and revise them when needed to produce clear, natural queries that reflect RAG-relevant information needs (e.g., realistic, retrieval-oriented queries).

3.5 Cross-Lingual Query Translation

To enable cross-lingual retrieval evaluation, we translate the finalized English queries into German, French, and Luxembourgish using a controlled LLM prompt that enforces semantic preservation and prevents paraphrasing or additional explanations. This produces aligned query sets across languages, where each translated query is intended to preserve the same retrieval intent as the original English query. We use Gemini-3-Flash for German and French, and Gemini-3-Pro for Luxembourgish. This choice was made after manually inspecting a small set of trial translations: Gemini-3-Pro produced better Luxembourgish translations, while Gemini-3-Flash was sufficient for German and French.

The final benchmark contains 579 QA pairs over 908 document pages. Table 1 summarizes the benchmark statistics, including the distribution between automatically generated text-focused QA pairs and manually authored visually grounded QA pairs. All prompts used during benchmark construction are available in the accompanying repository.

Benchmark statistic	**Value**
Number of pages	908
Number of queries	579
Avg. query length	19.5
Avg. corpus item length	722.5
QA type distribution	
Auto-generated text-focused	530
Manual visually grounded	49

Table 1: Overview statistics of the *LëtzCross* evaluation benchmark. Query and corpus item lengths are reported in tokens.

4 Methods

This section presents our method for investigating cross-lingual retrieval over Luxembourgish document pages as a low-resource case study using late-interaction page-image retrievers.

4.1 Training Data Construction

We construct the training data by scraping Luxembourgish Wikipedia articles and converting them into PDF documents. We first apply size-based filtering with minimum and maximum page thresholds to exclude trivial or overly long documents (e.g., pages containing only footnotes). QA pairs are then generated from 4,977 PDFs with 11,833 pages using the same automatic text-focused QA generation procedure as in the benchmark (Section 3.3). This resulted in 22,028 QA pairs. In addition, we construct 450 QA pairs from the same Wikipedia-derived pipeline and reserve them as a held-out development split for fine-tuning. To extend the training set for cross-lingual retrieval, we translate queries into German and French, and a 5,000-query subset into Luxembourgish, following the same query translation procedure as in the benchmark (Section 3.5).

4.2 Retriever and Training Objective

We fine-tune a late-interaction page-image retriever based on ColPali (Faysse et al., 2025) using the training setup provided by the underlying framework. Each document page is treated as an image and split into visual patches, which are encoded into contextualized image patch embeddings, while the query text is tokenized and encoded into query token embeddings. Both are represented in a shared latent space, and relevance is computed through late interaction between query tokens and page patches.

Let $q_i = \{q_{i1}, \ldots, q_{i|q_i|}\}$ denote the query token embeddings of query i, and let $d_j = \{d_{j1}, \ldots, d_{j|d_j|}\}$ denote the document image patch embeddings of page j. The score between query i

and page j is:

$$\text{score}(q_i, d_j) = \sum_{t=1}^{|q_i|} \max_{u=1,\ldots,|d_j|} \left(q_{it}^{\top} d_{ju} \right) \quad (1)$$

We use the original ColBERT (Khattab and Zaharia, 2020) contrastive objective with in-batch negatives:

$$\mathcal{L} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp\left(\text{score}(q_i, d_i)/\tau\right)}{\sum_{j=1}^{B} \exp\left(\text{score}(q_i, d_j)/\tau\right)} \quad (2)$$

where τ is a temperature hyperparameter. Each training instance consists of a query paired with its corresponding document page image as the positive example, while the remaining $B-1$ pages in the batch serve as in-batch negatives. We do not use external hard negatives or additional negative sampling.

4.3 Fine-Tuning Configurations

We consider two fine-tuning settings. In the *per-language* setting, the retriever is fine-tuned separately on English, French, and German queries. In the *multilingual* setting, a single model is trained on mixed-language data, using either EN+FR+DE or EN+FR+DE+LB. In all cases, document pages remain in Luxembourgish, and only the query language varies. To analyze visual adaptation, we also compare the default setting, which updates both textual and visual components, with a text-component-only variant. The EN, FR, and DE settings each use 22,028/450 train/dev QA pairs, while EN+FR+DE and EN+FR+DE+LB use 21,000/1,350 and 20,000/1,800 train/dev samples, respectively.

5 Experimental Setup

Models. Based on the input type, we consider two families of retrievers in our experiments: text-only retrievers and late-interaction page-image retrievers. The text-only baselines are bge-m3 (Chen et al., 2024), Qwen3-Embedding-0.6B (Zhang et al., 2025), multilingual-e5-large (Wang et al., 2024), and jina-embeddings-v4 (Günther et al., 2025)[1]. The late-interaction page-image retrievers are colSmol-256M, colSmol-500M, colpali-v1.3, and colqwen2.5-v0.2 (Faysse et al., 2025), and colnomic-embed-multimodal-3b (Nomic AI, 2025).

[1]Although `jina-embeddings-v4` is a multimodal embedding model, we use it here in a text-only setting.

Text Extraction Pipeline. The text-only retrievers described above use page-level text extracted with `unstructured[pdf]` v0.18.15 under the `hi_res` strategy, combining Poppler page rendering, YOLOX layout detection, Luxembourgish Tesseract OCR (`ltz`), and table-structure extraction. Existing PDF text was incorporated when available, and all PDFs were processed using the same pipeline. We consider this a practical system-level comparison between text-based and page-image retrieval, rather than an evaluation of PDF extraction methods.

Fine-Tuning Details. We fine-tune the retriever using the ColBERT late-interaction contrastive loss defined in Equation (2) with temperature $\tau = 0.02$ and LoRA adapters (`r`=32, `alpha`=32, dropout $= 0.1$). In the text-component-only setting, LoRA targets the query encoder's attention projections (`q_proj`, `k_proj`, `v_proj`, and `o_proj`), feed-forward projections (`down_proj`, `gate_proj`, and `up_proj`), and `custom_text_proj`; the vision encoder remains frozen. In the text+visual setting, the same attention and feed-forward projections are adapted in both the textual and visual transformer blocks, where the visual component denotes the blocks encoding page-image patches. All untargeted parameters remain frozen. Training is performed for 3 epochs with a learning rate of 5×10^{-5} and 100 warmup steps, using a per-device batch size of 16 on 4 NVIDIA L40S GPUs and no gradient accumulation, resulting in an effective batch size of 64. All runs use bfloat16 precision. Each fine-tuning configuration is repeated with three seeds (42, 43, and 44), and results are reported as mean $\pm$ standard deviation.

Evaluation Setup and Metrics. We evaluate retrieval at the page level: for each query, the retriever ranks all document pages in the benchmark against the gold relevant page. Results are reported separately for English, French, German, and Luxembourgish queries. Our main metric is nDCG@10; for analyses across retrieval depths, we additionally report nDCG@k for $k \in \{1, 3, 5, 10\}$. Indexing time per page and query latency per query are aggregated across the four query languages and reported as mean $\pm$ standard deviation over three runs. All evaluation runs, including effectiveness and efficiency measurements, use one NVIDIA L40S GPU and eight CPU cores.

Retriever	Query Language				Avg.	Index (ms/page) ↓	Latency (ms/query) ↓
	EN	FR	DE	LB			
TEXT-ONLY RETRIEVERS							
bge-m3	72.48	71.63	72.47	76.12	73.18	21.714 ± 0.015	3.859 ± 0.027
Qwen3-Embedding-0.6B	73.02	71.63	75.21	67.97	71.96	42.863 ± 0.115	6.935 ± 0.008
multilingual-e5-large	75.96	75.02	77.24	68.32	74.14	11.068 ± 0.070	3.858 ± 0.012
jina-embeddings-v4	76.51	75.39	77.30	73.86	75.76	113.007 ± 0.287	26.271 ± 0.336
MULTI-VECTOR PAGE-IMAGE RETRIEVERS							
colSmol-256M	71.61	59.45	52.76	55.11	59.73	367.108 ± 1.415	3.914 ± 0.035
colSmol-500M	73.90	67.56	65.92	63.87	67.81	361.103 ± 11.169	4.125 ± 0.017
colpali-v1.3	77.03	77.43	78.46	75.78	77.18	61.801 ± 0.110	2.997 ± 0.007
colqwen2.5-v0.2	78.07	**78.71**	**79.92**	**78.60**	**78.83**	133.876 ± 0.813	4.892 ± 0.113
colnomic-embed-multimodal-3b	**78.73**	78.44	79.59	77.89	78.66	145.193 ± 0.277	4.981 ± 0.067

Table 2: Cross-lingual retrieval over Luxembourgish document pages. Query-language columns report nDCG@10 where best results are shown in **bold**, and second-best results are underlined. Efficiency values are mean ± standard deviation over three runs on the same hardware.

6 Results and Analysis

Our experiments evaluate cross-lingual retrieval over Luxembourgish document pages. We first compare text-only and page-image retrievers, then analyze query-language-specific fine-tuning, including an ablation on the visual component, and finally evaluate multilingual fine-tuning.

6.1 Page-Image vs. OCR-Based Text Retrieval over Luxembourgish Documents

We compare multilingual OCR-based text-only embedding baselines, which index page-level text produced by the OCR, layout-analysis, and table-extraction pipeline described in Section 5, with late-interaction page-image retrievers ranging from lightweight models below 500M parameters to models of up to 3B parameters. Table 2 shows that the strongest page-image retrievers outperform the evaluated text-only baselines across all query languages. Since the two families differ in input representation, extraction pipeline, and model architecture, we interpret this as a practical system-level comparison rather than a controlled modality ablation.

Performance remains strongly model-dependent. The smaller colSmol models underperform strong multilingual text-only retrievers such as jina-embeddings-v4. In contrast, colqwen2.5-v0.2 and colnomic-embed-multimodal-3b outperform all evaluated text-only models across every query language, while colpali-v1.3 also performs strongly on English, French, and German. This shows that sufficiently capable multi-vector page-image retrievers can outperform multilingual text embedding models in cross-lingual retrieval. The best page-image retriever achieves an average nDCG@10 of 78.83, improving by 3.07 points over the strongest text-only baseline. A more customized OCR and layout-detection pipeline could nevertheless affect this gap.

These results are notable because the text-only baselines are explicitly designed for multilingual retrieval, whereas the page-image retrievers primarily target visual-document retrieval. Their strong zero-shot cross-lingual performance may partly reflect the multilingual VLM backbones used by several models.

Table 4 further examines performance by QA source. The page-image retrievers outperform the strongest text-only baseline on both the text-focused and visually grounded subsets, with larger gains on the visually grounded questions. This provides additional evidence that page-image retrieval is beneficial beyond the predominantly text-focused portion of the benchmark. However, the visually grounded subset contains only 49 QA pairs and yields higher scores for all retrievers, so these results remain descriptive and should be interpreted cautiously.

Efficiency. Table 2 shows clear efficiency differences across retrievers. Among text-only models, multilingual-e5-large indexes fastest, while jina-embeddings-v4 is slower, likely due to its larger multimodal backbone. Among page-image retrievers, colpali-v1.3 is the fastest, possibly because of its fixed-resolution processing and fixed visual-token budget. The slower ColSmol indexing further shows that runtime depends on visual-token count

Training setup	Query language				
	EN	FR	DE	LB	Avg.
Baseline	78.73	78.44	79.59	77.89	78.66
FT (EN)	**81.66** ± 0.33 (+2.93)	81.34 ± 0.40 (+2.89)	**81.84** ± 0.11 (+2.24)	78.48 ± 0.65 (+0.60)	80.83 ± 0.23 (+2.17)
FT (FR)	81.11 ± 0.57 (+2.38)	**81.50** ± 0.49 (+3.06)	81.76 ± 0.40 (+2.17)	**79.45** ± 0.75 (+1.57)	**80.96** ± 0.55 (+2.29)
FT (DE)	79.48 ± 0.27 (+0.75)	80.51 ± 0.88 (+2.07)	81.56 ± 0.34 (+1.97)	79.34 ± 0.33 (+1.45)	80.22 ± 0.14 (+1.56)
Avg. FT	80.75 ± 0.15	81.12 ± 0.22	81.72 ± 0.08	79.09 ± 0.34	80.67 ± 0.11

Table 3: nDCG@10 by query language (columns) for the baseline model and models fine-tuned on different training languages (rows). Values are reported as mean ± standard deviation over three runs; values in parentheses indicate gains over the baseline on the same query language. Best values are shown in **bold**.

Retriever	Text-focused	Visually grounded	Overall
jina-embeddings-v4	75.27	81.10	75.76
colqwen2.5-v0.2	78.42	83.23	78.83
colnomic-embed-multimodal-3b	77.91	86.82	78.66

Table 4: Descriptive nDCG@10 split by QA source, averaged across query languages.

and implementation choices, not only parameter count. Query latency remains similar across most page-image models because page representations are precomputed.

For the text-only retrievers, page parsing introduces an additional shared preprocessing cost. Using the extraction pipeline described in Section 5, parsing takes 5.382 ± 0.054 s/page across three runs. Including this cost, the end-to-end preprocessing and indexing time ranges from 5.393 ± 0.054 s/page for multilingual-e5-large to 5.495 ± 0.054 s/page for jina-embeddings-v4. Thus, for text-only retrieval, document parsing dominates the overall offline preprocessing cost.

6.2 Cross-Lingual Transfer of Query-Language-Specific Fine-Tuning

This subsection studies how fine-tuning transfers across evaluation languages for the late-interaction page-image retrievers considered in this work. We fine-tune the retriever separately on English, French, and German queries, and evaluate each variant on all four query languages. We use colnomic-embed-multimodal-3b for this analysis, as both the baseline model and the backbone for fine-tuning, because its performance remains very close to the top model in our experiments, while ViDoRe leaderboard (Loison et al., 2026) results suggest it is a strong page-image retriever overall.

Table 3 shows that all language-specific fine-tuning settings improve over the baseline on every query language. FT (FR) achieves the highest overall average at 80.96, closely followed by FT (EN) at 80.83, while FT (DE) reaches 80.22. The largest language-specific gains are obtained by FT (EN) on English (+2.93) and FT (FR) on French (+3.06). Cross-lingual transfer is also substantial: FT (EN) achieves the highest German score at 81.84 (+2.24), while FT (FR) performs best on Luxembourgish at 79.45 (+1.57).

These results show that the strongest configuration does not always correspond to matching the fine-tuning and evaluation languages. In particular, English fine-tuning transfers strongly to German, while French fine-tuning provides the strongest Luxembourgish performance. German fine-tuning improves all four languages but produces the lowest overall average among the three variants. Overall, query-language-specific fine-tuning consistently improves multilingual retrieval, while the magnitude of transfer depends on the training and evaluation language pair.

Figure 2 shows that these trends are broadly maintained across retrieval cutoffs. FT (EN) remains strongest for English queries, whereas the fine-tuned variants are much closer for French and German. For Luxembourgish, FT (FR) and FT (DE) generally perform better than FT (EN). The uncertainty bands also overlap in several cases, indicating that small differences between fine-tuning languages should not be over-interpreted. Overall, the improvements persist across different values of k, showing that the cross-lingual benefits of fine-tuning are not limited to nDCG@10.

Effect of Updating the Visual Component. In all previous fine-tuning experiments, we use a setting in which both the textual and visual components of the page-image retriever are updated. To isolate the contribution of visual adaptation, we compare this setting with the corresponding text-

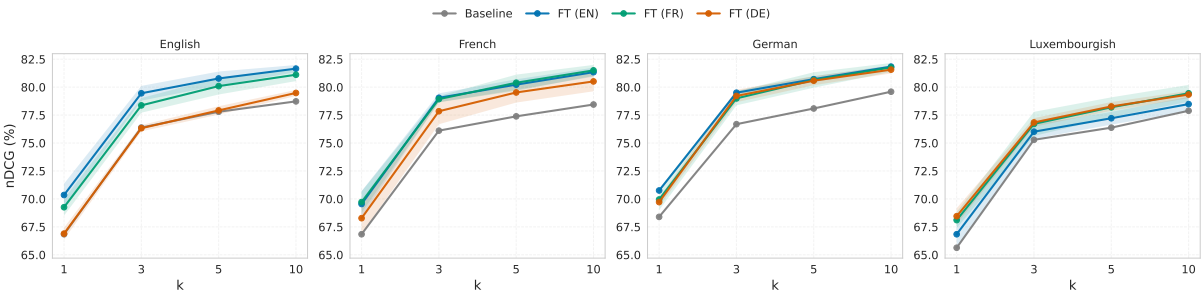


Figure 2: Multilingual retrieval performance (nDCG@k) across query languages. Shaded bands indicate ± 1 standard deviation across three fine-tuning runs.

Training setup	Query language			
	EN	FR	DE	LB
FT (EN)	1.69 ± 0.57	1.49 ± 0.53	1.41 ± 0.12	0.21 ± 0.38
FT (FR)	2.27 ± 0.16	1.32 ± 0.34	**1.78** ± 0.44	**1.27** ± 0.98
FT (DE)	**2.49** ± 1.25	**2.53** ± 1.46	1.34 ± 1.07	0.82 ± 0.48

Table 5: Difference in nDCG@10 between text+visual and text-component-only fine-tuning across query languages. Values are reported as mean ± standard deviation over three runs; positive values indicate gains from updating the visual component. Best values are shown in **bold**.

component-only fine-tuning variant. Table 5 reports the paired difference in nDCG@10 between the two settings across three runs.

The mean differences are positive for every training and evaluation language combination, indicating that updating the visual component provides an additional benefit beyond text-component-only adaptation. The magnitude of this gain varies substantially across settings, ranging from 0.21 points for FT (EN) evaluated on Luxembourgish to 2.53 points for FT (DE) evaluated on French. The larger gains observed for some FT (DE) configurations are accompanied by higher run-to-run variability, while several EN and FR configurations show smaller but more stable improvements. Overall, these results suggest that visual adaptation provides complementary gains, although its contribution depends on both the fine-tuning language and the evaluation language.

6.3 Multilingual Fine-Tuning for Improved Cross-Lingual Retrieval

We further investigate whether multilingual fine-tuning improves cross-lingual retrieval beyond the single-language fine-tuning studied earlier. We compare the best single-language setting, FT (FR), with two multilingual variants: one fine-tuned on English, French, and German, and another fine-tuned on English, French, German, and Luxembourgish. We again use colnomic-embed-multimodal-3b as both the baseline model and the backbone used for fine-tuning.

As shown in Table 6, multilingual fine-tuning on English, French, and German underperforms FT (FR) across all query languages, with an average decrease of 1.03 nDCG points. The largest decrease occurs on Luxembourgish queries (-1.97). This degradation may partly reflect the smaller number of training examples available per language in the multilingual setting compared with single-language FT (FR), which can reduce language-specific adaptation. The absence of Luxembourgish supervision may further contribute to the larger drop observed on Luxembourgish queries.

Adding Luxembourgish substantially changes this pattern. The EN+FR+DE+LB setting reaches an average nDCG@10 of 80.94, essentially matching FT (FR) at 80.96. Relative to EN+FR+DE, adding Luxembourgish improves performance across all query languages, with the largest increase on Luxembourgish queries, from 77.48 to 80.02 ($+2.54$ points). Compared with FT (FR), the multilingual setting remains close on English, French, and German, with differences of -0.08, -0.46, and -0.10, respectively, while improving Luxembourgish retrieval by $+0.57$.

Since the corpus pages are in Luxembourgish, direct exposure to Luxembourgish queries may help align query representations with the corresponding document pages. However, the multilingual settings also differ from single-language fine-tuning in the amount and distribution of training data per language, so the effect cannot be attributed to Luxembourgish supervision alone. Overall, the results show that multilingual fine-tuning including Luxembourgish can recover the performance lost in the EN+FR+DE setting, match the strongest

Training setup	Query language				
	EN	FR	DE	LB	Avg.
Baseline	78.73	78.44	79.59	77.89	78.66
FT (FR)	**81.11 ± 0.57**	**81.50 ± 0.49**	**81.76 ± 0.40**	79.45 ± 0.75	**80.96 ± 0.55**
Multilingual FT					
EN+FR+DE	80.47 ± 0.39 (-0.63)	80.46 ± 0.16 (-1.04)	81.30 ± 0.26 (-0.46)	77.48 ± 0.48 (-1.97)	79.93 ± 0.12 (-1.03)
EN+FR+DE+LB	81.03 ± 0.51 (-0.08)	81.04 ± 0.44 (-0.46)	81.66 ± 0.44 (-0.10)	**80.02 ± 0.56** (+0.57)	80.94 ± 0.41 (-0.02)

Table 6: nDCG@10 across query languages for the baseline, the best single-language fine-tuning setting, and multilingual fine-tuning variants. Results are reported as mean ± standard deviation over three runs. Values in parentheses indicate changes relative to FT (FR); the best value is shown in **bold** and the second-best is underlined.

single-language configuration overall, and provide the strongest Luxembourgish retrieval performance. Given the variability across three runs, the small differences between FT (FR) and EN+FR+DE+LB should be interpreted as comparable overall performance rather than a clear advantage for either setting.

7 Conclusion

Luxembourgish provides a representative low-resource use case for studying cross-lingual retrieval, given the scarcity of dedicated resources and retrieval benchmarks for the language. In this work, we introduced *LëtzCross*, a cross-lingual benchmark for page-level retrieval over Luxembourgish PDF documents in a page-image RAG setting. The benchmark combines automatically generated text-focused QA pairs with manually authored visually grounded QA pairs. Our results show that late-interaction page-image retrievers outperform the evaluated OCR-based text-only baselines in a system-level comparison, and that fine-tuning improves retrieval with gains that transfer across query languages. Among the evaluated single-language settings, French fine-tuning yields the highest mean performance on Luxembourgish queries, while among the multilingual configurations, including Luxembourgish produces the strongest results and substantially improves Luxembourgish-query retrieval. Together, these findings provide a first benchmark and empirical study for cross-lingual page-level retrieval over Luxembourgish document pages.

Limitations

This work has several limitations. First, *LëtzCross* is relatively small, with a limited manually authored visually grounded subset compared with the automatically generated text-focused QA pairs. Consequently, the evaluation is dominated by text-focused queries, and results on the visually grounded subset should be considered complementary rather than representative of the benchmark. Second, the comparison between text-only and page-image retrievers is system-level: the text-only baselines rely on document parsing with OCR, layout analysis, and table extraction, while page-image retrievers operate on rendered pages. Alternative parsing or hybrid pipelines could therefore affect the observed gap. Third, all non-English queries are translated from English rather than independently authored, which may introduce translation artifacts or English-oriented formulations. Moreover, query translation and automatic QA generation rely on LLMs, although we apply validation and manual review to improve consistency and quality. Finally, our findings are specific to Luxembourgish and should be validated on other low-resource languages before broader generalization.

Acknowledgments

This research was funded in whole or in part by the Luxembourg National Research Fund (FNR), grant reference NCER22/IS/16570468/NCER-FT. We acknowledge Google.org for providing Gemini credits used in this work. We also thank the anonymous reviewers for their careful reading and constructive feedback.

References

Voyage AI. 2026. Voyage-4-nano. https://huggingface.co/voyageai/voyage-4-nano. State-of-the-art text embedding model with 32k token context length.

Laura Bernardy. 2022. A Luxembourgish GPT-2 Approach Based on Transfer Learning. Master's thesis, University of Trier.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2025. Colpali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations*, volume 2025, pages 61424–61449.

Daniela Gierschek. 2022. *Detection of Sentiment in Luxembourgish User Comments*. Ph.D. thesis, University of Luxembourg.

Peter Gilles, Nina Hosseini Kivanani, and Léopold Edem Ayité Hillah. 2023. Lux-asr: Building an asr system for the luxembourgish language. In *Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE.

Ping Guo, Yubing Ren, Yue Hu, Yanan Cao, Yunpeng Li, and Heyan Huang. 2024. Steering large language models for cross-lingual information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 585–596.

Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Sedigheh Eslami, Scott Martens, Bo Wang, Nan Wang, and Han Xiao. 2025. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval.

Chao-Wei Huang, Chen-An Li, Tsu-Yuan Hsu, Chen-Yu Hsu, and Yun-Nung Chen. 2024. Unsupervised multilingual dense retrieval via generative pseudo labeling. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 736–746.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 6769–6781.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Hynek Kydlíček, Guilherme Penedo, and Leandro von Werra. 2025. Finepdfs. https://huggingface.co/datasets/HuggingFaceFW/finepdfs.

António Loison, Quentin Macé, Antoine Edy, Victor Xing, Tom Balough, Gabriel Moreira, Bo Liu, Manuel Faysse, Céline Hudelot, and Gautier Viaud. 2026. Vidore v3: A comprehensive evaluation of retrieval augmented generation in complex real-world scenarios.

Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Lisa Veiber, Tegawende Bissyande, Jacques Klein, Andrey Boytsov, Clément Lefebvre, and Anne Goujon. 2022. LuxemBERT: Simple and Practical Data Augmentation in Language Model Pre-Training for Luxembourgish. In *Proceedings of LREC*.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024. Unifying multimodal retrieval via document screenshot embedding. *arXiv preprint arXiv:2406.11251*.

Nomic AI. 2025. Colnomic embed multimodal 3b. https://huggingface.co/nomic-ai/colnomic-embed-multimodal-3b. Hugging Face model card.

Fred Philippy, Shohreh Haddadan, and Siwen Guo. 2024. Forget NLI, Use a Dictionary: Zero-Shot Topic Classification for Low-Resource Languages with Application to Luxembourgish. In *Proceedings of SIGUL (LREC-COLING)*.

Alistair Plum, Caroline Döhmer, Emilia Milano, Anne-Marie Lutgen, and Christoph Purschke. 2024. LuxBank: The First Universal Dependency Treebank for Luxembourgish. In *Proceedings of TLT*.

Alistair Plum, Tharindu Ranasinghe, and Christoph Purschke. 2025. Text Generation Models for Luxembourgish with Limited Data: A Balanced Multilingual Strategy. In *Proceedings of VarDial (COLING)*.

Christoph Purschke. 2020. Attitudes toward multilingualism in luxembourg. a comparative analysis of online news comments and crowdsourced questionnaire data. *Frontiers in Artificial Intelligence*, 3.

Tharindu Ranasinghe, Alistair Plum, Christoph Purschke, and Marcos Zampieri. 2023. Publish or hold? automatic comment moderation in Luxembourgish news articles. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 968–978, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Joshgun Sirajzade, Daniela Gierschek, and Christoph Schommer. 2020. An Annotation Framework for Luxembourgish Sentiment Analysis. In *Proceedings of SLTU-CCURL (LREC)*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Improving OCR for a Latvian Pronunciation Dictionary

Viesturs Jūlijs Lasmanis
University of Latvia, Institute of Mathematics and Computer Science
Raiņa bulvāris 29, Riga, Latvia
viesturs.lasmanis@lumii.lv

Abstract

This paper presents evaluation and data augmentation methods for improving OCR of the Latvian Language Spelling and Pronunciation Dictionary (LVPPV), a specialised dictionary containing numerous non-standard diacritic symbols used to transcribe Latvian pronunciation. To reduce the amount of manual correction required for OCR output, two new Tesseract v5 models were trained: one using a pre-existing annotated dataset and another using the same dataset supplemented with artificially generated training data. To enable more task-oriented model evaluation, two metrics were defined based on agreement between extracted lexeme–pronunciation pairs and two Latvian pronunciation resources: the Modern Latvian Language Dictionary and a rules-based phonetic transcriber. Both newly trained models outperform the previously published OCR model on these evaluation metrics, increasing transcription match rate from 71.81% to 87.56%. While synthetic data augmentation yields only limited improvements in overall transcription match rate, it substantially improves recognition of placenames containing uppercase letters, resulting in an 81% increase in correctly extracted placename pronunciations compared to the non-augmented model.

1 Introduction

The Latvian Language Spelling and Pronunciation Dictionary – "Latviešu valodas pareizrakstības un pareizrunas vārdnīca", LVPPV (Ceplītis et al., 1995) is a published dictionary that contains approximately 80 000 lexemes with a detailed transcription of pronunciation data, including fully annotated intonations, which are not reflected in Latvian orthography (see Figure 1). This dictionary also has a high-quality digital scan that can be used for information extraction using an OCR model.

dvest, dvešu, dves, dveš, dvešam, dvešat, *pag.* dvesu, dvesām, dvesāt; *pav.* dves, dvesiet; dvesdams [dvęzdàms], dvestu [dvęstu]
dvie/lis [dviēlis], -ļa
dvinga [dviŋga]
dvingains [dviŋgaîns]
dvīņi [dvìņi], *vsk.* dvī/nis, -ņa; *s.* **-ne**, -ņu
dvīņkristāls [dvìņkristāls]
dvīņledājs [dvìņļedāįs]
dvīņogas [dvìņuôgas], *vsk.* dvīņoga
dvīņsēkle/nis [dvìņsēklenis], -ņa
dvīņser/de [dvìņserde], -žu

Figure 1: Excerpt from the Latvian Language Spelling and Pronunciation Dictionary (LVPPV).

Although most of the dictionary is written in a typed font, it uses a varied list of combining diacritic marks not present in standard Latvian orthography, and there exist some handwritten corrections to the said diacritic marks, which provide an additional complication to the training of an accurate OCR model. The full list of additional symbols is depicted in Figure 2 with non-observed but theoretically possible symbols depicted in gray.

The pronunciation transcription used in the dictionary is the standard Latvian phonetic alphabet "Latviešu valodas fonētiskais alfabēts" (LFA) used in Latvian linguistic research, and defined within LVPPV (Ceplītis et al., 1995). However, this format is ill-suited for digital use due to the extensive use of diacritics, where many possible character-diacritic combinations do not result in a single Unicode character. To address this issue, a Latvian version of the SAMPA notation was introduced (Auziņa, 2005), which has since been adapted and used in the LVMLA (Latvian language machine-

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 295–303
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_030

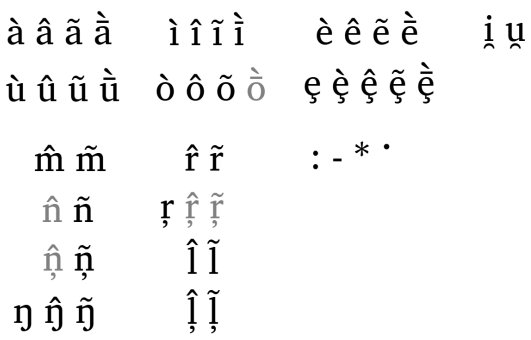


Figure 2: Non-standard characters used in LVPPV phonetic notation.

readable alphabet) notation. As shown in Table 1 all diacritic marks and symbols not present in the standard Latvian alphabet are replaced with symbols that can be easily typed and are not used elsewhere in the Latvian phonetic alphabet.

Lexeme (orthographic)	LFA	LVMLA	IPA
ezers	ęzęrs	EzErs	æzærs
sāls	sà̄ls	sāls	sɑːls
augs	aûks	auqks	ɑu̯ks
ola	uõla	uo=la	uo̯lă

Table 1: Comparison of LFA, LVMLA, and IPA notations

Therefore, in order to effectively use the pronunciation data provided by the LVPPV dictionary, it is important not only to have an accurate OCR model for transcribing the digital copy of the dictionary, but also to have a method to extract the lexemes and convert phonetic transcriptions from LFA notation into LVMLA notation.

2 Related work

The article "OCR Challenges for a Latvian Pronunciation Dictionary" by Strankale and Paikens (2020) introduces an approach to solve the OCR problem of this exact dictionary by utilising version 4 of Tesseract – an open-source model, which they argue is well suited for this specific task.

However, due to a high error-rate present within this optically recognised dataset, the dataset has yet to be meaningfully utilised in linguistic or speech synthesis research. While the precise evaluation set WER is not stated within the article, the graph of the 'verification set WER' indicates it being slightly above 10% (Strankale and Paikens, 2020), making it challenging and time-consuming to manually review. The authors identify diacritic recognition and character duplication as the primary sources of OCR errors.

Examining other articles related to diacritic recognition in non-standard language, Tesseract remains a reliable open-source choice. An article by Ilić et al. (2022) examines diacritic detection in Serbian dialectal dictionaries using Tesseract OCR, and an article by Do (2024) uses it for Vietnamese old text recognition.

Ilić et al. (2022) opt for a slightly different approach to diacritic detection in Serbian dialectal dictionaries – creating a pipeline that does text recognition and character bound box detection with Tesseract, but a separate CNN model that is responsible for diacritic detection. Although this approach is promising and has high accuracy for Serbian dictionary recognition, it is unclear how this pipeline-based approach compares in accuracy to the method of fine-tuning or retraining the top layers of the Tesseract model.

According to a study by Hegghammer (2022) for historical documents of English and Arabic Google DocumentAI and Amazon Textract models outperform fine-tuned Tesseract models. However, in the study by Do (2024) it was found that for dense book texts in Vietnamese, fine-tuned Tesseract OCR models outperform LLM based approaches such as GPT-4o, Azure DocAI and GoogleAI. The case of Vietnamese is also more relevant to the task at hand, because, while part of the standard orthography, the Vietnamese language utilises many diacritic combinations and is less resourced than English or Arabic texts. This does seem to suggest that fine-tuned Tesseract models are more suited for niche orthographies, such as the one used in LVPPV.

Existing synthetic data augmentation methods for OCR primarily focus on improving robustness across various different document collections by introducing variation in fonts, layouts, and imaging conditions (Jaderberg et al., 2014; Journet et al., 2017; Belval, 2019). Furthermore, many synthetic data generation strategies rely on post-OCR correction assisted by external vocabulary lists or language corpora, which relies on the existence of standard-language resources (Guan and Greene, 2024). In contrast, the OCR model developed in this study is optimized for the Latvian Lan-

guage Spelling and Pronunciation Dictionary, a single highly standardized resource with largely uniform typography and document structure. The core challenge therefore, is not the adaptation to various different documents, but instead the limited character combinations present within the available data. Therefore, the data augmentation strategy of image generation examined within this study primarily focuses on increasing the amount of under-represented character–diacritic combinations within the training dataset, rather than increasing robustness to changes in font or document style.

3 Methodology

Within the scope of this study, a new Tesseract v5 OCR model was trained by optimising the maximum iterations count. Two new models were trained using this principle – one using only the dataset used by Strankale and Paikens (2020), and another which was augmented with artificial data.

Models were trained using the Tesseract v5 training pipeline with the Latvian (lav) pretrained model as the starting model, network specification [Lfx256 O1c105], and a maximum of 300,000 training iterations. Training was performed on a workstation equipped with an NVIDIA GeForce RTX 4070 GPU (12 GB VRAM).

In order to compare the original Strankale and Paikens (2020) model results with the newly trained models, two different evaluation metrics based on available Latvian pronunciation resources were defined, implemented and utilised. The first compares phonetic transcriptions with those available in the MLVV dictionary (Zuicena et al., 2025), while the second performs a more limited comparison using a rules-based transcriber.

3.1 Datasets for model training

3.1.1 LVPPV dataset

The LVPPV dataset utilised in this article was created and manually annotated by Strankale and Paikens (2020). While the annotated dataset is not publicly available, the dataset and corresponding code-base was provided by the authors for further research. The full dataset consists of 42 box files, which includes 23 fully and 19 partially annotated pages.

The original annotation was created for training Tesseract version 4 model and thus contains box files denoting the pixel positions of individual characters within the scanned page images. However, since Tesseract version 5 the training method of the OCR model has changed to require line-level inputs as opposed to character-level inputs (Tesseract, 2025).

This necessitated the transformation of the existing dataset by merging all characters that appear in the same row to create a text file containing the annotation for the respective line. After format conversion, the final processed dataset consisted of 3520 lines of text from the LVPPV dictionary.

3.1.2 Artificially augmented dataset

Analysis of the character–diacritic combinations present in the prepared dataset revealed that some combinations have low prevalence in the annotated LVPPV dataset as seen in Table 2. It is also important to note, that the list of characters depicted in the article by Strankale and Paikens (2020) is not exhaustive and there are some rare characters, which may appear in a select few lexemes such as the character ‘ŋ̂’ as shown in Figure 3.

Symbol	Count	Symbol	Count
l̃	248	l̂	82
ļ̃	9	ļ̂	6
r̃	91	r̂	43
m̃	68	m̂	1
ñ	188	n̂	0
ņ̃	3	ņ̂	0
ŋ̃	15	ŋ̂	0

Table 2: Intonation marked consonant symbol frequencies in annotated LVPPV dataset.

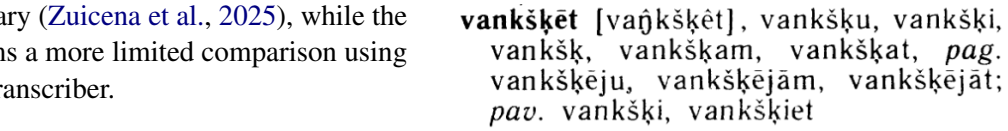


Figure 3: Lexeme containing character-diacritic combination not present in annotated dataset.

Augmenting the dataset with artificially generated line images was explored as a means of increasing the representation of rare character–diacritic combinations in the training data. In an initial experiment, the generated strings consisted of random sequences of Latvian orthographic symbols followed by randomly generated LFA symbol sequences enclosed in square brackets, ensuring that both common and rare symbols were represented in the training set.

However, because Tesseract v5 is based on an LSTM architecture (Tesseract, 2025), it is able to utilise contextual information during recognition. Consequently, training on completely arbitrary symbol sequences may limit the usefulness of this contextual information. Therefore, a string generation method based on Latvian lexical forms was developed in order to produce more linguistically plausible character sequences.

Sequences were generated using a list of Latvian words from Tēzaurs.lv online dictionary open data (Spektors et al., 2025) and transforming them by randomly appending intonation marking symbols in viable positions. These results were then structured as shown in Figure 4, which are theoretically possible sequences of LFA characters, although they do not correspond to the actual pronunciation of the respective word.



Figure 4: Artificial line image generated for data augmentation.

Furthermore, upon manually examining the output of Strankale and Paikens (2020) model, it was observed that the model made mistakes when trying to recognise the uppercase letters of the orthographic spelling of placenames in pages 878–942 of the LVPPV. This is most likely due to removing the top layers of the OCR model for retraining purposes and the absence of uppercase letters within the annotated Strankale and Paikens (2020) dataset. Therefore, while conducting the artificial augmentation of the dataset it was decided to re-introduce a limited amount of uppercase letters by randomly capitalising some of the orthographic words in order to potentially resolve this issue.

These sequences were then passed to TextRecognitionDataGenerator (*trdg*) python library to be generated into PNG images (Belval, 2019). Since the original version of trdg does not support many of the necessary combining diacritic symbols a slightly modified version was used. Additionally, to ensure that the combining diacritics were properly rendered, Charis version 7.000 font[1] was utilised, as it is designed for broad multi-lingual use, has great support for combining diacritic marks and is a serif font similarly to the font used within the LVPPV dictionary.

In order to test the potential of data augmentation, the aforementioned LVPPV dataset consisting of 3520 lines of annotated text was supplemented with 5000 artificially generated image files and their respective text files.

3.2 Evaluation metrics and data comparisons

Traditional OCR evaluation metrics such as Character Error Rate (CER) and Word Error Rate (WER) measure the fidelity of the reconstructed text. However, the intended downstream task of this work is extraction of lexeme–pronunciation pairs rather than reconstruction of dictionary pages. Character errors after the listed pronunciation may leave the pronunciation data usable, while an incorrect diacritic can invalidate the pronunciation despite having little impact on CER. Therefore, agreement between extracted lexeme–pronunciation pairs and external pronunciation resources provides a more task-relevant measure of OCR quality for the intended application.

3.2.1 Comparison with MLVV

Another pronunciation resource for the Latvian language is the MLVV dictionary, which is available digitally. However, this source is limited in the amount of lemmas with phonetic guides – 6117 as of Autumn 2025 version (Zuicena et al., 2025) – compared to the LVPPV dictionary. While Strankale and Paikens (2020) do mention the possibility of using the data from this dictionary for model evaluation in their article and the aforementioned code-base contains some code for realising this comparison, this is not explored within the published article.

The main advantage to this dataset is that the pronunciation information fully annotates intonation data and is stored within a structured XML file and can be easily accessed together with the corresponding lemma. This means that it is simple to extract a list of lexeme and phonetic transcription pairs, which can then be used to look for alignment with the OCR models predicted LVPPV text.

However, there are a few caveats to address regarding this approach. While the LFA phonetic notation is standardised in terms of character representation, the phonetic transcription used in MLVV is not fully phonetic as it does not represent consonant and consonant voicing assimilation, instead opting to preserve orthographic nuances. Table 3 illustrates some example words extracted from the

[1] https://software.sil.org/charis/release-7-000/

LVPPV dictionary by the OCR model and manually checked, do not match the MLVV form due to differences in the usage of the LFA notation.

Orthography	LVPPV	MLVV
apdzīt	abdzĩt	apdzĩt
jokdaris	juõk*daris	juõkdaris
plecs	pl̦ec:	pl̦ecs
spožs	spuôš:	spuôžs

Table 3: Comparison of phonetic transcription differences between LVPPV and MLVV dictionaries.

This approach of comparing extracted OCR results to MLVV lexemes creates a non-standard metric, which while cannot determine the overall accuracy of the model, can be utilised effectively in order to compare different models in their effectiveness at extracting reliable lexeme-pronunciation pairs. This information arguably may be more useful in determining which model is better for lexeme-pronunciation pair extraction as CER may not be fully indicative of the OCR models effectiveness for the intended task.

3.2.2 Comparison with rules-based transcription

The comparison with MLVV lexemes is very useful in determining the models' ability to correctly extract lexemes and their pronunciations with correctly determined intonation diacritic symbols. However, the MLVV set of lexemes with phonetic information is relatively small and cannot be used to examine a broader scope of lexemes.

In order to determine whether the trained models retain their accuracy on the entire LVPPV dictionary a comparison with rules-based generated phonetic transcriptions can be utilised. Latvian language pronunciation, except for a few nuances, is generally predictable from orthography and positional information, thus allowing rules-based transcriptions to be a viable comparison.

First, a list of Latvian language lexemes was extracted from the Tēzaurs.lv online dictionary dataset (Spektors et al., 2025) and then further transformed using a Latvian rules-based G2P transcriber (UL IMCS, 2025), which applies positional rules present in the Latvian language to transcribe the pronunciation. It should be noted that this transcriber is limited in the ability to assess correct LVPPV transcriptions, as this evaluation cannot determine the correct intonation, pronunciations of some vowels (/e/, /ē/, /o/) and if the word has non-standard stress.

By generating all possible vowel pronunciations and ignoring intonation and stress it is possible to compare whether the OCR recognised phonetic transcription aligns with the rules-based transcription on base character level. That is to say that, while these rules-based transcriptions cannot be used to evaluate models on their diacritic recognition accuracy, these transcriptions can therefore be used to evaluate the amount of other mistakes caused by such errors as – character doubling and other base character recognition mistakes.

This evaluation method can be of particular use in analysing the OCR model's ability for transcribing pages 878–942 of the LVPPV, which contain placenames with uppercase letters. For the OCR data extracted by the model of Strankale and Paikens (2020) it can be observed that in most cases the uppercase letters are skipped or predicted as completely different characters, with most correct predictions happening for uppercase letters that are similar in shape to their lowercase counterparts. These two types of errors are ones that can be detected using only the rules-based approach, as this doesn't involve diacritic detection.

4 OCR post-processing

In order to more accurately evaluate model performance, the LVPPV predictions and MLVV transcriptions were both post-processed normalising combining diacritic characters and applying Latvian consonant and consonant voicing assimilation rules where applicable. Diacritic character normalisation is absolutely necessary due to the combining diacritic characters potentially being recognised and applied in a different order compared to the MLVV or vice versa, resulting in the string being encoded in Unicode in differing ways, despite representing identical LFA symbols.

Since both LVPPV and MLVV dictionaries use the LFA transcription format with slight differences, the comparison was done in the LFA format. To ensure that the Strankale and Paikens (2020) model and the models trained within the scope of this article are comparable, each model's OCR output files without post-processing were used as the input data and then post-processing rules were applied to all files to ensure that all post-processing and rules for lexeme-transcription pair extraction were identical. Identical methods were used since

various post-processing methods can significantly increase the amount of lexemes and transcriptions that are correctly recognised, and since it is known that the same dataset was used in model training this means that there isn't a risk in some post-processing rules applying to only a certain model.

Finally, in addition to the post-processing steps necessary for model evaluation, it is necessary to transform the Latvian phonetic alphabet (LFA) notation into LVMLA notation. This format is not only the desired storage format for the extracted data, but is also the format that is used in the rules-based phonetic transcriber. Therefore, a format converter is necessary in order to apply the rules-based metric described in Subsection 3.2.2.

For the purposes of transforming between these formats a Python program was written, which can in almost all cases correctly convert between the formats using simple rule-based transformations of the LFA transcription.

However, the inspection of the dictionary output revealed that a small number of entries require orthographic or morphological information for correct LVMLA conversion, particularly in compound words. Figure 5 shows one such example, where the expected vowel length cannot be determined from the LFA transcription alone.

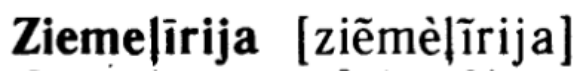


Figure 5: Lexeme where orthographic information is necessary to determine the correct LVMLA conversion.

5 Results and Evaluation

After the training process of the Tesseract v5 model, it is possible to run the in-built model evaluation function and get comparative CER and WER data (Tesseract, 2025). It is possible to observe from the plotted data (Figure 6) that the training and evaluation losses of this model at model checkpoints does not differ greatly from the model trained by Strankale and Paikens (2020) with the CER remaining at around 2%, while the WER at around 10%.

However, the models differ more substantially when evaluated using the task-specific metrics described in Subsection 3.2. Compared with the model of Strankale and Paikens (2020), the newly trained Tesseract v5 model increases the MLVV transcription match rate from 71.81% to 87.56% (Table 4) and the rules-based transcription match

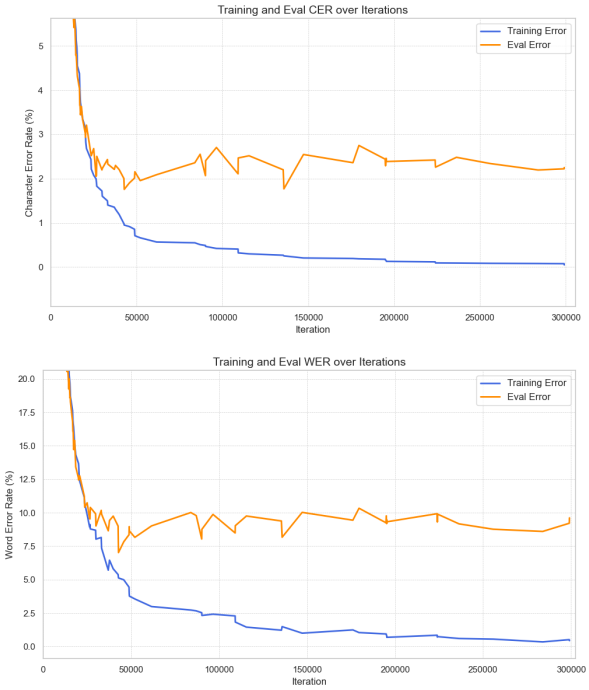


Figure 6: Tesseract v5 model checkpoint CER and WER for training and evaluation datasets depending on iteration count.

rate from 84.53% to 89.29% (Table 5).

Although the MLVV and rules-based evaluation metrics rely on different sources and therefore may have different potential causes of transcription mismatch, the most significant difference of the two methods is the fact that the MLVV metric includes diacritic information, while the rules-based method does not. The substantially larger improvement observed for the MLVV comparison (15.76 percentage points; Table 4) than for the rules-based comparison (4.76 percentage points; Table 5) therefore suggests that improved diacritic recognition contributes to the performance gain of the newly trained Tesseract v5 model.

Model	Lexeme & MLVV transcription match	Lexeme match	Transcription match rate
Strankale and Paikens (2020)	3571	4973	71.81%
Tesseract v5	**4413**	5040	**87.56**%
Tesseract v5 Artificially augmented	4356	**5045**	86.34%

Table 4: Comparison of MLVV lexeme transcription matches of trained OCR models.

The match rate metrics depicted in Tables 4, 5, 6 should not be interpreted as a standard OCR accuracy, since disagreements may arise from the LFA notation differences between LVPPV and MLVV or from other source data limitations rather than

Model	Lexeme & rule-based transcription match	Lexeme match	Transcription match rate
Strankale and Paikens (2020)	29815	35269	84.53%
Tesseract v5	31695	35498	**89.29**%
Tesseract v5 Artificially augmented	**31783**	**35656**	89.13%

Table 5: Comparison of rules-based transcription matches of trained OCR models.

recognition errors. Furthermore, for all of the models the rules-based extracted lexeme count depicted in Table 5 is significantly lower than that claim of 80 000 lexemes contained in the LVPPV (Ceplītis et al., 1995). However, it should be noted that many of the lexemes present in the LVPPV dictionary do not have a corresponding pronunciation guide, this being the case in situations when the pronunciation transcription in LFA is identical to the standard Latvian orthography.

For the most part the data augmented model achieves similar results to the base Tesseract v5 model. It achieves slightly lower agreement with the MLVV transcriptions (Table 4), while yielding a slightly larger number of matches according to the rules-based comparison (Table 5). Given the small magnitude of these differences and the fact that each model configuration was trained once, these results should be interpreted cautiously, as run-to-run variation was not evaluated. Therefore, the existing results do not provide evidence that synthetic augmentation improves overall transcription matching.

In contrast, the most notable improvement from artificially augmenting the dataset can be observed for placenames containing uppercase letters (Table 6). The number of matched placename transcriptions increased from 615 for the base Tesseract v5 model to 1,113 for the augmented model, with the augmented model also achieving a match rate closer to that observed in the main body of LVPPV. This substantial improvement suggests that the lack of uppercase characters in the original training data was a contributing factor to recognition errors in this section of the dictionary.

Model	Lexeme & rule-based transcription match	Lexeme match	Transcription match rate
Strankale and Paikens (2020)	412	524	78.63%
Tesseract v5	615	782	78.64%
Tesseract v5 Artificially augmented	**1113**	**1294**	**86.01**%

Table 6: Matches with rules-based transcriptions of placenames (LVPPV pages 878–942).

Overall, these results suggest that synthetic data is most effective for addressing coverage deficiencies in the training set rather than uniformly improving recognition accuracy. While the augmentation substantially improved recognition of placenames containing uppercase letters (Table 6), it produced only a small increase in the absolute number of correctly matched transcriptions in the main body of the dictionary (Tables 4, 5), while the overall transcription match rate remained nearly unchanged. These findings suggest that synthetic augmentation is most beneficial when it increases coverage of under-represented patterns, but does not necessarily improve precision for patterns already represented in the training data.

5.1 Error Analysis

Manual inspection of the OCR output indicates that the most prevalent OCR errors reported by Strankale and Paikens (2020), namely character duplication and symbol confusion, remain present in the Tesseract v5 models. Table 7 illustrates examples of these and other recognition errors observed in the model output. The improved agreement with the rules-based transcriptions (Table 5) suggests a reduction in base-character-level transcription errors, a category that includes character duplication and base character recognition errors. As these individual error categories were not quantified separately, their relative frequency cannot be determined from the rules-based metric alone. Nevertheless, manual examination of the OCR output suggests that character duplication remains a recurring error, particularly for characters with non-standard diacritics marking intonation.

6 Conclusion

The precise accuracy of the models remains difficult to ascertain due to the target material being unique in character usage, and due to the annotated data being limited in scope. However, the alternate metrics defined and applied within this article can be successfully utilised to compare model performances. These metrics are important to successfully train new models which are capable of more accurately transcribing the Latvian Language Spelling and Pronunciation Dictionary, thereby reducing the amount of OCR output requiring manual correction.

The Tesseract v5 models trained within the scope of this article achieve better results under both metrics, suggesting improvements in both base-

Error type	Ground truth	OCR output
Character duplication	dzirks/ts [dzìrksc]	dzirks/ts [dziìrksc]
Diacritic recognition	vankšķēt [vaŋkšķêt]	vankšķēt [vaŋkšķêt]
Base character recognition	aplūkojums [aplũkuõjùms]	aplūkojums [apìũkuõjùms]
Uppercase recognition	Adeleida [adele-ida]	mndeleida [adele-ida]
Bracket recognition	suverēns [suverẽns]	suverēns [iverēns

Table 7: Examples of the OCR error types observed in the non-augmented Tesseract v5 model output. Entries are shown in LFA notation and truncated to the segments relevant for lexeme–pronunciation pair extraction.

character-level and diacritic-sensitive recognition.

Despite the implementation of data augmentation in order to improve diacritic detection accuracy, the effects on overall model match rate from training Tesseract with the augmented dataset are minimal. However, the implemented data augmentation method does successfully improve the recognition of placenames containing uppercase letters within the dictionary, resulting in an 81% increase in correctly extracted placename pronunciations compared to the non-augmented model. Compared to the original OCR model of Strankale and Paikens (2020), the final augmented model more than doubles the number of correctly extracted placename pronunciations, highlighting the practical benefit of the improvements introduced in this work.

Although the methods developed in this study target LVPPV, the proposed task-oriented evaluation approach may also be applicable to other historical or linguistic resources where complete ground-truth transcriptions are unavailable but related lexical resources can provide partial external validation. Similarly, targeted synthetic augmentation may be useful for resources in which recognition difficulties arise from sparsely represented symbols or notation rather than document-level visual variation.

Further work should explore more effective data augmentation strategies and investigate whether varying the proportion of synthetic data in the training dataset yields better results. The robustness of the observed effects of data augmentation should also be evaluated across multiple training runs with different random seeds. Beyond model evaluation, future work should examine whether improvements in transcription matching translate into measurable reductions in manual post-editing effort.

References

Ilze Auziņa. 2005. *Latviešu valodas izrunas datormodelēšana*. Ph.D. thesis, University of Latvia.

Edouard Belval. 2019. Textrecognitiondatagenerator. https://github.com/Belval/TextRecognitionDataGenerator.

L. Ceplītis, A. Miķelsone, T. Porīte, and S. Raģe. 1995. *Latviešu valodas pareizrakstības un pareizrunas vārdnīca*. Avots.

Thao Do. 2024. Reference-based post-ocr processing with llm for diacritic languages. *arXiv e-prints*, pages arXiv–2410.

Shuhao Guan and Derek Greene. 2024. Advancing post-OCR correction: A comparative study of synthetic data. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6036–6047, Bangkok, Thailand. Association for Computational Linguistics.

Thomas Hegghammer. 2022. Ocr with tesseract, amazon textract, and google document ai: a benchmarking experiment. *Journal of Computational Social Science*, 5(1):861–882.

Velibor Ilić, Lenka Bajčetić, Snežana Petrović, and Ana Španović. 2022. Scydia–ocr for serbian cyrillic with diacritics. In *Dictionaries and Society. Proceedings of the XX EURALEX International Congress, 12-16 July 2022, Mannheim, Germany*, pages 387–400. Mannheim: IDS-Verlag.

Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*.

Nicholas Journet, Muriel Visani, Boris Mansencal, Kieu Van-Cuong, and Antoine Billy. 2017. Doccreator: A new software for creating synthetic ground-truthed document images. *Journal of Imaging*, 3:62.

Andrejs Spektors, Lauma Pretkalniņa, Normunds Grūzītis, Pēteris Paikens, Laura Rituma, Baiba Saulīte, Gunta Nešpore-Bērzkalne, Ilze Lokmane, Agute Klints, Madara Stāde, Mikus Grasmanis, Ilze Auziņa, Artūrs Znotiņš, Roberts Darģis, and Guntis Bārzdiņš. 2025. Tēzaurs.lv 2025 (autumn edition) (2025-09-22). CLARIN-LV digital library at IMCS, University of Latvia.

Laine Strankale and Pēteris Paikens. 2020. Ocr challenges for a latvian pronunciation dictionary. In *Human Language Technologies–The Baltic Perspective*, pages 199–206. IOS Press.

Tesseract. 2025. *Tesseract User Manual*. Available at https://tesseract-ocr.github.io/tessdoc/.

UL IMCS. 2025. Parruna latvian g2p transcriber. Online pronunciation transcription tool for Latvian.

Ieva Zuicena, Ieva Auziņa, Santa Briede, Irēna Ilga Jansone, Ieva Kuplā, Gunta Lejniece, Ilga Migla, Laimdota Oldere, Ārija Ozola, Vija Požarnova, Sanda Rapa, Anitra Roze, Imants Šmidebergs, Dorisa Šnē, Māra Šnē, Agris Timuška, Mikus Grasmanis, Lauma Pretkalniņa, and Artūrs Znotiņš. 2025. Dictionary of contemporary latvian language (MLVV) (2025-06-21). CLARIN-LV digital library at IMCS, University of Latvia.

LLM-based conversational systems for Sinhala: Investigating model limitations and exploring a framework for agentic retrieval

Malithi Alahapperuma
Department of Information Studies
University College London
malithi.alahapperuma.20@ucl.ac.uk

Andreas Vlachidis
Department of Information Studies
University College London
a.vlachidis@ucl.ac.uk

Antonis Bikakis
Department of Information Studies
University College London
a.bikakis@ucl.ac.uk

Abstract

Large Language Models (LLMs) have introduced significant shifts to the way we communicate and seek information. However, LLMs perform suboptimally in low-resource languages. Given their unprecedented capabilities, LLMs provide a unique opportunity for under-served linguistic communities to access knowledge and skills that were hitherto inaccessible. Therefore, finding practical and realistic ways for low-resource language users to access LLM-based conversational systems can have far-reaching societal and economic benefits. In this study, we characterise three limitations that reduce the usability of LLM-based conversational systems for Sinhala speakers and conceptually explore the design of an LLM agent-based retrieval framework.

1 Introduction

The introduction of Large Language Models (LLMs) for wider public use has created a paradigm shift in the way we communicate and seek information. However, LLMs perform poorly in low-resource languages (Kshetri, 2024; Zhong et al., 2024). A language is considered low-resource when it lacks data, artefacts, resources, and institutional support (Nigatu et al., 2024). Due to their immense language processing capacity, LLMs provide a hitherto unavailable opportunity to support low-resource languages. Providing equitable access to language technology is an opportunity to correct the historical inequalities faced by under-represented linguistic communities across the globe (Smart et al., 2024). Against this backdrop, this study aims to bring together advances in LLM workflows to improve LLM-based conversational systems for Sinhala. We specifically focus on workflows that can be practically deployed to non-expert users in low-resource contexts. The work discussed in this paper is exploratory and is intended to provide a conceptual foundation upon which to develop further system specifications.

2 Setting the landscape

There are several threads of work in progress to improve the performance of LLMs for low-resource languages. Two key approaches are model fine-tuning and developing LLM-based workflows. In fine-tuning, LLMs are further trained on language data from a given low-resource language. In the language context of this study, Sinhala, we come across SinLlama, an LLM pretrained and fine-tuned for three tasks, writing-style classification, news categorisation and sentiment analysis (Aravinda et al., 2025). The authors of the study show how this targeted fine-tuning significantly improves the performance of the LLM for the selected tasks (Aravinda et al., 2025). Similar approaches to language-specific adaptations of language models are seen across other low-resource language contexts (Joshi et al., 2024; Tonja et al., 2024).

However, improving underlying model performance does not entirely solve the issues encountered in low-resource contexts. In order to address these problems, LLM-based workflows with additional features and functionalities have been explored. Retrieval Augmented Generation (RAG) was conceptualised to provide non-parametric memory (external data sources) to LLMs, with the goal of improving the retrieval performance of LLM-based systems (Lewis et al., 2020). RAG has been used in low-resource contexts to improve LLM-based systems. Ndimbo et al. (2025) have built a RAG-based conversation system for Swahili, and Christian et al. (2025) have built an adaptive

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 304–308
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_031

RAG system for Indonesian. Alshammary et al. (2025) present a RAG framework for Arabic named ARCADIA (Arabic Retrieval Contextual Analysis Data-Intensive Applications). Alshammary et al. (2025) note that capturing complex nuances and the rich morphology of Arabic is a particular challenge in working with Arabic, which we particularly note as relevant to our study.

Given the focus of our work on developing LLM-based conversational systems that can be practically deployed in the Sinhala-speaking context, the feasibility of model fine-tuning as a solution is limited. This is particularly due to constraints in data and financial and technical resources. Given the scalability, adaptability, and reusability of modular LLM-based workflows, we believe a system that brings together RAG and agentic capabilities is best suited for further exploration.

3 Investigating model behaviour

Prior to conceptualising the system, we use a set of experiments to investigate the specifics of the challenges encountered when using LLM-based conversational systems in Sinhala. The experiments focus on three tasks: (1) text generation, (2) translation, and (3) information retrieval. These tasks are specifically selected given their relevance for practical use by non-expert users. For similar reasons, the queries used in the experiments are framed to represent the kind of questions a non-expert user would ask an LLM during a daily interaction. For each query, we generate responses using three widely available LLM conversational tools: ChatGPT (based on GPT 5.5), Claude (based on Sonnet 5 Medium) and DeepSeek (based on DeepSeek-V4). ChatGPT and Claude were selected because they are commonly used tools. DeepSeek was included to examine whether an open-weights multilingual model would provide different insights. Close reading of the generated texts is used to draw initial observations and trends. We do not use any computational analysis at this point, but rather aim to uncover broad themes in the limitations observed. The findings of the experiments are discussed below under each task category. Altogether, we used five queries under each task, with three responses generated per query from the different models. In order to best utilise the space available, we only provide a few salient examples in the discussion. All LLM output, prompts used, and further discussion are available in the project GitHub repository, available at: https://github.com/malithia/diagnostic-bench-llms-sinhala/.

Experiment 01: Text generation

In this task, we asked the models to generate texts for various scenarios. We select two examples to compare here: (1) A cover letter for a job application for the role of a driver, and (2) a Facebook post to advertise a clothing business. In evaluating the responses, we note that for the cover letter, which is a formal piece of writing, the LLMs performed relatively well. The tone of the text was appropriate and the content of the letter was accurate and suitable. There were no glaring errors in grammar and syntax, but, with some variability across models, the choice of words was sometimes unsuitable. However, the output of the Facebook post was noticeably poor across all three models. The responses contained grammatical and syntactic errors. The most prominent drawback was the tone, which was a mix of formal and informal and did not sound like human speech. On a more positive note, the actual content of the Facebook post was still relevant to the task. This trend was noticed across all queries [1].

Experiment 02: Translation

In this section, we particularly reflect on two samples of text translated from English to Sinhala: (1) a newspaper notice for a road closure, and (2) an opinion piece from a newspaper about addressing psychosocial risks in the workplace. English newspaper articles from a Sri Lankan newspaper were specifically selected to ensure that the text is set in the Sri Lankan context. This is to control for any errors that may creep in from mere context mismatch. The trends in the Sinhala output of this experiment closely followed those of the text generation task. The road closure notice, which took a formal tone and contained direct factual statements, was appropriately translated. The grammar, syntax, and tone were suitable for the text. The opinion piece, which contained complex syntax, metaphorical speech, and conversational tone, was translated into Sinhala in a stilted and formulaic tone. The limitations persisted as a trend across all queries and models used [2].

[1] Please see 'text-gen/text-gen-1.txt' in the GitHub repository for specifics on the Facebook post and 'text-gen/text-gen-2.txt' for specifics on the cover letter. All queries used for testing text generation are available in the 'text-gen' folder in the GitHub repository.

[2] Please see 'translation/translation-1.txt' in the GitHub repository for specifics on the opinion piece, and 'translation/translation-2.txt' for specifics on the road closure

Experiment 03: Information retrieval

For this experiment, we prompt the models to find information on topics that are relevant in the Sri Lankan context. All models were enabled to conduct a web search. We use two specific examples here for the discussion: (1) how to apply for a national identity card from a specific city, and (2) whether taxes must be paid to import a business sample to Sri Lanka. Across the models, responses were comprehensible, despite some errors in grammar and syntax. In the first query about applying for a national identity card, all model output contained hallucinated steps in the process. In the second query, the answer to the specific question was accurate (leaving aside all other contextual information that LLMs often provide), when compared with the information provided by the Sri Lanka Customs website as the source of truth. The information on the Sri Lanka Customs website is clearly provided in a Frequently Asked Questions page, under the question "Can I import a sample for my business without taxes?". In this task category, we notice a two-pronged challenge: (1) LLMs provide hallucinated information as facts; (2) the information space itself can be limited and poorly organised, preventing not only LLMs but even humans from finding the necessary information [3].

In the next section, we synthesise the findings of the above experiments to formulate the research question.

4 Problem statement and research question

Based on our experiments, we identify the following three key limitations in the practical usage of LLM-based conversational tools in Sinhala. Note that these are generalised from the output we have obtained from our experiments. Further discussion regarding this generalisation is provided in section 6.

Limitation 01: LLM conversational tools produce acceptable responses in formal Sinhala, but not in informal Sinhala.

Limitation 02: LLM conversational tools perform poorly when they do not have access to data from the information space relevant to Sinhala speakers.

Limitation 03: LLM conversational tools make errors in grammar, syntax, spelling, and writing.

Based on these limitations observed, improvements are needed in the linguistic capabilities of conversational systems and the information space that the systems can access.

As such, the research question we aim to answer is:

How can advancements in LLMs and RAG workflows be brought together to improve the practical value and usability of LLM-based conversational systems for Sinhala speakers?

In section 5, we discuss how we envision an agentic RAG system to address the research question and the limitations we identified.

5 Exploring agentic retrieval as a solution

In order to optimise LLM-based conversational systems for low-resource languages, we propose an agentic RAG framework that holistically brings together LLMs, RAG, and natural language processing.

The agentic RAG system contains three key components, as detailed below:

5.1 User interface

This is the interface where the user inputs their query in their chosen natural language. The interface will closely resemble an LLM-based chat interface, because this reduces the cognitive load on the user in using the system.

5.2 LLM agent

This is an LLM agent that receives the user query and decides the best possible mechanism to formulate a response. Depending on the query, the agent may respond with trained (parametric) knowledge, or may resort to utilising additional tools from the LRL-Toolkit outlined in section 5.3.

5.3 LRL-Toolkit

The LRL-Toolkit (Low-Resource Language Toolkit) is a set of natural language processing and LLM-based tools. At this stage of the study, we specifically envision three tools that address the three limitations that we identified through the experiments presented in section 3. The LRL-Toolkit includes a **retrieval tool** with access to external knowledge, addressing

notice. All queries used for testing translation are available in the 'translation' folder in the GitHub repository.

[3] Please see 'retrieval/retrieval-2.txt' in the GitHub repository for specifics on the identity card-related query and 'retrieval/retrieval-4.txt' for specifics on the taxes-related query. All queries used for testing information retrieval are available in the 'retrieval' folder in the GitHub repository.

the limitation of sparse parametric knowledge in LLMs for low-resource languages. We have considered a variety of underlying information architectures to support this tool, including semantic databases, knowledge graphs, and ontologies. We are particularly keen to investigate incorporating knowledge graphs and ontologies, as their applicability in low-resource contexts is discussed in detail in the literature (Kaffee et al., 2023; Vats et al., 2023; Lignos et al., 2022). We believe that building domain-specific repositories of knowledge (for example, government processes or legal or medical information) would provide a way to scale the system meaningfully. The **context tool** provides additional linguistic and cultural nuances to the LLM agent to decide how best to formulate the final output. For instance, diglossia is a feature of Sinhala, where a high variety of the language is used for formal and written Sinhala, while a low variety of the language is used in informal and conversational settings. As we noted in Limitation 01, LLMs tend to misrepresent the tone of speech when responding to diverse conversational queries. The purpose of the context tool is to ensure that such contextual information is available to the agent when drafting the response. The **orthography tool** supports the agent in drafting responses that align with the grammatical rules and conventions of writing in the language. We envision using rule-based natural language processing approaches, as we aim to incorporate more deterministic linguistic knowledge into probabilistic large language models.

Figure 1 is a diagrammatic representation of the conceptualised framework.

A key differentiator of this system, compared to conventional agentic systems used with higher resourced languages, is the context and orthography tools, which aim to ground the output in the cultural and linguistic nuances of the language.

5.4 System evaluation

We aim to evaluate the final implementation of the system with human evaluators, using question-answer pairs that can particularly test the three limitations we identified in section 4. Using human evaluators enables us to test for limitations related to diglossia, which is where conventional models have been observed to fail. We are cautious of using emerging approaches such as LLMs-as-a-judge, because evaluating subtle nuances in the tone of speech is not a task that we believe an LLM can reliably carry out in Sinhala.

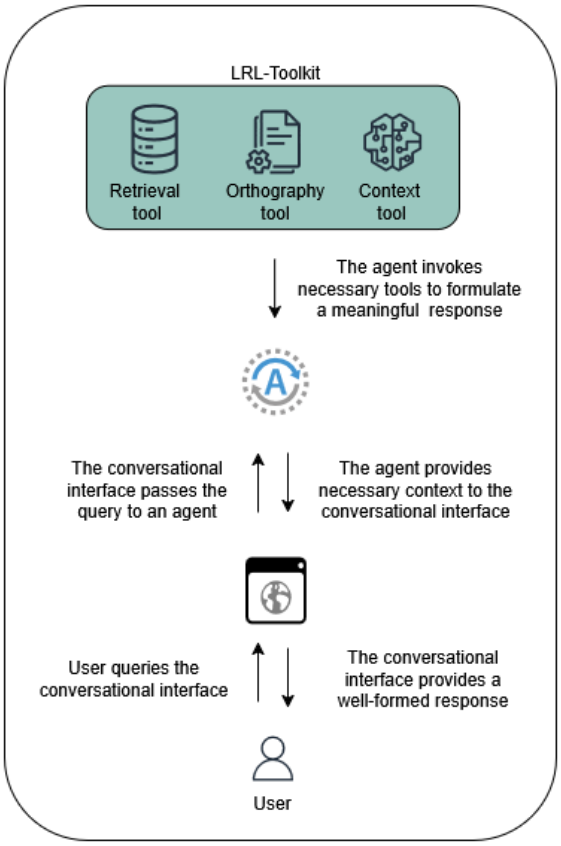


Figure 1: Diagrammatic representation of the proposed agentic RAG system

In our evaluations, we use the same question-answer pairs with a select sample of widely available LLM conversational tools and the implementation of the proposed system, where performance on widely available tools forms the baseline. The evaluation rubrics include factual accuracy, comprehensiveness, and precision to measure the quality of information retrieval, and grammar, syntax, lexical choice, and orthography to measure linguistic performance.

6 Limitations and mitigation strategies

In this section, we discuss two limitations of our work, and lay out the planned way forward.

In the experimental setup of section 3, we have used a select set of examples from three widely available LLM-based conversational tools to derive our findings. These limitations broadly align with challenges noted in the literature and the practical and anecdotal evidence provided by Sinhala speakers. However, we have not conducted a sys-

tematic evaluation with multiple annotators. While we believe our findings at this stage are sufficient to conceptualise the limitations and the retrieval framework, as the study progresses, we aim to do wider benchmarking to further refine the problem.

As this study and the system are in initial exploratory stages, the proposed system conceptualisation is subject to change to adapt to the evolving needs of the work, as well as to integrate the latest developments in LLMs. As a result, we may need to modify the proposed components as we develop and evaluate our work. Presenting this exploratory work for early feedback helps us shape our future work, and makes us aware of any potential drawbacks that we may have overlooked in our conceptualisation of system design.

References

Mitha Alshammary, Md Nahiyan Uddin, and Latifur Khan. 2025. Arcadia: A scalable arabic rag framework with hierarchical and context-aware retrieval. In *2025 IEEE 8th Congress on Information Science and Technology (CiSt)*, pages 238–243.

H.W.K. Aravinda, Rashad Sirajudeen, Samith Karunathilake, Nisansa de Silva, Rishemjit Kaur, and Surangika Ranathunga. 2025. Sinllama - a large language model for sinhala. In *2025 Moratuwa Engineering Research Conference (MERCon)*, page 617–622. IEEE.

William Christian, Daniel Adamlu, Adrian Yu, and Derwin Suhartono. 2025. Bridging language gaps with adaptive rag: Improving indonesian language question answering.

Shreyas Joshi, Muhammad Shahnawaz Khan, Aditya Dafe, Kavita Singh, Vedant Zope, and Tanish Jhamtani. 2024. Fine tuning llms for low resource languages. In *2024 5th International Conference on Image Processing and Capsule Networks (ICIPCN)*, pages 511–519.

Lucie-Aimée Kaffee, Russa Biswas, C. Maria Keet, Edlira Kalemi Vakaj, and Gerard de Melo. 2023. Multilingual Knowledge Graphs and Low-Resource Languages: A Review. *Transactions on Graph Data and Knowledge*, 1(1):10:1–10:19.

Nir Kshetri. 2024. Linguistic challenges in generative artificial intelligence: Implications for low-resource languages in the developing world. *Journal of Global Information Technology Management*, 27(2):95–99.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *ArXiv*, abs/2005.11401.

Constantine Lignos, Nolan Holley, Chester Palen-Michel, and Jonne Sälevä. 2022. Toward more meaningful resources for lower-resourced languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 523–532, Dublin, Ireland. Association for Computational Linguistics.

Edmund V. Ndimbo, Qin Luo, Gimo C. Fernando, Xu Yang, and Bang Wang. 2025. Leveraging retrieval-augmented generation for swahili language conversation systems. *Applied Sciences*, 15(2).

Hellina Hailu Nigatu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury. 2024. The Zeno's paradox of 'low-resource' languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17753–17774, Miami, Florida, USA. Association for Computational Linguistics.

Andrew Smart, Ben Hutchinson, Lameck Mbangula Amugongo, Suzanne Dikker, Alex Zito, Amber Ebinama, Zara Wudiri, Ding Wang, Erin van Liemt, João Sedoc, Seyi Olojo, Stanley Uwakwe, Edem Wornyo, Sonja Schmer-Galunder, and Jamila Smith-Loud. 2024. Socially responsible data for large multilingual language models.

Atnafu Lambebo Tonja, Bonaventure F. P. Dossou, Jessica Ojo, Jenalea Rajab, Fadel Thior, Eric Peter Wairagala, Anuoluwapo Aremu, Pelonomi Moiloa, Jade Abbott, Vukosi Marivate, and Benjamin Rosman. 2024. Inkubalm: A small language model for low-resource african languages.

Preeti Vats, Nonita Sharma, and Deepak Kumar Sharma. 2023. Hkg: A novel approach for low resource indic languages to automatic knowledge graph construction. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

Tianyang Zhong, Zhenyuan Yang, Zheng Liu, Ruidong Zhang, Yiheng Liu, Haiyang Sun, Yi Pan, Yiwei Li, Yifan Zhou, Hanqi Jiang, Junhao Chen, and Tianming Liu. 2024. Opportunities and challenges of large language models for low-resource languages in humanities research. *ArXiv*, abs/2412.04497.

Clinical Entity Recognition from Electronic Health Records and Linking to Biomedical Knowledge Bases in Low-resource Language Settings

Eno-Martin Lotman
Tallinn University of Technology and
North Estonia Medical Centre
Tallinn, Estonia

David Hübner and **Kris Collins**
Averbis GmbH
Freiburg, Germany

Svetla Boytcheva
Sofia, Bulgaria

Ivelina Nikolova-Koleva
Ontotext and
Bulgarian Academy of Sciences
Sofia, Bulgaria

Abstract

Clinical Natural Language Processing (NLP) in low-resource languages is severely constrained by strict privacy regulations and a scarcity of localized biomedical ontologies. In Estonian, for instance, the localized SNOMED CT translation covers less than 10% of the international target concept space. In this paper, we present a unified framework for clinical Named Entity Recognition (NER) and Entity Linking (EL) in low-resource settings, evaluated on real-world clinical narratives in Estonian, German, and Dutch. For NER, we deploy a hybrid architecture combining fine-tuned multi-task Transformers, a few-shot generative LLM (Qwen2.5-14B), and dictionary-based lookups. For EL, we implement a two-stage pipeline using deterministic gazetteers alongside SapBERT encoders cross-lingually augmented via SNOMED CT terms and multi-ontology mappings. On clinical datasets, the NER pipeline achieved F1-scores of 0.87 (Dutch), 0.84 (Estonian), and 0.67 (German). However, direct cross-lingual performance comparisons are confounded by substantial variations in corpus size. Compact SapBERT models deliver superior EL precision (0.76) compared to larger multilingual baselines. This work aims to inform future NLP projects by identifying both limitations and promising pathways for advancing low-resource language technologies in biomedical informatics and beyond.

1 Introduction

In natural language processing, a low-resource language is defined not by the size of its speaker base, but by the scarcity of the linguistic infrastructure required for computational work, including annotated corpora, digital lexicons, domain-specific datasets, and pre-trained models. Estonian is considered a low-resource language due to the limited available digital datasets and resources compared to more widely spoken languages. Clinical documentation in low-resource languages faces unique challenges stemming from both technological and ethical constraints. Privacy regulations and strict data protection laws severely limit the availability and public sharing of authentic clinical language samples, making it more difficult to develop sophisticated NLP tools for these languages. Regulations typically mandate clinical documentation in local languages, reinforcing the need for robust local-language NLP capabilities for improved patient care and safety, but also for maintaining linguistic diversity and precise domain-specific vocabulary into the future. Named Entity Recognition (NER) and Entity Linking (EL)—or entity normalization—are essential for interpreting and utilizing unstructured electronic health records (EHR). By identifying medical terms and mapping them to standardized ontologies, these tasks ensure high data accuracy, semantic interoperability, and effective secondary reuse.

2 Related Work

Clinical NLP in Low-Resource Settings While clinical NLP has achieved remarkable performance in high-resource languages like English, adaptation to lower-resource or morphologically complex languages remains a critical bottleneck. Early clinical systems relied heavily on rule-based patterns and localized dictionaries, which fail to capture natural syntactic variations. The advent of transformer-based models domain-adapted for medicine, such as SapBERT (Liu et al., 2021), revolutionized concept extraction and entity linking (Gallego et al., 2024). However, because these models are predominantly pre-trained on English corpora (like MIMIC-IV or PubMed), direct zero-shot transfer to low-resource languages typically results in steep performance degradation. Consequently, recent

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 309–317
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_032

efforts have shifted toward cross-lingual transfer, machine translation of localized ontologies, and parameter-efficient fine-tuning on small, manually annotated clinical datasets (Vassileva et al., 2024).

Estonian General NLP and NER As a Finno-Ugric language characterized by rich morphology, extensive compounding, and free word order, Estonian poses distinct challenges for tokenization and syntactic parsing. The computational foundation for the language was significantly advanced by the introduction of `EstNLTK` (Orasmaa et al., 2016), a Python-based library tailored to handle Estonian-specific morphological complexities.

The Named Entity Recognition in Estonian began with the work of Tkachenko et al. (2013), who developed the first Estonian NER corpus (EstNER) using news articles annotated with classic semantic types (person, location, organization) using Conditional Random Fields. This corpus was later expanded and re-annotated to suit neural architectures. The transition to deep learning in Estonian NLP was accelerated by the development of `EstBERT` (Tanvir et al., 2021), a monolingual pretrained transformer model that consistently outperformed multilingual models, such as mBERT (Devlin et al., 2018) and XLM-RoBERTa (Conneau et al., 2020), across various syntax and semantic tasks, including general-domain NER.

Clinical NLP and Concept Mapping in Estonian Compared to general-domain NLP, clinical text processing in Estonian is still in its infancy, largely due to strict privacy regulations governing the Estonian National Health Information System (*Tervise Infosüsteem*). However, a few notable initiatives have paved the way for structured clinical extraction.

Researchers have increasingly focused on processing unstructured clinical narratives, such as epicrises, using specialized machine learning workflows. For instance, Šuvalov et al. (2023) successfully built an information extraction pipeline for breast cancer clinical records in Estonian, utilizing active learning and human-in-the-loop workflows to train a localized BERT model for TNM clinical staging extraction. Perli (2022) demonstrated that the general-purpose EstBERT model cannot match the clinical domain accuracy achieved by a BERT model pre-trained from scratch on clinical notes. Furthermore, the challenge of mapping localized clinical texts to standardized terminologies has been explored through cross-lingual resource alignment. Since the official Estonian translation of SNOMED CT covers only a fraction of the international version, researchers have focused on mapping local terminologies, ICD-10 classifications, and pharmaceutical registries to global standards like the OMOP Common Data Model (Uusküla et al., 2023) to enable semantic interoperability.

Entity linking Entity linking is the task of connecting named entities mentioned in text to their corresponding resources in a knowledge base or database (e.g., linking "Diabetes" in a sentence to its identifier in a knowledge base like UMLS). This involves assigning a unique database identifier to a text snippet that represents the entity. Several architectures have been proposed for clinical entity linking for low-resource languages. For example, ClinLinker (Gallego et al., 2024) combines a two-stage pipeline with initial candidate retrieval using a SapBERT-based bi-encoder and subsequent re-ranking with a cross-encoder, following a contrastive learning strategy tailored to medical concepts in Spanish. ClinLinker outperformed multilingual SapBERT by a substantial margin. The SympTEMIST (Vassileva et al., 2024) multilingual linking approach fine tuned a RoBERTa-based token-level classifier with Bidirectional LSTM and CRF layers on an augmented train set. EL was performed via a hybrid approach with dictionaries, generating candidates from a knowledge base using the cross-lingual SapBERT and reranking the top candidates using GPT-3.5.

The state-of-the-art EL approaches usually employ two-stage architectures combining candidate generation with transformer-based ranking (Wang et al., 2020). BERT-based and domain-adapted models, such as SapBERT and PubMedBERT have become foundational, with domain-specific pretraining providing gains of 6.5 points over general BERT initialization (Zhang et al., 2022).

Our work builds on these efforts by evaluating a unified, hybrid NER and EL pipeline designed specifically to tackle the extreme data scarcity of specialized clinical concepts in Estonian.

2.1 Overcoming the Resource Restrictions

Mitigating data scarcity in low-resource clinical NLP requires targeted, privacy-preserving strategies. Modern approaches leverage local language models to generate expert-reviewed synthetic text, alongside hybrid de-identification pipelines to safely anonymize patient data before annotation.

Engaging clinicians and native speakers ensures local linguistic and domain accuracy, while federated learning enables multi-institutional model training directly on local datasets without exposing raw patient data.

2.2 Motivation

The project of which this work is a part employs an array of artificial intelligence tools—including data onboarding assistants, multilingual NLP models (developed in Estonian, Dutch, and German), and entity linking frameworks—combined with an advanced human-in-the-loop curation pipeline. Central to the project is the transformation of disparate patient health data into an integrated Personal Health Knowledge Graph aligned with international standards such as SNOMED CT and HL7 FHIR[1], thereby ensuring semantic interoperability and sustainable data reuse. By automating data quality assessment and annotation, and by publishing reusable benchmark datasets and tools, the project not only addresses technical challenges but also lays the groundwork for the wider adoption of safe, efficient, and standardized health data management in both high- and low-resource language contexts. Consequently, the NLP tools employed to structure health data into a knowledge graph play a major role in driving this transformation.

3 Multilingual Clinical NLP Pipeline

3.1 NER and Temporal Information Extraction

Named Entity Recognition is a fundamental task in Natural Language Processing (NLP) wherein certain spans of text are identified and classified into semantic categories, such as people, locations, or organizations. In the clinical domain, this entails the recognition and classification of disorders, procedures, medications, laboratory values and more.

Although NLP has profited from the deployment of large annotated training corpora and the advent of pre-trained Transformer models, clinical NLP presents unique challenges through specialized language, including ambiguous abbreviations, Latin terminology, institution-specific conventions and domain-specific ontologies. We developed a hybrid NER architecture to address these challenges. Rather than relying on a single approach, we combine three complementary components: fine-tuned multitask Transformer models for high-precision extraction of well-defined entity types, a generative LLM (Qwen2.5-14B) deployed in a few-shot regime to extend coverage to entities lacking annotated training data, and a dictionary-based lookup for specific, well-enumerated entities linked to SNOMED CT codes.

3.1.1 Hybrid NER Architecture

The NER module is part of a broader clinical NLP pipeline developed for two primary use cases: Breast Cancer and Cardiovascular Disease. The architecture shown on Fig. 1 is designed to recognize spans belonging to SNOMED CT semantic categories, primary among them Body Structure, Disorder, Finding, Medicinal Product, Morphological Abnormality, Observable Entity, Procedure, Product, Qualifier Value and Substance. This system is designed to achieve both high precision on core entity types and adequate coverage across all semantic categories.

Fine-Tuned Transformer Models The foundation of the NER module is a set of multitask Transformer models fine-tuned on annotated clinical data. We utilize a pre-trained multilingual BERT model, fine-tuned for parallel sequence tagging of all semantic tags. Models are trained for 27 epochs (with early stopping) using a batch size of 32, a maximum sequence length of 128 tokens (with an overlap window of 14 tokens to accommodate sliding window inference across longer narratives), and a learning rate of 3×10^{-5} optimized against validation loss.

Generative LLM with Few-Shot Prompting To extract sparse or out-of-vocabulary entities, we employ few-shot prompting with Qwen2.5-14B. The model identifies five core semantic types (Disorder, Finding, Observable Entity, Procedure, Substance) and temporal expressions (in a different request). The 3-shot NER prompt is structured as follows:

> **System:** You are an expert AI assistant for medical information extraction. Your task is to analyze a clinical document (in German, Dutch, or Estonian) and extract key medical concepts.
>
> For each concept you identify, provide a JSON object with the following features:
>
> - `coveredText`: The exact, verbatim text of the concept as it occurs in the document. It should be concise.
> - `SemanticTag`: The concept's category, chosen from the five options below.
>
> **Semantic Tag Definitions (based on SNOMED CT)**
>
> - **disorder**: A disease, diagnosis, or pathological condition. E.g., *Hypertonie* (DE), *hartfalen* (NL, heart failure), *diabeet* (ET).
> - **finding**: A sign, symptom, or a specific lab/test result, representing evidence or an observation about a condition. E.g., *Fieber* (DE, fever), *pijn op de borst* (NL, chest pain), *köha* (ET, cough).

[1] http://hl7.org/fhir/R4/

- **observable entity**: A characteristic, quality, or state that can be measured or assessed. This represents the parameter being measured. E.g., *Körpertemperatur* (DE, body temperature), *bloeddruk* (NL, blood pressure), *südame löögisagedus* (ET, heart rate).
- **procedure**: A diagnostic or therapeutic action, test, surgery, or the act of making a measurement. E.g., *EKG* (DE), *operatie* (NL, operation), *röntgenuuring* (ET, x-ray), *Blutdruckmessung* (DE, blood pressure measurement).
- **substance**: A drug, medication, or chemical. Only extract the name, no auxiliary information. E.g., *Ibuprofen* (DE), *paracetamol* (NL), *insuliin* (ET, insulin).

Extraction Rules

- **Be Precise**: The `coveredText` must be the exact text. For Diagnose: akute Bronchitis', extract akute Bronchitis'.
- **Separate Listed Items**: If a text lists multiple items like 'Hypertonie und Diabetes', extract them as two separate concepts.
- **No Concepts**: If no valid concepts are found, return nothing.

[Few-shot examples are truncated to save space ...]

Dictionary-based Lookup For entities which have limited lexical variation and well-defined format, like measurements, we use a dictionary-based lookup. This method matches surface strings in the document against a curated dictionary and assigns the corresponding SNOMED CT code. Dictionary lookup is precise, interpretable, and computationally efficient.

Post-Processing: Filtering and Deduplication Due to the nature of this hybrid approach, the pipeline may produce overlapping, duplicate or conflicting annotations. To minimize these unwanted results, we incorporate two harmonization steps before the final results are output.

- Confidence-Based Filtering: Each annotation coming from the Transformer model has a confidence score. To optimize for precision, annotations with confidence scores which don't pass our threshold are discarded. LLM predictions are likewise filtered based on confidence scores produced during Entity Linking.
- Deduplication: Overlapping and redundant annotations produced by two different components are resolved through rule-based logic.

3.1.2 NER Results

As summarized in Table 1, the NER pipeline achieved F1-scores of 0.84 on a large Estonian corpus, 0.87 for Dutch, and 0.67 for German. Direct cross-lingual comparisons are confounded by structural differences in the datasets (Tables 1 and 5). Notably, the Dutch corpus contains only 708 entities compared to 11,523 for Estonian, and the Estonian test set has a near-total concept overlap with its training data (215 of 216 concepts).

A recognized limitation of our current study is the absence of an ablation analysis to isolate the performance gains of each distinct NER module. Unfortunately, strict institutional data protection protocols meant our access to the raw clinical narratives was revoked upon completion of the initial evaluation runs, preventing any subsequent component-wise testing.

3.2 Entity Linking

In alignment with the recent advances in EL, here we present the Estonian EL pipeline in detail and only refer to the German and Dutch ones. The goal is to resolve entities to 207,799 SNOMED CT concepts. Although the approach is concerned only with a SNOMED CT subset, the complexity of the EL task is high as it constitutes a classification problem across more than 207,000 candidate classes.. This scale requires a machine learning approach and a significant amount of training data.

3.2.1 Data

During training, we teach the model that the different labels of the same concept are closer to each other in the model space than the labels of other concepts. For this reason we provide mapping for each of these labels, to the concept ID and feed them to the model. E.g. in the Estonian version of SNOMED CT, the concept with identifier 44054006 has 11 labels:

- 5 labels in Estonian: "*2. tüüpi diabeet*", "*2. tüüpi diabetes mellitus*", "*2. tüüpi melliitdiabeet*", "*2. tüüpi suhkurtõbi*", "*Insuliinisõltumatu suhkurtõbi*" and
- 6 labels in English: "*Diabetes mellitus type 2 (disorder)*", "*Type 2 diabetes mellitus*", "*Diabetes mellitus type 2*", "*Diabetes mellitus type II*", "*T2DM - diabetes mellitus type 2*", "*Type II diabetes mellitus*".

At training time we create exhaustive list of combinations of size 2 from these labels and map them to the concept ID, 44054006 in this case, and supply these mappings as training samples to the model.

Although the project's target concepts span over 20 semantic types, the majority of entities in both SNOMED CT and our training dataset are covered by just four high-priority classes: Measurement (snomed:363787002), Drug (snomed:410942007),

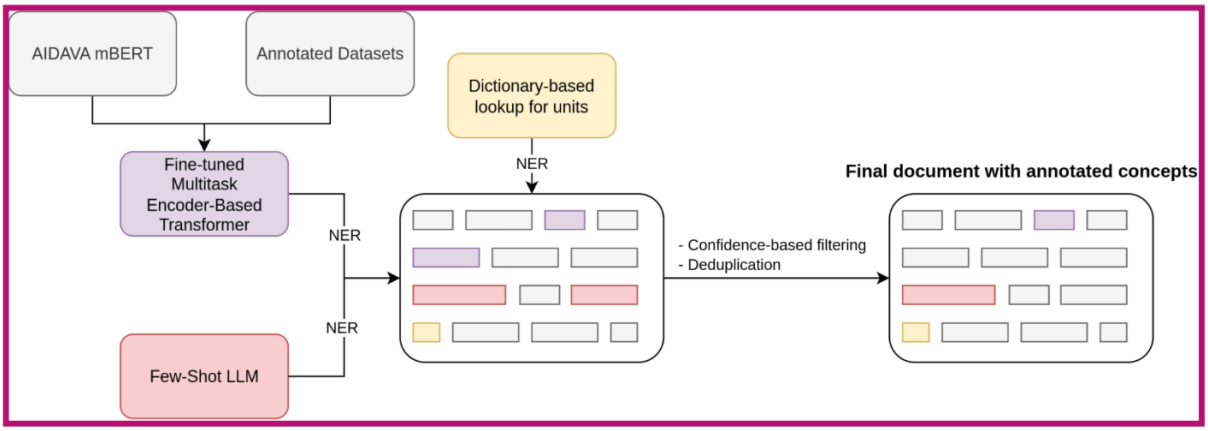


Figure 1: The hybrid architecture for Named Entity Recognition (NER). The system combines three parallel approaches: a fine-tuned Transformer model (mBERT trained on project-annotated data) for high-precision extraction, a few-shot LLM for broad coverage of less common entities, and a dictionary-based lookup for unambiguous units. The outputs are merged and then refined through confidence-based filtering and deduplication to produce the final annotated document.

Table 1: Overview of Named Entity Recognition (NER) performance evaluation across the three target languages at Medical University of Graz (MUG), University of Maastricht (UM) , and North Estonia Medical Centre (NEMC).

Site (lang.)	F1	Precision	Recall	TPs	FPs	FNs	Entities
MUG (de)	0.67	0.64	0.69	624	353	274	898
UM (nl)	0.87	0.92	0.83	588	49	120	708
NEMC (et)	0.84	0.84	0.84	9665	1791	1858	11,523

ProblemCondition (snomed:404684003), and Procedure (snomed:71388002).

Terminological sources A natural source of mappings between concept identifiers (IDs) and their corresponding labels is the international edition of SNOMED CT, given that it serves as the target KB for the EL task and offers multiple lexicalizations for many concepts. Among the three target languages in the project, the Dutch translation of SNOMED CT is the most comprehensive. In contrast, the German and Estonian releases contain a significantly limited number of translated concepts, with the Estonian version being the most constrained. Alongside Estonian descriptors, English labels are integrated into the training data. This inclusion enables the processing of narratives that may retain English terminology.

In addition to $< label-to-concept\_id >$ mappings from SNOMED, more resources are used to extend the coverage of the model. These are included in the training set through inferred mappings between SNOMED codes and UMLS codes, SNOMED and ICD classifications (ICD-9, ICD-10 and ICD-O), SNOMED and ATC. By applying these mappings, the training set for Estonian was augmented with concepts and labels from ICD-10, EMA drugs register; Estonian Ravimiregister; and MedDRA. Fig. 2 shows an example how the label "Lahtine haav" originating from MedDRA has been added to the Estonian training set.

The contribution of each of the resources to the final training set is shown on Table 2.

Table 2: Model 1 training data.

Source dataset	#concepts	#labels
SNOMED-EN	207,798	647,315
SNOMED-ET	8,980	12,974
ICD-10	5,211	5,283
EMA-DRUGS	161	372
Ravimiregister	534	2,781
MedDRA	12,074	17,479
Machine-translated	194,010	579,389
Manual annotaion	264	7,506
Total	**208,106**	**4,488,782**

Estonian SNOMED CT scarcity It is worth mentioning that Estonian is less represented in SNOMED than the other project languages - the Estonian SNOMED contains only about 5% of the concepts in the Dutch version and less than 50% of the concepts in the German and Austrian SNOMED together. As shown in Table 3, despite being the

From the inferred mappings:
MEDDRA ID *10058834* == UMLS CUI *C0332798* == SNOMED ID *125643001*
–>
"***Lahtine haav***" (MEDDRA ID *10058834*) is a label of the SNOMED concept with ID ***125643001*** *which is not available in the localized ET version of SNOMED therefore it contributes to the training set.*

Figure 2: MEDDRA-UMLS-SNOMED mappings

primary source of localized labels, the Estonian edition of SNOMED is highly limited. Indeed, the number of translated concepts represents only 9% of the target concepts. Furthermore, nearly 10% of these share labels with other concepts, introducing significant ambiguity and making model training more challenging.

Table 3: Dataset Statistics: Concept and Label Ambiguity Analysis

Concept Statistics	
Total number of concepts	17,749
Clean concepts (no ambiguous labels)	16,188
Concepts with ambiguous labels	1,561
Label Statistics	
Number of unambiguous labels	16,966
Number of ambiguous labels	783
Ambiguity Distribution	
Max. concepts per ambiguous label	4
Labels mapping to exactly 2 concepts	753
Labels mapping to exactly 3 concepts	29
Labels mapping to exactly 4 concepts	1

Machine-translated Labels To overcome this problem, all concepts in the international version of SNOMED (in English), which were among the project-related concepts were translated automatically to Estonian, using EasyNMT engine[2] and the opus-mt model[3] for English-Estonian. The translation was done locally to preserve privacy signed with SNOMED CT. These labels are marked as Machine Translated on Table 2 and Table 4 visualizing the training data in the EL experiments. As a result more than 194 000 concepts and 579 000 labels were added to the training set of Model 1 and 117 concepts and 426 labels to Model 2, respectively. It is important to mention that these translated labels were manually checked (a subset) and were not included in the indexed KB, which is used for the resolution task. They were only for training.

Manually Annotated Data Manually annotated data is of great importance to creating a good quality training dataset, as it represents the concepts' usage in reality with their specific lexicalizations originating from project-relevant documents. For each of the three languages, a manually annotated corpus was provided. The entities in the documents are manually tagged with a semantic type and the most relevant SNOMED concept ID. Table 5 shows the statistics on the manually annotated data. Each annotation is a mention/label of a given concept. Since a concept may have more than one label and a label may occur more than once in the dataset, the number of annotations exceeds the number of labels and concepts. The Estonian dataset has the largest number of documents; however, it covers the lowest number of concepts among the three language datasets, and it's the only dataset whose training set covers more than 99% of the concepts in the test set. This is a prerequisite for good results on the test set and a low degree of generalization to unseen data/concepts.

The quality and relevance of the training data has a direct impact on the model's ability to make correct predictions. However, due to the sensitivity of the data in the project domain - healthcare, only limited access to such data was available at development time. It was granted in compliance with the regulations and ethical policies established in the project and in the respective countries.

The contribution of each dataset to the training data for Model 1 in terms of labels and concepts in Estonian is shown in Table 2, and the resources attributed to Model 2 are shown in Table 4.

Table 4: Model 2 training data

Source dataset	#concepts	#labels
SNOMED-EN	183	780
SNOMED-ET	58	84
ICD-10	13	13
MedDRA-ET	30	68
Machine-translated ET	117	426
Manual annotation	259	7,425
MA-context	264	7,506
Total	**264**	**16,302**

[2]https://easynmt.net/demo/
[3]https://github.com/Helsinki-NLP/OPUS-MT-train

Table 5: Distribution of the annotations in the manually annotated data.

Site (lang.)	Split	# annotations	# labels	# labels in common with train	# concepts	# concepts in common with train
DE	train	4,087	1,210	—	575	—
	test	415	201	132	142	132
	validation	483	245	152	169	145
ET	train	76,382	7,147	—	264	—
	test	9,688	2,004	1,504	216	215
	validation	9,639	2,157	1,599	220	219
NL	train	5,127	2,106	—	912	—
	test	740	487	206	308	183
	validation	745	447	231	278	198

3.2.2 Approach

The input for the EL module consists of an entity string and its corresponding semantic type. The objective is to return the most appropriate SNOMED concept ID for that entity. Two types of annotators are employed for this task: gazetteers and machine learning models. Gazetteers are tailored exclusively for unambiguous entities; consequently, their confidence score is very high, and they are executed up-front. If a gazetteer fails to find a match, the entity string is passed to a transformer model to generate an embedding. This embedding is subsequently used to search for the nearest concepts within an indexed knowledge base.

Gazetteer The gazetteer comprises all labels from the complete training dataset (Table 2) that are associated with a single concept identifier. Ambiguous labels are strictly excluded from the gazetteer. Consequently, the accuracy of this annotator is exceptionally high—approaching 1.

Models Following on the related approaches, the machine learning model we chose for solving this problem is a transformer model with applied Self-Aligned Pretraining (SapBERT). This is a fine-tuning method that leverages synonymy sets extracted from UMLS to improve BERT's ability to act as a biomedical entity encoder. It has proven good performance and SotA results on the EL task.

Experiments The final training setup was made after a series of experiments with subsets of training data, split by semantic type, data source and language. Two models were fine-tuned for Estonian: big (model 1) and small (model 2). Both are trained in 3 epochs with 128 batch size, maximum lenght of 25 and optimized against multi-similarity loss.

Model 1 (big): SapBERT-UMLS-2020AB-all-lang-from-XLMR fine-tuned with the whole training data listed on Table 2.

Model 2 (small): SapBERT-UMLS-2020AB-all-lang-from-XLMR tuned only on the subset of concepts available in the manually annotated data for the corresponding language. This resulted in a smaller dataset, shown on Table 4.

3.2.3 EL Results

The EL module performance is tested on 2 datasets: the test (unseen) and validation (used for development) splits of the Manually annotated dataset. Table 6 evaluates the contribution of gazetteers to the models and compares performance between the large and small model variants. The baseline model—a multilingual SapBERT—is trained on SNOMED CT labels across all three languages, supplemented by Estonian translation equivalents and German labels extracted from hospital patient records. In contrast, the subsequent models are additionally trained on manually annotated data.

The evaluation on the test and validation splits shows similar results; they are both randomly selected from the manually annotated documents and contain similar terminology. The results are comparable on both datasets, except for Dutch, where the results on the validation dataset were higher by more than 10 points. The small models for Estonian and German outperformed the big models (Table 6 Row 2 and 3) – 0.758 vs. 0.27 for Estonian and 0.874 vs. 0.79 for German. For Dutch, however, the big model performed slightly better than the small one (0.59 vs. 0.55 in terms of Precision). The performance of Dutch is surprising, it is worth seeing the number of concepts overlapping in the training and test set in this respect. For Dutch there is the lowest overlap.

Gazetteer impact: it was difficult to estimate the

Table 6: Model Performance Evaluation Across Data Splits and Languages

Data split	Model	Semantic types	ET		DE		NL	
			model + gaze	model only	model + gaze	model only	model + gaze	model only
test	baseline	all	—	0.119	—	0.157	—	0.177
	big	all	0.270	0.270	0.790	0.790	0.590	0.590
	small	all	0.758	0.758	0.874	0.874	0.552	0.543
validation	small	all	0.762	0.762	0.858	0.858	0.6875	0.6807

exact gazetteer contribution to the model performance, because the concepts in the test are almost fully covered by the training set. Dutch is an exception, as the set of concepts in the Netherlands test dataset is covered by the train set only in about 2/3 and probably that's why the gazetteer shows impact – it succeeds in recognizing some of the remaining 1/3 of the concepts which were not seen by the model. To further prove the gazetteer impact, additional experiments with the Dutch models were carried out on the least represented concepts dataset and the most represented concepts dataset to test the models' performance on more challenging datasets, which are outside the small model's training data. On these datasets, the experiments were run with the small and the big model, with and without gazetteers and show that where models are troubled with unseen or less/over-represented concepts, the gazetteers raise the precision with 0.13 to 0.35 points for the big model and between 0.35 and 0.96 points for the small one.

4 Discussion and Conclusion

Here we describe the project's NER and EL pipelines for the low-resource language Estonian in comparative evaluation with Dutch and German from clinical documentation, which yielded several important findings. (*i*) Although a low-resource language, the models show good performance for Estonian. (*ii*) Annotations were performed by common annotation guidelines and task complexity was similar across the three languages. (*iii*) This study offers valuable insights into the practical implications of working with low-resource languages in clinical NLP, presenting notable strengths and limitations. Among its strengths, the project's cross-linguistic design, consistent methodology, and deployment across real-world clinical environments allowed for a robust comparison between Estonian, Dutch, and German — languages differing significantly in their NLP resource availability. (*iv*) Utilizing harmonized annotation protocols and evaluation criteria minimized bias and enhanced the comparability of results, while the integration of human-in-the-loop curation captured both quantitative and qualitative measures of system performance. (*v*) Multilingual models trained on the three languages reached comparative results with models fine-tuned on a specific language.

However, there are inherent weaknesses. The project was limited by the scale of data available, potentially affecting the generalizability of findings to broader clinical language use. Additionally, some annotation and evaluation steps relied on domain experts, which, though necessary, introduced subjectivity and potential inter-annotator variability. The marked disparity in digital clinical ontologies and language tools—a defining trait of low-resource status—likely affected downstream processes such as knowledge graph construction and IPS generation, reinforcing the critical bottleneck imposed by limited language resources.

A methodological limitation of our evaluation is that while data access was available, NER and EL were evaluated independently on gold annotations, meaning downstream error propagation from imperfect NER predictions to the EL module remains unmeasured. Additionally, while experiments represent single runs, variance was minimized by employing deterministic transformer-encoder models and setting the LLM generation temperature to zero.

An important next step involves evaluating the developed NLP modules on external datasets to assess their ability to generalize beyond the project corpus.

Acknowledgments

This work is partially supported by European Union's Horizon Europe research and innovation programme project AIDAVA [Grant Agreement No.101057062][4].

[4]https://www.aidava.eu

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Multilingual BERT (mBERT). https://github.com/google-research/bert/blob/master/multilingual.md.

F. Gallego, G. López-García, L. Gasco-Sánchez, M. Krallinger, and F. J. Veredas. 2024. Clinlinker: Medical entity linking of clinical concept mentions in spanish. *Lecture Notes in Computer Science*, pages 266–280.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.

Siim Orasmaa, Timo Petmanson, Alexander Tkachenko, Sven Laur, and Heiki-Jaan Kaalep. 2016. EstNLTK - NLP toolkit for Estonian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2460–2466, Portorož, Slovenia. European Language Resources Association (ELRA).

Meelis Perli. 2022. Representation learning on free text medical data. Master's thesis, University of Tartu, Tartu, Estonia.

Hendrik Šuvalov, Sven Laur, Tanel Laisaar, and Alan Altraja. 2023. Information extraction from medical texts with BERT using human-in-the-loop labeling. In *Caring is Sharing – Exploiting the Value in Data for Health and Innovation*, volume 305 of *Studies in Health Technology and Informatics*, pages 831–835. IOS Press.

Hasan Tanvir, Claudia Kittask, Sandra Eiche, and Kairit Sirts. 2021. EstBERT: A pretrained language-specific BERT for Estonian. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 11–19.

Alexander Tkachenko, Timo Petmanson, and Sven Laur. 2013. Named entity recognition in Estonian. In *Proceedings of the 4th Workshop on Balto-Slavic Natural Language Processing*, pages 78–83.

Anneli Uusküla, Heti Pisarev, Kaja-Triin Laisaar, Sven Laur, Kärt Võilo, Helen Task, Reelika Tamm, and Tarmo Meister. 2023. Transforming Estonian health data to the Observational Medical Outcomes Partnership (OMOP) Common Data Model: lessons learned. *Database*, 2023:baad080.

S. Vassileva, G. Grazhdanski, I. Koychev, and S. Boytcheva. 2024. Transformer-based approach for symptom recognition and multilingual linking. *Database*, 2024.

Yizhe Wang, Beatrice Daille, and Nabil Hathout. 2020. A study of semantic projection from single word terms to multi-word terms in the environment domain. In *Proceedings of the 6th International Workshop on Computational Terminology*, pages 50–54, Marseille, France. European Language Resources Association.

S. Zhang, H. Cheng, S. Vashishth, C. Wong, J. Xiao, X. Liu, T. Naumann, J. Gao, and H. Poon. 2022. Knowledge-rich self-supervision for biomedical entity linking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*.

Beyond RAG: A Multi-Graph, Multi-Agent, Recursive Retrieval Architecture for Traceable and Source-Grounded Legal Answers

Assia Bouamir[1], **Marie Bonnin**[2], **Youssef AL Mouatamid**[1,3], **Jihad Zahir**[1,4]
[1]LISI, Faculty of Sciences Semlalia, Cadi Ayyad University, Morocco
[2]IRD, Univ Brest, CNRS, Ifremer, LEMAR & Université Cadi Ayyad, Morocco
[3]LEMAR, Univ Brest, Plouzané, France
[4]UMMISCO-IRD

Abstract

Classical Retrieval-Augmented Generation (RAG) systems frequently generate legally relevant answers that lack accurate source citations, undermining their reliability for professional legal practice. This limitation is further compounded by document mismatch retrieval (DMR) — a systematic failure mode in which retrieved passages belong to structurally similar but legally distinct documents — which significantly reduces precision in complex multi-jurisdictional corpora. We present a novel architecture organized around three complementary knowledge layers: structural lexical graphs that capture document hierarchy, citation graphs enabling deterministic cross-reference resolution, and definition graphs for grounding formal legal terminology. Orchestrated by a multi-agent pipeline employing hybrid BM25 and vector fusion retrieval, the system enforces citation discipline at inference time and substantially improves answer accuracy. Evaluated on a novel benchmark in African francophone statutory law, consisting of 263 expert-validated question-answer pairs across two thematic domains—whale hunting regulation and hydrocarbon discharge control—our system achieves an average BERTScore F1 of 0.845, compared to 0.378 for a standard RAG baseline. Citation correctness reaches **94.0%** on average (versus **25.9%** for the baseline), as confirmed by expert ground truth and an independent human evaluation panel.

1 Introduction

The rapid adoption of large language models (LLMs) in the legal domain has opened new possibilities for automated legal question answering, contract analysis, and regulatory compliance checking. Retrieval-Augmented Generation (Lewis et al., 2020) has emerged as the dominant paradigm. However, classical RAG architectures often struggle to reproduce accurate source references. Legal practitioners require every claim to be traceable to a specific article of law. We identify four root causes: single-pass flat retrieval that ignores citation links, lack of architectural constraints for source attribution, absence of quality-control loops, and document mismatch retrieval (DMR) that surfaces jurisdictionally or temporally incorrect documents. This paper presents an architecture that addresses these limitations via the tight co-design of three complementary knowledge layers, hybrid retrieval, and multi-agent orchestration, focusing on African francophone civil-law statutes—a low-resource setting under-served by existing benchmarks. The multi-agent layer, based on a LangGraph state machine with five specialized agents, dynamically builds complete legal dossiers via iterative reasoning and citation traversal, enforcing strict citation discipline at inference time. Concretely, we contribute: (i) a multi-graph knowledge layer; (ii) a hybrid retrieval engine with recursive citation traversal; (iii) a multi-agent orchestration layer; and (iv) a thorough evaluation on a new African francophone legal benchmark.

2 Related Work

RAG for Legal NLP. Legal question answering has attracted growing attention (Zhong et al., 2020; Chalkidis et al., 2022). Several studies identify retrieval quality as the main bottleneck in statutory legal QA (Louis et al., 2022). Dhakal et al. (2026) further demonstrate that a hybrid BM25 + dense + reranking approach yields a 41% improvement in Precision@10, motivating the BM25-dominant weighting in our fusion strategy. Existing benchmarks such as LegalBench (Guha et al., 2023) and CLERC (Hou et al., 2025) focus primarily on US common-law settings, offering limited coverage of statutory citation-chain retrieval in non-English, low-resource civil-law jurisdictions—a gap our work directly addresses.

Graph-based and Citation Retrieval. GraphRAG (Edge et al., 2024), HippoRAG

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 318–325
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_033

(Jiménez Gutiérrez et al., 2024), and RAPTOR (Sarthi et al., 2024) explore entity-level or hierarchical graphs to improve retrieval depth. In the legal domain, citation graphs have been used for precedent retrieval in common-law systems (Dadgostari et al., 2021). Our architecture adapts citation-graph traversal to *statutory law* and integrates it with recursive multi-agent orchestration for deterministic cross-reference resolution.

Multi-hop and Iterative Retrieval. IRCoT (Trivedi et al., 2023) and Self-RAG (Asai et al., 2024) interleave retrieval and reasoning in iterative loops but rely on learned or probabilistic chaining. Our system instead leverages the explicit, deterministic citation topology of legal statutes for more reliable recursive traversal.

Multi-Agent Systems and Citation Grounding. We build upon LangGraph (LangChain, Inc., 2024) to implement a controlled recursive retrieval loop with strict citation enforcement, extending recent advances in multi-agent frameworks (Hong et al., 2024). Generating accurate citations remains a persistent challenge (Gao et al., 2023); prompt-based mitigations (Liu et al., 2024) address only surface symptoms. Our architecture tackles the root cause by enforcing retrieval and verification *before* answer generation.

3 Proposed Architecture

3.1 System Overview

Our system comprises four tightly integrated layers: (1) document parsing and data processing, (2) a multi-graph knowledge layer, (3) a hybrid retrieval engine, and (4) a multi-agent orchestration layer. As illustrated in Figure 1, each layer operates sequentially to produce citation-grounded legal answers.

3.2 Document Parsing and Processing

Raw legal PDFs present significant challenges, including multi-column layouts and inconsistent formatting. We employ LlamaParse (LlamaIndex, 2024) for its superior structural fidelity, enforcing a strict JSON schema that preserves article hierarchy and extracts footnotes. Among the parsers evaluated, LlamaParse was the only candidate to consistently maintain article hierarchy and schema integrity—a finding corroborated by ParseBench, in which LlamaParse Agentic achieves the highest overall score (84.9%) across five agent-critical dimensions on ∼2,000 human-verified pages (Zhang et al., 2026). Following parsing, a normalization pipeline transforms raw output into `DocumentNode` objects, one per legal article, preserving clause-level retrieval granularity. Processing comprises: (i) NFKD Unicode normalization; (ii) metadata extraction (official titles, ISO 8601 dates, jurisdiction); (iii) context enrichment with a unique identifier of the form `{Country}_{Year}_{LawName}_{ClauseID}`.

3.3 Multi-Graph Knowledge Layer

To facilitate recursive reasoning and deterministic retrieval, we model the legal corpus as three complementary knowledge layers. The *structural lexical graph* preserves the logical hierarchy of each law using directed `contains` edges linking articles to parent sections, chapters, and titles. The *citation graph* enables deterministic resolution of legal dependencies by extracting explicit cross-references (regex patterns such as *conformément à l'article*) and creating directed `references` edges. The *definitions graph* grounds formal legal terminology by linking technical terms (e.g., *cétacé*) to their authoritative definitions through `means` edges. Upon a query, the orchestrator enriches context in three steps. The definitions graph first injects authoritative definitions (e.g., "hydrocarbure", "navire"). The citation graph then recursively fetches cross-referenced articles. Finally, the lexical graph supplies hierarchical parent-chapter context, clarifying scope and producing a complete, self-contained legal dossier for the answering agent.

3.4 Hybrid Retrieval Engine

To ensure both semantic breadth and legal precision, we employ a multi-stage retrieval pipeline comprising hybrid fusion, graph-based boosting, and precision reranking. **1. Score Fusion.** Three independent retrievers are invoked simultaneously and fused via:

$$s_{\text{fused}} = 0.35\, s_{\text{Dense}} + 0.50\, s_{\text{BM25}} + 0.15\, s_{\text{Graph}} \quad (1)$$

The BM25 component (weight 0.50) is intentionally prioritized for exact-term matching on article numbers and decrees, while dense vector retrieval (`bge-m3`, weight 0.35) captures semantic intent. The graph component (weight 0.15) leverages the Lexical Graph to reinforce contextually related provisions. This bias toward BM25 is motivated by the heavy reliance of legal corpora on precise alphanumeric identifiers that dense embeddings often miss (Dhakal et al., 2026). Weights were confirmed by a two-phase grid search: a full 3D

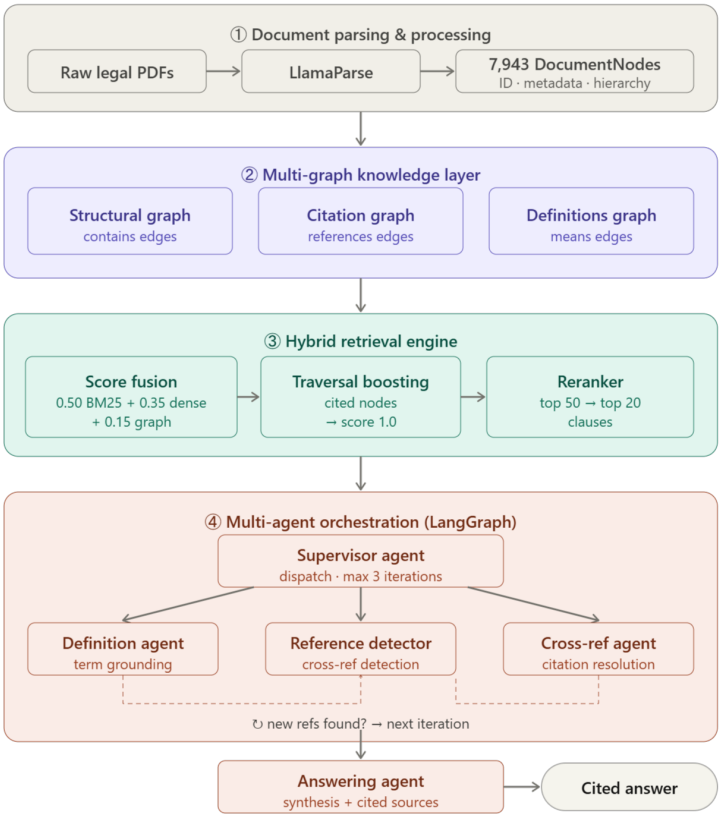


Figure 1: Overview of the proposed pipeline

search over 27 combinations identifies the graph signal optimum at 0.15; a subsequent 2D search on a 10% validation sample confirms the final weights, achieving F1 $= 0.870$ vs. best alternative $(0.40, 0.45, 0.15)$ at 0.868. **2. Traversal Boosting.** After initial retrieval, the engine performs recursive traversal on the Citation Graph. Any `DocumentNode` explicitly referenced by a top-retrieved `DocumentNode` is automatically promoted to a relevance score of 1.0, ensuring that legally dependent clauses are never excluded. **3. Precision Reranking.** We apply a two-stage Retrieval Funnel:

- Broad Retrieval: The hybrid engine and graph traversal aggregate a pool of up to 50 candidate clauses.
- Cross-Encoder Reranking: These clauses are scored by a Cross-Encoder (`ms-marco-TinyBERT-L-2-v2`). Only the top 20 clauses are passed to the answering agent.

3.5 Multi-Agent Orchestration

The system employs a `LangGraph StateGraph` to coordinate five specialized agents within a cyclic, stateful workflow orchestrated by a central `Supervisor_agent`, which dynamically dispatches agents based on dossier completeness.

1. `Supervisor_agent`: Central orchestrator that evaluates dossier state after each step, dispatches the next agent, and enforces a hard termination ceiling of 3 iterations.
2. `Definition_agent`: Queries the definitions graph to inject authoritative groundings for unresolved legal terms (e.g., *hydrocarbure*, *cétacé*).
3. `Reference_detector_agent`: Scans retrieved clauses for explicit cross-references and registers unresolved citation targets.
4. `Cross_reference_resolution_agent`: Traverses the citation graph to resolve pending targets, promoting referenced `DocumentNodes` to a relevance score of 1.0. The supervisor then re-evaluates; if new references or definitions surface, another cycle is triggered; otherwise control passes to the answering agent.
5. `Answering_agent`: Synthesizes the final citation-grounded response once the supervisor certifies dossier completeness, including a mandatory "Sources consulted" section.

4 Baseline

Our baseline shares the same `bge-m3` embeddings, cross-encoder reranker, and LLM as the proposed system. Score fusion follows:

$$s_{\text{fused}} = 0.55\, s_{\text{Dense}} + 0.45\, s_{\text{BM25}} \quad (2)$$

The baseline performs flat, single-pass retrieval with no graph layer, no recursive citation traversal, no iterative multi-agent refinement, and no citation enforcement at inference time. This design isolates the contribution of the multi-graph knowledge layer and agentic orchestration. To ensure a fair comparison, both systems use the same underlying LLM and receive the top-20 reranked clauses. The key difference lies in the prompting strategy. The baseline prompt is deliberately minimal:

> *"You are a legal assistant specialized in African francophone environmental law. Answer the user's question based solely on the provided context. Be concise, accurate, and professional."*

In contrast, the `Answering_agent` of the multi-agent system receives an explicit citation-enforcement instruction:

> *"You are a legal assistant. Answer strictly on the basis of the retrieved legal dossier. Every claim must be supported by an explicit reference (law name + article number). End your response with a mandatory section entitled "Sources consulted" listing all cited articles. Do not answer if the dossier is incomplete."*

This controlled difference in prompting isolates the effect of the multi-graph + multi-agent architecture rather than differences in surface-level instructions.

5 Experimental Setup

5.1 Corpus

The corpus consists of French-language statutory and regulatory texts (lois, décrets, ordonnances) governing environmental protection, particularly whale hunting regulation, marine birds protection, and hydrocarbon discharge control. These documents originate from African francophone jurisdictions and the international MARPOL convention. All texts were processed natively in French. These jurisdictions share a common civil-law tradition inherited from the French legal system, with substantial lexical and structural overlap in environmental statutes, while still presenting important national variations in transposition of international conventions such as MARPOL (e.g., differing sanction regimes). This multi-jurisdictional setting makes citation-chain retrieval and document-mismatch issues particularly acute, which is why we deliberately selected 17 countries rather than focusing on a single jurisdiction. The knowledge base comprises 50 unique PDFs spanning 17 jurisdictions. After parsing and normalization, the corpus yields 7,943 qualified clauses (DocumentNodes). Detailed statistics by theme and jurisdiction are provided in Tables 1 and 2.

Theme	Documents	Clauses
Whale Hunting	23	2,080
Marine Birds	14	957
Hydrocarbon Discharge	21	4,906
Total (unique PDFs)	50	7,943

Table 1: Corpus statistics by theme. Some documents cover multiple themes. All documents are in French.

Jurisdiction	Documents	Clauses
Algeria	4	621
Benin	3	418
Cameroon	4	593
Comoros	2	284
Congo	2	310
Côte d'Ivoire	3	452
Djibouti	2	195
France	3	489
Gabon	4	601
Guinea	2	263
Madagascar	3	374
MARPOL	3	712
Mauritania	2	297
Morocco	4	538
Senegal	4	612
Togo	2	287
Tunisia	3	397
Total	50	7,943

Table 2: Per-jurisdiction document and clause counts. All texts are in French. MARPOL documents include the original convention and national transpositions.

5.2 Evaluation Dataset and Benchmark Construction

The evaluation uses 263 expert-curated question-answer pairs covering multi-hop legal queries. The benchmark was constructed by three legal experts specializing in African francophone statutory and environmental law (practitioners and academics with 5–15 years of experience). Inter-annotator agreement on a subset of 40 questions yielded $\kappa = 0.79$. For each question, experts provided a ground-truth answer together with the authoritative article(s) required to support it. "Exact" citation is defined as matching the law name and

article number (not the full verbatim text). Marine bird protection documents are included in the retrieval corpus to preserve ecological coverage and test cross-theme robustness; they are excluded from the QA benchmark because expert annotation for this sub-domain was not completed within the scope of the current study. The evaluation thus covers 263 question-answer pairs across two thematic benchmarks (131 whale hunting, 132 hydrocarbon discharge).

5.3 Evaluation Framework

We evaluate the system across three complementary dimensions. **RAGAS Metrics.** We use the RAGAS framework to assess five key dimensions: faithfulness, answer relevancy, context recall, semantic similarity, and answer correctness (all scores range from 0 to 1). Context precision is excluded, as it requires chunk-level ground-truth labels not available in the current annotation protocol; this is an acknowledged limitation (Section 7.1). **BERTScore.** We compute pairwise F1 scores between the standard RAG baseline, our multi-agent system, and the expert ground truth using `bge-m3` embeddings. Additional verification with `multilingual-e5-large` rules out embedding-induced inflation. **Binary Citation Correctness.** This metric measures whether the generated response cites the exact law and article number matching the expert ground truth. It serves as a critical hard evaluation in legal applications. Statistical significance is assessed via 95% bootstrap confidence intervals and paired Wilcoxon signed-rank tests ($\alpha = 0.05$), appropriate for non-normally distributed paired scores. Cross-contamination between themes is prevented through metadata-based partitioning of the indices at query time.

5.4 LLM and Embedding Configuration

Component	Model / Library
Agent LLM	Groq – LLaMA 3.1
Embeddings	HuggingFace `bge-m3`
RAGAS Judge	Llama-3.1-8B-Instruct
Orchestration	LangGraph

Table 3: System configuration.

6 Results and Analysis

6.1 RAGAS Evaluation

Tables 4 and 5 present the five RAGAS metrics for both thematic benchmarks. All scores are on a 0–1 scale; higher is better. Each cell reports the mean score ± standard deviation. The baseline's low context recall (0.22 / 0.19) confirms that single-pass retrieval fails to surface enough legally relevant articles. Recursive citation traversal is the primary driver of the performance gains. Higher Answer Correctness in the Hydrocarbon theme (0.90) compared to Whale Hunting (0.78) is attributed to more constrained, citation-dense regulatory texts.

System	Faith.	Ans.R	C.R	A.Cor	A.Sim
Naive RAG	$0.41_{\pm.03}$	$0.51_{\pm.04}$	$0.22_{\pm.03}$	$0.14_{\pm.02}$	$0.18_{\pm.02}$
Multi-Agent	$0.83_{\pm.02}$	$0.84_{\pm.03}$	$0.86_{\pm.03}$	$0.78_{\pm.03}$	$0.85_{\pm.02}$

Table 4: RAGAS scores (±SD) — Theme 1: *Whale Hunting* ($n = 131$). Faith.= Faithfulness; Ans.R= Answer Relevancy; C.R= Context Recall; A.Cor= Answer Correctness; A.Sim= Answer Similarity.

System	Faith.	Ans.R	C.R	A.Cor	A.Sim
Naive RAG	$0.39_{\pm.05}$	$0.48_{\pm.06}$	$0.19_{\pm.04}$	$0.12_{\pm.03}$	$0.16_{\pm.03}$
Multi-Agent	$0.88_{\pm.03}$	$0.87_{\pm.04}$	$0.89_{\pm.04}$	$0.90_{\pm.03}$	$0.80_{\pm.04}$

Table 5: RAGAS scores (±SD) — Theme 2: *Hydrocarbon Discharge* ($n = 132$).

6.2 BERTScore Evaluation

Tables 6 and 7 present the pairwise BERTScore F1 results. The Multi-Agent system achieves 0.882 and 0.807 against the human reference, versus baseline 0.376 and 0.379. **Pooled analysis.** Weighted by sample size, the aggregated BERTScore F1 for the Multi-Agent system is:

$$F1_{\text{pooled}} = \frac{(131 \times 0.882) + (132 \times 0.807)}{263} \approx 0.844$$

compared to the pooled baseline mean of 0.377. We report this figure only as a convenience summary; the per-theme results constitute the primary analysis. **Component ablation.** Table 8 isolates the individual contributions within the *Whale Hunting* theme. The results demonstrate a super-additive interaction: removing the graph retrieval layer drops BERTScore to 0.559, and removing agentic orchestration yields 0.579. Only the full integration achieves peak performance. Although a finer-grained ablation isolating each graph type and each of the five agents was not feasible within the ex-

Comparison	BERTScore F1	95% CI
m1 (Baseline) vs m3 (Ref.)	0.376	[0.358, 0.394]
m2 (Multi-Agent) vs m3 (Ref.)	0.882	[0.861, 0.903]
m1 vs m2	0.451	[0.432, 0.470]

Table 6: BERTScore F1 — Theme 1: *Whale Hunting* ($n = 131$). Differences are statistically significant (Wilcoxon, $p < 0.001$).

Comparison	BERTScore F1	95% CI
Baseline vs Ref.	0.379	[0.341, 0.417]
Multi-Agent vs Ref.	0.807	[0.773, 0.841]
Baseline vs Multi-Agent	0.393	[0.357, 0.429]

Table 7: BERTScore F1 — Theme 2: *Hydrocarbon Discharge* ($n = 132$).

Mode	BERTScore F1	95% CI	Cit. Acc. (%)
No-Graph	0.559	[0.530, 0.588]	71.97
No-Agents	0.579	[0.551, 0.607]	60.61
Full System	0.882	[0.861, 0.903]	89.3

Table 8: BERTScore ablation — Theme 1: *Whale Hunting*. Non-overlapping CIs confirm statistical significance.

perimental budget, the combined ablations already demonstrate that both the multi-graph knowledge layer and the agentic orchestration contribute substantially and interact positively. A single-agent variant with the same tools is left for future work (see Section 7.1).

6.3 Binary Citation Correctness

For each query, we evaluate whether the system correctly cites the authoritative legal article. Table 9 reports the citation accuracy.

Theme	Multi-Agent		Naive RAG	
	Correct	%	Correct	%
Whale Hunting ($n = 131$)	126/131	96.2	36/131	27.5
Hydrocarbon Discharge ($n = 132$)	121/132	91.7	32/132	24.2
Pooled	247/263	93.9	68/263	25.9

Table 9: Document-level citation correctness. The pooled accuracy of 93.9% (247/263) is verified by an independent human evaluation panel (Section 6.4).

6.4 Human Evaluation Panel

To independently validate the citation correctness results, we conducted a held-out human evaluation on a stratified random sample of 60 questions (30 per thematic benchmark). Two legal annotators with expertise in African francophone statutory law independently assessed each system response, blinded to system identity. Inter-annotator agreement was $\kappa = 0.81$. Table 10 reports the human-validated citation accuracy (Naive RAG 26.7%, Multi-Agent 93.3%), consistent with the automatic pooled accuracy of 93.9%. We acknowledge that a larger panel would further strengthen confidence.

6.5 Legal RAG Error Taxonomy

The failure cases motivate a structured taxonomy of legal RAG error modes across three layers. **Retrieval-level:** *ERR-1.1 (Implicit Reference)*—implicit statutory references without citable identifiers, mitigated by `bge-m3` semantic fallback; *ERR-1.2 (Isolated Article)*—isolated clause without parent definitional chapter, mitigated by hierarchical graph expansion; *ERR-1.3 (Jurisdiction Contamination)*—cross-jurisdiction civil-code overlap, mitigated by metadata filtering. **Reasoning-level:** *ERR-2.1 (Species–Class Mismatch)*—species/statutory-class mismatch, mitigated by taxonomical query expansion; *ERR-2.2 (Temporal Asymmetry)*—retrieval of superseded legislation (e.g., Law 82-10 vs. Law 04-07 in Algeria), the primary unresolved error; *ERR-2.3 (Ghost Exception)*—LLM-assumed implicit exceptions, mitigated by negative-constraint prompting. **Systemic:** *ERR-3.1 (Supervisor Pruning)*—context-window overflow, mitigated by ranked pruning; *ERR-3.2 (Segmentation Noise)*—OCR-broken article identifiers, mitigated by LlamaParse's OCR pipeline. **Error analysis.** Qualitative analysis of the Multi-Agent system failures reveals three primary modes: (1) hallucinated decree numbers, (2) temporal precedence errors, and (3) cross-jurisdictional leakage. Table 11 illustrates these. While ERR-2.2 remains the main unresolved challenge, hierarchical graph expansion, metadata filtering, and negative-constraint prompting reduce the other error modes.

System	Correct / Total	Accuracy (%)
Naive RAG	16/60	26.7
Multi-Agent	56/60	93.3

Table 10: Human evaluation panel — citation correctness on a 60-question held-out set.

Query	Response	Failure Mode
Cetaceans (Algeria)	"Loi 82-10 du 26 août 1982"	Correct
Sanctions (Gabon)	"Art. 23 du Décret 03-1374"	Hallucinated decree
Declaration (Benin)	"Art. 14 de la loi 98-030"	Temporal error
MARPOL (Gabon)	"...Art. 12 [Cameroun]"	Cross-jurisdiction leakage

Table 11: Representative Multi-Agent failure cases. The temporal error row illustrates ERR-2.2.

6.6 Citation Depth Ablation

Citation depth—the number of unique, relevant articles cited per response—increases from 1–2 (baseline) to 2–3 (+hybrid), 5–8 (+recursive link following), and 7–10 (full system). This demon-

strates the system's transition from index lookup to structured legal synthesis.

6.7 Score Fusion Weight Robustness

The fusion weights $(0.35, 0.50, 0.15)$ are a domain-motivated choice. A grid search over 18 weight combinations on a 10% held-out validation sample ($n_{\text{val}} = 13$) shows the selected weights achieve F1 $= 0.870$ vs. the best alternative $(0.40, 0.45, 0.15)$ at 0.868 ($\Delta = 0.002$), well within confidence interval overlap. This confirms robustness: the BM25-dominant signal drives performance regardless of minor weight variations. The baseline $(0.55, 0.45)$ was tuned independently on the same validation sample, achieving F1 $=$ 0.843.

6.8 Qualitative Examples

Table 12 contrasts the baseline and the proposed system on representative queries.

Query	Naive RAG	Multi-Agent
Sanctions for cetacean hunting in Benin?	Cetacean hunting is regulated by Beninese environmental law.	Article 97 of Law 87-014 provides for a fine of 500 000 FCFA and a prison sentence of 1 to 3 years. *Sources: Law 87-014, Art. 97; Decree 2000-278, Art. 4.*
Decree governing hydrocarbon discharges in Gabon?	Discharges are governed by MARPOL legislation.	Decree 0163/PR/MEPN of 3 February 2010, Article 12, establishes authorized thresholds, based on Article 23 of the Fisheries Code. *Sources: Decree 0163, Art. 12; Gabon Fisheries Code, Art. 23.*

Table 12: Representative query–response pairs. The Multi-Agent system provides precise legal references with verifiable sources.

7 Discussion

7.1 Limitations

Computational overhead. Each query requires 2–5 LLM calls vs. 1 for the baseline (Table 13). Single-pass mode returns a response in ≈3.0 s; complex queries take a mean of 8.1 s. The multi-agent design enforces explicit separation of definition grounding, reference detection and resolution; a single-agent variant is a natural direction for reducing LLM calls while preserving most gains. **Scalability.** Validated on 50 documents / 7,943 clauses, the system is constrained by a 6,000-token context budget on the consumption side (dossier of up to 10 nodes + system prompt). Truncation occurs in ≈18% of complex multi-jurisdictional queries. **Residual error modes and generalizability.** Implicit citations account for ≈2.4% of retrieval failures. More critically, the system does not yet resolve temporal precedence (ERR-2.2). The restriction to two thematic domains and the inclusion of marine-bird documents only in retrieval (not QA) limit claims of broad generalizability. Finally, *context_precision* is excluded from RAGAS as chunk-level ground truth is unavailable.

Configuration	LLM calls	Latency (mean)
Single-pass RAG (baseline)	1	1.8 s
Multi-Agent, 1 pass	2	3.0 s
Multi-Agent, 2–3 passes	3–5	8.1 s

Table 13: Empirical latency (Llama-3.1-8B-Instruct). Wall-clock time from query receipt to full response.

7.2 Future Work

Primary targets: (1) temporal precedence module for ERR-2.2; (2) semantic context compression; (3) single-agent tool-using variant and distillation for sub-2 s latency; (4) expansion of human evaluation to ≥200 questions and additional themes + Portuguese-language African jurisdictions.

8 Conclusion

We presented a citation-disciplined architecture designed to bridge the gap between general-purpose RAG and the traceability requirements of professional legal practice, focusing on African francophone civil-law statutes. By co-designing a multi-graph knowledge layer with a recursive multi-agent orchestration loop, we have successfully addressed the critical failure modes of flat, single-pass retrieval systems. Our empirical validation on a novel African Francophone legal benchmark demonstrates transformative gains: BERTScore F1 of 0.882 and 0.807 across two thematic domains, and a citation accuracy of 93.9% confirmed by independent human evaluation—a prerequisite for deployment in high-stakes legal environments. Our structured error taxonomy provides a diagnostic roadmap, with residual challenges in temporal precedence and context-window truncation forming the basis of ongoing research. This work demonstrates that the future of legal AI lies in

"citation-first" architectures: by moving beyond text-matching toward recursive, source-grounded synthesis, we provide a concrete blueprint for trustworthy legal AI.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330.

Faraz Dadgostari, Neel Guha, et al. 2021. Modelling and analysis of legal citation networks. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law*.

Aashish Dhakal, Nabin Chaulagain, and Shashank Shree Neupane. 2026. A comparative study of information retrieval models: BM25 versus hybrid retrieval on the Cranfield collection. *International Journal of Engineering Technology Research & Management (IJETRM)*, 10(01):164–167.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph RAG approach to query-focused summarization. In *arXiv preprint arXiv:2404.16130*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, et al. 2023. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 44123–44279.

Sirui Hong, Mingchen Zhuge, Jiaqi Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.

Abe Bohan Hou, Orion Weller, Guanghui Qin, Eugene Yang, Dawn Lawrie, Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2025. CLERC: A dataset for U.S. legal case retrieval and retrieval-augmented analysis generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7913–7928.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems*, volume 38.

LangChain, Inc. 2024. LangGraph: Building stateful, multi-actor applications with LLMs.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

LlamaIndex. 2024. LlamaParse: Advanced PDF parsing for RAG.

Annie Louis, Ari Holtzman, Aditya Siddhant, Jonathan Clark, and Sebastian Riedel. 2022. Statutory reasoning assessment with a challenging dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10507–10527.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037.

Boyang Zhang, Sebastián G. Acosta, Preston Carlson, Sacha Bron, Pierre-Loïc Doulcet, Daniel B. Ospina, and Simon Suo. 2026. ParseBench: A document parsing benchmark for AI agents.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyu Zhang, Zhiyuan Liu, and Maosong Sun. 2020. JEC-QA: A legal-domain question answering dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9498–9505.

FLICK: Few-Label Incremental Learning for Low-Resource Dialects

Ali Almutairi[1], **Abdullah Alsuhaibani**[2], **Shoaib Jameel**[3],
Aditya Joshi[1], Gelareh Mohammadi[1], Imran Razzak[4]
[1]University of New South Wales, Australia, [2]University of Technology Sydney, Australia
[3]University of Southampton, UK, [4]MBZUAI, UAE
ali.almutairi@unsw.edu.au

Abstract

Low-resource (LR) languages continue to face persistent challenges in natural language processing due to limited labelled datasets, scarce online textual resources, and the lack of NLP tools. Pre-trained language models and fine-tuning offer a promising starting point, yet their performance remains constrained by the scarcity of reliable annotations. Recent works have explored intermediate training with clusters and pseudo-labels to mitigate the limitation of few-label in fine-tuning. However, relying on cluster assignments or repeated can degrade performance, particularly for linguistically complex languages such as Arabic. To address these challenges, we propose FLICK, a framework for intermediate fine-tuning in few-label learning. FLICK first fine-tunes a language-specific encoder on a held-out split of pseudo-labelled data and retains only the Top-R clusters whose labels are reliably predicted; then, we train the model on these refined pseudo-labels. The resulting model is then transferred to the final few-label fine-tuning stage. Experiments across seven Arabic datasets spanning sentiment, sarcasm, dialect, and offensive language detection show that FLICK achieves the highest Macro-F1 scores among the evaluated Arabic BERT baselines and state-of-the-art few-label methods.

1 Introduction

Modern natural language processing (NLP) techniques have achieved notable success when trained on a large volume of labelled data (Mirończuk and Protasiewicz, 2018), typically comprising tens of thousands of examples. However, the limited availability of both labelled and unlabeled data remains a major barrier to applying these models effectively, particularly in the case of low-resource languages. This data scarcity further exacerbates linguistic inequality in terms of access to technology for many language communities. While there is sufficient data for some well-studied tasks in some low-resource languages, many tasks still suffer from poor data availability. When labelled data are scarce, these complex neural networks tend to underperform (LaBonte et al., 2024). Arabic is a representative and challenging case: even large language models underperform on Arabic relative to English, owing to limited training data, wide variation between Modern Standard Arabic and regional dialects, and complex morphology (Al-Matham et al., 2025); dialectal Arabic in particular is classified as a low-resource setting (Khered et al., 2025; Alshabanah and Annavaram, 2025). These constraints make the few-label setting, in which only a small set of labels is available alongside a large pool of unlabelled text, both for Arabic and for low-resource text classification more broadly. This has resulted in semi-supervised learning (SSL) approaches, which either augment existing labelled data or employ unlabeled data to alleviate these limitations, as unlabeled data is generally easier and more cost-effective to obtain.

SSL methods often rely on Neural Machine Translation (NMT) systems for data augmentation through back-translation (Rastgoo et al., 2021), where sentences in the training set are translated into another language and then back into the original language to create a dataset with variations in text. However, these approaches heavily rely on quality translation models, which are often limited or do not perform well for many low-resource languages. Moreover, if the data distribution of the specific task differs from that used to pre-train the NMT system, the quality of the generated sentences may suffer (Shen et al., 2021). Furthermore, the phenomenon of 'translationese' underscores the ways in which translated texts may occasionally lose cultural nuances (Gellerstam, 1986), leading to a linguistic representation that can appear less natural and culturally embedded compared to the

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 326–338
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_034

original language. This divergence from the source material raises important considerations regarding the fidelity and authenticity of translated works in conveying the richness of cultural context. A recent study shows that such cases affect the training model stage (Yu et al., 2022).

An alternative set of approaches involves few-shot learning (FSL), which involves selecting k samples from each class (*N-way-K-shot*) (Wang et al., 2020a) and providing them as input to decoder models. FSL has been found to be useful when labelled data is scarce (Lu et al., 2023; Alsuhaibani et al., 2024; Xu et al., 2025). However, using FSL in inference models, particularly with BERT-based models or LLMs, presents certain challenges. The former is often limited by its smaller model size and parameters, while the latter, despite its greater knowledge, incurs high computational costs. Furthermore, both types of models may struggle with LR languages, as they have primarily been developed based on high-resource (HR) languages.

Nevertheless, fine-tuning Bidirectional Encoder Representations from Transformers (BERT), such as the BERT model (Devlin et al., 2019a) or AraBERT (Antoun et al., 2020a) [1], helps address the limited label challenge, but its performance remains constrained (Goyal et al., 2023). An alternative approach is self-training, where pseudo-labels (PL) are generated for unlabeled data, which are then treated as labelled data for subsequent training. Traditional self-training methods typically combine PLs noise with real labels in iterative training. To address this, (Shnarch et al., 2022) propose $Cluster\&Tune$, introducing an intermediary task to enhance the fine-tuning performance of BERT. Their approach involves leveraging unsupervised learning, which generates PL that are subsequently utilised as inputs for fine-tuning. Another notable contribution in this area is IDoFew (Alsuhaibani et al., 2024), which incorporates dual clustering, with the second cluster serving as a correction mechanism. However, despite the elegance and efficacy of these methodologies, their applicability to LR languages remains largely unexplored. In particular, the effectiveness of fine-tuning with a limited number of labelled examples, especially in the context of the Arabic language, remains an open question. Moreover, these methods improve cold-start performance but were designed for English and rely on multiple, pre-set clusters, which adds cost without addressing the central weakness of inter-training in low-resource settings, the noise in the PL themselves.

Our starting observation is that not all PL clusters are equally trustworthy, and that using them indiscriminately is actively harmful. In our experiments, fine-tuning on every pseudo-labelled cluster yields a worse model than skipping the intermediate stage and fine-tuning directly on the 100 real labels. The problem is therefore not whether to use PL, but which PL to use. FLICK addresses this with a PL-refinement component that acts as a quality gate. From an initial K-means clustering of unlabelled text (Stage 1), FLICK fine-tunes a language-specific encoder (ARBERTv2; Abdul-Mageed et al., 2021) on a held-out split of the PL and ranks clusters by how reliably that encoder predicts their labels, retaining only the Top-R (Stage 2). The retained clusters train an intermediate model whose weights initialise the final few-label fine-tuning (Stage 3). Crucially, FLICK filters clusters by a learnability signal rather than adding clusters, distinguishing it from the multi-cluster cascades of prior work.

Our key contributions are as follows: **(1)** Our key contributions are as follows. (1) We introduce FLICK, a three-stage few-label framework whose core is a predictability-based Top-R filter: clusters are retained only if a language-specific encoder can reliably reproduce their pseudo-labels. FLICK therefore removes unreliable clusters rather than adding clustering stages, which is what distinguishes it from the multi-cluster cascades. **(2)** We present a systematic study of cluster-based intermediate learning for Arabic and its dialects, spanning seven datasets and six task types, with a matched-protocol comparison against nine BERT-family backbones, two SOTA few-label pipelines re-run with an Arabic backbone, and five LLMs including a fine-tuned GPT-4o. **(3)** We report ablations that isolate every component, including the finding that inter-training on all pseudo-labelled clusters is marginally worse than skipping inter-training entirely, and a selection-criterion ablation showing that predictability outperforms random, centroid-distance, cluster-size and entropy selection by 5.7–9.9 mean percentages. We offer this as a template for evaluating multi-stage few-label frameworks. **(4)** FLICK (small encoder) surpasses

[1] A language representation model for Arabic, based on Google's BERT

every open-source LLM baseline in our experiment and exceeds fine-tuned GPT-4o on both dialect-grounded pragmatic tasks, with no API dependency and on a single consumer GPU. These findings validate the effectiveness of our PL refinement component in mitigating PL noise and achieving robust classification performance where traditional multi-cluster approaches falter due to data scarcity and linguistic peculiarities (See Tables 2-5).

2 Related Work

Despite its ubiquity, data is often unlabelled or sparsely labelled, which limits the effectiveness of fully supervised learning approaches. To address this challenge, various methods have been developed that leverage a small number of labels from diverse perspectives relevant to the specific task (Alsuhaibani et al., 2024). This is known as the 'FL' setting, and implies that a small portion of the dataset is labelled for a given task (Wang et al., 2020b), such as text classification (Liang et al., 2024). The concept of leveraging the "few labels" method has proven advantageous in addressing different issues such as cyberbullying, cybercrimes, hate speech, spam, phishing, and offensive content detection , primarily due to the scarcity of annotated labels in these domains and the fast-evolving nature of the issues themselves. Consequently, the Few-label (FL) techniques have gained widespread adoption across various domains and data types, encompassing both image and text data.

Researchers have addressed the scarcity of fully annotated datasets challenge through various methods, including synthetic data generation (Jiang et al., 2023; Xu et al., 2025), semi-supervised learning (Lee et al., 2024; Das et al., 2025), transfer learning (Pan et al., 2019; Pham et al., 2024), and few-shot learning (Liu and Yang, 2024; Wang et al., 2025), each of which has shown promising results. However, employing these methods individually has revealed inherent limitations, thus creating ample space for improved solutions.

Although synthetic data methods can produce PLs, they may introduce artificial noise. Additionally, semi-supervised learning relies on the entire set of generated PLs, potentially leading to inefficient training by incorporating classes with limited observations, thus compromising the model's effectiveness. Furthermore, the main challenge with transfer learning in FL low-resource settings is that fine-tuning foundational models on extremely limited labelled data often leads to overfitting and poor generalisation. Lastly, FSL may not leverage the latent knowledge in unlabeled data and inherit the limitations posed by BERT-based models and LLMs.

Intermediate/Incremental Learning (IL) has emerged as a promising methodology for addressing the challenges of FL Arabic text classification tasks. IL offers various approaches to tackle this issue. Shnarch et al. (2022) introduced the Cluster and Tune framework, which has been recognised as SOTA in the scientific community. Additionally, Alsuhaibani et al. (2024) developed an optimised version of the previous architecture called IDoFew, which surpassed the results of their predecessor on the same benchmarks. However, their use of multiple clusters slowed down the model and incurred additional costs.

Many of the above techniques are primarily used for HR languages, although some have been adapted for LR languages such as Arabic and its widely spoken dialects. (AlYami and Al-Zaidy, 2022) employed a semi-supervised technique to tackle the limited labelled data issue within the context of Arabic dialects, utilising two datasets: (1) Arabic Dialect Short Text and (2) Arabic Dialect Dictionary. They obtained significant results by incorporating the AraRoBERTa model. Additionally, (Ameur et al., 2023) demonstrated that a semi-supervised approach achieved 91% accuracy in sarcasm detection using a limited labelled dataset. However, accuracy alone may not suffice as an evaluation metric, especially in cases of dataset imbalance. Furthermore, IL could potentially enhance the current state, particularly in LR languages like Arabic.

Our approach markedly diverges from the aforementioned techniques. The Cluster&Tune framework, which has become acknowledged as a prominent method in the academic field. However, it's worth noting that they employed a fixed value for K across all tasks. This may not be adequate for other languages such as Arabic, as varying characteristics and the embedding space, which K relies on, can differ significantly. Moreover, the IDoFew framework (Alsuhaibani et al., 2024) increase the number of models used during inter-training and introduces a fraction of training instances in the subsequent stage, thus boosting performance and contributing to SOTA outcomes. Nevertheless, their inter-training still involved complex methods such

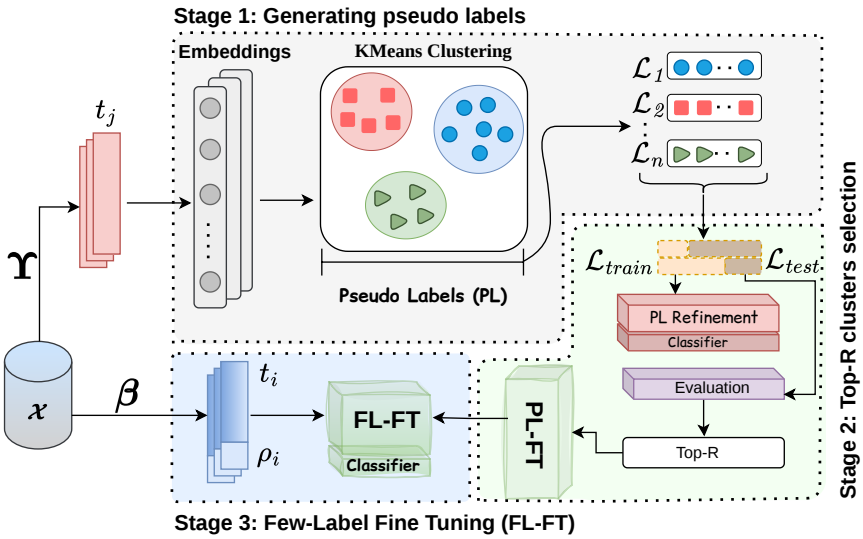


Figure 1: Our novel FLICK framework.

as the use of several clusters and was limited to English data. Our method, however, is designed not only for effectiveness but also for speed, potentially outpacing the IDoFew model. Moreover, we have refined the concept of fraction random sampling by selecting the top pseudo-labelled samples for ongoing learning.

Furthermore, compared to previous studies, we focused on LR languages such as Arabic, which are more difficult to model than English. Some challenges include script complexity, morphological complexity, ambiguity and homography, data scarcity and quality, and dialect variation. We enhance the IDoFew architecture by improving clusters by integrating a PL Refinement component, functioning as a BERT enhancer, implemented through the utilisation of language-related BERT models such as the ARBERTv2 model to select higher quality PLs. Unlike semi-supervised labelling, we select only the top PL instead of the entire set. Furthermore, by training the top PL in the PL-FT step on the base BERT, such as ARBERTv2 and transferring this knowledge to the FL-FT model, we eliminate the assumptions of traditional transfer learning and enhance the original BERT's knowledge for the specific task at hand.

3 FLICK Framework

We tackle the issue of learning with FL, especially in LR languages such as Arabic, in a text classification setting. In doing so, we not only confront the difficult task of modelling with a limited number of labels but also take into account datasets with scarce resources. This presents a greater challenge compared to previous works such as (Alsuhaibani et al., 2024), given the nature of the dataset. For example, employing machine learning to model the Arabic language introduces unique challenges due to its complex linguistic structure and cultural subtleties. Arabic is a richly inflected language, where words acquire new meanings by adding prefixes, suffixes, and internal modifications. Arabic also exhibits significant diversity in terms of the dialects of Arabic. FL learning necessitates a minimal amount of labelled data, which poses a challenge for Arabic, given the scarce annotated datasets. The limited data may also suffer from an imbalanced label distribution, with certain classes having far fewer examples than others, impeding the model's learning effectiveness. Additionally, the multitude of Arabic dialects adds complexity to FL learning, as the model must be able to generalise to unfamiliar dialects with scant training data.

In our setting, the dataset for training a model contains significantly fewer labelled examples than ideal. This usually can lead to suboptimal performance due to insufficient information to learn patterns and relationships. Our FLICK framework addresses this by employing a novel three-stage approach: (1) PL generation and initial Clustering (Stage 1), where we harness unlabeled data to create initial semantic groupings; and (2) PL refinement and intermediate fine-tuning (Stage 2), where higher-quality PL are distilled and used to train a

Algorithm 1 Top-R Extraction and Selection

Input: Pseudo labels clusters
Output: Top pseudo label clusters
Divide: The input into training and test datasets:
    Let $\mathcal{L}$ the set of data points with pseudo labels.
    Training set $\mathcal{L}_{\text{train}} = \{\iota_1, \ldots, \iota_m\}$
    Testing set $\mathcal{L}_{\text{test}} = \{\iota_{m+1}, \ldots, \iota_n\}$
Feed: $\mathcal{L}_{\text{train}}$ into the language-related BERT model for initial training:
    Optimize the model parameters θ by minimizing the loss function ℓ, over $\mathcal{L}_{\text{train}}$
Feed: $\mathcal{L}_{\text{test}}$ into the trained language-related BERT model to refine and evaluate:
    Compute predictions $\hat{y}_i = \text{BERT}_\theta(\iota_n)$ for each ι_n in $\mathcal{L}_{\text{test}}$
    Evaluate model performance based on metrics such as accuracy and loss
Identify: top clusters from $\mathcal{L}_{\text{test}}$:
    Rank clusters based on correctly predicted culsters
    Select top performing clusters, denoted as Top-R
Return: Top-R as the final output

powerful intermediate classifier. This intermediate learning phase is crucial for enhancing the model's generalisability and enriching the base knowledge of the fine-tuned model before exposure to a limited set of real labels. We have depicted the high-level overview of our model in Figure 1.

As shown in the Figure 1, the input is a text dataset whose semantic representations are obtained using a language model. Specifically, each sentence is encoded into a low-dimensional embedding representation that captures its semantic content. These representations are then passed into a K-means clustering model that semantically groups them into distinct clusters, from which PL are generated. These PL are then used to fine-tune an existing language-specific pre-trained classifier, PL-FT. The classifier is designed to predict the correct PL for new, unseen data. The classifier's performance is continuously evaluated and improved. This involves adjusting the model's parameters to enhance its accuracy. Initially, K-means clustering (with K fine-tuned per dataset) groups these embeddings into distinct pseudo-labelled clusters. FLICK then employs a novel PL refinement component (detailed in Algorithm 1) where the classifier's performance is evaluated per cluster. Based on this evaluation, the model's predictions are refined by training the model on the samples from the Top-R clusters. Finally, the PL-FT learnt weights are then used to initialise FL-FT, which will be fine-tuned on the real labels.

The mathematical formulation is as follows. Given a dataset $\mathcal{X}$ defined by two subsets: β and Υ. The subset β is a supervised set $\{(t_i, \rho_i)\}_{i=1}^{\varepsilon}$, where $t_i = \{t_1, t_2, t_3, \ldots, t_\varepsilon\}$ represents the text, and $\rho_i \in \{1, 2, 3, \ldots, Z\}$ denotes the corresponding labels for each t_i. Each text t_i is described by λ-dimensional vectors. The labels ρ_i can represent either binary or multiclass classification. The subset Υ represents the unlabeled data, denoted by $\{t_j\}_{j=1}^{\nu}$, containing text instances with no labels. It is worth mentioning ε is much less than ν. Our objective is to harness the potential of unlabeled datasets through the implementation of the incremental method, resulting in a viable solution for LR languages and FL fine-tuning challenges.

3.1 Architecture

FLICK employs two distinct algorithms to generate and refine PL. This approach minimises the risk of inadequate pseudo-labelling, which is particularly pertinent for Arabic texts where semantic content can be highly ambiguous. For example: Input Text [2]: لا يشيخ قبل الكوره ما تطلع من رجله ; Label: تهكم and Input Text [3]:لا يشيخ لاوالله قهرك عشان اخذها ; Label: ليست تهكم . The first example is ironic because it mocks the player's poor performance, whereas the second example expresses a straightforward, literal explanation with no sarcasm or hidden criticism.

Existing methods such as Cluster&Tune (Shnarch et al., 2022) and IDoFew (Alsuhaibani et al., 2024) misclassify both examples as **لا يشيخ** . FLICK involves two key training stages: (1) PL Generation: where the unlabeled dataset is clustered (e.g., into 20 initial clusters) to generate PLs; (2) PL Refinement and Intermediate Fine-Tuning, PL-FT: Given that noisy PL or an excessive number of unreliable clusters can compromise performance, this stage focuses on identifying and leveraging the top-performing clusters as a highly representative subset of the dataset using a language-specific BERT model. This process enables the intermediate classifier PL-FT to acquire robust representations and is detailed in the following. Our training protocols consist of three key stages. Stage 1, Stage 2 (described in Algorithm 1), and Stage 3.

Stage 1: Generating pseudo labels: Our goal is to generate distinct clusters of PL with consistent patterns from unlabelled data using embedding and k-means clustering algorithms, which yield more reliable PL for subsequent analysis. The unlabelled data is fed into the BERT-based model known as paraphrase-multilingual-MiniLM-L12-v2 for vector representation. The paraphrase-

[2]Text: "Bro, before the ball even leaves his foot, he already messes it up.", Label: *"Ironic"*

[3]Text: "Bro, no, I swear he annoyed you because he took the ball.", Label: *"Not Ironic"*

Dataset	Train	Test	C	Task
ArBNTopic	15,827	1,583	14	Topic
Ar-Sentiment	8,437	2,110	3	Sentiment
Ar-Sarcasm	8,437	2,110	2	Sarcasm
Sa'7r	3,959	1,187	2	Irony
DART	972	455	5	Dialect
Ar-Dialects	8,437	2,110	5	Dialect
MPOLD	3,400	600	2	Offensive

Table 1: Arabic evaluation datasets. C = number of classes.

multilingual-MiniLM-L12-v2 model, accessed via the sentence-transformers library (Reimers and Gurevych, 2019), is capable of producing embeddings for sentences in multiple languages, resulting in dense vector representations. We use this encoder because it produces sentence-level embeddings for Arabic without task supervision and is cheap enough to run over the full unlabelled pool; FLICK places the language-specific model in Stage 2, so Stage 1 needs only to produce a partition whose parts are worth ranking. We compared k-means against various clustering algorithms and found it consistently strongest; it is also the choice made by Cluster&Tune and IDoFew, which keeps our comparison on common ground.

In the K-means algorithm, when the K centroids converge to the nearest point within the clusters, PL are assigned to each cluster. We refine this process by selecting the Top-R performing clusters and discarding the rest, which results in more accurate PL and the removal of noisy clusters. The cluster count k is set per dataset and does not assume knowledge of the number of task classes, which is unavailable in a genuine few-label setting; Stage 2 therefore operates in the cluster space, and the mapping to real classes is learned only in Stage 3.

Stage 2: Top-R clusters extraction and selection: This stage, encompassing the core intermediate learning of FLICK, refines the initial PL generated in Stage 1 and trains an intermediate classifier. The process involves two key steps:

- From the pseudo-labelled data (generated in Stage 1), we create an internal training set $\mathcal{L}_{\text{train}}$ and an internal testing set $\mathcal{L}_{\text{test}}$. A vanilla language BERT model (e.g., ARBERTv2 for Arabic) is initially fine-tuned on $\mathcal{L}_{\text{train}}$. This model then is evaluated on $\mathcal{L}_{\text{test}}$, and the clusters are ranked based on the Top-R correctly predicted clusters on $\mathcal{L}_{\text{test}}$. The Top-R performing clusters, denoted as Top-R, are then identified as the most reliable subset of pseudo-labelled data.
- The selected Top-R pseudo-labelled clusters are then used to further fine-tune a language-specific BERT model (e.g., ARBERTv2 for Arabic) to create our PL-FT model. This training step may solidify the model's understanding of the semantic space represented by higher-quality PLs, enhancing its generalisability. This is the continual learning phase where the model significantly augments its knowledge base. The weights learned by the PL-FT model serve as crucial initialisation for the final FL fine-tuning.

Stage3: Few-Label Fine-Tuning (FL-FT): The knowledge acquired during the earlier learning PL-FT phase has allowed for effectively utilising a small set of labelled data in resource-constrained environments. We employed multiple datasets from various domains and fine-tuned the model separately on each dataset for their corresponding downstream tasks. Finally, the weights obtained from the previous stage (PL-FT model) are used in the FL-FT model to further fine-tune on a few true labels.

4 Experiments and Results

4.1 Experimental Setup

We evaluate FLICK on seven publicly available Arabic classification datasets, which include (1) ArBNTopic (Albared et al., 2023) (2) Ar-Sentiment (Abu Farha and Magdy, 2020) (3) Ar-Sarcasm (Abu Farha and Magdy, 2020) (4) The Saudi Dialect Irony (Sa'7r) (AlMazrua et al., 2022) (5) Dialectal Arabic Tweets (DART) (Alsarsour et al., 2018) (6) Ar-Dialects (Abu Farha and Magdy, 2020) (7) Multi Platforms Offensive Language Dataset (MPOLD) (Chowdhury et al., 2020). The Sentiment, Sarcasm, and Dialects datasets were derived from the ArSarcasm dataset. The datasets were deliberately selected to evaluate FLICK across diverse classification settings, encompassing topical and non-topical tasks, varying dataset sizes, different dialectal phenomena, binary and multi-class prediction problems, and varying levels of class imbalance. This diversity enables us to assess the robustness and generality of the method under a broad range of realistic Arabic NLP conditions. These datasets span topic, sentiment, sarcasm, dialect, and offensive-language detection (Table 1). Because several datasets are class-imbalanced, we report Macro-F1. To simulate the few-label (FL)

Model	ArBNTopic	Ar-Sent.	Ar-Sarc.	Sa'7r	DART	Ar-Dial.	MPOLD	Mean
Zero-shot								
ArabicBERT	3.27	24.68	45.77	32.63	5.30	2.72	46.12	22.93
ARBERTv2	2.18	11.41	47.56	32.92	10.77	7.88	45.57	22.61
AraBERT-twitter	4.14	25.02	46.76	38.52	7.40	10.14	50.51	26.07
MARBERTv2	5.05	20.17	50.40	51.74	10.77	6.03	45.66	27.12
BERT-Based Models (supervised fine-tuning (SFT) on 100 labels)								
ArabicBERT	25.47	52.43	58.52	65.83	32.92	23.80	56.39	45.05
ARBERTv2	45.17	61.00	58.33	66.28	37.66	36.80	62.65	52.56
AraBERT-twitter	37.01	54.07	53.29	66.36	47.53	26.00	50.14	47.77
AlcLaMv2	1.01	38.78	54.38	42.13	7.65	16.02	65.34	32.19
CamelBERT-mix	25.78	57.10	53.59	62.41	22.41	26.98	49.61	42.55
CamelBERT-MSA-mix	20.96	50.54	55.24	61.05	43.08	30.05	64.27	46.46
MARBERTv2	29.78	46.97	56.80	65.18	33.97	25.02	55.56	44.75
BERT	3.13	41.02	53.67	53.48	24.03	20.61	57.87	36.26
DistilBERT	3.43	40.53	49.86	56.83	7.05	21.25	57.87	33.83
SOTA Few-Label Methods (SFT on 100 labels)								
Cluster&Tune	1.82	40.73	44.86	46.45	14.32	11.92	45.59	29.38
Cluster&Tune$^{*}_{\text{ARBERTv2}}$	7.48	39.79	49.09	62.97	26.24	8.41	45.42	34.20
IDoFew	4.52	42.26	49.81	64.28	13.53	21.04	55.96	35.91
IDoFew$^{*}_{\text{ARBERTv2}}$	56.22	61.06	64.84	65.92	60.77	42.62	54.82	58.04
FLICK (ours)	**61.52**	**65.02**	**69.24**	**67.96**	**64.70**	**44.33**	**74.46**	**63.89**

Table 2: Macro-F1 (%) against BERT-based models and SOTA few-label methods on seven Arabic datasets under low-resource settings, including a zero-shot (no fine-tuning) reference for four Arabic backbones. **Bold** = best, underline = second-best per column. * = re-run under our matched protocol with the ARBERTv2 backbone. ARBERTv2 is the FLICK backbone.

Model	Params	ArBNTopic	Ar-Sent.	Ar-Sarc.	Sa'7r	DART	Ar-Dial.	MPOLD	Mean
LLMs (SFT on 100 labels)									
ALLaM-7B-Instruct	7B	46.96	65.34	67.38	63.03	17.98	37.72	71.77	52.88
LLaMA 3.1-8B	8B	1.00	14.00	46.00	34.46	5.23	16.02	13.98	18.67
AceGPT-16B	16B	22.92	50.07	57.33	59.08	27.63	28.60	58.69	43.47
Qwen2.5-72B-Instruct	72B	1.24	22.54	14.05	34.11	1.71	7.40	13.81	13.55
GPT-4o	N/A	**68.79**	**72.47**	67.14	65.18	**66.49**	**48.63**	**81.00**	**67.10**
FLICK (ours)	$\approx$ 110M	61.52	65.02	**69.24**	**67.96**	64.70	44.33	74.46	63.89

Table 3: Macro-F1 (%) against LLMs. **Bold** = best, underline = second-best per column.

regime, we retain only 100 real labels per dataset for the final fine-tuning stage; all remaining training text is treated as unlabelled and used solely for pseudo-label generation. For more details on the datasets, see C. Similarly, we evaluate across a diverse set of BERT-based models (BBMs) and LLMs backbones, including Arabic-specific, multilingual, and English-centric models, allowing us to examine whether FLICK's benefits are consistent across models with different pre-training corpora, vocabulary coverage, parameter scales, and levels of Arabic-language specialisation. All BBMs and FLICK use an identical protocol: the same optimiser (Adam), learning rate (3e$-$5), batch size, epoch schedule, label split, and random seed. The only parameters exclusive to FLICK are the cluster count K, the retained-cluster count R (Top-R), and the internal PL train/test split. For the SOTA methods (Cluster&Tune, IDoFew), we report their original published configurations and variants re-run under our matched protocol with the ARBERTv2 backbone, so that gains cannot be attributed to backbone or tuning effort. All open-source LLM baselines are fine-tuned under an identical protocol via direct supervised fine-tuning on the same 100 labels; GPT-4o is fine-tuned through its provider API.

In Appendix B, we report the shared and per-dataset hyperparameters used in the FLICK matched-parameter experiments (Table 6).

4.2 Results and Analysis

Table 2 illustrate results against BBMs and SOTA few-label methods. FLICK attains the highest

Variant	ArBNTopic	Ar-Sent.	Ar-Sarc	Sa'7r	DART	Ar-Dial	MPOLD	Mean
FLICK (full)	**61.52**	**65.02**	**69.24**	**67.96**	**64.70**	**44.33**	**74.46**	**63.89**
w/o Top-R selection	56.56	58.26	60.47	48.68	40.79	33.49	68.58	52.40
w/o PL-FT	45.17	61.00	58.33	66.28	37.66	36.80	62.65	52.56
w/o FL-FT	3.55	13.99	45.48	48.00	8.86	10.23	45.64	25.11
w/o PL-FT & FL-FT	2.18	11.41	47.56	32.92	13.18	7.88	45.57	22.96

Table 4: Component ablation (Macro-F1 %): w/o Top-R selection (all clusters), w/o PL-FT (direct SFT on 100), w/o FL-FT (PL-FT only, no real labels), and w/o PL-FT & FL-FT (zero-shot)

Selection criterion	ArBNTopic	Ar-Sent.	Ar-Sarc.	Sa'7r	DART	Ar-Dial.	MPOLD	Mean
Ours	**61.52**	**65.02**	**69.24**	**67.96**	**64.70**	**44.33**	**74.46**	**63.89**
Random Top-R	54.98	57.99	64.45	66.68	51.98	37.29	73.87	58.18
Nearest-to-centroid	54.61	60.88	59.14	65.49	51.49	39.71	66.15	56.78
Largest cluster size	53.81	62.62	58.81	66.27	27.68	38.01	71.04	54.03
Lowest entropy	55.74	60.62	61.00	66.39	26.72	37.89	69.60	53.99
No selection	56.56	58.26	60.47	48.68	40.79	33.49	68.58	52.40

Table 5: Top-R selection-criterion ablation (Macro-F1 %).

Macro-F1 on all seven datasets and mean (63.89). To isolate the contribution of the FLICK pipeline, we compare against its own backbone fine-tuned directly on the 100 real labels (ARBERTv2, 52.56), which is also the strongest BBM by mean. FLICK improves this backbone by +11.3 mean points, with all per-dataset deltas measured against the same baseline: the largest gains are on the dialect task DART (+27) and the 14-class ArBNTopic (+16.4), followed by MPOLD (+11.8), Ar-Sarcasm (+10.9), Ar-Dialects (+7.5), Ar-Sentiment (+4), and Sa'7r (+1.7). Task-specialised backbones do not close this gap: MARBERTv2 and AraBERT-twitter, despite in-domain social-media pretraining, trail FLICK on every task. English-only BERT/DistilBERT collapses on the 14-class ArBNTopic (3.1/3.4) indicates that cross-lingual transfer without Arabic pretraining is inadequate for multi-class Arabic classification.

Against SOTA few-label methods, the English-backbone Cluster&Tune (29.4) and IDoFew (35.9) fall below every Arabic BBM, showing that English-designed pipelines do not transfer to Arabic without a language-appropriate backbone. Re-equipped with ARBERTv2 under our matched protocol, IDoFew rises to its strongest configuration (58.0), yet FLICK still exceeds it on all seven datasets, by (+5.9) mean points and by (+19.6) on MPOLD. Because FLICK and this variant share the same backbone, embeddings, and tuning budget, the improvement is attributable to the PL-refinement component, not to the encoder.

Table 3 situates FLICK against five LLMs fine-tuned on the same 100 labels. FLICK (110M) surpasses all open-source LLMs by wide margins: +11.0 over the strongest open Arabic model ALLaM-7B (52.9), +20.4 over AceGPT-16B, and +45-50 over LLaMA 3.1-8B and Qwen2.5-72B, whose Macro-F1 collapses on several tasks (e.g. ArBNTopic 1.0/1.2). Parameter count is thus a poor predictor of few-label Arabic performance: a compact, language-specific encoder with targeted PL refinement outperforms a 72B general model by nearly 50 points. Only fine-tuned GPT-4o exceeds FLICK on average (67.1 vs. 63.9). Crucially, GPT-4o's edge might be confined to tasks with comparatively explicit semantic cues (topic, sentiment, offensive language), whereas FLICK beats GPT-4o on the two culturally-embedded pragmatic tasks: Ar-Sarcasm (+2.1, predominantly Egyptian dialect) and Sa'7r (+2.8, Saudi dialect irony). This pattern suggests that dialect-specific pragmatic inference, underrepresented in proprietary pretraining, is precisely where a small, language-aligned model with refined in-domain PL remains competitive with the frontier, while requiring no API access, no per-query cost, and a single consumer GPU.

5 Ablation Study

5.1 Does each stage contribute?

Table 4 dismantles FLICK stage by stage. The zero-shot ARBERTv2 backbone scores 23.0. Direct supervised fine-tuning on the 100 real labels (w/o PL-FT) raises this to 52.6, the natural strong baseline. Full FLICK reaches 63.9, a (+11.3) gain over direct SFT that is consistent across all seven tasks, confirming that the intermediate PL stage

contributes beyond what real labels alone provide.

The most informative row is w/o Top-R: applying PL-FT over all clusters yields 52.4, marginally below direct SFT (52.6). Training indiscriminately on every pseudo-labelled cluster might not merely be unhelpful; it produces a slightly worse initialisation than skipping the stage. The transition from 52.4 (unfiltered) to 63.9 (Top-R) is the single largest jump in the pipeline.

5.2 Why predictability-based selection?

Table 5 provides the empirical test. We replace predictability-based Top-R with four alternative criteria: random selection, nearest-to-centroid, largest cluster size, and lowest prediction entropy, holding K, R, backbone, and budget fixed. Predictability-based selection (63.9) outperforms every alternative by 5.7-9.9 mean points: random 58.2, centroid 56.8, size 54.0, entropy 54.0. The decisive observation is that the entropy criterion, which selects the lowest uncertainty, is the lowest of the learned criteria. We conclude that what a language-specific encoder captures, whether a cluster's pseudo-labels are consistently learnable by an Arabic model, might be a substantively different and stronger signal than confidence, geometry, or cluster mass. We interpret this as the encoder admitting clusters whose PL fall in linguistically coherent regions of the representation space, rather than clusters that are merely compact, large, or low-entropy.

6 Conclusion

We introduced a novel FLICK framework designed to support more inclusive technologies for low-resource language/dialect communities and address linguistic inequality caused by limited labelled data via continual label learning with a particular focus on Arabic and its dialects data. FLICK employs selective pseudo-labels to mitigate noise, focusing on choosing the most pertinent examples. Our extensive experiments show a marked enhancement in performance across low-resource settings that feature a dearth of labelled data, spanning diverse classes, tasks, data sizes, and both topical and non-topical settings.

Limitations

All results come from a fixed seed, shared by FLICK and every baseline; comparisons are therefore seed-matched, but we do not characterise the variance of the pipeline itself. Moreover, the quality and quantity of available Arabic datasets critically affect model performance, potentially introducing biases or impeding generalisation to new domains, and the opaqueness of deep learning models together with the risk of propagating training data biases necessitates careful consideration of fairness and responsible application.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *COLING*, pages 7088–7105, Online.

Ibrahim Abu Farha and Walid Magdy. 2020. From Arabic sentiment analysis to sarcasm detection: The ArSarcasm dataset. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 32–39, Marseille, France.

Murtadha Ahmed, Saghir Alfasly, Bo Wen, Jamal Addeen, Mohammed Ahmed, and Yunfeng Liu. 2024. AlclaM: Arabic dialect language model. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 153–159.

Rawan Al-Matham, Kareem Darwish, Raghad Al-Rasheed, Waad Alshammari, Muneera Alhoshan, Amal Almazrua, Asma Al Wazrah, Mais Alheraki, Firoj Alam, Preslav Nakov, Norah Alzahrani, Eman Albilali, Nizar Habash, et al. 2025. BALSAM: A platform for benchmarking Arabic large language models. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, pages 258–277. Association for Computational Linguistics.

Doha Albared, Hadi Hamoud, and Fadi Zaraket. 2023. Arabic topic classification in the generative and AutoML era. In *Proceedings of ArabicNLP 2023*, pages 399–404.

Halah AlMazrua, Najla AlHazzani, Amaal AlDawod, Lama AlAwlaqi, Noura AlReshoudi, Hend Al-Khalifa, and Luluh AlDhubayi. 2022. Sa'7r: A saudi dialect irony dataset. In *Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 60–70.

Israa Alsarsour, Esraa Mohamed, Reem Suwaileh, and Tamer Elsayed. 2018. DART: A large dataset of dialectal Arabic tweets. In *LREC 2018*.

Abdulla Alshabanah and Murali Annavaram. 2025. On using Arabic language dialects in recommendation systems. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2178–2186, Albuquerque, New Mexico. Association for Computational Linguistics.

Abdullah Alsuhaibani, Hamad Zogan, Imran Razzak, Shoaib Jameel, and Guandong Xu. 2024. Idofew: Intermediate training using dual-clustering in language models for few labels text classification. In *WSDM*, page 18–27.

Reem AlYami and Rabah Al-Zaidy. 2022. Weakly and semi-supervised learning for Arabic text classification using monodialectal language models. In *WANLP*, pages 260–272.

Asma Ameur, Sana Hamdi, and Sadok Ben Yahia. 2023. Domain adaptation approach for arabic sarcasm detection in hotel reviews based on hybrid learning. *Procedia Computer Science*, 225:3898–3908. KES.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020a. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020b. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham Abdullah Alyahya, Sultan AlRashed, Faisal Abdulrahman Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Saad Amin Hassan, Dr. Majed Alrubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, Ahmed Abdelali, Jeril Kuriakose, Abdalghani Abujabal, Nora Al-Twairesh, Areeb Alowisheq, and Haidar Khan. 2025. ALLam: Large language models for arabic and english. In *The Thirteenth International Conference on Learning Representations*.

Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Abdelali, Soon-gyo Jung, Bernard J. Jansen, and Joni Salminen. 2020. A multi-platform Arabic news comment dataset for offensive language detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6203–6212.

Trisha Das, Afrah Shafquat, Mandis Beigi, Jacob Aptekar, and Jimeng Sun. 2025. *secret*: Semi-supervised clinical trial document similarity search. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5278–5291, Vienna, Austria. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. Bert: Pre-training of deep bidirectional transformers for language understanding.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

Martin Gellerstam. 1986. Translationese in swedish novels translated from english.

Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. 2023. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *CVPR*, pages 19338–19347.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104.

Songhao Jiang, Yan Chu, Zhengkui Wang, Tianxing Ma, Hanlin Wang, Wenxuan Lu, Tianning Zang, and Bo Wang. 2023. Explainable text classification via attentive and targeted mixing data augmentation. In *IJCAI*, pages 5085–5094.

Abdullah Khered, Youcef Benkhedda, and Riza Batista-Navarro. 2025. A multi-task learning approach to dialectal Arabic identification and translation to modern standard Arabic. In *Proceedings of the First Workshop on Advancing NLP for Low-Resource Languages (LowResNLP 2025)*, pages 21–31, Varna, Bulgaria. INCOMA Ltd.

Tyler LaBonte, Vidya Muthukumar, and Abhishek Kumar. 2024. Towards last-layer retraining for group robustness with fewer annotations. *Neurips*, 36.

Ju-Hyoung Lee, Joonghyuk Hahn, Hyeon-Tae Seo, Jiho Park, and Yo-Sub Han. 2024. SuperST: Superficial self-training for few-shot text classification. In *LREC-COLING*, pages 15436–15447, Torino, Italia.

Wenxin Liang, Tingyu Zhang, Han Liu, and Feng Zhang. 2024. SELP: A semantically-driven approach for separated and accurate class prototypes in few-shot text classification. In *ACL*, pages 9732–9741.

Jinshuo Liu and Lu Yang. 2024. Knowledge-enhanced prompt learning for few-shot text classification. *Big Data and Cognitive Computing*, 8(4).

Jiang Lu, Pinghua Gong, Jieping Ye, Jianwei Zhang, and Changshui Zhang. 2023. A survey on machine learning from few samples. *Pattern Recognition*, 139:109480.

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date.

Marcin Michał Mirończuk and Jarosław Protasiewicz. 2018. A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications*, 106:36–54.

OpenAI. 2025. GPT-4.1 (April 14 version). Large language model.

Chongyu Pan, Jian Huang, Jianxing Gong, and Xingsheng Yuan. 2019. Few-shot transfer learning for text classification with lightweight word embedding based models. *IEEE Access*, 7:53296–53304.

Trinh Pham, Khoi Le, and Anh Tuan Luu. 2024. UniBridge: A unified approach to cross-lingual transfer learning for low-resource languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3168–3184, Bangkok, Thailand. Association for Computational Linguistics.

Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, and Mohammad Sabokrou. 2021. Sign language production: A review. In *CVPR*, pages 3451–3461.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online).

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Jiajun Shen, Peng-Jen Chen, Matthew Le, Junxian He, Jiatao Gu, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2021. The source-target domain mismatch problem in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1519–1533, Online. Association for Computational Linguistics.

Eyal Shnarch, Ariel Gera, Alon Halfon, Lena Dankin, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2022. Cluster & tune: Boost cold start performance in text classification. In *ACL*, pages 7639–7653.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Shaobo Wang, Xiangqi Jin, Ziming Wang, Jize Wang, Jiajun Zhang, Kaixin Li, Zichen Wen, Zhong Li, Conghui He, Xuming Hu, and Linfeng Zhang. 2025. Data whisperer: Efficient data selection for task-specific LLM fine-tuning via few-shot in-context learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23287–23305, Vienna, Austria. Association for Computational Linguistics.

Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. 2020a. Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv.*, 53(3).

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020b. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34.

Hongling Xu, Yice Zhang, Qianlong Wang, and Ruifeng Xu. 2025. DS^2-ABSA: Dual-stream data synthesis with label refinement for few-shot aspect-based sentiment analysis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15460–15478, Vienna, Austria. Association for Computational Linguistics.

Sicheng Yu, Qianru Sun, Hao Zhang, and Jing Jiang. 2022. Translate-train embracing translationese artifacts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 362–370, Dublin, Ireland. Association for Computational Linguistics.

Jianqing Zhu, Huang Huang, Zhihang Lin, Juhao Liang, Zhengyang Tang, Khalid Almubarak, Mosen Alharthi, Bang An, Juncai He, Xiangbo Wu, Fei Yu, Junying Chen, Zhuoheng Ma, Yuhao Du, Yan Hu, He Zhang, Emad A. Alghamdi, Lian Zhang, Ruoyu Sun, Haizhou Li, Benyou Wang, and Jinchao Xu. 2024. Second language (arabic) acquisition of llms via progressive vocabulary expansion.

A Comparative Models

BERT models: We have compared FLICK with popular publicly available language models that include BERT (Devlin et al., 2019b) and DistilBERT (Sanh et al., 2020), all of which are built on the transformer architecture.

Arabic-BERT models: We have also used the AraBicBERT in our study. These models are: AraBicBERT (Safaya et al., 2020), ARBERTv2 (Abdul-Mageed et al., 2021), AraBERT-Twitter (Antoun et al., 2020b), AlcLaMv2 (Ahmed et al., 2024), CamelBERT-Mix (Inoue et al., 2021), CamelBERT-MSA-Mix (Inoue et al., 2021), and MARBERTv2 (Abdul-Mageed et al., 2021).

LLMs: We compare FLICK with fine-tuned LLMs via Supervised Fine-Tuning (SFT), given their demonstrated proficiency in handling complex tasks. For this purpose, we select AllAM-7B (Bari et al., 2025), Llama3.1-8B (Meta, 2024), AceGPT (Zhu et al., 2024), Qwen-72B-instruct (Team, 2024), and GPT-4o (OpenAI, 2025) as representative models in our comparison.

Intermediate models: To validate the effectiveness of our intermediate stage, we compared FLICK performance with Cluster & Tune (Shnarch et al., 2022) and IDoFew (Alsuhaibani et al., 2024), which both make use of an inter-training task. We

Dataset	K	Top-R	TS (%)	BS
ArBNTopic	15	10	25	16
Ar-Sentiment	50	10	15	32
Ar-Sarcasm	20	5	85	32
Sa'7r	30	15	15	16
DART	20	10	15	16
Ar-Dialects	15	5	85	16
MPOLD	20	5	85	64

Table 6: Per-dataset FLICK hyperparameters. *K*: initial K-means cluster count; *Top-R*: retained clusters after selection; *TS*: percentage of pseudo-labelled data used as the internal training split for Top-R ranking (Algorithm 1), the remainder forming the internal test split; *BS*: batch size.

also substitute BERT with language-specific BERT, such as ARBERTv2, to ensure a fair comparison for Arabic datasets, and a similar approach for the other languages.

B FLICK Settings

This appendix details the training configuration of FLICK. The per-dataset hyperparameters (K, Top-R, TS, BS) are listed in Table 6; the remaining settings are shared across all datasets and baselines and held fixed throughout: PL-FT epochs $= 1$, refinement and FL-FT epochs $= 10$, label budget $= 100$, learning rate $=$ 3e-5, $\varepsilon =$ 1e-6, the Adam optimiser, and seed 0. The same shared settings were applied to all baselines, except for the original SOTA method versions, which retain their own published configurations.

C Datasets statistics

We report dataset sizes and token statistics across all splits in Tables 7–13.

Split	#Rows	Tok. Mean	Tok. Med.	Tok. Max	Tok. Total	Word Mean	Char Mean
Pseudo-Labeled	11008	30.08	26	793	331079	30.08	177.90
Real-Labeled	100	28.07	28	89	2807	28.07	165.18
Eval	2390	30.04	26	392	71789	30.04	176.98
Test	4703	29.98	26	264	140998	29.98	176.79

Table 7: Statistics of the ArBNTopic dataset splits and stats.

Split	#Rows	Tok. Mean	Tok. Med.	Tok. Max	Tok. Total	Word Mean	Char Mean
Pseudo-Labeled	1205	61.58	59	144	74207	22.07	143.86
Real-Labeled	100	60.62	57	101	6062	22.08	143.13
Eval	173	40.76	40	78	7051	14.76	85.32
Test	345	37.46	35	91	12923	13.53	78.11

Table 8: Statistics of the DART dataset splits and stats.

Split	#Rows	Tok. Mean	Tok. Med.	Tok. Max	Tok. Total	Word Mean	Char Mean
Pseudo-Labeled	5805	66.72	68	185	393979	21.89	157.82
Real-Labeled	100	67.87	68	94	6787	22.59	159.27
Eval	2532	46.21	48	156	116993	14.83	103.47
Test	2110	45.63	47	135	96272	14.82	103.56

Table 9: Statistics of the Dialects dataset splits and stats.

Split	#Rows	Tok. Mean	Tok. Med.	Tok. Max	Tok. Total	Word Mean	Char Mean
Pseudo-Labeled	3958	50.67	41	175	200571	19.49	111.88
Real-Labeled	100	50.31	42	151	5031	19.87	113.85
Eval	3458	34.05	25	143	117751	14.27	79.59
Test	1186	34.59	26	181	41019	14.31	80.62

Table 10: Statistics of the Sa'7r dataset splits and stats.

Split	#Rows	Tok. Mean	Tok. Med.	Tok. Max	Tok. Total	Word Mean	Char Mean
Pseudo-Labeled	5905	66.50	68	182	392659	21.68	157.65
Real-Labeled	100	65.71	67	163	6571	21.50	154.83
Eval	2432	46.10	48	156	112109	14.80	104.04
Test	2110	45.63	47	135	96272	14.82	103.56

Table 11: Statistics of the Sarcasm dataset splits and stats.

Split	#Rows	Tok. Mean	Tok. Med.	Tok. Max	Tok. Total	Word Mean	Char Mean
Pseudo-Labeled	5835	65.91	68	184	384561	21.28	158.79
Real-Labeled	100	63.23	67	91	6323	20.45	152.93
Eval	2502	46.21	48	156	115627	14.73	103.48
Test	2110	45.63	47	135	96272	14.82	103.56

Table 12: Statistics of the Sentiment dataset splits and stats.

Split	#Rows	Tok. Mean	Tok. Med.	Tok. Max	Tok. Total	Word Mean	Char Mean
Pseudo-Labeled	2730	23.24	15	628	63456	23.24	129.67
Real-Labeled	100	24.40	17	134	2440	24.40	136.69
Eval	390	22.07	15	240	8607	22.07	124.21
Test	780	21.57	14	283	16828	21.57	121.68

Table 13: Statistics of the MPOLD dataset splits and stats.

Linguistic Proximity Enables Pivot-Based Machine Translation for an Under-Resourced Tribal Language

Pooja Singh[1*] **M Kaab Bin Shahid**[2*] **Atai Waris Khan**[3]
Aryan Kumar Jha[4] **Sandeep Kumar**[1]

[1]Indian Institute of Technology Delhi [2]University of Stuttgart
[3]NIT Jalandhar [4]Indian Institute of Technology Delhi, Abu Dhabi, UAE

{eez228470,ksandeep}@ee.iitd.ac.in st200141@stud.uni-stuttgart.de
ataiwk.ai.24@nitj.ac.in aryankumarjha2007@gmail.com

Abstract

For more than 10.4 million speakers of Bhili, an under-resourced tribal Indo-Aryan language of central and western India, limited machine translation support restricts digital accessibility and language inclusion. Beyond data scarcity, this challenge is intensified by the linguistic distance between English and Bhili. We examined whether this gap can be reduced through linguistically motivated pivot translation via Hindi and Marathi, two better-resourced languages that are more closely related to Bhili. Computational analysis shows substantially higher character-level overlap with Hindi and Marathi (67.6% and 57.5%) than with English (1.9%), providing an empirical basis for pivot selection. In supervised experiments, cascaded fine-tuning via pivot languages improves the strongest direct translation of NLLB-200 baseline from 7.87 to 9.46 BLEU for English→Bhili and from 25.23 to 29.84 BLEU for Bhili→English, with corresponding ChrF++ gains of +1.91 and +5.18. Pivot-based prompting further shows that LLMs can benefit from related-language cues: Gemini 2.5 Flash improves by +20.09 ChrF++ for English→Bhili using Hindi as the pivot, while GPT-4.5 reaches 46.87 ChrF++ for Bhili→English. We release **BhilSetu**, a five-language parallel corpus of 1,33,000 sentences and the first Bhili FLORES-200-aligned benchmark. Our results suggest that leveraging linguistic proximity can improve translation quality while supporting broader efforts toward digital inclusion and preservation of under-resourced indigenous languages.

1 Introduction

Neural Machine Translation (NMT) has advanced rapidly for high-resource languages through large-scale parallel corpora and multilingual pre-trained models (Devlin et al., 2019; Faheem et al., 2024). However, this progress remains uneven: many under-resourced languages, including Bhili, Santali, and Bodo, are still largely absent from translation systems that mediate access to education, governance, healthcare, and digital communication. Even commercial systems cover only a small fraction of the world's 7,000+ languages (Haddow et al., 2022), leaving many linguistic communities without basic translation support (Fan et al., 2021).

Bhili occupies a particularly urgent position among these excluded languages. Spoken by more than 10.4 million[1] people across Madhya Pradesh, Gujarat, and Maharashtra, it is one of India's largest tribal languages, yet it has limited presence in formal education, scarce digitized resources, and almost no dedicated translation infrastructure. Since many Bhili speakers belong to Scheduled Tribe communities facing long-standing marginalization, the absence of translation technologies for Bhili represents not merely a technical gap but a question of linguistic justice and cultural survival. It also affects access to public services and mother-tongue information, while language shift toward dominant regional languages threatens indigenous knowledge embedded in agricultural practices, medicinal taxonomies, kinship systems, ecological memory, and oral traditions (Joshi et al., 2020).

However, developing NMT for Bhili remains challenging even with the available parallel data. Although the BHEPC corpus (Singh et al., 2025) provides Bhili–English–Hindi parallel data, models fine-tuned directly on English↔Bhili produce substantially degraded translations (Section 6.2), while LLMs also perform poorly in zero-shot settings (Table 3), despite stronger performance reported for other low-resource languages (Elsner

*Equal contribution.

[1]https://en.wikipedia.org/wiki/Bhil_languages

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 339–349
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_035

et al., 2024). One factor contributing to this difficulty is the linguistic distance between Germanic English and Indo-Aryan Bhili: models must learn correspondences between structurally different linguistic systems (Littell et al., 2017), a challenge compounded by Bhili's extremely limited representation in model pre-training.

Pivot-based translation provides a promising alternative (Utiyama and Isahara, 2007; Wu and Wang, 2009): rather than directly bridging this typological gap, translation is routed through intermediate languages that are structurally closer to Bhili (Zhu et al., 2025). Its effectiveness, however, depends on selecting pivots that are both linguistically close to Bhili and sufficiently well-resourced to support reliable translation. Hindi, Marathi, and Gujarati are natural candidates given their shared Indo-Aryan ancestry (Masica, 1991), geographical proximity, and long-standing contact with Bhili-speaking communities. Rather than relying solely on these intuitions, we evaluate their suitability empirically by quantifying morphological, lexical, and syntactic proximity in Section 3.

Building on this analysis, we present the first comprehensive analysis of pivot-based Bhili translation for English↔Bhili, covering both cascaded fine-tuning of multilingual models (mT5 (Xue et al., 2021), NLLB (Costa-Jussà et al., 2022), Llama-3.1-8B (Grattafiori et al., 2024), and BLOOMZ-7B1 (Muennighoff et al., 2023)) and zero-shot pivot prompting with six contemporary LLMs (GPT-4.5 (Achiam et al., 2023), Llama-3.1-8B, Gemma-2-9B (Team et al., 2024), BLOOMZ-7B1, Mistral-7B-v0.1 (Jiang et al., 2023), and Gemini-2.5 Flash (Comanici et al., 2025)).

Our contributions are:

1. **Linguistic Analysis:** First systematic computational analysis of Bhili's structural relationship to surrounding Indo-Aryan languages, quantifying morphological, lexical, and syntactic proximity to Hindi, Gujarati, and Marathi, establishing theoretical foundations for pivot selection .
2. **Multilingual Corpus:** We introduce **BhilSetu**, the largest 1,33,000-sentence, five-language parallel corpus covering Bhili, English, Hindi, Gujarati, and Marathi.
3. **Bhili FLORES-200 benchmark:** We create the first Bhili-aligned FLORES-200 test set with 997 sentences across all five languages.

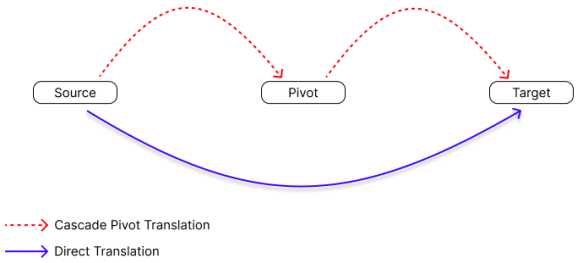


Figure 1: Direct and cascade pivot translation pathways. The blue arrow denotes direct source-to-target translation, while the red dashed arrows denote cascade translation through a pivot language.

4. **Cascaded Fine-tuning:** Cascaded fine-tuning strategies with sequential source-to-pivot and pivot-to-target training, demonstrating substantial improvements over direct translation.
5. **Pivot-based prompting:** We systematically evaluate whether LLMs benefit from translating through intermediate Indo-Aryan languages under zero-shot prompting.

2 Related Work & Background

Translation between high resource and low-resource endangered languages has remained a persistent challenge in NMT (Zhu et al., 2024). Several frameworks have been proposed to adapt existing NMT models for low-resource settings (Sennrich et al., 2016), utilizing data augmentation (Elmadani and Buys, 2024), transfer learning (Zoph et al., 2016), and pre-training strategies (Elmadani and Buys, 2024). Multilingual transformer-based architectures such as mBERT (Pires et al., 2019), and IndicTrans 2 (Gala et al., 2023) have further enhanced cross-lingual transfer capabilities for Indian languages. However, despite notable progress for major Indian languages, extremely low-resource languages like Bhili remain largely absent from these multilingual models.

Pivot-based translation strategies offer an effective alternative when the linguistic distance between the high- and low-resource language is large (as quantified in Section 3). Existing work on Indic pivoting, however, has two recurring limitations. First, it predominantly relies on *high-resource* pivots such as English to bridge gaps between Indic languages (Zhang et al., 2024), and where multiple pivots are combined, they are typically ensembled to boost coverage rather than selected for structural proximity to the target (Gaik-

wad et al., 2024). Second, pivot choice is rarely justified empirically; linguistically related *regional* languages are seldom evaluated as candidates. For instance, Gujarati and Marathi share substantial overlap with Bhili but have not been systematically explored in the literature. Our work departs from both practices: rather than defaulting to English or to convenience-driven selection, we choose pivots (Hindi, Marathi) through a quantitative linguistic-proximity analysis and route through related Indo-Aryan intermediaries.

In parallel, LLMs have introduced new paradigms in low-resource translation through in-context learning (ICL) and multi-lingual prompting (Chitale et al., 2024; Brown et al., 2020). Studies have shown that LLMs can effectively generalize translation patterns even from limited examples, leveraging linguistic cues embedded in prompts (Cahyawijaya et al., 2024; Zhang et al., 2023). However, their effectiveness for tribal and endangered languages remains underexplored. Linguistic studies suggest that translation accuracy improves with greater relatedness between source and pivot languages (Siripragada et al., 2020), but empirical evidence for this in the Indian tribal context is scarce.

3 Linguistic Analysis and Pivot Selection

Motivation and Analytical Framework. The selection of pivot languages in low-resource machine translation is not a neutral decision. It is a linguistically consequential choice whose implications compound across both legs of the translation pipeline. A poorly chosen pivot, one that is typologically distant from the target language, amplifies errors at each stage (Dabre et al., 2017), ultimately producing translations that fail both lexically and structurally. For Bhili, identifying optimal pivot languages requires systematic analysis of its structural relationships to candidate intermediaries.

We evaluate three candidate pivot languages: Hindi, Marathi, and Gujarati. Yet demographic proximity does not guarantee linguistic commensurability, and the question of which language constitutes the most reliable structural bridge between English and Bhili demands empirical rather than intuitive answers.

To ground this choice quantitatively, we conducted a comprehensive computational analysis of the BhilSetu parallel corpus (1,33,000 sentence pairs), examining six complementary metrics across lexical, morphological, and syntactic dimensions (Lin et al., 2019). Specifically, we measured: *character-level overlap* (Jaccard similarity over character sets, capturing orthographic affinity); *word-level overlap* (Jaccard similarity over token sets, capturing direct lexical sharing); *sentence length correlation* (Pearson r, capturing syntactic complexity parallelism); *normalized edit distance* (mean Levenshtein distance normalized by maximum sentence length, capturing surface form divergence); *character trigram similarity* (Jaccard over character 3-grams, capturing subword morphological patterns); and *heuristic cognate density* (proportion of word pairs with normalized Levenshtein similarity ≥ 0.6, serving as a proxy for shared etymological vocabulary (Inkpen et al., 2005)). Together, these metrics capture complementary aspects of cross-linguistic proximity that are theoretically linked to pivot translation quality: lexical overlap (Mohammadshahi et al., 2024) predicts vocabulary preservation, morphological similarity facilitates inflection transfer, and syntactic correlation reduces reordering errors across the two translation stages, computed from the BhilSetu corpus using the six metrics defined above (Table 1).

Hindi and Marathi: Devanagari-Normalized Evidence. Hindi and Marathi, sharing Devanagari script with Bhili, show consistent proximity across all six metrics as shown in Table 1. Hindi is the closest computational neighbor (character overlap 0.676, word overlap 0.150, length correlation $r = 0.945$, normalized edit distance 0.472, trigram similarity 0.247, cognate density 2.71%), reflecting both genetic relatedness and parallel orthography and syntax. Marathi ranks second (character overlap 0.575, word overlap 0.042, $r = 0.848$), maintaining strong structural similarity despite more divergent vocabulary due to Sanskrit *tatsama* influence and southern contact.

The Gujarati Paradox: Sociolinguistic Proximity vs. Computational Distance. Gujarati illustrates a gap between sociolinguistic reality and computational evidence. In Bhili speech communities, Gujarati is omnipresent: speakers in the eastern Gujarat tribal belt (Dahod, Panchmahal, Chhota Udaipur) code-switch fluidly with Gujarati. Yet our analysis yields very low Gujarati scores

Table 1: Computational linguistic similarity metrics between Bhili and candidate pivot languages. ↑ = higher is more similar; ↓ = lower is more similar.

Language	Char.↑	Word↑	Len. Corr.↑	Edit Dist.↓	3-gram↑	Cognate%↑
Hindi	0.676	0.150	0.945	0.472	0.247	2.71%
Marathi	0.575	0.042	0.848	0.689	0.114	1.23%
Gujarati	0.012	0.003	0.886	0.891	0.003	0.04%

across all character-level metrics. This is because, for example, *ghar* (house) is written घर in Bhili, Hindi, and Marathi (Devanagari) but ઘર in Gujarati script, thus distinct Unicode code points encoding the same phoneme. Although Gujarati shares syntactic complexity patterns with Bhili ($r = 0.886$), two languages may share deep lexical and morphological ancestry yet exhibit near-zero character overlap simply because they employ different writing systems.

Pivot Selection. Based on the computational evidence and sociolinguistic context (Table 1), we select **Hindi and Marathi** as pivot languages. Hindi is justified by its dominance across all valid metrics and its role as the primary contact language in Bhili speakers' educational and civic lives. Marathi is justified by its consistent secondary proximity and its geographic overlap with Bhili communities in northern Maharashtra.

4 Dataset

4.1 Corpus Construction & Processing

We extend the Bhili–Hindi–English Parallel Corpus (BHEPC) by adding parallel Marathi and Gujarati sentences and also added 23,000 human-translated high-quality sentence pairs across the five languages, resulting in a final corpus **BhilSetu** of 1,33,000 parallel sentences. This makes **BhilSetu** the first large-scale, human-curated, gold-standard dataset for Bhili MT across five languages. The additional 23,000 Hindi source sentences were primarily drawn from the Bharat Parallel Corpus Collection (BPCC) (Gala et al., 2023), government speeches (Ramesh et al., 2022; Kunchukuttan et al., 2018), and NCERT textbooks[2]. All translations were created and validated by native-speaking linguists to ensure contextual fidelity and naturalness. The dataset statistics are summarized in Table 2.

[2]https://ncert.nic.in/textbook.php

Table 2: Statistics of the corpus and FLORES-200 test set.

Dataset	Train	Test (FLORES)	Langs
BHEPC	1,10,000	-	3
Bhili-FLORES	–	997	5
BhilSetu	1,33,000	997	5

To ensure data quality, we applied preprocessing steps, including normalization of Bhili homophones, removal of extraneous symbols, lowercasing of English text, and strict de-duplication using cosine-similarity filtering. Sentences shorter than 6 or longer than 80 words were discarded. Throughout this work, we use `bhb` to denote Bhili, following its ISO 639-3 language identifier.

4.2 Evaluation Datasets & Metrics

For evaluation, we use the *FLORES-200* benchmark dataset (Costa-Jussà et al., 2022), a multilingual corpus created to assess translation quality across 200 languages. We rely on its *devtest* split, which contains 997 parallel sentences per language pair, allowing standardized evaluation of multilingual and low-resource MT systems. *FLORES devtest* is translated into Bhili with the help of language experts, and the pivot languages, Hindi, Gujarati and Marathi to assess the reliability of pivot-based translation across two directions. Performance is measured using BLEU (Papineni et al., 2002) and ChrF++ (Popović, 2017), which evaluate n-gram precision and character-level similarity respectively, both suitable for morphologically rich Indian languages. All scores are computed with SacreBLEU (Post, 2018) and NLTK (Bird et al., 2009) for reproducibility.

4.3 Model Training & Hyperparameters

All models are trained under a consistent setup to enable fair comparison. For Llama-3.1-8B and Bloomz-7B1 we experiment with Parameter-Efficient Fine-Tuning (PEFT) using LoRA (Hu et al., 2022), rank 8. Key hyperparameters include

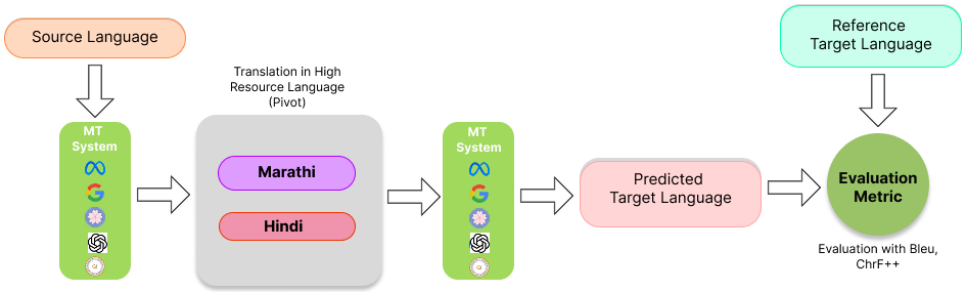


Figure 2: Illustration of the Pivot-Based Translation Framework.

Adam optimizer, learning rate 5e-5, batch size 16, weight decay 0.01, and early stopping patience of 10 based on validation chrF++ scores. For prompting, all models are evaluated with temperature 0 and the same fixed prompt template. Five-shot examples are sampled once from the training set and kept identical across all models. Experiments are conducted on a NVIDIA A100 GPUs, with training times ranging from 10 to 72 hours.

5 Methodology

Pivot-based translation addresses the fundamental challenge of translating between a high-resource and an extremely low-resource language by routing through one or more intermediate languages that maintain closer linguistic affinity to the target. As established in Section 3, direct translation between English and Bhili is hampered not only by data scarcity but by the wide typological gap separating a Germanic language from a tribal Indo-Aryan one. Even with the 1,10,000-sentence BHEPC corpus, models trained directly on English–Bhili pairs produce degraded output, motivating the pivot architectures described here (see Table 3). Figures 1 and 2 illustrate the direct and cascaded pivot pipelines.

Within this work, we explore two distinct pivoting strategies, **1. Cascading Pivoting** and **2. Pivot Prompting**, evaluated across two translation directions: English→Bhili and Bhili→English, using Hindi and Marathi as pivot languages (see Section 3 for linguistic justification).

5.1 Cascading Pivoting

The cascading approach decomposes translation into two sequential, independently trained stages: source-to-pivot (S→P) and pivot-to-target (P→T). Each stage leverages the same underlying multilingual model architecture, ensuring consistent vocabulary representations and embedding spaces across both directions. The shared model parameters promote stable lexical and syntactic mappings, reducing the distributional drift that commonly afflicts cascade pipelines operating with heterogeneous model pairs. Let $\mathbf{x}$, $\mathbf{z}$, and $\mathbf{y}$ denote the source, pivot, and target sequences, respectively. Ideally, the final translation can be obtained by marginalizing over all possible pivot sequences:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \sum_{\mathbf{z}} P(\mathbf{y} \mid \mathbf{z}) P(\mathbf{z} \mid \mathbf{x}). \tag{1}$$

Since exact marginalization over all pivot sequences is computationally intractable, we approximate this process using a standard two-step cascade. First, the source sentence is decoded into the most likely pivot sentence:

$$\hat{\mathbf{z}} = \arg\max_{\mathbf{z}} P(\mathbf{z} \mid \mathbf{x}), \tag{2}$$

and the resulting pivot sentence is then decoded into the final target sentence:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} P(\mathbf{y} \mid \hat{\mathbf{z}}). \tag{3}$$

For training, the S→P and P→T models are optimized separately using parallel source–pivot and pivot–target sentence pairs. Given triplets $(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}, \mathbf{y}^{(i)})$, the training objective is the sum of the two negative log-likelihood losses:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \Big[\log P(\mathbf{z}^{(i)} | \mathbf{x}^{(i)}) + \log P(\mathbf{y}^{(i)} | \mathbf{z}^{(i)}) \Big] \tag{4}$$

This sequential decomposition provides modularity: the S→P and P→T components can be trained, evaluated, and updated independently, and existing high-resource translation models can be reused for the first stage without re- training from scratch. Concretely, we fine-tune a single multilingual checkpoint on both the source–pivot and

pivot–target directions (Eq. 4). At inference, the model is used twice: first to generate the pivot hypothesis $\hat{\mathbf{z}}$ from the source, and then to generate the target $\hat{\mathbf{y}}$ from $\hat{\mathbf{z}}$ (Eqs. 2–3). Using the same model for both steps keeps the vocabulary and embedding space consistent and helps reduce the distributional drift that can arise when different models are used for the two stages of the cascade.

5.2 Pivot Prompting

While cascading relies on separately fine-tuned models for each translation stage, the pivot prompting approach leverages a single multilingual LLM to perform zero-shot translation across pivot pathways within a unified prompt. Rather than producing an explicit intermediate output, the model receives the source sentence alongside a pivot language in its context and is instructed to translate directly into the target while implicitly reasoning through the specified intermediate steps.

This design allows the model to internally synthesize lexical and structural cues from pivot languages simultaneously, without committing to a single intermediate output as cascading requires. Conceptually, the pivot formulation models the conditional distribution:

$$P(\mathbf{y} \mid \mathbf{x}, P_1) \tag{5}$$

where $\{P_1\}$ is the auxiliary pivot language incorporated through the prompt context. Unlike cascading, which explicitly produces a pivot sequence $\mathbf{z}$ as a discrete intermediate artifact, pivot prompting treats pivot-level reasoning as an implicit latent process encoded within the LLM's internal cross-lingual representations. The model is not required to externalize pivot-language text; instead, it leverages its multilingual pretraining to bridge linguistic gaps internally.

This helps particularly in zero-shot settings where fine-tuning data is unavailable, and where the target language (Bhili) may have minimal direct representation in the model's pretraining corpus. By prompting through linguistically proximate pivots; Hindi and Marathi, the model is guided toward a region of its representation space closer to Bhili, improving the fidelity of the final output even without Bhili-specific supervision.

5.3 Experimental Setup

We compare three settings across English→Bhili (`eng`→`bhb`) and Bhili→English (`bhb`→`eng`):

Table 3: Direct translation baseline results for English↔Bhili without pivot languages. Fine-tuned models report BLEU / ChrF++. Direct prompting models report ChrF++ for 0-shot / 5-shot prompting. Best results within each block and metric are shown in **blue**.

Model	eng→bhb	bhb→eng
Fine-tuned models: BLEU / ChrF++		
mT5 Base	7.13 / 35.63	19.76 / 46.45
NLLB-200	**7.87** / **37.18**	**25.23** / **54.34**
Llama-3.1-8B	5.12 / 22.23	12.10 / 30.59
BLOOMZ-7B1	4.28 / 21.43	9.12 / 29.78
Direct prompting: 0-shot / 5-shot ChrF++		
Llama-3.1-8B	0.55 / 13.50	12.00 / 16.50
BLOOMZ-7B1	0.81 / 11.12	19.34 / 21.02
Mistral-7B-v0.1	0.76 / 10.87	12.89 / 13.25
Gemma-2-9B	5.27 / 13.53	30.50 / 23.25
Gemini 2.5 Flash	19.35 / 21.30	37.56 / 38.89
GPT-4.5	**23.35** / **25.67**	**42.16** / **43.78**

- **Direct Baseline**: the source sentence is translated directly into the target language without any pivot.
- **Cascaded Pivot**: the source sentence is translated into Hindi or Marathi, and the pivot output is then translated into the target language.
- **Pivot Prompting**: a single LLM is prompted to translate the source sentence into the target language using Hindi or Marathi as an implicit pivot cue.

6 Experimental Results and Analysis

6.1 Baseline Models

We evaluate encoder-decoder and decoder-only models to benchmark direct and pivot-based Bhili translation. For supervised experiments, we fine-tune **mT5 Base** (Xue et al., 2021), **NLLB-200 (600M)** (Costa-Jussà et al., 2022), **Llama-3.1-8B** (Grattafiori et al., 2024), and **BLOOMZ-7B1** (Muennighoff et al., 2023) under both direct English↔Bhili translation and cascaded pivot translation through Hindi and Marathi. For zero-shot prompting, we evaluate **Llama-3.1-8B**, **BLOOMZ-7B1**, **Mistral-7B-v0.1** (Jiang et al., 2023), **Gemma-2-9B** (Team et al., 2024), **Gemini 2.5 Flash** (Comanici et al., 2025), and **GPT-4.5** (Achiam et al., 2023) under both direct and pivot-based prompting. These models cover multilingual pre-training, instruction tuning, and different translation regimes.

Table 4: Cascading pivot results for English↔Bhili translation on the Flores-200 benchmark. Scores are BLEU / ChrF++. Δ reports the difference compared with the corresponding direct fine-tuned baseline for the same model. Best pivot result for each model and direction is shown in **blue**.

Model	Pivot	eng→bhb	Δ	bhb→eng	Δ
mT5 Base	Hindi	**8.71 / 37.85**	+1.58 / +2.22	**22.57 / 49.91**	+2.81 / +3.46
	Marathi	7.57 / 36.84	+0.44 / +1.21	21.22 / 47.14	+1.46 / +0.69
NLLB-200 (600M)	Hindi	**9.46 / 39.09**	+1.59 / +1.91	**29.84 / 59.52**	+4.61 / +5.18
	Marathi	8.51 / 37.97	+0.64 / +0.79	27.58 / 55.17	+2.35 / +0.83
Llama-3.1-8B	Hindi	**6.12 / 26.20**	+1.00 / +3.97	14.54 / 33.05	+2.44 / +2.46
	Marathi	5.96 / 23.02	+0.84 / +0.79	**15.63 / 34.01**	+3.53 / +3.42
BLOOMZ-7B1	Hindi	**7.58 / 27.96**	+3.30 / +6.53	**10.85 / 31.03**	+1.73 / +2.82
	Marathi	5.38 / 22.86	+1.10 / +1.43	9.37 / 28.42	+0.25 / +0.21

6.2 Direct Translation Baseline

Before assessing the contribution of pivot strategies, we establish direct translation baselines for `eng`→`bhb` and `bhb`→`eng` without any intermediate pivot language. These results quantify what supervised and zero-shot models can achieve when forced to bridge the full English–Bhili typological gap in a single step, providing the reference point against which all pivot-based improvements are measured. Table 3 shows that direct translation performance is consistently low across model types, confirming the central difficulty of the English–Bhili translation problem. Among fine-tuned models, NLLB-200 achieves the strongest direct baseline in both directions, while direct prompting shows that larger instruction-tuned models perform better but still struggle substantially when Bhili is the target language.

6.3 Cascading Pivot Results

Table 4 presents the results of cascaded pivot translation using Hindi and Marathi as intermediate languages. Pivot-based routing consistently improves over direct translation across all fine-tuned models and both translation directions. For English→Bhili, the strongest result is obtained by NLLB-200 using Hindi as the pivot, improving from 7.87 / 37.18 to 9.46 / 39.09 (+1.59 BLEU, +1.91 ChrF++). Similar gains are observed for mT5 Base, Llama-3.1-8B, and BLOOMZ-7B1, with Hindi generally outperforming Marathi. For Bhili→English, improvements are even more pronounced, with NLLB-200 reaching 29.84 / 59.52 using Hindi as the pivot and Llama-3.1-8B achieving its best performance through Marathi (15.63 / 34.01). Overall, results indicate that linguistically related pivots reduce translation difficulty and provide consistent benefits over direct fine-tuning.

Pivot Effectiveness: Across models, Hindi emerges as the most reliable pivot language, yielding the best performance in the majority of settings. Marathi also provides positive gains, although its improvements are generally smaller and less consistent. These findings align with the linguistic similarity analysis in Section 3, where Hindi exhibits the strongest structural and lexical proximity to Bhili.

6.4 Pivot Prompting Results

Table 5 evaluates pivot prompting under a zero-shot setting. Unlike cascaded fine-tuning, pivot prompting produces more modest gains and exhibits a clear directional asymmetry. For English→Bhili, improvements are generally small for most models, suggesting that current LLMs still struggle to generate fluent Bhili without supervised exposure. The largest gain is observed for Gemini 2.5 Flash, which improves from 19.35 to 39.44 ChrF++ using Hindi as the pivot. In contrast, Bhili→English benefits more consistently from pivot prompting, with GPT-4.5 improving from 42.16 to 46.87 ChrF++ and Gemini 2.5 Flash reaching 44.75 ChrF++. Overall, pivot prompting helps LLMs interpret Bhili inputs but cannot reliably generate fluent Bhili outputs, whereas routing through Hindi or Marathi markedly improves English generation.

6.5 Pivot Contribution: Direct vs. Pivot Comparison

Tables 4 and 5 show that pivot-based translation consistently improves over direct translation (Table 3), but the gains depend strongly on direction, model type, and strategy. Cascaded fine-tuning delivers stable improvements in both directions for all fine-tuned models, with Hindi providing the

Table 5: Full pivot-prompting results for English↔Bhili translation. Scores are ChrF++. Δ reports the gain of the better pivot over the corresponding direct zero-shot baseline. Best pivot result within each model and direction is highlighted in **blue**.

Model	English→Bhili				Bhili→English			
	Direct	Hindi	Marathi	Δ	Direct	Hindi	Marathi	Δ
Llama-3.1-8B	0.55	**1.44**	1.39	+0.89	12.00	**13.82**	10.99	+1.82
BLOOMZ-7B1	0.81	**3.76**	2.00	+2.95	19.34	19.56	**19.76**	+0.42
Mistral-7B-v0.1	0.76	**1.94**	1.65	+1.18	12.89	**14.40**	13.68	+1.51
Gemma-2-9B	5.27	**6.31**	4.51	+1.04	30.50	**33.86**	32.31	+3.36
Gemini 2.5 Flash	19.35	**39.44**	38.39	+20.09	37.56	**44.75**	42.05	+7.19
GPT-4.5	23.35	**24.56**	23.76	+1.21	42.16	**46.87**	43.24	+4.71

strongest overall pivot in most settings. Translating through a well-represented Indo-Aryan intermediate language improves source interpretability and reduces the structural mismatch before Bhili generation, leading to consistent gains over direct English↔Bhili translation.

In contrast, pivot prompting remains asymmetric and more model-dependent: gains are clearer for `bhb→eng`, where models can generate English more reliably, but are smaller for `eng→bhb`, where limited Bhili generation capacity remains a bottleneck in zero-shot LLMs. While pivot cues improve source understanding, they do not fully compensate for the absence of supervised Bhili training signals. Thus, supervised cascading is preferable when target-language generation must be reliable, whereas pivot prompting mainly benefits translation *from* low-resource Bhili into English.

7 Conclusion

This work presents a linguistically grounded pivot-based approach for English↔Bhili machine translation, targeting an extremely under-resourced tribal language of India. Our results show that direct translation remains difficult across both fine-tuned and prompted models, especially when Bhili is the target language, while routing through related Indo-Aryan pivots provides consistent improvements over direct baselines. Hindi emerges as the most reliable pivot in most settings, with Marathi also offering useful gains for selected models and directions. Our computational linguistic analysis quantifies lexical, morphological, and structural proximity between Bhili and candidate pivot languages, providing an empirical basis for pivot selection. We also release **BhilSetu**, a 1,33,000-sentence five-language parallel corpus covering Bhili, English, Hindi, Gujarati, and Marathi, along with the first Bhili-aligned FLORES-200 benchmark of 997 sentences. By grounding pivot selection in linguistic proximity rather than resource availability alone, this work offers a practical and reproducible path toward extending machine translation to other marginalized indigenous languages.

Limitations

Although pivot-based training improves Bhili translation performance, error propagation through intermediate pivot languages may still affect final outputs, particularly in morphologically rich and low-resource settings. Future work will explore extending the approach to additional under-resourced languages, improving robustness through synthetic and monolingual data augmentation, and combining Hindi and Marathi through multi-source pivot ensembling and adaptive pivot selection.

Ethical Considerations

All translations and annotations were carried out by qualified native speakers and linguists under informed consent procedures and with compensation provided in accordance with government norms. Publicly available datasets and the newly created **BhilSetu** corpus were used and curated following research ethics and open-data practices.

Acknowledgments

We gratefully acknowledge the Ministry of Tribal Affairs, Government of India, and the Adi Vaani Project, together with the Tribal Research Institute, Bhopal, which enabled the human resources, cloud infrastructure, and expert translators essential to this work. We also thank all translators for their efforts in creating the BhilSetu.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. Llms are few-shot in-context low-resource language learners. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 405–433.

Pranjal Chitale, Jay Gala, and Raj Dabre. 2024. An empirical study of in-context learning in llms for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7384–7406.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Raj Dabre, Tetsuji Nakagawa, and Hideto Kazawa. 2017. An empirical study of language relatedness for transfer learning in neural machine translation. In *Proceedings of the 31st Pacific Asia conference on language, information and computation*, pages 282–286.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Khalid N Elmadani and Jan Buys. 2024. Neural machine translation between low-resource languages with synthetic pivoting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12144–12158.

Micha Elsner et al. 2024. Shortcomings of llms for low-resource translation: Retrieval and understanding are both the problem. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1332–1354.

Mohamed Atta Faheem, Khaled Tawfik Wassif, Hanaa Bayomi, and Sherif Mahdy Abdou. 2024. Improving neural machine translation for low resource languages through non-parallel corpora: a case study of egyptian dialect to modern standard arabic translation. *Scientific Reports*, 14(1):2265.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Pranav Gaikwad, Meet Doshi, Raj Dabre, and Pushpak Bhattacharyya. 2024. How effective is multi-source pivoting for translation of low resource indian languages? *arXiv preprint arXiv:2406.13332*.

Jay P Gala, Pranjal A Chitale, Varun Gumma, Sumanth Doddapaneni, Kumar M Aswanth, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, et al. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Diana Inkpen, Oana Frunza, and Grzegorz Kondrak. 2005. Automatic identification of cognates and false friends in french and english. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, volume 9, pages 251–257.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel,

Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6282–6293.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The iit bombay english-hindi parallel corpus. In *proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, et al. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135.

Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14.

Colin P Masica. 1991. *The indo-aryan languages*. Cambridge University Press.

Alireza Mohammadshahi, Jannis Vamvas, and Rico Sennrich. 2024. Investigating multi-pivot ensembling with massively multilingual machine translation models. In *Proceedings of the Fifth Workshop on Insights from Negative Results in NLP*, pages 169–180.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4996–5001.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the third conference on machine translation: Research papers*, pages 186–191.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan Ak, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, et al. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 86–96.

Pooja Singh, Shashwat Bhardwaj, Vaibhav Sharma, and Sandeep Kumar. 2025. Leveraging the cross-domain & cross-linguistic corpus for low resource nmt: A case study on bhili-hindi-english parallel corpus. *arXiv preprint arXiv:2511.00486*.

Shashank Siripragada, Jerin Philip, Vinay P Namboodiri, and CV Jawahar. 2020. A multilingual parallel corpora collection effort for indian languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3743–3751.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491.

Hua Wu and Haifeng Wang. 2009. Revisiting pivot language approach for machine translation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 154–162.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 483–498.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *International conference on machine learning*, pages 41092–41110. PMLR.

Zhihan Zhang, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. 2024. Plug: Leveraging pivot language in cross-lingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7025–7046.

Enchang Zhu, Yuxin Huang, Yantuan Xian, Junguo Zhu, Minghu Gao, and Zhiqiang Yu. 2025. Enhancing distant low-resource neural machine translation with semantic pivot. *Alexandria Engineering Journal*, 116:633–643.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the association for computational linguistics: NAACL 2024*, pages 2765–2781.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1568–1575.

Cross-Lingual Hate Speech Detection in Low-Resource Languages: The Case of Persian and Kurdish

Shahin Yousefi
Independent Researcher, Iran
sh.yousefi72@gmail.com

Ernesto Luis Estevanell-Valladares
University of Alicante, Spain
University of Havana, Cuba
ernesto.estevanell@ua.es

Ruslan Mitkov
University of Alicante, Spain
ruslan.mitkov@gcloud.ua.es

Abstract

Hate speech detection in low-resource languages remains severely constrained by data scarcity and linguistic complexity. This study presents a comprehensive investigation of Persian-to-Kurdish cross-lingual transfer, evaluating four methodological paradigms: rule-based, traditional machine learning, transformer-based, and large language models, across monolingual and cross-lingual settings. Through a two-phase experimental design, we demonstrate that Persian data can substantially enhance Kurdish detection performance. Notably, models trained solely on Persian data frequently outperform the Kurdish (only), with ParsBERT achieving the strongest overall results. While translation-based data augmentation yields limited gains and introduces noise, model-level cross-lingual transfer proves highly effective. These findings highlight the practical value of leveraging related higher-resource languages for low-resource hate speech detection and provide actionable insights for building more equitable content moderation systems in underrepresented linguistic communities.

1 Introduction

The rapid expansion of social media has amplified both the benefits and perils of online communication. Among these perils, hate speech stands out as a particularly corrosive force, capable of inciting violence, deepening societal divisions, and marginalising vulnerable communities (Poletto et al., 2021). In multilingual and geopolitically complex regions such as Iran and Kurdistan, where Persian (Farsi) and Kurdish serve as dominant languages for millions of users, the challenge is compounded by linguistic nuances, dialectal variations, and script differences. Despite the colours of local context, automated hate speech detection systems for these languages remain underdeveloped, largely due to the scarcity of annotated resources and the inherent difficulties of low-resource NLP settings (Das et al., 2024; Poletto et al., 2021).

This scarcity is particularly acute for Persian and Kurdish, both Indo-Iranian languages that share historical roots yet face distinct resource constraints. Persian benefits from a growing but still limited set of datasets and models (Mozafari et al., 2024; Kebriaei et al., 2024), while Kurdish varieties, especially Sorani, suffer from even fewer annotated corpora and pretrained resources (Saeed et al., 2022). Surveys of hate speech detection consistently highlight that while high-resource languages like English have seen substantial progress through large-scale benchmarks and sophisticated models, low-resource settings continue to rely on fragmented efforts, often with modest performance gains from traditional approaches (Das et al., 2024; Poletto et al., 2021). In the case of Persian, recent efforts have introduced datasets focussed on offensive and hateful content from social media. Yet, challenges such as cultural specificity, code-switching, and evolving online slang persist (Mozafari et al., 2024; Bokaei et al., 2026). For Kurdish, early studies have collected Facebook comments and applied basic classifiers, but systematic evaluations remain rare (Saeed et al., 2022).

Cross-lingual transfer learning offers a promising pathway to address these gaps, particularly for closely related language pairs like Persian and Kurdish (Jiang and Zubiaga, 2026). Recent comprehensive reviews emphasise the transformative role of transformer-based architectures in advancing hate speech detection, while underscoring the persis-

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 350–360
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_036

tent challenges in low-resource and multilingual contexts (Das et al., 2024; Ramos et al., 2024). By leveraging model-level adaptations (such as multilingual transformers) and data-level strategies (including dataset merging and translation), it becomes possible to enhance hate speech detection in the lower-resourced language. This study adopts a two-phase experimental design. First, a separate rule-based approach is evaluated as a simple pattern-matching baseline. Phase I then establishes performance across traditional ML, transformer-based, and LLM paradigms. Phase II examines whether augmenting the data with original Persian data and its Kurdish translation meaningfully improves hate speech detection performance on Kurdish. In doing so, it contributes to the broader understanding of how transfer techniques can support more equitable and effective content moderation tools in underrepresented linguistic communities (Das et al., 2024; Bokaei et al., 2026; Saeed et al., 2022).

1.1 Motivation

Persian and Kurdish, as closely related Indo-Iranian languages, offer a valuable opportunity for cross-lingual transfer in hate speech detection. While Persian has benefited from a growing number of datasets and models, Kurdish varieties, particularly Central Kurdish (Sorani), remain severely low-resourced, with limited annotated corpora and pre-trained models available (Ahmadi et al., 2024; Das et al., 2024). This resource imbalance is despite the fact that both languages are widely used in online discourse across Iran and Kurdistan, where hate speech frequently draws on shared cultural and political contexts. Previous research has shown that cross-lingual transfer learning can effectively improve hate speech detection in low-resource settings by leveraging data and models from related higher-resource languages (Bigoulaeva et al., 2021). Motivated by this, the present study investigates whether augmenting Kurdish data with Persian resources, through both model-level (e.g., multilingual and language-specific transformers) and data-level strategies (dataset combination and machine translation), can yield meaningful performance gains, while establishing strong baselines using, traditional ML, and LLM approaches.

This study contributes to low-resource hate speech detection by presenting a comprehensive evaluation across rule-based, traditional ML, transformer, and LLM paradigms; systematically comparing model- and data-level cross-lingual transfer strategies for Persian-Kurdish; and demonstrating the benefits of integrating multiple Persian datasets to address Kurdish data scarcity.

2 Related Work

This section reviews existing research on hate speech detection in low-resource languages, cross-lingual transfer learning, and various detection methods, providing the foundation for the present study on Persian and Kurdish.

2.1 Low-Resource Languages in Hate Speech Detection

Low-resource languages, lack sufficient digital resources, such as corpora, annotated datasets, and linguistic tools, for effective NLP model training, despite often having large speaker populations. Pakray et al. (2025) further elaborate on the challenges and opportunities in developing NLP applications for these languages, emphasising the critical gap between speaker population and available computational resources. Hate speech detection in low-resource languages remains a significant challenge due to limited annotated datasets, linguistic diversity, and cultural specificities. Das et al. (2024) provide a dedicated survey on automatic online hate speech detection in low-resource languages, highlighting the rapid growth of toxic content on social media and the persistent gaps in resources for non-English languages. Similarly, Sharma et al. (2025) survey hate speech detection research across South Asian languages, documenting tasks, datasets, and methods while emphasising the under-representation of many Indo-Aryan and Iranic varieties.

Poletto et al. (2021) offer a systematic review of resources and benchmark corpora for hate speech detection, revealing a heavy bias toward high-resource languages and calling for more diverse datasets. In the context of Persian, Mozafari et al. (2024) explore offensive language detection as a use case for this low-resource language, while Sheykhlan et al. (2023) investigate Pars-HaO using Machine Learning and deep learning on Persian tweets. Kebriaei et al. (2024) further advance Persian offensive language detection through detailed modelling.

For Kurdish, Badawi (2024) applies deep learning to cyberbullying detection, and Ahmadi et al.

(2024) develop language and speech technology resources for Central Kurdish varieties, including corpora from transcribed media as an alternative to traditional fieldwork. Roberts (2024) addresses automated hate speech detection in a low-resource environment, while Gutha et al. (2023) tackle multilingual hate speech and offensive language detection specifically for low-resource languages. Bokaei et al. (2026) benchmark offensive language detection in Persian and Pashto, evaluating zero-shot, fine-tuned, and cross-lingual settings. Fetahi et al. (2025) enhance social media hate speech detection in low-resource languages using transformers and explainable AI. Mahmud et al. (2023) review cyberbullying detection for low-resource languages and dialects. Nkemelu et al. (2023) emphasise the role of context experts in tackling hate speech in low-resource settings like Myanmar.

2.2 Cross-lingual Transfer Learning

Transfer learning and cross-lingual approaches have become central to addressing data scarcity in hate speech detection. Ruder et al. (2019) provide a foundational tutorial on transfer learning in Natural Language Processing, outlining methods that leverage pre-trained models across tasks and languages. It is shown that cross-lingual transfer learning for hate speech detection can be achieved by training on high-resource data (English) and adapting to low-resource targets (German) using bilingual embeddings and data bootstrapping (Bigoulaeva et al., 2021). Jiang and Zubiaga (2026) present a systematic review of cross-lingual offensive language detection, analysing datasets, approaches, and challenges across instance, feature, and parameter transfer strategies. Researchers in (Hämmerl et al., 2024) survey understanding of cross-lingual alignment, while Pallucchini et al. (2026) examine methods for cross-lingual alignment of contextualised representations. Al-Badani et al. (2025) discuss transfer learning and cross-linguistic generalisation in multilingual hate speech detection, reviewing approaches and challenges.

Practical applications in (Firmino et al., 2024) improve hate speech detection using cross-lingual learning from English and Italian to Portuguese, achieving state-of-the-art results. Shi et al. (2022) propose cross-lingual offensive speech identification with transfer learning and adversarial methods for low-resource languages. de Oliveira et al. (2023) use a multilingual approach in cross-lingual transfer learning for hate speech detection. Usman et al. (2025) explore multilingual hate speech detection using translation-based approaches with Large Language Models. Research in (Mollanorozy et al., 2023) specifically investigates cross-lingual transfer learning with Persian as a source language, revealing task-dependent benefits. Narula and Chaudhary (2024) provide a comprehensive review of hate speech detection for multilingual data.

2.3 Hate Speech Detection Methods and Resources

The evolution of hate speech detection methods spans rule-based, traditional ML, transformers, and Large Language Models. Albladi et al. (2025) offer a comprehensive review of hate speech detection using Large Language Models. Ramos et al. (2024) survey automatic hate speech detection in the age of transformers. Lee and Ram (2024) propose a personality-based deep learning approach for hate speech detection.

For Persian and related languages, work (Nasiri and Analoui, 2022) apply transfer learning and data augmentation to Persian stance detection, with implications for hate speech tasks. Pakray et al. (2025) discuss Natural Language Processing applications for low-resource languages, while a survey by Joshi et al. (2025) reviews the development of LLMs for European languages, highlighting datasets and resource challenges in low-resource regional dialects. Ali and Pyysalo (2024) survey Large Language Models for European languages, providing a broader context for multilingual capabilities.

These works collectively underscore the shift toward multilingual models, the value of language-specific resources, and the need for hybrid approaches combining classical methods with modern transformers and LLMs. The current study builds upon this foundation by conducting a three-way comparative evaluation focused on Persian and Kurdish hate speech detection.

3 Data and Preprocessing

This section presents the datasets employed in the study, describes the standardisation procedures applied to ensure cross-lingual compatibility, and details the preprocessing pipeline used for all experiments.

Dataset	Original Size	Cleaned Size
Kurdish	6,882	5,738
Naseza	5,719	5,162
PHATE	8,727	6,999
PHICAD	124,100	99,318

Table 1: Data statistic

3.1 Datasets Description

This study utilises four publicly available datasets to support hate speech detection research in Persian and Kurdish. These datasets were selected for their complementary characteristics in terms of size, annotation quality, and domain coverage.

The *Kurdish Hate Speech Dataset (KHSD)* (Saeed et al., 2022) consists of Central Kurdish (Sorani) texts written in the Arabic-based script. No Latin transliteration was applied; only standard Arabic/Persian character normalisation was performed as part of the unified pre-processing pipeline. This dataset contains 6,882 manually annotated Facebook comments in a binary hate/non-hate scheme. It represents one of the few available resources for the Kurdish language. For Persian, three datasets were employed: *Naseza* (Amirivojdan, 2025), a substantial Telegram-based dataset of 5,719 entries focused on political and social discussions; *PHATE* (Delbari et al., 2024), containing 8,727 tweets with multi-label hate annotations and valuable annotator rationales; and *PHICAD* (Davardoust et al., 2024), a large-scale collection of more than 124,100 Instagram comments labelled for harmful content. Table 1 summarises the statistics of the original datasets used in this study.

The combination of these datasets allows for robust evaluation in Phase I (language-specific baselines) (see section 5.2). It enables data-level cross-lingual experiments in Phase II (see section 5.3) by merging Persian and Kurdish resources.

3.2 Data Standardisation

To ensure consistency across all datasets for both language-specific and cross-lingual experiments, a unified format consisting of three columns, *id* (starting from 0), *text*, and *label* (1 for hate, 0 for not hate), was adopted. For the **Kurdish Hate Speech Dataset (KHSD)**, an *id* column was added, and the original Class labels were mapped to binary values (hate $\rightarrow$ 1, not hate $\rightarrow$ 0). The **Naseza** dataset required only remapping of its *label* column to the 0/1 scheme while resetting the *id*. In the **PHATE** dataset, the *HateSpeech* column was used as the binary label, the *text* column was retained, and a new sequential *id* was assigned. For **PHICAD**, rows labelled as *spam* or *spamobscene* were removed, the *comment_normalised* column was selected as *text*, and the *class* column was mapped to binary labels (hate and hateobscene $\rightarrow$ 1; obscene and clean $\rightarrow$ 0). This standardisation process enabled seamless integration of Persian and Kurdish data in the subsequent experimental phases.

3.3 Preprocessing Pipeline

All datasets underwent a uniform preprocessing pipeline. First, empty rows and exact duplicate texts were removed to prevent data leakage. Texts shorter than three words were discarded to retain only samples with sufficient semantic context. Subsequently, emojis were removed, repeated spaces collapsed, and Persian/Arabic character variants were normalised to a standard representation. Diacritics were stripped, newline and carriage return characters removed, and common social media noise, including URLs, user mentions, hashtags, excessive punctuation, textual emoticons, and numeric digits, was systematically cleaned. This deterministic pipeline was applied identically across all Persian and Kurdish datasets before splitting and model training, transforming raw social media content into clean, standardised texts while preserving original labels (0 for not hate, 1 for hate).

4 Methodology

This study adopts a multi-paradigm approach to hate speech detection, comparing four distinct methodological frameworks across Persian and Kurdish datasets. The experiments are structured into two phases: Phase I (see section 5.2) evaluates each language independently on smaller, language-specific datasets, while Phase II (see section 5.3) examines the impact of cross-lingual data augmentation by incorporating Persian data (raw or translated) into Kurdish training sets.

4.1 Paradigm A: Rule-based Approach

Rule-based methods serve as baseline systems that rely on linguistic patterns without requiring training data. These approaches include keyword matching using curated lists of hate-related terms in Persian and Kurdish, pattern matching through regular expressions to capture common hate speech structures (e.g., slurs combined with specific targets),

phrase matching for multi-word expressions, and basic negation handling to reduce false positives. While simple and interpretable, these methods are expected to provide lower performance in morphologically rich and context-dependent languages like Persian and Kurdish.

4.2 Paradigm B: Traditional ML

Traditional Machine Learning models were implemented using TF-IDF vectorisation for feature extraction. Four classifiers were evaluated: Logistic Regression, Linear Support Vector Machine (LinearSVC), NuSVC, and Random Forest. These models were trained on the same standardised datasets as other paradigms. Hyperparameters were tuned using grid search with 5-fold cross-validation on the validation set, optimising for macro-F1 score. This paradigm provides a strong classical baseline against which the benefits of deep learning and transfer learning can be measured.

4.3 Paradigm C: Transformer-based Models

Five transformer models were fine-tuned to capture contextual representations, which are particularly suited for low-resource and cross-lingual settings. These include the multilingual models mBERT (Wu and Dredze, 2020) `bert-base-multilingual-cased`, XLM-RoBERTa (Conneau et al., 2020) `xlm-roberta-base`, and mDeBERTa-v3 (He et al., 2023) `mdeberta-v3-base`, as well as language-specific models: ParsBERT (Farahani et al., 2021) `bert-base-parsbert-uncased` for Persian and KuBERT (Awlla et al., 2025) `KuBERT-Central-Kurdish-BERT-Model` for Central Kurdish. All models were fine-tuned with a maximum sequence length of 128 tokens, a learning rate of 2e-5, a batch size of 16 for training and 32 for evaluation, a weight decay of 0.01, and class-weighted cross-entropy loss to address label imbalance. Training ran for up to five epochs with early stopping based on validation macro-F1. The best checkpoint was selected according to validation performance. This setup allows for a direct comparison of general multilingual capabilities against language-specific pre-training.

4.4 Paradigm D: Large Language Models (LLMs)

Large Language Models were evaluated in zero-shot and few-shot settings without any fine-tuning. The models included `Llama-3.1-8B-Instruct` (Team, 2024), `Qwen3-8B` and `Qwen3-32B` (Team, 2025b), and `Gemma-12B-Instruct` (Team, 2025a). For few-shot prompting, eight balanced examples (four hate, four not hate) were randomly selected from the training set and included in the prompt for each test instance. Prompts were designed to be concise and consistent across languages. Greedy decoding was used to ensure deterministic outputs. Open-weight models were employed; this paradigm assesses the out-of-the-box cross-lingual understanding and reasoning capabilities of modern LLMs in low-resource hate speech detection.

All experiments used the same data splits and preprocessing pipeline to ensure a fair comparison. Evaluation was performed on fixed test sets using Accuracy, macro-F1, per-class F1, and confusion matrices, with macro-F1 serving as the primary metric.

5 Experimental Setup

This section describes the experimental configuration across all four paradigms. It covers the rule-based setup, data splitting for Phase I and Phase II, and the evaluation metrics used to compare model-level versus data-level transfer.

5.1 Rule-based Approach

Paradigm A serves as the non-learning baseline for the study. A unified rule-based system was developed and applied consistently across all datasets without any dataset-specific tuning. The system combines four components: curated keyword lists (hate-related terms in Persian and Kurdish), regular expression patterns to capture common hate structures, phrase matching for multi-word expressions, and basic negation handling to mitigate false positives.

A fixed detection threshold of 1.0 was used, meaning a single match against any rule triggers a hate prediction. This conservative setting prioritises precision over recall and ensures full reproducibility, as the approach is entirely deterministic with no randomness or training involved. The results reveal significant variation across languages

and datasets. The system performed markedly better on Persian datasets (especially Naseza) than on Kurdish, highlighting the challenges of applying simple pattern matching to morphologically complex and lower-resource languages like Kurdish. Detailed confusion matrix analysis confirmed a strong bias towards predicting the non-hate class, particularly in the Kurdish dataset, where hate recall was only 0.216.

5.2 Phase I: Resource-Specific Pipelines and Evaluation

Phase I aims to establish strong language-specific baselines for each language and dataset before investigating cross-lingual effects. This phase evaluates all paradigms (Traditional ML, Transformers, and LLMs) independently on three smaller but representative datasets: the original Kurdish dataset (Kurdish), and two Persian datasets (Naseza and PHATE).

For each dataset, fixed train-validation-test splits were used consistently across all models to ensure fair and reproducible comparisons. The Kurdish test set contains 1,148 samples, while the Persian test sets include 1,033 samples for Naseza and 1,400 samples for PHATE. These splits maintain the original class distribution of each dataset, with hate speech comprising approximately 56% of the Kurdish test set and around 43–45% of the Persian test sets. In this phase, models are trained and tested within the same language and domain. This allows us to understand the inherent difficulty of hate speech detection for each resource and provides a clear reference point against which the gains (or losses) from cross-lingual approaches in Phase II can be measured. All preprocessing steps, including emoji removal, Persian/Arabic script normalisation, diacritic stripping, and whitespace collapsing, were applied uniformly.

5.3 Phase II: Cross-Lingual Transfer and Data-vs-Model Level Enhancement

This section addresses the core research question of this study: to *what extent can Persian data enhance hate speech detection performance in Kurdish through cross-lingual transfer?* This phase systematically compares model-level transfer (using multilingual models on raw data) with data-level transfer (augmenting the training data with Persian content, either in its original or translated form). Only the training data composition differs across the configurations described below.

Three primary training configurations were examined:

- **Persian-only** (Combined_Persian): Models trained solely on the combined Persian datasets (Naseza + PHATE + PHICAD) to establish a pure cross-lingual baseline.
- **Raw cross-lingual transfer** (Combined_Kurdish_Persian): Persian data added directly (without translation) to the original Kurdish training set.
- **Translation-based data augmentation** (MT_Kurdish_Aug): Persian data first translated into Central Kurdish using Google Translate and then combined with the original Kurdish data.

To disentangle the contribution of different components, three additional setups were evaluated on the identical Kurdish test set: **Persian-only cross-lingual** (Persian_xling), **translated Persian only** (MT_Kurdish), and the **fully mixed translated** set (Combined_KP_MT). The Persian-to-Kurdish translation was performed using the Google Translate API. After automatic translation, all outputs were manually reviewed and edited by a native Kurdish speaker to ensure linguistic accuracy and cultural appropriateness. This human post-editing step significantly improved the quality of the augmented data.

Special attention was paid to data splits in Phase II. The validation and test sets were constructed as the union of Phase I validation and test sets, with any overlapping instances carefully removed from the training data.

This design eliminates data leakage and enables a clean, apples-to-apples comparison between Phase I (language-specific) and Phase II (cross-lingual) results. By comparing these configurations, the experimental setup allows us to distinguish between model-level transfer (leveraging multilingual model capabilities on raw data) and data-level transfer (augmenting training data via translation), directly supporting the three-way analysis (Baselines vs Model-level vs Data-level Transfer) of the study.

6 Results and Discussion

This section presents the experimental results from both Phase I (language-specific baselines) and Phase II (cross-lingual transfer), followed by a detailed analysis and comparison across all paradigms.

Dataset	Model	Best macro-F1
Kurdish	LinearSVC	0.609
Naseza	LinearSVC	0.801
PHATE	LogisticRegression	0.699

Table 2: Best macro-F1 scores for traditional ML TF-IDF-based models (Paradigm B - Phase I)

Dataset	Model	Best macro-F1
Kurdish	ParsBERT	0.630
Naseza	ParsBERT	0.893
PHATE	ParsBERT	0.791

Table 3: Best macro-F1 scores for Transformer-based models (Paradigm C - Phase I)

6.1 Evaluation of Rule-based Approach

The rule-based approach (Paradigm A), which relied on keyword lists, regex patterns, phrase matching, and basic negation handling, achieved a macro-F1 of 0.495 on Kurdish, 0.797 on Naseza, and 0.629 on PHATE (see Figure 1). This highlights the challenges of pattern-based methods in morphologically rich languages like Kurdish, where recall for the hate class was particularly low (0.216).

6.2 Evaluation of Language-Specific Baseline (Phase I)

Phase I established strong language-specific baselines across the three datasets. As shown in Figure 2, performance varied significantly by language and dataset.

Traditional ML Models (Paradigm B) using TF-IDF features showed clear improvements over the rule-based baseline. LinearSVC consistently performed best, reaching macro-F1 scores of 0.609 on Kurdish, 0.801 on Naseza, and 0.699 with Logistic Regression on PHATE (see table 2). These results confirm that learning-based shallow models provide substantial gains, especially on the more challenging Kurdish (+0.114) and PHATE (+0.070) datasets, while offering only marginal improvement on Naseza, where the rule-based system was already strong.

Transformer-based models (Paradigm C) further advanced performance. ParsBERT emerged as the strongest model overall in Phase I, achieving 0.630 on Kurdish, 0.893 on Naseza, and 0.791 on PHATE. XLM-RoBERTa and mBERT also delivered competitive results, while mDeBERTa-v3 struggled with training stability on these datasets. Language-specific models (ParsBERT and KuBERT) generally outperformed multilingual ones on their respective languages, though the gap was smaller than expected (see Table 3).

In the LLM paradigm (Paradigm D), few-shot prompting with models like Gemma and Qwen3 showed promising results, particularly on Persian datasets (see Table 4). However, zero-shot performance was generally lower, and all LLMs trailled the best fine-tuned transformers.

Overall, Phase I demonstrates that Persian hate speech detection is relatively more mature, while Kurdish remains challenging even for advanced models.

6.3 Impact of Cross-lingual Transfer Strategies (Phase II)

Phase II directly tested the central hypothesis: whether adding Persian data enhances Kurdish hate speech detection. All models were evaluated on the same Kurdish test set for fair comparison.

The most striking finding is the strong performance of pure cross-lingual transfer. Training solely on combined Persian data (Combined_Persian) and testing on Kurdish yielded excellent results with traditional ML (LinearSVC: macro-F1 0.741), significantly outperforming Kurdish baseline (0.609). This suggests substantial overlap in hate speech patterns between Persian and Kurdish. In contrast, mixing raw Persian with original Kurdish led to a drop in performance. The small proportion of Kurdish data (3%) appears to have introduced noise rather than a helpful signal. Table 5 summarises the experiments described in this section.

Translation-based augmentation produced mixed outcomes. The MT_Kurdish_Aug setting (translated Persian + original Kurdish) achieved a modest improvement for some models (e.g., ParsBERT reached 0.640), but overall gains were limited compared to Persian-only cross-lingual transfer. Translation-only data (MT_Kurdish) performed worse than the Kurdish baseline in traditional ML, indicating that current machine translation quality still introduces noticeable noise. Despite manual post-editing by a native speaker, translation-based configurations consistently underperformed pure Persian-only transfer. We attribute this to residual translationese, loss of culture-specific hate expressions, and the morphological complexity of Kur-

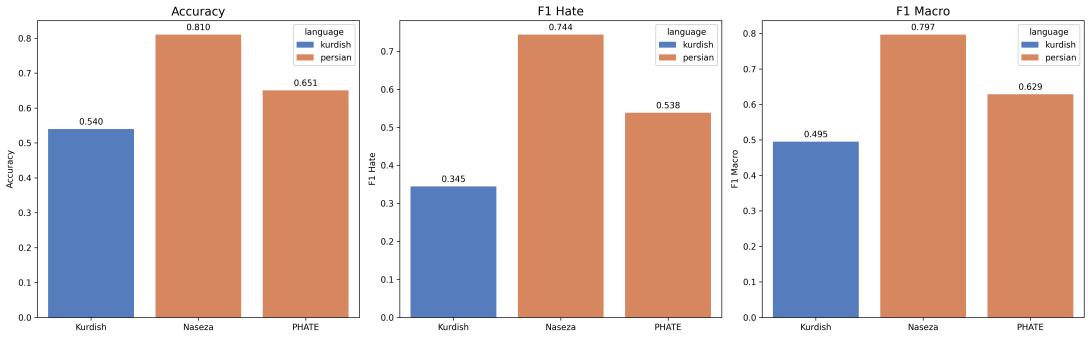


Figure 1: Rule-based approach (Paradigm A) performance across small datasets

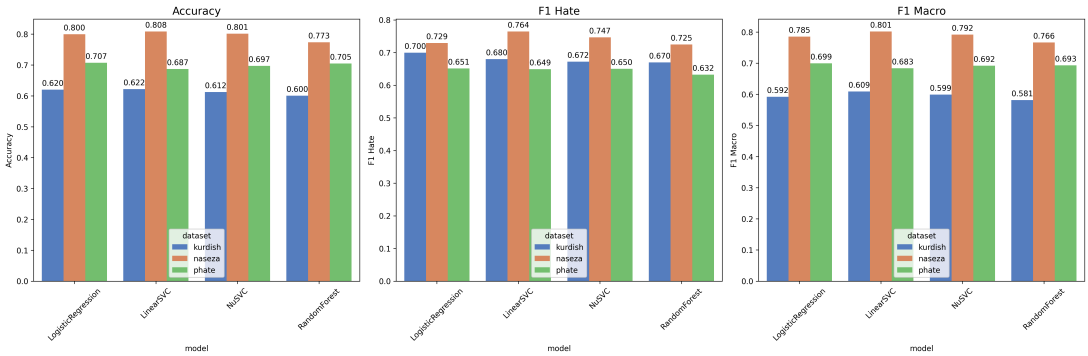


Figure 2: Traditional ML TF-IDF-based models (Paradigm B) - Phase I

Dataset	Model	Best macro-F1	Prompt
Kurdish	Qwen3-8B / Gemma-12B	0.525, 0.499	Few-shot
Naseza	Gemma-12B	0.863	Few-shot
PHATE	Gemma-12B	0.770	Few-shot

Table 4: Best macro-F1 scores for LLM paradigm (Paradigm D - Phase I)

Dataset	Training Data Composition	Best Model	Best macro-F1
combined_persian	Persian data only (Naseza + PHATE + PHICAD)	LinearSVC	**0.741**
combined_kp	Persian + original Kurdish combined	LinearSVC	0.690
mt_kurdish_aug	Original Kurdish + Persian-to-Kurdish translation	RandomForest	0.608
mt_kurdish	Persian-to-Kurdish translation only	RandomForest	0.572
combined_kp_mt	Full combination (Persian + original Kurdish + translation)	LogisticRegression	0.561

Table 5: Best macro-F1 scores for traditional ML TF-IDF-based models (Paradigm B - Phase II)

dish that current machine translation systems (even after post-editing) fail to capture fully. This finding suggests that for closely related languages, model-level transfer on original data can be more robust than imperfect data-level augmentation.

The full combination setting (Combined_KP_MT) yielded the lowest performance among Phase II configurations, further confirming that unrefined addition of translated data can be counterproductive for some model types.

In Phase II, LLM results mirrored transformer trends. Gemma (few-shot) achieved the best performance on the augmented dataset (MT_Kurdish_Aug) with a macro-F1 of 0.642. Persian-only cross-lingual transfer also performed strongly, while translation-heavy configurations

(MT_Kurdish and Combined_KP_MT) generally underperformed, signalling that raw Persian data provides a more reliable signal than current machine-translated data for Kurdish hate speech detection.

7 Conclusion and Future Work

This study systematically examined hate speech detection in Persian and Kurdish through multiple paradigms. Phase I established language-specific baselines, while Phase II demonstrated that Persian data can effectively support Kurdish detection via cross-lingual transfer. Notably, training on Persian data alone often outperformed Kurdish baselines, with ParsBERT consistently delivering the strongest results. In line with recent studies, the best performance was achieved by domain-specific deep learning models (transformers), followed by traditional ML models, both of which outperformed LLMs. Translation-based augmentation provided limited gains and sometimes introduced noise.

Future Directions. Future work should focus on developing better Persian-to-Kurdish translation models, exploring advanced cross-lingual techniques such as meta-learning and adapters, creating larger native Kurdish datasets, and incorporating multimodal signals for more robust real-world moderation systems.

Limitations. The study used a single random seed, and translation quality remained a bottleneck for data-level transfer. Domain differences across datasets and limited tokeniser coverage for Kurdish (especially in mBERT) may also affect generalisability. We did not explore re-balancing the Persian–Kurdish ratio through down-sampling of Persian or up-sampling of Kurdish; future work should investigate optimal mixing ratios. Similarly, the underperformance of LLMs may stem from tokeniser limitations on Kurdish morphology and script, domain mismatch with social-media hate speech, and prompt sensitivity, avenues we leave for future investigation.

Ethics Statement. This work focuses on hate speech detection in Persian and Kurdish, languages spoken in politically sensitive regions. We acknowledge the dual-use risk of such systems for censorship or disproportionate targeting of ethnic or political groups. All datasets used are publicly available and were collected for research purposes. We recommend that any deployment of these models incorporate human-in-the-loop moderation, regular bias audits, and alignment with local cultural and legal definitions of hate speech rather than solely Western frameworks.

Acknowledgments

This research has been partially funded by the University of Alicante, the University of Havana, the Generalitat Valenciana, and the Spanish Ministry of Science and Innovation. At the regional level, the Generalitat Valenciana (Conselleria d'Educació, Investigació, Cultura i Esport) and the European Regional Development Fund (ERDF) provided funding through CIDEGENT (CIDEXG/2023/13). At the national level, support was provided through the HEART-NLP project (PID2024-156263OB-C22).

References

Sina Ahmadi, Daban Q. Jaff, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2024. Language and speech technology for Central Kurdish varieties. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10034–10045, Torino, Italia. ELRA and ICCL.

Ghadeer Al-Badani, Muneer Hazaa Alsurori, and Akram Alsubari. 2025. Transfer learning and cross-linguistic generalization in multilingual hate speech detection: Approaches and challenges. In *2025 5th International Conference on Emerging Smart Technologies and Applications (eSmarTA)*, pages 1–8.

Aish Albladi, Minarul Islam, Amit Das, Maryam Bigonah, Zheng Zhang, Fatemeh Jamshidi, Mostafa Rahgouy, Nilanjana Raychawdhary, Daniela Marghitu, and Cheryl D. Seals. 2025. Hate speech detection using large language models: A comprehensive review. *IEEE Access*, 13:20871–20892.

Wazir Ali and Sampo Pyysalo. 2024. A survey of large language models for european languages. *CoRR*, abs/2408.15040.

Ahmad Amirivojdan. 2025. Naseza: A large-scale dataset for persian hate speech and offensive language detection.

K.M. Awlla, H. Veisi, and A.A. Abdullah. 2025. Sentiment analysis in low-resource contexts: Bert's impact on central kurdish. *Language Resources & Evaluation*, 35(1).

Soran Badawi. 2024. Deep learning-based cyberbullying detection in kurdish language. *Comput. J.*, 67(7):2548–2558.

Irina Bigoulaeva, Viktor Hangya, and Alexander Fraser. 2021. Cross-lingual transfer learning for hate speech

detection. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 15–25, Kyiv. Association for Computational Linguistics.

Zahra Bokaei, Bonnie Webber, and Walid Magdy. 2026. Benchmarking offensive language detection in Persian and Pashto. In *The Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family*, pages 13–23, Rabat, Morocco. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Susmita Das, Arpita Dutta, Kingshuk Roy, Abir Mondal, and Arnab Mukhopadhyay. 2024. A survey on automatic online hate speech detection in low-resource languages. *CoRR*, abs/2411.19017.

Hadi Davardoust, Hadi Zare, and Hossein RafieeZade. 2024. The dark side of instagram: A large dataset for identifying persian harmful comments. *SoCal NLP Symposium 2024*.

Zahra Delbari, Nafise Sadat Moosavi, and Mohammad Taher Pilehvar. 2024. Spanning the spectrum of hatred detection: A persian multi-label hate speech dataset with annotator rationales. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17889–17897. AAAI Press.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2021. Parsbert: Transformer-based model for persian language understanding. *Neural Process. Lett.*, 53(6):3831–3847.

Endrit Fetahi, Arsim Susuri, Mentor Hamiti, Zenun Kastrati, Ercan Canhasi, and Arta Misini. 2025. Enhancing social media hate speech detection in low-resource languages using transformers and explainable AI. *Soc. Netw. Anal. Min.*, 15(1):82.

Anderson Almeida Firmino, Cláudio de Souza Baptista, and Anselmo Cardoso de Paiva. 2024. Improving hate speech detection using cross-lingual learning. *Expert Systems with Applications*, 235:121115.

Abhinav Reddy Gutha, Nidamanuri Sai Adarsh, Ananya Alekar, and Dinesh Reddy. 2023. Multilingual hate speech and offensive language detection of low resource languages. In *Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation (FIRE-WN 2023), Goa, India, December 15-18, 2023*, CEUR Workshop Proceedings, pages 445–458. CEUR-WS.org.

Katharina Hämmerl, Jindrich Libovický, and Alexander Fraser. 2024. Understanding cross-lingual alignment - A survey. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Findings of ACL, pages 10922–10943. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Aiqi Jiang and Arkaitz Zubiaga. 2026. Cross-lingual offensive language detection: A systematic review of dataset, approach and challenge. *ACM Comput. Surv.*, 58(10).

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2025. Natural language processing for dialects of a language: A survey. *ACM Comput. Surv.*, 57(6).

Emad Kebriaei, Ali Homayouni, Roghayeh Faraji, Armita Razavi, Azadeh Shakery, Heshaam Faili, and Yadollah Yaghoobzadeh. 2024. Persian offensive language detection. *Mach. Learn.*, 113(7):4359–4379.

Kyuhan Lee and Sudha Ram. 2024. Deep learning for hate speech detection: A personality-based approach. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 1667–1671, New York, NY, USA. Association for Computing Machinery.

Tanjim Mahmud, Michal Ptaszynski, Juuso Eronen, and Fumito Masui. 2023. Cyberbullying detection for low-resource languages and dialects: Review of the state of the art. *Inf. Process. Manag.*, 60(5):103454.

Sepideh Mollanorozy, Marc Tanti, and Malvina Nissim. 2023. Cross-lingual transfer learning with Persian. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 89–95, Dubrovnik, Croatia. Association for Computational Linguistics.

Marzieh Mozafari, Khouloud Mnassri, Reza Farahbakhsh, and Noel Crespi. 2024. Offensive language detection in low resource languages: A use case of persian language. *PLoS one*, 19(6):e0304166.

Rachna Narula and Poonam Chaudhary. 2024. A comprehensive review on detection of hate speech for multi-lingual data. *Soc. Netw. Anal. Min.*, 14(1):244.

Homa Nasiri and Morteza Analoui. 2022. Persian stance detection with transfer learning and data augmentation. In *2022 27th International Computer Conference, Computer Society of Iran (CSICC)*, pages 1–5.

Daniel Nkemelu, Harshil Shah, Michael Best, and Irfan Essa. 2023. Tackling hate speech in low-resource languages with context experts. In *Proceedings of the 2022 International Conference on Information and Communication Technologies and Development*, ICTD '22, New York, NY, USA. Association for Computing Machinery.

Aillkeen Bezerra de Oliveira, Cláudio de Souza Baptista, Anderson Almeida Firmino, and Anselmo Cardoso de Paiva. 2023. Using multilingual approach in cross-lingual transfer learning to improve hate speech detection. In *ICEIS (1)*, pages 374–384.

Partha Pakray, Alexander Gelbukh, and Sivaji Bandyopadhyay. 2025. Natural language processing applications for low-resource languages. *Natural Language Processing*, 31(2):183–197.

Filippo Pallucchini, Lorenzo Malandri, Fabio Mercorio, and Mario Mezzanzanica. 2026. Lost in alignment: A survey on cross-lingual alignment methods for contextualized representation. *ACM Comput. Surv.*, 58(5):116:1–116:34.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Lang. Resour. Evaluation*, 55(2):477–523.

Gil Ramos, Fernando Batista, Ricardo Ribeiro, Pedro Fialho, Sérgio Moro, António Fonseca, Rita Guerra, Paula Carvalho, Catarina Marques, and Cláudia Silva. 2024. A comprehensive review on automatic hate speech detection in the age of the transformer. *Soc. Netw. Anal. Min.*, 14(1):204.

Ethan Roberts. 2024. Automated hate speech detection in a low-resource environment. *Journal of the Digital Humanities Association of Southern Africa (DHASA)*, 5(1).

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Ari M. Saeed, Aso N. Ismael, Danya L. Rasul, Rayan S. Majeed, and Tarik A. Rashid. 2022. Hate speech detection in social media for the kurdish language. In *Proceedings of the ICR'22 International Conference on Innovations in Computing Research, Athens, Greece, 29-31 August, 2022*, Advances in Intelligent Systems and Computing, pages 253–260. Springer.

Deepawali Sharma, Tanusree Nath, Vedika Gupta, and Vivek Kumar Singh. 2025. Hate speech detection research in south asian languages: A survey of tasks, datasets and methods. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 24(3):25:1–25:44.

Mohammad Karami Sheykhlan, Jana Shafi, Saeed Kosari, Saleh Kheiri Abdoljabbar, and Jaber Karimpour. 2023. Pars-hao: Hate and offensive language detection on persian tweets using machine learning and deep learning.

Xiayang Shi, Xinyi Liu, Chun Xu, Yuanyuan Huang, Fang Chen, and Shaolin Zhu. 2022. Cross-lingual offensive speech identification with transfer learning for low-resource languages. *Computers and Electrical Engineering*, 101:108005.

Gemma Team. 2025a. Gemma 3 technical report. *CoRR*, abs/2503.19786.

Llama Team. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Qwen Team. 2025b. Qwen3 technical report. *CoRR*, abs/2505.09388.

Muhammad Usman, Muhammad Ahmad, M Shahiki Tash, Irina Gelbukh, Rolando Quintero Tellez, and Grigori Sidorov. 2025. Multilingual hate speech detection in social media using translation-based approaches with large language models. *arXiv preprint arXiv:2506.08147*.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? In *Proceedings of the 5th Workshop on Representation Learning for NLP, RepL4NLP@ACL 2020, Online, July 9, 2020*, pages 120–130. Association for Computational Linguistics.

Is Overconfidence Language-Specific? Cross-Lingual Calibration and Recalibration Transfer in Base and Instruction-Tuned LLMs

Divya Gupta
Lancaster University
divya@lancaster.ac.uk

Ruslan Mitkov
Lancaster University
r.mitkov@lancaster.ac.uk

Abstract

Large language models are often treated as if the probability assigned to an answer were a reliable measure of confidence. Most evidence for this assumption, however, comes from English. This paper asks two related questions: whether instruction tuning changes calibration across languages, and whether a calibration correction learned in one language transfers to others. We compare base and instruction-tuned versions of Qwen2.5-7B and Llama-3.1-8B on Global-MMLU and Belebele, covering 42 languages across high-, mid-, and low-resource groups. Confidence is measured from the normalized next-token probabilities of the four answer options, and calibration is evaluated using expected calibration error , overconfidence error, and the multiclass Brier score. Instruction-tuned models show higher ECE than their base counterparts in 154 of 168 model–benchmark–language comparisons at an unadjusted $p < 0.05$ threshold, and in 135 of 168 after Holm–Bonferroni correction for multiple comparisons (see Section 3.6), while accuracy changes much less consistently. The increase is largest in lower-resource languages. Temperature scaling fitted separately for each language substantially improves calibration across all three resource groups. A temperature fitted on English transfers reasonably well to high- and mid-resource targets, but leaves a larger residual error for low-resource targets. These results suggest that multilingual overconfidence is partly language-specific and that an English-only post-hoc correction may be least effective where labelled calibration data are scarcest. Because base and instruction-tuned models are evaluated in their respective released prompting formats , we describe this as a property of the released systems rather than an isolated causal effect of the instruction-tuning stage itself. The findings apply to multiple-choice confidence and should not yet be generalized to open-ended generation.

1 Introduction

An accurate answer is not the same as a trustworthy confidence estimate. This distinction matters whenever a language model must decide whether to answer, abstain, retrieve additional information, or defer to a person. A model is calibrated when predictions made with confidence p are correct approximately a proportion p of the time (Naeini et al., 2015; Guo et al., 2017). For example, among answers assigned 80% confidence, roughly 80% should be correct. If the model is wrong far more often, its confidence can encourage unsafe decisions even when its overall accuracy appears acceptable.

Instruction tuning and reinforcement learning from human feedback make models easier to use, and prior work has found that this alignment process is often accompanied by increased overconfidence (Desai and Durrett, 2020; Tian et al., 2023; OpenAI, 2023). Most of this evidence is based on English. That leaves a practical multilingual question unanswered: does this pattern hold in the same way across languages, including those with much less training and evaluation data?

Recent work by Huang et al. (2026) reports that instruction tuning can inflate confidence in lower-resource languages without producing comparable accuracy gains. A key contribution of the present study is that it independently corroborates this pattern in a controlled base-versus-instruction comparison using two different model families and two parallel multilingual benchmarks. This is important because it tests whether the pattern generalizes across model families and evaluation setups, although as we discuss in Section 3 prompt format necessarily differs between base and instruction-tuned variants, so our results describe the behaviour of released systems in their intended prompting format rather than isolating a causal effect of the

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 361–369
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_037

tuning stage itself.

The standard post-hoc correction is temperature scaling (Guo et al., 2017). It is inexpensive and requires only one fitted parameter, but it still needs labelled calibration examples. A separate labelled set for every language is unrealistic in many settings. We therefore construct a cross-lingual recalibration matrix: for each source language, we fit a temperature and apply it to every target language. This makes it possible to see whether the correction transfers or whether languages require meaningfully different temperatures.

The study seeks to answer several connected questions within a single narrative. It establishes how well calibrated base language models are compared with their instruction-tuned counterparts across languages of differing resource levels. It then investigates whether instruction-tuned models are more overconfident per language, whether this is accompanied by comparable accuracy gains, and whether calibration error systematically grows as language resource level decreases. Finally, it asks a deployment-oriented question: how much calibration quality is lost when a temperature fitted on one language, especially English, is transferred to another?

We also introduce and analyse the transfer matrix as a compact way to study multilingual recalibration. This extends closely related prior work on cross-lingual post-hoc calibration transfer (Ahuja et al., 2022; Jiang et al., 2022) to a controlled base-versus-instruction comparison spanning two model families and two parallel benchmarks, giving one of the more detailed views to date of how a one-language post-hoc calibration correction behaves across a broad set of target languages. This matters for practical multilingual systems because English is often the easiest language in which to collect validation data, while the greatest calibration risks may occur elsewhere.

Our results point to a consistent pattern. Instruction-tuned models are more overconfident than their base counterparts in most comparisons, and the gap is generally larger in lower-resource languages. Language-specific temperature scaling improves calibration substantially, but a temperature learned on English leaves more residual error for low-resource targets. The result is not that multilingual calibration is impossible. Rather, the easiest global correction does not work equally well for every language.

The rest of the paper is structured as follows. Section 2 reviews related work on calibration, multilingual evaluation, and multilingual confidence estimation. Section 3 describes the benchmarks, languages, models, confidence extraction procedure, calibration metrics, and cross-lingual temperature-scaling setup. Section 4 presents the main empirical results. Section 5 discusses their implications for multilingual deployment, Section 6 summarizes limitations and open methodological questions, Section 7 provides additional experimental details, and Section 8 concludes.

2 Related Work

Calibration in neural language models. Modern neural networks can produce probabilities that do not match their empirical accuracy, even when their classifications are correct frequently (Guo et al., 2017). Expected calibration error (ECE) is a widely used summary of this mismatch (Naeini et al., 2015). Within natural language processing, Desai and Durrett (2020) studied the calibration of pretrained transformers, while Jiang et al. (2021) examined whether question-answering systems can recognize when they are likely to be wrong. Work on aligned language models has similarly found that useful instruction-following behaviour does not guarantee reliable confidence (Kadavath et al., 2022; Tian et al., 2023).

Multilingual evaluation. Language-model performance varies considerably across languages, reflecting uneven training data, benchmark coverage, and research investment (Joshi et al., 2020; Ahuja et al., 2023). Parallel benchmarks help separate language effects from task effects by presenting equivalent content in multiple languages. Global-MMLU provides multilingual versions of broad knowledge questions (Singh et al., 2024), while Belebele offers parallel reading-comprehension tasks built on FLORES-200 passages (Bandarkar et al., 2024).

Multilingual confidence. Evidence on multilingual calibration remains limited. Xue et al. (2025) finds that confidence estimation is stronger in English than in many other languages. Huang et al. (2026) reports that instruction tuning can increase confidence in lower-resource languages without a matching accuracy improvement and proposes label smoothing as a training-time response. Closer to the deployment question we study, Ahuja et al.

(2022) directly compare a temperature fitted on English against a target-language "Self-TS" correction, and Jiang et al. (2022) study how a post-hoc calibration correction fitted only on a source language transfers across target languages. Our work extends this line of investigation with a base-versus-instruction comparison across two model families and two parallel multilingual benchmarks, and, unlike this prior work, examines whether an English-only correction is differentially effective by target-language resource level.

3 Methodology

3.1 Benchmarks and Languages

We use two parallel multiple-choice benchmarks with gold answers. Global-MMLU tests broad subject knowledge across translated versions of the same multiple-choice questions. It is evaluated with five-shot demonstrations drawn from its development split, so the prompt includes examples in the target evaluation format before the test item. Belebele tests reading comprehension using parallel passages derived from FLORES-200, with each example consisting of a passage, a question, four answer options, and one correct answer. Belebele is evaluated zero-shot because it does not provide a development split.

For each benchmark–language pair, we use up to 500 gold-labelled test items. The analysis data are therefore item-level records containing the benchmark, language, model family, model variant, predicted answer option, correctness, and the four answer-option logits used to compute confidence. Keeping the answer space fixed to the letters {A, B, C, D} lets us compare calibration across languages without introducing language-specific verbal answer forms.

The evaluation covers 42 languages: 17 in the reported high-resource group, 13 in the mid-resource group, and 12 in the low-resource group. We also retain the original six-level resource class from Joshi et al. (2020) for correlation analyses, where 0 denotes the lowest-resource class and 5 the highest. The three descriptive groups and the six-level classes should be treated as related but distinct variables; the exact assignments used in the experiments are listed in Section 7.1.

3.2 Models

We compare base and instruction-tuned variants from two families: Qwen2.5-7B (Qwen Team, 2025) and Llama-3.1-8B (Grattafiori et al., 2024). Within a family, the models share a pretraining lineage but differ in their instruction-tuning stage. Base models receive completion-style prompts, while instruction-tuned models use their published chat templates. Models run in half precision at batch size 1 when confidence is recorded.

This is a controlled within-family comparison, but it is not a pure intervention on model weights: the prompt format and chat template necessarily differ between base and instruction-tuned variants, so model weights, prompt formatting, and answer cues all change simultaneously between conditions. We therefore report our findings as a comparison of the released base and instruction-tuned systems under their respective intended prompting setups, rather than as evidence that the instruction-tuning stage in isolation causes the observed calibration differences. Establishing a strictly causal claim would require identical-prompt or template ablations, which we leave to future work.

3.3 Prediction and Confidence

Every item has four answer options and one gold answer. We read the next-token logits, restrict them to the answer letters {A, B, C, D}, and normalize those four values with a softmax. The option with the largest probability is the prediction, and its normalized probability is used as confidence.

This design gives an identical answer space across scripts and avoids probability being spread over multiple verbal forms of the same answer (Holtzman et al., 2021). It also means that confidence is conditional on the four answer letters. It is not the model's probability of producing a valid answer in unrestricted generation, so the results should be interpreted as multiple-choice calibration.

3.4 Calibration Metrics

Let $\hat{p}_i$ be the confidence for item i and let $c_i \in \{0, 1\}$ indicate whether the prediction is correct. With $M = 10$ equal-width confidence bins B_m, expected calibration error is

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|. \quad (1)$$

Lower ECE is better. We also report overconfidence error (OE), which counts only bins in which

mean confidence exceeds mean accuracy,

$$\mathrm{OE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \max\left(\mathrm{conf}(B_m) - \mathrm{acc}(B_m), 0\right), \tag{2}$$

and the multiclass Brier score. Accuracy is always shown alongside calibration because a uniformly uncertain but unhelpful model can still appear well calibrated.

3.5 Temperature Scaling and Cross-Lingual Transfer

Temperature scaling divides the option logits by a positive scalar T before the softmax. For each language, T is fitted by minimizing negative log-likelihood on a seeded 30% calibration split. Calibration is then evaluated on that language's held-out 70% split.

Let T_S be the temperature fitted on source language S, and let $\mathrm{ECE}_G(T)$ be the held-out ECE of target language G after applying T. We define the transfer difference

$$D_{S \to G} = \mathrm{ECE}_G(T_S) - \mathrm{ECE}_G(T_G). \tag{3}$$

A value near zero means the source temperature performs similarly to the target language's own fitted temperature. A positive value means transfer leaves additional ECE. A negative value is possible because temperatures are fitted to negative log-likelihood rather than directly to ECE and because both are estimated from finite samples; it means the transferred temperature happened to obtain lower held-out ECE. For this reason, we refer to D as a difference rather than assuming it is always a penalty.

The English-source row represents a realistic one-language correction. Averaging $D_{\mathrm{English} \to G}$ by target resource group shows whether this correction becomes less effective for lower-resource languages.

3.6 Uncertainty and Statistical Testing

Metrics use 1,000 bootstrap resamples to obtain 95% confidence intervals. Base-versus-instruction comparisons use paired resampling over the shared items. Because 168 paired significance tests are run, we apply a Holm–Bonferroni correction to control the family-wise error rate at $\alpha = 0.05$. Of the 154 comparisons significant at the unadjusted $p < 0.05$ threshold, 135 remain significant after correction, all in the direction of higher ECE for the instruction-tuned model. The direction of the effect is consistent across the vast majority of comparisons even where individual significance does not survive correction. We also relate per-language ECE to the six-level resource class using Spearman and Pearson correlations. Random operations use seed 42.

4 Results

4.1 Instruction-Tuned Models Show Higher Calibration Error

Instruction-tuned models show higher calibration error than base models in nearly every part of the evaluation. On Global-MMLU, mean ECE for the base models is 0.06, 0.10, and 0.14 across the high-, mid-, and low-resource groups. For the instruction-tuned models, the corresponding values rise to 0.26, 0.31, and 0.35. Belebele shows the same ordering: 0.05, 0.06, and 0.12 for the base models, compared with 0.14, 0.18, and 0.36 after instruction tuning.

Before correction for multiple comparisons, the increase is significant in 154 of 168 paired language-level comparisons at $p < 0.05$; after Holm–Bonferroni correction, 135 of 168 remain significant at the family-wise level (see Section 3.6). The calibration change is much larger and more consistent than the accuracy change. For Qwen2.5-7B on Yoruba, for example, accuracy moves from 0.30 to 0.33 while ECE rises from 0.15 to 0.50. On Hindi, the same model's accuracy changes from 0.46 to 0.43 while ECE rises from 0.14 to 0.49. These examples illustrate why accuracy alone cannot describe whether confidence is trustworthy. As noted in Section 3, because prompt format differs between the base and instruction-tuned conditions, we interpret this as a property of the released systems in their intended prompting formats rather than isolated evidence that the instruction-tuning stage alone is the causal mechanism.

4.2 Calibration Error and Resource Level

The relationship between resource level and calibration is visible across both benchmarks, though its strength varies by model and benchmark. On Global-MMLU, the Spearman correlation between resource class and ECE is $\rho = -0.69$ for base models ($p < 10^{-12}$) and $\rho = -0.30$ for instruction-tuned models ($p < 0.01$). On Belebele, the corresponding correlations are -0.50 and -0.68. Negative values mean that calibration error tends to increase as the resource class falls.

The pattern is stronger for Qwen2.5-7B-Instruct than for Llama-3.1-8B-Instruct in the Global-MMLU visualization ($\rho = -0.83$ versus -0.33). Given this range from a moderate association ($\rho \approx -0.30$ to -0.33) to a strong one ($\rho \approx -0.69$ to -0.83) depending on the specific model–benchmark combination we treat the general claim that calibration error grows as resources fall as a pattern that holds most clearly for some model–benchmark pairs (notably Qwen2.5-7B on Global-MMLU) and more weakly for others (notably Llama-3.1-8B-Instruct on Global-MMLU), rather than a single uniform trend across all settings. Resource level is not a complete explanation, either. Hindi is assigned Joshi class 4 and grouped as high-resource here (Table 4) but still shows a large calibration deterioration for Qwen. Resource labels summarize broad data conditions; they do not determine the behaviour of every model–language pair.

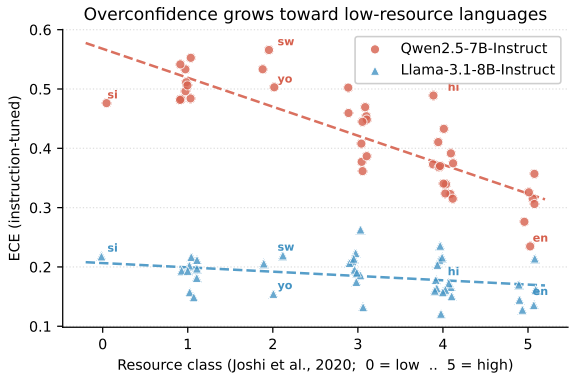


Figure 1: Instruction-tuned ECE against the six-level Joshi resource class on Global-MMLU. Each point is one language. Lower-resource languages tend to have higher ECE, with a steeper relationship for Qwen2.5-7B-Instruct ($\rho = -0.83$) than for Llama-3.1-8B-Instruct ($\rho = -0.33$).

4.3 Language-Specific Scaling

Fitting a temperature separately for each language substantially reduces ECE in every resource group. On Belebele, ECE falls from 0.140 to 0.064 for the high-resource group, from 0.177 to 0.063 for the mid-resource group, and from 0.361 to 0.070 for the low-resource group. Global-MMLU shows a similar result: ECE falls from 0.251, 0.302, and 0.356 to 0.066, 0.063, and 0.060.

The fitted temperature also increases as resource level falls. On Global-MMLU, the mean fitted value rises from 5.82 in the high-resource group to 7.14 in the low-resource group. On Belebele, it rises from 4.03 to 6.35. Because larger temperatures flatten the predicted probability distribution, these values provide another indication that the lower-resource predictions begin more overconfident.

Model	Var.	Acc.	ECE	OE	Brier
Qwen2.5-7B English					
Base		0.73	0.02	0.01	0.14
Instruct		0.73	0.24	0.24	0.24
Qwen2.5-7B Hindi					
Base		0.46	0.14	0.14	0.28
Instruct		0.43	0.49	0.49	0.47
Qwen2.5-7B Yoruba					
Base		0.30	0.15	0.15	0.23
Instruct		0.33	0.50	0.50	0.51
Llama-3.1-8B English					
Base		0.64	0.05	0.02	0.19
Instruct		0.65	0.15	0.14	0.23
Llama-3.1-8B Hindi					
Base		0.40	0.11	0.11	0.26
Instruct		0.43	0.18	0.18	0.28
Llama-3.1-8B Yoruba					
Base		0.33	0.13	0.13	0.27
Instruct		0.32	0.16	0.15	0.25

Table 1: Selected Global-MMLU results. Lower ECE, OE, and Brier scores are better.

Bench.	Group	$T^\star$	ECE bef.	ECE aft.
Belebele	High	4.03	0.140	0.064
	Mid	4.56	0.177	0.063
	Low	6.35	0.361	0.070
Global-MMLU	High	5.82	0.251	0.066
	Mid	6.57	0.302	0.063
	Low	7.14	0.356	0.060

Table 2: Temperature scaling by benchmark and reported resource group. ECE is evaluated on held-out data.

4.4 English Calibration Transfers Unevenly

The central deployment result concerns the difference between language-specific and transferred scaling. A temperature fitted on English performs nearly as well as a target-specific temperature for the high-resource group. The mean English-to-target ECE difference is approximately 0.00 on Global-MMLU and 0.01 on Belebele. For the mid-resource group, it is 0.02 and 0.01. For the low-resource group, however, it rises to 0.05 on Global-MMLU and 0.08 on Belebele.

The full transfer matrix shows that the effect is not explained by English alone. Temperatures learned from different sources produce different outcomes across targets, and the largest positive

Target group	Target ECE	GMMLU D	Belebele D
High	0.07	≈ 0.00	0.01
Mid	0.06	0.02	0.01
Low	0.06	0.05	0.08

Table 3: Held-out ECE after target-specific temperature scaling and the ECE difference when the English-fitted temperature is used instead. Values are averaged across the two model families.

differences are concentrated around several lower-resource languages. Figure 2 summarizes the transfer difference aggregated by resource tier for Qwen2.5-7B-Instruct on Global-MMLU: a temperature fitted on a high-resource source leaves the most residual ECE on low-resource targets, and vice versa, while same-tier transfer is close to zero. The full 42-language matrix is shown in Figure 3 (Appendix 7), where the diagonal is zero by definition because each language is compared with its own temperature; the remaining three model–benchmark matrices are provided in the released artifact (Section 5).

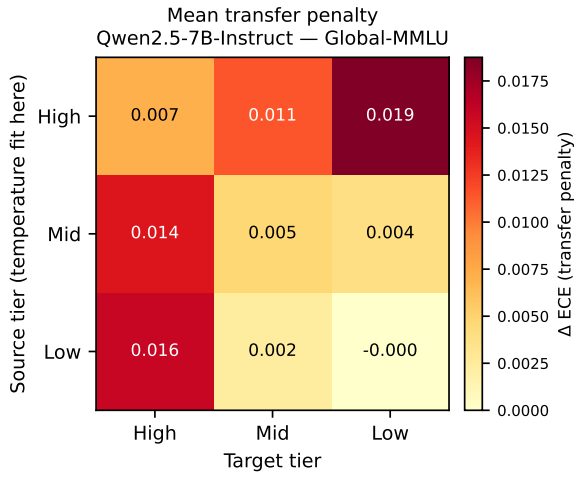


Figure 2: Cross-lingual recalibration difference $D_{S \to G}$ aggregated by resource tier (source tier $\times$ target tier) for Qwen2.5-7B-Instruct on Global-MMLU. Each cell averages all source–target language pairs falling in the corresponding tiers. Positive values mean the source-tier temperature leaves more held-out ECE than the target's own temperature. These tier-averaged values are computed over all source languages and differ from the English-only, family-averaged values in Table 3.

5 Discussion

The experiments reveal two connected patterns. First, the released instruction-tuned systems we evaluate are less well calibrated than their base counterparts under their respective intended prompting formats, and this difference in confidence is often much larger than the corresponding difference in accuracy. Second, this pattern is not uniform across languages: lower-resource languages generally show greater calibration error and are less completely corrected by a temperature learned on English.

This matters for systems that use confidence operationally. Suppose a model answers a question in Yoruba with 90% confidence. If comparable predictions are correct only half the time, a policy that accepts all answers above an 80% threshold will fail. Temperature scaling can reduce this mismatch without retraining the model, but the present results suggest that one English-derived threshold or temperature should not automatically be treated as multilingual.

The results also expose a data problem. A target-specific temperature uses little data compared with model training, but it still requires labelled examples. Those examples are hardest to obtain for the languages with the largest calibration gaps. A training-time intervention, such as the label-smoothing approach explored by Huang et al. (2026), may therefore be attractive when per-language calibration sets are unavailable. This is a practical implication rather than proof that a particular training-time method will solve the problem across all settings.

The findings should be interpreted alongside accuracy. Good calibration does not make a weak model useful, and poor accuracy can interact with ECE in unintuitive ways. Likewise, a high-level resource label cannot explain every language. Script, tokenizer coverage, benchmark translation, prompt format, and the distribution of pretraining data may each contribute. Future work should attempt to separate these mechanisms rather than treating "resource level" as a causal variable, and, more generally, should use identical-prompt or template ablations to more directly isolate the causal contribution of the instruction-tuning stage itself from the accompanying prompt-format change.

Data and code availability. We release the evaluation code, the per-language calibration/held-out splits, the item-level logit records, and the analysis and figure-generation scripts needed to reproduce every reported ECE, OE, Brier, temperature-scaling, and transfer-matrix result.

6 Limitations and Open Methodological Questions

Multiple-choice scope. Confidence is calculated from a softmax restricted to four answer letters. This creates a clean cross-lingual comparison, but it does not measure confidence in open-ended generation. The findings should be tested using free-form answers, verbalized confidence, and selective prediction before being generalized to conversational systems.

Benchmark construction. Global-MMLU contains machine-translated material. Translation quality may vary across languages and may contribute to both accuracy and calibration differences. Human-verified subsets should be reported separately where possible.

Prompt and format adherence. Base models sometimes fail to follow the answer-letter format, particularly in non-Latin scripts. This can reduce accuracy and alter the relationship between accuracy and calibration. We therefore treat format adherence as a possible source of residual variation that should be measured explicitly in future work.

Parallel-item splitting. The experiments use a 30% calibration split for temperature fitting and a 70% held-out split for evaluation. A stricter multilingual protocol would ensure that the same item identifiers are used for the calibration split in every translation of a benchmark, preventing translated versions of a calibration item from appearing in another language's held-out evaluation.

Statistical multiplicity and ECE sensitivity. As reported in Section 3.6, we apply a Holm–Bonferroni correction to the 168 paired significance tests; 135 of the 154 comparisons significant at the unadjusted threshold survive correction, all in the same direction. Because ECE depends on the number and placement of bins, robustness checks with adaptive bins, different bin counts, and a proper scoring rule would also strengthen the conclusion.

Confidence normalisation. Confidence is defined via a softmax over the four answer-letter logits (Section 3), and we have not tested sensitivity to this choice against alternative normalisations or temperature-independent metrics; we leave this robustness check to future work.

Models and alignment methods. The study covers two families at the 7–8B scale and focuses on their released base and instruction-tuned variants. Larger systems (e.g., 70B-parameter variants) and models trained with different mixtures of supervised fine-tuning and reinforcement learning, or additional model families beyond Qwen and Llama, may behave differently; we leave this scaling and cross-family generalisation question to future work.

Resource categories. The three broad groups are necessarily coarse and do not map one-to-one onto the six-level Joshi classes used in the correlation analysis. We document the exact assignments in Section 7.1. More languages, including Marathi and Punjabi, would improve coverage of Indic language families.

7 Additional Experimental Details

To make the experimental setup self-contained, we include the language/resource assignments and per-tier summary results as part of the main paper rather than as separate supplementary material. These tables document the experimental setup and make the resource-group comparisons easier to inspect directly.

7.1 Languages and Reported Resource Groups

Table 4 reports the descriptive resource groups and six-level Joshi classes used in the experiments. The descriptive groups do not correspond exactly to fixed ranges of the six-level class, so we report both variables explicitly. Language codes follow ISO 639-1; see the legend below the table for the full language names.

7.2 Per-Tier Accuracy and Calibration

Table 5 reports mean accuracy and ECE broken down by benchmark, model family, resource group, and variant, summarizing the per-tier patterns discussed above.

7.3 Full Cross-Lingual Transfer Matrix

Figure 3 shows the complete 42×42 recalibration transfer matrix for Qwen2.5-7B-Instruct on Global-MMLU, with rows ordered from high- to low-resource. Each cell is the held-out ECE difference $D_{S \to G}$ between applying the source-language temperature and the target's own temperature. The diagonal is zero by construction. The three remaining model–benchmark matrices are provided in the released artifact (Section 5).

L.	Grp	Cl	L.	Grp	Cl
en	High	5	he	Mid	3
es	High	5	ro	Mid	3
fr	High	5	sr	Mid	3
de	High	5	sv	Mid	4
zh	High	5	uk	Mid	3
ru	High	4	ms	Mid	3
pt	High	4	fil	Mid	3
it	High	4	lt	Mid	3
ja	High	5	te	Mid	1
ar	High	4	am	Low	1
hi	High	4	ha	Low	2
ko	High	4	ig	Low	1
nl	High	4	ky	Low	1
pl	High	4	mg	Low	1
tr	High	4	ne	Low	1
vi	High	4	ny	Low	1
id	High	3	si	Low	0
bn	Mid	3	sn	Low	1
cs	Mid	4	so	Low	1
el	Mid	3	sw	Low	2
fa	Mid	4	yo	Low	2

Table 4: Languages, reported descriptive groups, and six-level resource classes. L. = ISO 639-1 code, Grp = descriptive resource group, Cl = six-level Joshi class. *Legend:* en=English, es=Spanish, fr=French, de=German, zh=Chinese, ru=Russian, pt=Portuguese, it=Italian, ja=Japanese, ar=Arabic, hi=Hindi, ko=Korean, nl=Dutch, pl=Polish, tr=Turkish, vi=Vietnamese, id=Indonesian, bn=Bengali, cs=Czech, el=Greek, fa=Persian, he=Hebrew, ro=Romanian, sr=Serbian, sv=Swedish, uk=Ukrainian, ms=Malay, fil=Filipino, lt=Lithuanian, te=Telugu, am=Amharic, ha=Hausa, ig=Igbo, ky=Kyrgyz, mg=Malagasy, ne=Nepali, ny=Chichewa, si=Sinhala, sn=Shona, so=Somali, sw=Swahili, yo=Yoruba.

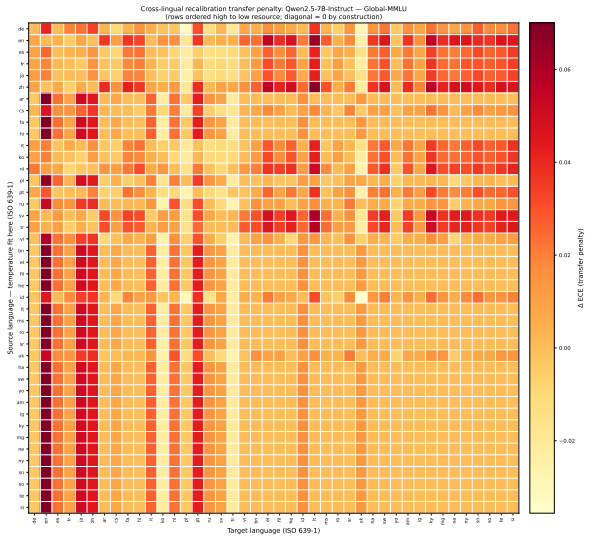


Figure 3: Full cross-lingual recalibration transfer matrix for Qwen2.5-7B-Instruct on Global-MMLU. Rows are source languages (temperature fitted here), columns are target languages, both ordered high- to low-resource by ISO 639-1 code. Positive (darker) values mean the source temperature leaves more held-out ECE than the target-specific temperature; small negative values can occur because of finite-sample variation and the difference between the fitting (NLL) and evaluation (ECE) objectives.

Bench.	Group	B.Acc	B.ECE	I.Acc	I.ECE
Llama-3.1-8B					
Belebele	High	0.78	0.06	0.69	0.14
Belebele	Mid	0.73	0.07	0.66	0.14
Belebele	Low	0.41	0.11	0.42	0.19
GMMLU	High	0.53	0.06	0.51	0.17
GMMLU	Mid	0.45	0.09	0.44	0.19
GMMLU	Low	0.33	0.14	0.35	0.19
Qwen2.5-7B					
Belebele	High	0.82	0.04	0.85	0.14
Belebele	Mid	0.72	0.05	0.76	0.22
Belebele	Low	0.35	0.13	0.38	0.52
GMMLU	High	0.62	0.06	0.60	0.35
GMMLU	Mid	0.50	0.10	0.50	0.43
GMMLU	Low	0.33	0.14	0.32	0.51

Table 5: Mean accuracy and ECE by benchmark, model family, resource group, and variant. B. = Base, I. = Instruction-tuned.

8 Conclusion

Across two model families, two multilingual benchmarks, and 42 languages, instruction-tuned models are generally more overconfident than their base counterparts in their respective released prompting formats. The increase in confidence is often not matched by an increase in accuracy, and calibration error tends to be larger for lower-resource languages, though the strength of this relationship varies by model and benchmark. Fitting a temperature separately for each language substantially improves calibration, showing that the problem is repairable when labelled examples are available. A temperature learned only from English, however, leaves progressively more residual error for lower-resource targets.

The practical lesson is straightforward: confidence calibration should be evaluated in the language in which a model will be used. English calibration is a useful baseline, but the evidence here does not support treating it as a universal multilingual correction. Further work should test open-ended generation, use identical-prompt ablations to more directly isolate the causal role of instruction tuning, report corrected significance tests and format-adherence rates, and establish whether aligned training-time methods can reduce the need for separate calibration data in every language.

Acknowledgments

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, et al. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Kabir Ahuja, Sunayana Sitaram, Sandipan Dandapat,

and Monojit Choudhury. 2022. On the calibration of massively multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4310–4323, Abu Dhabi, United Arab Emirates.

Lucas Bandarkar, Davis Liang, Benjamin Muller, et al. 2024. The Belebele benchmark: A parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 749–775.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Aaron Grattafiori, Abhimanyu Dubey, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Jerry Huang, Peng Lu, Qiuhao Zeng, et al. 2026. Investigating the multilingual calibration effects of language model instruction-tuning. *arXiv preprint arXiv:2601.01362*.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Zhengping Jiang, Anqi Liu, and Benjamin Van Durme. 2022. Calibrating zero-shot cross-lingual (un-)structured predictions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2648–2674, Abu Dhabi, United Arab Emirates.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Saurav Kadavath, Tom Conerly, Amanda Askell, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Qwen Team. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, et al. 2024. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*.

Katherine Tian, Eric Mitchell, Allan Zhou, et al. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Boyang Xue, Hongru Wang, Rui Wang, et al. 2025. MlingConf: A comprehensive study of multilingual confidence estimation on large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*.

Knowledge Tracing for Early Childhood Learners with Minimal Telemetry from Low-Connectivity Environments: Evidence from India's *Anganwadi* Ecosystem

Badmavasan Kirouchenassamy[1] **Rahul Singh**[1]
Vishnu Dev[1] Meenakshi Garg[1]
Sukhna Sawhney[1]
Sudeep Gowrishankar[1]
[1]Rocket Learning
{badmavasan.k, rahul.s, vishnu.d,
meenakshi.g, sukhna.s, sudeep.g}@rocketlearning.org

Abstract

Many of the students who could benefit most from distance learning lack the infrastructure it assumes. Most knowledge-tracing research takes personal devices, stable internet, and rich interaction logs for granted, and it has mostly been developed and evaluated with older learners. Early childhood in low-connectivity environments is therefore still underexplored. We ask whether two established Knowledge Tracing (KT) models, Bayesian Knowledge Tracing (BKT) and the Dynamic Key-Value Memory Network (DKVMN), can accurately model early-childhood Hindi literacy in India's *Anganwadi* ecosystem. Both were designed for high-resource settings, whereas here browser-based Learning Games are played on shared phones over slow, intermittent connections. Using only a timestamped sequence of correct and incorrect responses from 210,500 question attempts by 7,547 learners aged 3 to 6, we validate the models' mastery estimates against offline field assessments of 552 children. DKVMN predicts the next answer well (AUC 0.858), on par with high-resource benchmarks, and its mastery estimates correlate significantly with field assessments from as few as three attempts. BKT is less accurate overall, but gives more reliable per-skill estimates, interpretable diagnostics, and greater robustness to submissions lost to poor connectivity. These results show that meaningful learner modeling is feasible for early childhood education in low-connectivity environments, and that evaluating such models requires external validation against offline assessments and robustness to data loss.

1 Introduction

The Anganwadi system is the world's largest publicly funded early-childhood care and development program, providing nutrition, health, daycare, and early-childhood education (ECE) through a nationwide network of Anganwadi Centers (AWCs). For many low-resource families, these centers are the primary point of access to ECE (ASER Centre, 2025). Rocket Learning (RL), an Indian nonprofit, partners with this system by training Anganwadi Workers (AWWs), building WhatsApp communities that connect workers and parents, and delivering play-based activities for AWCs and home. RL also sends hyperlinks to mobile browser-based Learning Games (LGs) over WhatsApp, letting children practice foundational skills without installing an app (Rocket Learning, 2024).

RL's model already drives large-scale engagement and learning gains across several Indian states, but evaluation at this scale still relies on periodic, sampled field measurements, which makes it hard to monitor every child's progress. Reliable learner-level estimates from LGs could instead flag children who need support, personalize learning, and speed up content iteration. We do not propose a new platform; we strengthen RL's intervention by enabling continuous learner-level measurement from minimal telemetry.

This is hard in RL's deployment environments. Dedicated devices and unrestricted data plans are rare, so games are delivered through lightweight web links that must work despite intermittent connectivity, shared phones, and limited bandwidth. As a result, only a little interaction data can be collected consistently. Yet most knowledge-tracing research assumes persistent connectivity, personal devices, and rich logs, so it is unclear whether these methods still help with minimal telemetry.

This gap motivates our **Research Question (RQ)**:

> *To what extent do standard Knowledge Tracing (KT) models, specifically*

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 370–377
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_038

classical Bayesian Knowledge Tracing (BKT) and neural Dynamic Key-Value Memory Networks (DKVMN), remain effective in Hindi early-childhood literacy education when learner data are limited to sparse, intermittent, and shared-device telemetry, and how can their performance be rigorously evaluated in such low-resource settings?

Knowledge tracing (KT) estimates a student's evolving mastery of skills from their sequence of right and wrong answers (Corbett and Anderson, 1994). It has been studied extensively in well-instrumented classrooms, universities, and online platforms (Piech et al., 2015; Zhang et al., 2017; Gervet et al., 2020), but it is relatively new in settings like ours. The closest precedents are adaptive learning deployments in South America and other low- and middle-income regions, especially for early childhood education (Araya and Diaz, 2020; Nye, 2015) (Section 2).

This paper makes three contributions.

1. **Established KT models transfer to low connectivity**: We show that minimal telemetry from mobile browser-based games, a timestamped (learner, skill, correctness) sequence, is enough to train both a classical (BKT) and a neural (DKVMN) KT model and to produce learner-level estimates useful for monitoring and personalization. DKVMN reaches a next-step AUC of **0.858** on held-out students, comparable to standard high-resource benchmarks, and BKT stays the more robust of the two when poor connectivity drops submissions (Section 7.5).

2. **An evaluation methodology for such environments**: Next-step prediction accuracy alone can overestimate model quality, because it rewards models that exploit game difficulty rather than track learner mastery. We therefore pair internal evaluation (student-level splits, cross-validation) with *external field validation*: mastery estimates are correlated against offline assessments of **552 children**, administered on paper by trained evaluators. This channel needs no connectivity of its own (Section 6).

3. **Reusable infrastructure and a useful baseline**: We treat interpretable BKT as a deliberate starting point rather than an exhaustive model comparison, use its learned parameters to diagnose game design and learning dynamics, and document where both models fail (ceiling effects, rare skills, and negative correlations). The same harness lets future work evaluate other BKT variants and newer KT models.

2 Background and Related Work

Educational technology in low-resource settings. Nye (2015) reviews intelligent tutoring systems in the developing world and catalogs the barriers that shape our design space: shared, mobile-first devices, unreliable electricity, data costs, and intermittent internet. Despite these barriers, well-designed systems work. Mindspark produced large test-score gains for secondary students in urban India (Muralidharan et al., 2019); ConectaIdeas improved mathematics in randomized trials in Chilean primary schools and was adapted in Peru (Araya and Diaz, 2020); and tablet programs such as onebillion in Malawi (Pitchford, 2015; Pitchford et al., 2019) and E-Learning Sudan (Stubbé et al., 2016) showed early-grade children learn from adaptive software with no schooling infrastructure at all. These deployments show that such *interventions* can improve outcomes at the population level, but comparatively little work models the trajectories of *individual students* in these settings. That is the gap we address, and tracking individual progress is a prerequisite for instruction that adapts to each learner.

Knowledge tracing. BKT models each skill as a two-state hidden Markov model: the student either knows the skill or does not, and four probabilities govern initial knowledge, learning, guessing, and slipping (Corbett and Anderson, 1994). Its interpretability has kept it in production for three decades, and a rich family of variants exists, including individualized parameters (Yudelson et al., 2013; Pardos and Heffernan, 2010) and contextual guess and slip (Baker et al., 2008); see Pelánek (2017) for a survey. Deep learning models replaced the hand-built state with learned representations: Deep Knowledge Tracing (Piech et al., 2015), the Dynamic Key-Value Memory Network (DKVMN) that maintains explicit per-concept memory (Zhang et al., 2017), and attention-based successors (Pandey and Karypis,

2019; Ghosh et al., 2020). Khajah et al. (2016) and Gervet et al. (2020) showed that the gap between classical and deep neural models depends heavily on the size and structure of the dataset. This is directly relevant here, where some skills have thousands of learners and others have just four. Recent work treats *low-resource knowledge tracing* as a problem in its own right, using pre-training to cope with scarce interactions (Zhang et al., 2024), but still evaluates on standard online-platform datasets. To our knowledge, ours is the first KT study trained on mobile browser-game telemetry from a low-resource, low-connectivity population and validated against offline field assessments.

Early literacy in Hindi. Our target skills focus on early reading in Hindi, written in the Devanagari script. Unlike alphabetic systems, Devanagari is an *akshara*-based orthography with a large symbol inventory, so letter knowledge is acquired more gradually (Nag, 2007). Foundational literacy thus depends on both phonological awareness (e.g., identifying initial sounds) and letter–sound knowledge. India's national priorities emphasize both (Government of India, 2020; ASER Centre, 2025), and both are reflected in the skills included in this study.

3 The Learning Framework

Rocket Learning's ECE content is layered. A **skill** in this study is defined as a *(learning module, learning competency)* pair such as `FL-3.5.1_FL-LC49`. A **Learning Domain (LD)** is a developmental area such as foundational literacy (`FL`); all our skills fall within the foundational literacy LD. A **Learning Module (LM)** is a themed content unit (e.g., `FL-3.5.1`), typically one game targeting a piece of curriculum. A **Learning Competency (LC)** is the observable ability it exercises (e.g., `FL-LC49`, "recognizes words that begin with the same sound"); there is a many-to-many mapping between LMs and LCs.

In this study, we focus on 32 skills that fall into four families: twelve modules of *beginning-sound recognition* (LC49), twelve modules of *beginning-sound differentiation* (LC50), and four modules of Devanagari letters (*ma*, *ka*, *cha*, *ba*), each at *basic* (LC34.1) and *advanced* (LC34.2) levels of letter-sound knowledge.

Learning Games (LGs) mainly reach children between 3 and 6 years of age, as personalized hyperlinks sent on WhatsApp to parents' phones (Rocket Learning, 2024). Each LG is a short sequence of questions answered through touch interactions, and each interaction is scored and logged. Crucially for this paper, the only telemetry we can rely on across devices and connection qualities is minimal: *(learner, timestamp, skill, isCorrect)*. Everything that follows is built from this four-tuple.

Skill family	Skills	Attempts	Learners	% cor.
Beginning sounds: recognition	12	100,866	7,140	86.9
Beginning sounds: differentiation	12	91,944	6,599	82.6
Letters: basic	4	8,876	1,214	35.3
Letters: harder	4	8,814	1,208	36.8
Total	32	210,500	7,547	80.8

Table 1: Training data by skill family (December 2025). "% cor." is the percentage of correct responses within each family.

4 Data

Training telemetry. The training set has 210,500 scored question attempts by 7,547 learners from December 2025 (Table 1); each attempt is binarized (a score above zero is correct). Where age is recorded (43% of attempts), learners are almost entirely 3- and 4-year-olds (61,859 and 27,652 attempts; only 146 from ages 5–6). Learners are identified by a persistent device-browser identifier. This is a constraint of the setting: digital literacy is low, and adding a login step causes large drop-off (Section 8). Engagement is short but meaningful: the median learner contributes 12 attempts (mean 27.9; top decile $>$60), and the median learner-skill sequence is 2 attempts (mean 3.1) across 68,356 sequences. About 27% of attempts fall on a later-day return. Any model used here must therefore say something useful from very few observations.

Assessed-cohort telemetry and field assessments. Separately, 9,275 attempts by 783 children were collected in April–May 2026 from a cohort that also took the *LCC field assessment*, a one-on-one, paper-based test administered offline by trained evaluators and covering 10 of the 32 skills in this study. For external validation, we matched child identifiers across the two datasets and kept only child–skill pairs with at least five question attempts. This yielded 552 overlapping children and 1,055 child–skill pairs.

5 Models

Why start with BKT. Our purpose is not to compare KT models exhaustively. It is to establish, with the most interpretable tool available, that student modeling is possible in this challenging environment, and to build infrastructure for richer comparisons in future work. We therefore begin with BKT, whose parameters have intuitive interpretations that curriculum designers can use when creating and revising content.

BKT. For each skill, we fit a two-state hidden Markov model (Corbett and Anderson, 1994) with four parameters, estimated by Baum–Welch expectation-maximization: $P(L_0)$, the probability that a child already knows the skill before playing; $P(T)$, the probability of *transitioning* to the known state after each practice opportunity (the learning rate); $P(G)$, the probability of *guessing* correctly while not knowing; and $P(S)$, the probability of *slipping* (answering wrongly despite knowing). After each observed attempt, the posterior probability of mastery $P(L_n)$ is updated by Bayes' rule. We train 32 independent models, one per skill.

DKVMN. As a strong learned counterpart, we train a single Dynamic Key-Value Memory Network (Zhang et al., 2017) over all 32 skills jointly. It maintains a key matrix of latent concepts and a value matrix of mastery, read and written through attention. This shares statistical strength across skills, which helps skills with sparse data but, as we will see, occasionally hurts them.

6 Methodology : Evaluating in a Low-Connectivity Environment

Figure 1 summarizes the pipeline. Our second contribution is the evaluation design itself, which we argue is the right template for settings like ours.

Internal evaluation: next-attempt prediction. We split *students*, not interactions: 85% train+validation / 15% test (6,414 vs. 1,133 learners), stratified by each student's most practiced skill, with stratified 5-fold cross-validation inside the training portion. Both models share identical splits. Splitting by student matters here: with shared devices and short sequences, splitting by interaction would leak a child's own history into training.

Model	Eval	AUC	Acc	F1	LL
BKT	CV	.654$_{\pm.005}$	.598	.529	.701
BKT	Test	.656	.609	.532	.711
DKVMN	CV	.852$_{\pm.002}$	.845	.715	.352
DKVMN	Test	**.858**	**.849**	**.716**	**.341**

Table 2: Internal evaluation: 5-fold cross-validation (mean±std) and held-out test students. LL = log-loss.

External validation: the field as ground truth. Internal metrics such as next-step prediction cannot tell us whether the predicted "mastery" reflects what a child actually knows or merely patterns in the game. To check this, we train the models on December 2025 data and use them to estimate mastery for the April–May 2026 assessment cohort after 3, 5, 10, and all 12 attempts. We then compare these estimates with the independent offline assessments and report Pearson correlations (r, 95% CI) per skill and globally.

7 Results

7.1 Internal: predicting the next attempt

Table 2 reports next-attempt prediction on held-out students across all skills. DKVMN clearly outperforms BKT (AUC 0.858 vs. 0.656), with little variance across folds. For context, this matches the high end of state-of-the-art KT on high-connectivity benchmarks: DKVMN reports AUCs of roughly 0.73–0.83 on ASSISTments and Statics (Zhang et al., 2017), and Gervet et al. (2020) place most models between 0.70 and 0.85 across nine datasets. Minimal browser-game telemetry thus supports next-step prediction *on par with standard benchmarks*, our first main finding. Part of DKVMN's edge is diagnostic rather than pedagogical: it learns that letter games are hard and sound games easy, something BKT's per-skill structure already encodes. This is what motivates the external validation below.

7.2 BKT's learned parameters, read aloud

Because BKT's parameters mean something, we can read them (Figure 2). For the sound families, the model explains near-ceiling performance as *fast learning with generous guessing*: children start out not knowing the skill ($P(L_0) \approx .01$) but transition almost immediately ($P(T) \approx .8$–$.9$) and guess often ($P(G) \approx .38$). For the letter families, the learning rate collapses to near zero. BKT assumes mastery can only increase, so a within-session accuracy *decline* cannot be reconciled with a

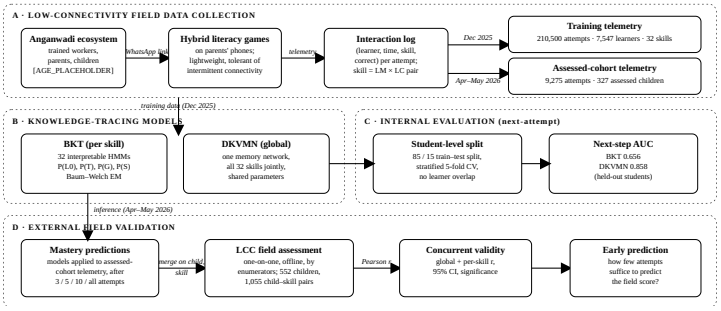


Figure 1: Evaluation pipeline. (A) Games reach families as web-links on WhatsApp; the only telemetry is a minimal four-tuple per attempt. (B) Two KT models train on December 2025 data. (C) Internal evaluation uses student-level splits (no child in both train and test). (D) External validation correlates predicted mastery with offline field assessments of the same children.

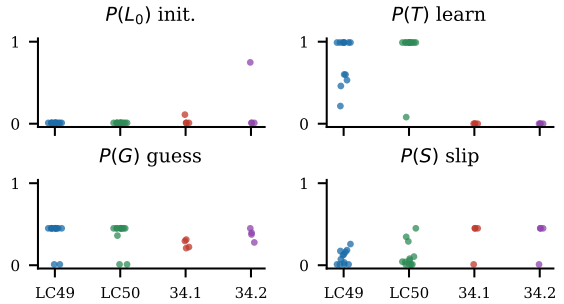


Figure 2: Learned BKT parameters for all 32 skills by family. Many sit at the bounds of their ranges (e.g. $P(T) = 0.99$, $P(G) = 0.45$), reflecting BKT's known identifiability limitation, where many parameter combinations explain the data equally well.

Model	N	r	95% CI	p
BKT	1,055	.061	[.001, .121]	.048
DKVMN	1,055	**.170**	[.111, .228]	2.7×10^{-8}

Table 3: Global correlation between predicted mastery and offline field assessment scores (552 children, 10 skills, 1055 child-skill pairs).

positive $P(T)$; instead a high slip rate ($P(S) \approx .34$) absorbs the within-game difficulty ramp. The fits also expose the model's limits: 30 of 32 skills sit at the $P(L_0)$ floor and many at the $P(T)$ or $P(G)$ ceiling (Figure 2), which is BKT's classic identifiability and degeneracy problem (Beck and Chang, 2007; Baker et al., 2008). Still, these interpretable parameters give curriculum designers a useful lever: flatten the within-game difficulty ramp to better match learners' abilities and improve model fit.

7.3 External: does predicted mastery match the field?

Globally (Table 3, Figure 3), both models correlate positively with the offline assessments, but

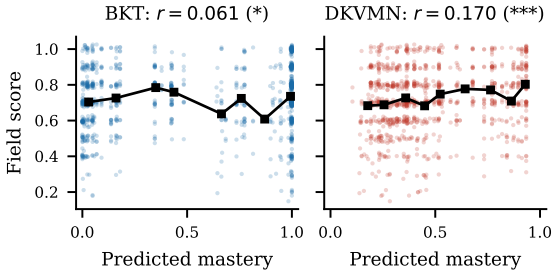


Figure 3: Predicted mastery vs. field assessment score for all 1,055 child–skill pairs (jittered), with binned means. Both relationships are positive but modest; DKVMN's is three times stronger.

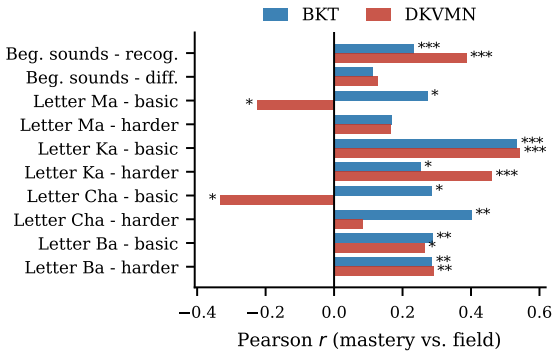


Figure 4: Per-skill correlation with the field assessment. BKT is significant on 8 of 10 skills and always positive; DKVMN is stronger where data is plentiful but significantly *negative* on two rare letter skills. Skills labelled "hard" are the advanced (LC34.2) variants.

modestly: BKT at $r = .061$ (barely significant) and DKVMN at $r = .170$ ($p < 10^{-7}$). Two constraints keep these numbers small. The assessment has to be brief given the children's young ages, which limits the reliability of any single score, and several skills vary little across learners, which limits what a correlation can reveal.

The per-skill view (Figure 4) is more instructive than the global number, and this is where the two

models trade places.

Where both models work. On the letter *ka*, the best-represented letter skill, both models reach $r \approx .54$ ($p < 10^{-7}$), and DKVMN reaches $r = .46$ on its advanced variant. On beginning-sound recognition, DKVMN reaches $r = .39$ versus BKT's .23. So for skills with enough data and enough variation, predicted mastery aligns well with field assessments.

Where BKT is safer. BKT is significant on 8 of 10 skills and *positive on all ten*. DKVMN is significant on 6 of 10, but two of those are significantly negative (letters *ma* and *cha*, basic; $r = -.23$ and $-.33$), so for these skills its predictions are actively misleading. These letters contribute only 1.7–4% of training attempts, while the sound games dominate (∼80% of attempts). DKVMN's shared parameters follow that majority and project its pattern, "children who play a lot succeed", onto skills where it does not hold. BKT trains one small model per skill, so it cannot be misled by the statistics of another skill. In low-resource contexts, where some skills will often be rare, this failure mode argues for keeping an interpretable per-skill model in the loop, whatever the global AUC says. Behavior also varies across skills within a single model, which suggests that neither of the two formulations is optimal here and that different skill families may need different modeling assumptions.

Where both fail: the ceiling. Neither model correlates on beginning-sound differentiation (LC50). Figure 2 gives the diagnosis: with 91–98% correct rates in those games, every child's mastery estimate saturates near 1.0, and a constant cannot correlate. The fix is pedagogical, not statistical.

7.4 Early prediction: how many attempts are needed for reliable prediction?

A program that must detect mastery early, so it can adapt games and flag children who need support, cares less about whether a model is eventually right than about how soon it is right. Following the truncated-history analysis of the DKVMN literature (Zhang et al., 2017), we re-estimate each child–skill mastery from only the first k attempts and re-correlate with field scores (Figure 5). DKVMN is already significant after **three** attempts ($r = .141$, $p < 10^{-5}$), about a minute of play, and gains little beyond ten ($r = .170$, its full-history value). BKT becomes only marginally significant once the full game is seen ($r = .024$ at k=3, .061 at k=12). Since the median training learner–skill sequence is just 2–3 attempts (Section 4), this is perhaps the most deployment-relevant result: with the right model, the very short sessions this environment produces already carry a meaningful signal about a child's mastery.

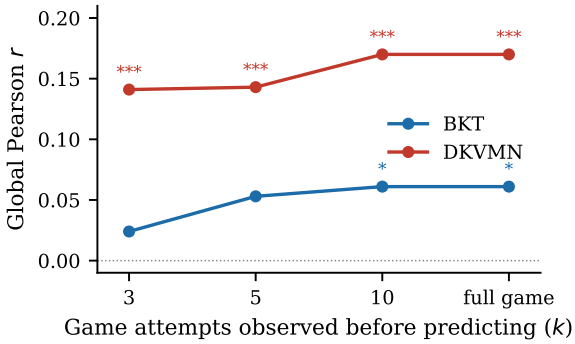


Figure 5: Global correlation with the field assessment when mastery is estimated after only the first k attempts per child–skill pair. DKVMN is significant from k=3; BKT needs the full game.

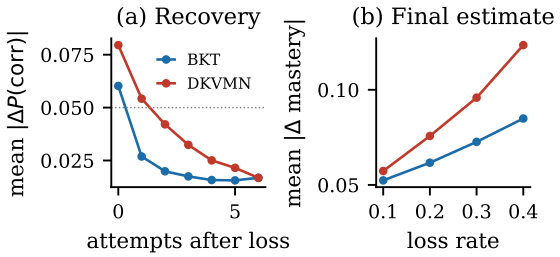


Figure 6: Robustness to submissions lost to poor connectivity. (a) Recovery after a single lost attempt. (b) Drift of the final mastery estimate under independent loss at rate ρ. BKT is more stable in both.

7.5 Robustness to lost submissions

Intermittent connectivity drops submissions before they reach the server, so a model updates on a partial sequence. On the assessed-cohort games, we take every child–skill sequence of at least five attempts, randomly drop attempts, and compare the model's belief on the thinned stream against its full-data belief, using the predicted probability of the next answer as a common signal. Two patterns emerge (Figure 6):

- *Recovery*: a single lost attempt perturbs the prediction, but the effect fades as more attempts are observed. BKT returns close to its original estimate after one further attempt, while DKVMN needs two to three and starts

from a larger perturbation.

- *Graceful degradation*: when submissions are dropped at rate ρ, the final mastery estimate moves less for BKT than for DKVMN (mean $|\Delta|$ 0.068 vs. 0.088 over $\rho \in [0.1, 0.4]$) and flips fewer master/non-master decisions (6.3% vs. 8.6%), with the gap widening as ρ increases.

A plausible explanation is that BKT's per-skill update is self-correcting and immune to other skills' data, whereas DKVMN's shared memory spreads each loss across concepts. This is another reason to keep the interpretable model in the loop where connectivity is unreliable.

8 Discussion and Limitations

What we claim. Minimal four-tuple telemetry from WhatsApp-delivered games supports (i) next-step prediction on par with high-connectivity benchmarks, (ii) significant agreement with offline field assessments on well-supported skills, (iii) usable signal within three attempts, and (iv) predictions that degrade gracefully when connectivity drops submissions, BKT most robustly. We do not claim the estimates are accurate for every skill or precise enough for individual personalization. We claim only that even sparse game telemetry carries enough signal to flag children who may need support. This establishes the feasibility of large-scale learner-level modeling in low-resource, low-connectivity early-childhood literacy, and lays a foundation for continuous progress monitoring, adaptive learning, and timely intervention.

Limitations. Learner identity is a device-browser identifier: siblings sharing a phone blur sequences, and a child who changes phones may appear as several learners. Scores are binarized, discarding partial credit. The assessed cohort is small for rare skills (N=46–90 per letter). Field ground-truth covers only 10 of the 32 skills; the rest are validated internally. Finally, games are played at home, where parental help varies between sessions and is unobserved.

Future work. The infrastructure now exists to make the next steps cheap: BKT variants designed for exactly our pathologies (individualized priors (Yudelson et al., 2013; Pardos and Heffernan, 2010), contextual guess/slip (Baker et al., 2008), forgetting (Khajah et al., 2016)), attention-based models (Pandey and Karypis, 2019; Ghosh et al., 2020), and low-resource pre-training (Zhang et al., 2024) can all be evaluated against the same student splits and field-validation harness. On the program side, the gains are pedagogical: flatten within-game difficulty ramps, harden the ceiling-bound sound games, log partial credit, add question-difficulty information, and adjust for parental help.

More broadly, the long-term value of KT here may lie less in personalization than in system-level program improvement. Instead of periodic sampled evaluations, operators could monitor every learner, spot under-performing content, and refine pedagogy and game design continuously. Future models should also combine complementary evidence, such as worksheets, classroom activities, engagement, and educator observations, for more robust mastery estimates.

9 Conclusion

The children of India's Anganwadi ecosystem will not soon have the devices, bandwidth, or rich logging that much knowledge-tracing research assumes. Our results show that meaningful student modeling can begin without them: a timestamped stream of right and wrong answers from lightweight browser-based games is enough to train models whose next-step accuracy matches high-connectivity benchmarks, whose mastery estimates agree with offline field assessments after as few as three attempts, and which stay usable even when connectivity drops submissions.

The interpretable BKT baseline also proved valuable as a diagnostic: it exposed steep within-game difficulty ramps and flagged skills where sparse data give unstable estimates. By showing both competitive predictive performance and measurable agreement with independent field assessments, this work establishes that learner-level modeling is feasible in low-resource, low-connectivity settings using only minimal telemetry, and it provides a practical foundation for future adaptive learning systems at scale.

Acknowledgements

We thank the Rocket Learning team including Azeez Gupta, Utsav Kheria, Namya Mahajan, and Siddhant Sachdeva for enabling this collaboration, and to Bilal Afroz, Pratik Purohit, Swati Pandey, Pratibha Rohal, Drishti Mrigwani, Simran Kaur,

and Asha Tiwari for their support in game design and field data collection. We also want to thank the Anganwadi Workers and parents who participated in the benchmarking study that enabled this research.

References

Roberto Araya and Karina Diaz. 2020. Implementing government elementary math exercises online: Positive effects found in RCT under social turmoil in Chile. *Education Sciences*, 10(9):244.

ASER Centre. 2025. Annual status of education report (ASER) 2024: National findings. Technical report, ASER Centre, Pratham, New Delhi, India.

Ryan S. J. d. Baker, Albert T. Corbett, and Vincent Aleven. 2008. More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian knowledge tracing. In *Proceedings of the 9th International Conference on Intelligent Tutoring Systems (ITS 2008)*, pages 406–415.

Joseph E. Beck and Kai-min Chang. 2007. Identifiability: A fundamental problem of student modeling. In *Proceedings of the 11th International Conference on User Modeling (UM 2007)*, pages 137–146.

Albert T. Corbett and John R. Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278.

Theophile Gervet, Ken Koedinger, Jeff Schneider, and Tom Mitchell. 2020. When is deep learning the best approach to knowledge tracing? *Journal of Educational Data Mining*, 12(3):31–54.

Aritra Ghosh, Neil Heffernan, and Andrew S. Lan. 2020. Context-aware attentive knowledge tracing. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2020)*, pages 2330–2339.

Government of India. 2020. National education policy 2020. Technical report, Ministry of Human Resource Development, Government of India.

Mohammad Khajah, Robert V. Lindsey, and Michael C. Mozer. 2016. How deep is knowledge tracing? In *Proceedings of the 9th International Conference on Educational Data Mining (EDM 2016)*, pages 94–101.

Karthik Muralidharan, Abhijeet Singh, and Alejandro J. Ganimian. 2019. Disrupting education? Experimental evidence on technology-aided instruction in India. *American Economic Review*, 109(4):1426–1460.

Sonali Nag. 2007. Early reading in Kannada: The pace of acquisition of orthographic knowledge and phonemic awareness. *Journal of Research in Reading*, 30(1):7–22.

Benjamin D. Nye. 2015. Intelligent tutoring systems by and for the developing world: A review of trends and approaches for educational technology in a global context. *International Journal of Artificial Intelligence in Education*, 25(2):177–203.

Shalini Pandey and George Karypis. 2019. A self-attentive model for knowledge tracing. In *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*, pages 384–389.

Zachary A. Pardos and Neil T. Heffernan. 2010. Modeling individualization in a Bayesian networks implementation of knowledge tracing. In *Proceedings of the 18th International Conference on User Modeling, Adaptation, and Personalization (UMAP 2010)*, pages 255–266.

Radek Pelánek. 2017. Bayesian knowledge tracing, logistic models, and beyond: An overview of learner modeling techniques. *User Modeling and User-Adapted Interaction*, 27(3):313–350.

Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J. Guibas, and Jascha Sohl-Dickstein. 2015. Deep knowledge tracing. In *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, pages 505–513.

Nicola J. Pitchford. 2015. Development of early mathematical skills with a tablet intervention: A randomized control trial in Malawi. *Frontiers in Psychology*, 6:485.

Nicola J. Pitchford, Antonie Chigeda, and Paula J. Hubber. 2019. Interactive apps prevent gender discrepancies in early-grade mathematics in a low-income country in sub-Saharan Africa. *Developmental Science*, 22(5):e12864.

Rocket Learning. 2024. How Rocket Learning is using WhatsApp to end learning poverty. https://rocketlearning.org/how-rocket-learning-is-using-whatsapp-to-end-learning-poverty/. Accessed June 2026.

Hester Stubbé, Aiman Badri, Rebecca Telford, Anja van der Hulst, and Wouter van Joolingen. 2016. E-learning Sudan: Formal learning for out-of-school children. *Electronic Journal of e-Learning*, 14(2):136–149.

Michael V. Yudelson, Kenneth R. Koedinger, and Geoffrey J. Gordon. 2013. Individualized Bayesian knowledge tracing models. In *Proceedings of the 16th International Conference on Artificial Intelligence in Education (AIED 2013)*, pages 171–180.

Hengyuan Zhang, Zitao Liu, Shuyan Huang, Chenming Shang, Bojun Zhan, and Yong Jiang. 2024. Improving low-resource knowledge tracing tasks by supervised pre-training and importance mechanism fine-tuning. arXiv preprint arXiv:2403.06725.

Jiani Zhang, Xingjian Shi, Irwin King, and Dit-Yan Yeung. 2017. Dynamic key-value memory networks for knowledge tracing. In *Proceedings of the 26th International Conference on World Wide Web (WWW 2017)*, pages 765–774.

Kuwain 1.5B: An Arabic SLM via Language Injection

Khalil Hennara
Misraj AI
Khobar, Saudi Arabia
hennara@misraj.ai

Sara Chrouf
Misraj AI
Khobar, Saudi Arabia
chrouf@misraj.ai

Mohamed Motasim Hamed
Misraj AI
Khobar, Saudi Arabia
hamed@misraj.ai

Zeina Aldallal
Misraj AI
Khobar, Saudi Arabia
aldallal@misraj.ai

Safwan AlModhayan
Misraj AI
Khobar, Saudi Arabia
safwan@misraj.ai

Abstract

Enhancing existing models with new knowledge is a crucial aspect of AI development. This paper introduces a novel method for integrating a new language into a large language model (LLM). Our approach successfully incorporates a previously unseen target language into an existing LLM without compromising its prior knowledge. We trained a tiny model with 1.5 billion parameters named ***Kuwain*** [‡] by injecting the Arabic language into a small open-source model mainly trained in English. Our method demonstrates significant improvements in Arabic language performance, with an average 8% improvement across various benchmarks, while retaining the model's existing knowledge with a minimum amount of the original model's data. This offers a cost-effective alternative to training a comprehensive model in both English and Arabic. The results highlight the potential for efficient, targeted language model expansion without extensive retraining or resource-intensive processes. To support the research community, we also *open-source Kuwain's model weights*, enabling further development for Arabic and other underrepresented languages.

[‡]**Kuwain:** is a diminutive form of the Arabic word (Kawn), which means "universe". So, "Kuwain" means "tiny universe" or "little cosmos". Kuwain is one of a series of Arabic-English multilingual Large Language Models (LLMs), embodying the concept of a vast amount of knowledge condensed into a compact, accessible form, much like a miniature universe

1 Introduction

LLMs have achieved significant advances in recent years, demonstrating impressive performance across a wide range of natural language processing (NLP) tasks [1]. However, a notable limitation of these models is their English-centric nature [2], with the majority focusing mainly on English tasks. This English bias makes language models (LMs) less effective in multilingual settings [3], particularly for languages with distinct linguistic characteristics and writing systems.

Despite the significant progress made by open-source models [1, 4, 5, 6, 7] in natural language processing, they exhibit inherent limitations in supporting tasks in languages like Arabic. This may result in misunderstandings and impede the effectiveness of LMs when trained on Arabic data.

In this paper, we present ***Kuwain 1.5B***, a compact model with enhanced capabilities for understanding and generating Arabic language, competing with much larger models, as illustrated in Figure 1. To mitigate the high computational costs associated with training large language models, we propose a more efficient alternative which includes expanding a monolingual Small Language Model (SLM) with new Arabic language capabilities by training only additional layers on top of the open-source model instead of retraining all model's pa-

of information. Model weights are available at https://huggingface.co/Misraj/kuwain-1.5B.

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 378–387
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_039

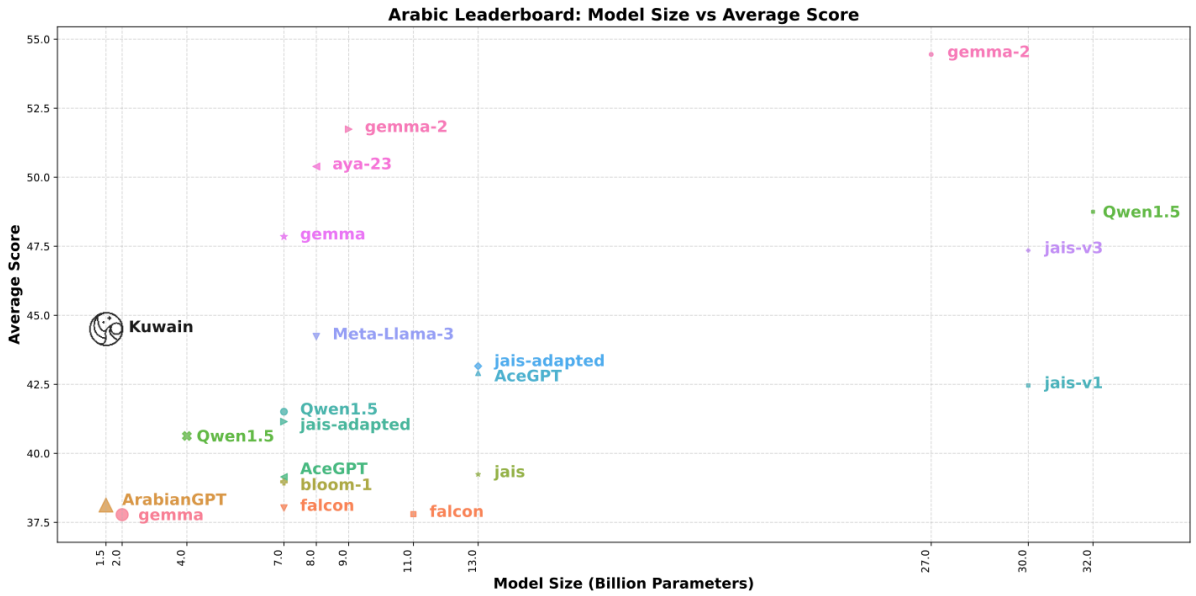


Figure 1: Model performance vs size visualization: Points represent models on the Arabic NLP benchmark, where point size indicates efficiency (average score divided by model size). Larger points show better performance-to-size ratio.

rameters. This research addresses two significant questions:

1. *Expanding Language Support: How can we effectively and efficiently expand a monolingual LLM to support new languages while minimizing cost?*

2. *Preserving Original Language Performance: How can we maintain the original language performance and knowledge of an LLM while incorporating new language abilities, without compromising either?*

Our contributions can be summarized as follows:

- We propose a novel method for injecting a new language (Arabic) into an English-centric monolingual language model without retraining from scratch.

- We present **Kuwain**, a compact open-source model for the Arabic-English language, designed to effectively process and understand a wide range of Arabic language tasks.

The remainder of this paper is organized as follows: In Section 2, we review the related work. The description of our training data is in section 3. Section 4 details our proposed method. In Section 5, we detail our experimental setup and results. Section 6 offers a comprehensive evaluation of our model's performance. Finally, in Section 7, we summarize our findings and suggest directions for future research.

2 Related work

This section explores the methodologies for acquiring new languages in LLMs and provides an overview of recent advancements in Arabic LLMs.

2.1 Multilingual LLMs Overview

Several studies [1, 4, 8, 9, 10] incorporated multilingual data during the pre-training phase to improve model alignment. As [11] observed, even the unintentional inclusion of a small amount of multilingual data during initial pre-training can significantly enhance multilingual performance. Training LLMs from scratch presents significant challenges due to the complexity of optimizing randomly initialized parameters [1, 7]. This optimization process demands extensive training data to guide the model toward effective performance [12, 13].

However, integrating new languages into existing LLMs is a promising area of exploration [14]. Continuous pre-training extends the pre-training process on already pre-trained models, reducing computational costs while keeping models up-to-date without full retraining [15]. [16] proposes a two-stage pre-training process that extends LLM vocabularies to accommodate new languages. Several models have been built this way:

LeoLM-7B [17] for German via continual pre-training on Llama2-7B, Swallow-MS-7b-v0.1 [18] for Japanese on Mistral-7B, and Meltemi-7B [19], the first open LLM for Greek, adapted from Mistral using re-warmed and re-decayed learning rates with monolingual data. However, these approaches consistently degrade base language performance: Meltemi-7B drops -6% on English [19], LeoLM-7B gains +4.8% on German but loses -2.6% on English [17], and Swallow gains +8% on Japanese but loses -5.3% on English [18]. To address such trade-offs, [20] leverages Mixture-of-Experts (MoE) by training separate models on distinct multilingual corpora subsets, reducing cross-language interference. Similarly, [21] proposes MoE-LPR, which uses language prior routing to preserve original language knowledge while improving expanded language performance.

2.2 Arabic-centric LLMs Overview

Recent Arabic Large Language Models (LLMs) address language-specific challenges through diverse strategies. Jais [22] is an open-source, bilingual model trained from scratch. However, its training data is predominantly English or translated Arabic, leading to high costs and English-centric cultural biases that overlook Arabic's literary heritage. To improve cultural alignment, AceGPT [23] and ArabianGPT [24] incorporate Arabic values. AceGPT adapts LLaMA2 via additional pre-training and RLAIF without vocabulary expansion. This lack of expansion limits performance, as the character-level LLaMA2 tokenizer inadequately covers Arabic morphology. ArabianGPT [24], based on GPT-2-small, utilizes the AraNizer tokenizer for optimized text segmentation. Yet, its Arabic-only focus lacks the cross-lingual knowledge transfer found in bilingual models like AceGPT or ALLaM [25]. Its smaller size and Modern Standard Arabic (MSA) bias also hinder its performance in dialectal and complex multilingual tasks. Conversely, the ALLaM series [25] achieves state-of-the-art results by combining vocabulary expansion with continued pre-training on 4 trillion English and 540 billion Arabic tokens. Despite these advancements, Arabic LLMs typically face three hurdles: (1) high training costs, (2) catastrophic forgetting of the original language, and (3) inefficient tokenization for morphologically rich text. Most models struggle to scale without extensive retraining.

Our work, *Kuwain*, provides a lightweight, cost-efficient alternative. Instead of full retraining, *Kuwain* extends an English-centric LLM using minimal resources while preserving original capabilities. By strategically expanding the tokenizer and adapting specific layers through continual pre-training, Kuwain yields significant Arabic performance gains without the typical trade-offs.

3 Data

3.1 Data Sources

Our training dataset comprises 110 billion tokens, with 90 billion tokens in Arabic and 20 billion in English, all derived from publicly available open-source sources: CulturaX [26], C4 [27], and ArabicText 2022[¶]. In addition to Modern Standard Arabic, the dataset also includes a portion of Arabic dialectal data sourced from Hugging Face repositories to preserve the linguistic richness and diversity of regional Arabic varieties. Importantly, none of the evaluation datasets were included in the training data, preventing data leakage and ensuring the integrity of our evaluation.

3.2 Data Cleaning

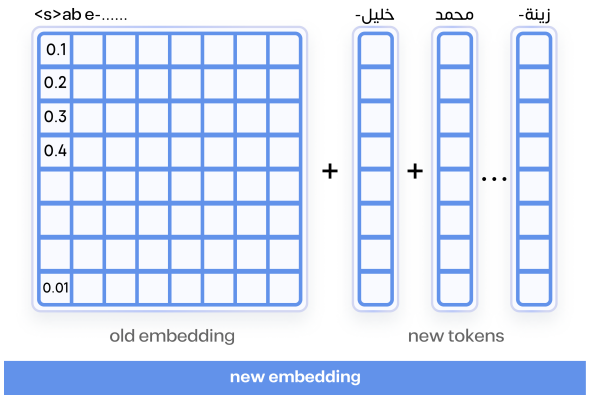
To improve the overall quality of the Arabic data, we applied a range of filtering and cleaning steps. These included removing corrupted or unreadable characters and repeated characters, stripping markup and elongation characters, and white space normalization, while retaining non-Arabic characters that may appear within the text. Additionally, we preserved Quranic symbols and other special characters commonly found in Arabic texts to maintain their integrity. In addition, we normalized encoding inconsistencies and unified orthographic variants (e.g., different forms of *baa*), Malformed examples and short text were also filtered out to ensure consistency and improve data quality. This cleaning process was implemented using a custom script, which we released to support reproducibility and future improvements in Arabic text pre-processing. The script includes both the specific cleaning steps used in this study and general-purpose Arabic text cleaning functions, which are fully configurable, making it suitable for

[¶]**ArabicText 2022:** an open-source Arabic collection prepared by the Beijing Academy of Artificial Intelligence (BAAI), which includes Arabic corpora such as ArabicWeb22-A, ArabicWeb16, OSCAR, ArabicWeb22-B, CC100-AR, and Arabic Tweets.

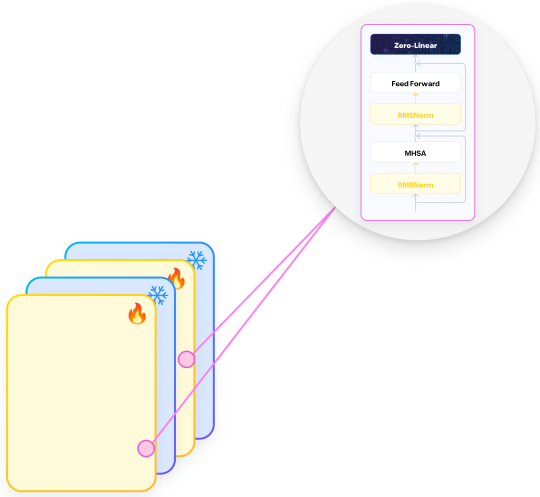
broader Arabic text pre-processing applications[§]. For English data, we adopted the filtering pipeline introduced by the BLOOM project [9], which promotes clean, diverse, and high-quality content.

4 Method

Our approach integrates a new language into an existing LLM while fully preserving its prior knowledge. This approach not only maintains the model's existing capabilities but also enriches and expands them. It is based on two key concepts: **(1)** extending the model architecture and **(2)** expanding the tokenizer's vocabulary.



(a)



(b)

Figure 2: (a) Vocabulary expansion by adding new Arabic tokens to the tokenizer. (b) Extension of model layers, where newly added layers were trainable while the original layers remained frozen.

[§]https://github.com/misraj-ai/Kuwain-Arabic-cleaner

4.1 Extending Model's Layers

Inspired by [28], which reaches state-of-the-art performance across a broad range of general, code, and math tasks, they explored extending the model's capabilities by injecting a new layer every n layers. We adopt a similar approach but with a broader experimental scope. The concept of layer extension, which incorporates an identity layer after a group of existing layers in the original model. This technique can be applied to various open-source models, such as Llama-2,3 [1, 7], Mistral [4], Phi-3 [5], and Gemma [6], which share similar decoder layer structures with minor variations. These minor modifications do not significantly alter the identity block concept as described in [28].

A typical decoder block in these models produces an output y from an input x as defined by the following equations:

$$x' = x + MHSA(NormLayer(x))$$

$$x'' = x' + FFN(NormLayer(x'))$$

Here, *MHSA* stands for Multi-Head Self-Attention, *FFN* is the Feed-Forward Network, and *NormLayer* is a normalization layer (e.g., LayerNorm or RMSNorm). The internal structures of *MHSA* and *FFN* are defined as:

$$\text{FFN}(x) = \text{Activation}(x, W^1)W^2W^3 \quad (1)$$

$$\text{MHSA}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O \quad (2)$$

Where W^1, W^2, W^3 are the learnable weight matrices in the *FFN*, and the activation function may vary depending on the model. For instance, LLaMA [1] uses SwiGLU [29], whereas Gemma-2 [6] uses a Linear activation. In the attention mechanism, W^O is the output projection matrix.

To extend the model architecture, we build upon this standard decoder block by introducing additional blocks while preserving the original model's behavior. Suppose the model consists of blocks $(\theta_0, \theta_1, ..., \theta_n)$. The extended model includes additional identity blocks θ_{id}, defined as:

$$\theta_{id}(x) = x$$

This ensures the input and output of the identity block are the same. The identity behavior can be achieved if the following conditions are satisfied:

$$\begin{aligned} \text{MHSA}(\text{NormLayer}(x)) &= 0, \\ \text{FFN}(\text{NormLayer}(x')) &= 0 \end{aligned} \tag{3}$$

These conditions can be met by initializing the additional blocks' weight matrices W^O and W^3 in Equations 1 and 2 to zero. This ensures that the new blocks initially perform no transformation on the input. While [28] focuses on enhancing model performance on tasks it has previously encountered, our work addresses a more challenging objective: adapting a model to the Arabic language, which is absent from its original pretraining data. In addition to this language adaptation, we extend the model's vocabulary by introducing several new tokens. Our full methodology is illustrated in Figure 2.

4.2 Vocabulary Expansion

Extending a tokenizer's vocabulary is a critical step when adapting a model to a new language particularly when the original tokenizer, such as a character-based one, is not well-suited for that language [16, 19, 30, 31]. Inspired by the above we conclude that vocabulary expansion improves tokenization efficiency, enabling the model to represent the language more effectively. This not only enhances performance but also reduces training costs by shortening the number of required training steps and increasing the effective context length.

The first step in extending a tokenizer's vocabulary is to analyze the new language's text data to assess how well the existing tokenizer covers it. This involves identifying gaps where the current vocabulary fails to capture important characters, words, or subwords. Many open-source models, such as **LLaMA-2** [1], **Phi-3** [5], and **Mistral** [4], are primarily designed for English and include only 28 Arabic tokens covering just the Arabic alphabet. This limited representation is inadequate for capturing the morphological richness and lexical diversity of Arabic. In contrast, the **Llama-3** [7] and **Gemma** [6] includes over 4,000 Arabic tokens, offering better coverage. However, even this may not fully reflect the complexity and nuance of Arabic [32, 33, 34].

In our work, we have trained a tokenizer using **SentencePiece** [35] with our training data, resulting in a vocabulary of **26K** tokens. The vocabulary from this tokenizer is then used to extend the vocabulary of the base model's tokenizer, resulting in 54K tokens. We conducted an extensive evaluation to ensure that these 26K tokens are optimized for our model; see section 5.2 for more details.

5 Experiments and Results

We conducted experiments using *TinyLlama* [36] as our base model to evaluate Arabic language integration capabilities. *TinyLlama*, a compact 1.1B language model pretrained on around 1 trillion tokens, demonstrating remarkable performance despite its size. The experimental architecture involved adding 8 new layers and extending the vocabulary with 26K Arabic tokens while keeping the original layers frozen. Training was performed on 8 A100 GPUs for 3 epochs, and the effective batch size is 1M tokens, focusing only on the newly added layers to preserve existing model knowledge while developing Arabic language capabilities.

5.1 Number and Placement of New Layers

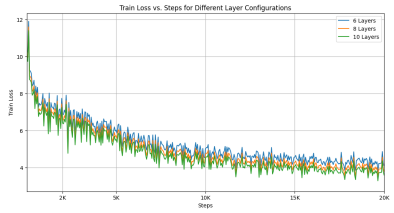
We conducted comprehensive experiments to evaluate both the number and positioning of the inserted layers, exploring configurations ranging from 6 to 10 layers. A subset of our dataset was used for this investigation, with all experiments related to layer count and placement performed over 20K training steps. Additionally, we assessed the model's language generation capabilities through a human evaluation of a 500-sentence completion task. Our findings reveal several key insights:

1. **Avoid Stacking:** Consecutive layer insertion causes training instability and performance degradation (Figure 3c).
2. **Trainable Last Layer:** The final encoder layer must remain trainable to prevent volatility. While [37] notes the importance of the final layer in pruning, we find its trainability critical for stability during expansion (Figure 3a).
3. **Placement Geometry:** Optimal placement does not require regular patterns or equal intervals between existing blocks.
4. **Layer Count:** Models with 8–10 new layers showed similar performance, but 8 layers (a $\sim$30% size increase) yielded the best results.

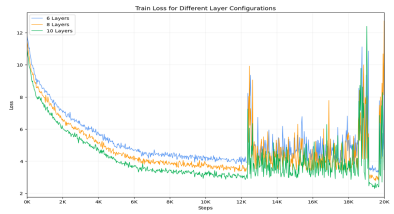
As illustrated in Figure 3b, a distributed 8-layer configuration with a trainable last layer emerged as the most stable and effective architecture for *Kuwain*



(a)



(b)



(c)

Figure 3: Training loss for *Kuwain*: (a) Freezing the last layer causes instability; (b) A trainable last layer and distributed (non-consecutive) insertion ensures stability; (c) Stacking layers consecutively degrades performance.

5.2 Tokenizer Comparison

We evaluate our tokenizer against *AraBERT* [38] and *Jais* [22] using two metrics:

1. **Vocabulary Size (VS):** The number of unique tokens. While larger vocabularies capture more linguistic nuance, they risk overfitting and computational inefficiency.

2. **Expansion Ratio (ER):** The ratio of tokenized length to original word length ($\frac{L_{\text{tokenized}}}{L_{\text{original}}}$). A lower ratio indicates higher tokenization efficiency [39].

Effective tokenization requires balancing a low expansion ratio with a manageable vocabulary size to ensure generalization and efficiency, particularly for morphologically rich languages like Arabic. Following the scaling analysis by [40], we fine-tuned multiple tokenizer parameters to optimize this trade-off for Arabic vocabulary expansion.

Table 1 compares our optimal configuration against established models based on over 1 million text examples.

Model	ER	VS
Kuwain Tokenizer	2.30	26K
AraBERT	2.51	54K
Jais	2.19	44K

Table 1: Comparison of tokenization metrics. Vocabulary size reflects the Arabic-specific portion of each model's full vocabulary.

While Jais [22] achieves the lowest expansion ratio, the *Kuwain* tokenizer offers superior efficiency regarding the ratio-to-vocabulary size trade-off. By achieving a competitive expansion ratio with a significantly smaller vocabulary (26K vs. Jais's 44K), our approach facilitates more efficient model training and deployment without sacrificing representation quality.

5.3 Data Proportion

Optimal data mixing is critical for multilingual adaptation. We found that *Kuwain* requires only 20% original English data to maintain performance while acquiring Arabic capabilities significantly less than the 50% ratio used in comparable approaches [16, 19, 25, 30]. This 20% proportion serves as a knowledge "anchor," preventing performance degradation without unnecessary training costs.

Table 2 illustrates the impact of this balance. Reducing English data below 20% (*Kuwain-lt-*ϕ) leads to a notable decline in benchmarks (Avg. 49.56 vs. 52.99). Conversely, the 20% threshold used in *Kuwain* preserves and occasionally enhances performance relative to *TinyLlama* (Avg. 53.28). These findings confirm that while a minimal amount of original-language data is essential to retain existing capabilities, exceeding this threshold provides diminishing returns for adaptation.

Dataset	TinyLlama	Kuwain	Kuwain-lt-ϕ
HellaSwag	**59.20**	57.79	55.79
Obqa	**36.00**	35.60	32.60
ARC-c	30.10	**30.29**	27.29
ARC-e	55.25	**56.31**	51.31
boolq	57.83	**60.43**	54.43
piqa	**73.29**	72.63	69.63
WinoGrande	59.12	**59.99**	55.99
Avg	52.99	**53.28**	49.56

Table 2: Performance on English benchmarks: Comparing TinyLlama, Kuwain (20% English data), and Kuwain-lt-ϕ (less that 20% English data).

5.4 Conventional Training (Without Layer Extension)

To assess the impact of conventional approaches in the absence of our proposed strategy namely, language extension through continued pre-training and vocabulary expansion without layer extension we trained the base model on the complete Arabic dataset, extending its vocabulary using the same tokenizer but without applying any layer extension. We refer to this baseline model as **Kuwain-Naive.**

While *Kuwain-Naive* demonstrated effective acquisition of new linguistic information, particularly in Arabic, it suffered a significant loss in its previously acquired knowledge base. This is evident in Table 3, where its performance on English benchmarks drastically declines compared to the original base model, *TinyLlama*.

In contrast, our proposed approach, denoted as *Kuwain*, applies layer extension and selective training (training only newly added layers), aiming to integrate Arabic while preserving the model's prior capabilities. As shown in Table 3 *Kuwain* maintains strong performance on English benchmarks, while also achieving Arabic performance comparable to *Kuwain-Naive* as demonstrated in 4

This contrast highlights the strength of our strategy: it enables the model to expand into a new language (Arabic) without catastrophic forgetting of previously learned knowledge (English), thereby supporting more robust multilingual development.

Dataset	TinyLlama	Kuwain-Naive
HellaSwag	59.20	45.35
Obqa	36.00	29.20
ARC-c	30.10	25.68
ARC-e	55.25	45.24
boolq	57.83	61.90
piqa	73.29	66.70
WinoGrande	59.12	53.91
Avg	**52.99**	46.85

Table 3: English evaluation of *TinyLlama* vs. *Kuwain-Naive* (standard pre-training).

Dataset	Kuwain-Naive	Kuwain
HellaSwag	32.59	33.20
Obqa	27.80	27.60
ARC-c	25.26	25.49
ARC-e	36.44	36.76
boolq	61.80	62.35
piqa	54.06	54.35
copa	57.30	56.17
Avg	**42.17**	42.27

Table 4: Arabic evaluation of *Kuwain-Naive* vs. our model *Kuwain*.

Dataset	TinyLlama	Kuwain
HellaSwag	29.90	**37.14**
Obqa	28.12	**29.20**
ARC-c	23.00	**28.15**
ARC-e	26.76	**40.10**
boolq	50.88	**62.04**
piqa	51.68	**56.42**
copa	48.31	**58.38**
Avg	36.95	**44.49**

Table 5: **Arabic** Benchmark Evaluation of Kuwain and TinyLlama

6 Evaluation

We present a comprehensive evaluation of the *Kuwain* model, highlighting its effectiveness in Arabic language tasks while maintaining robust English capabilities.

Table 5 shows that *Kuwain* significantly outperforms the base *TinyLlama* [36] in Arabic language understanding and generation. These gains reflect the model's successful adaptation to Arabic with limited supervision, demonstrating the strength of our multilingual training approach.

To contextualize these results, we evaluated *Kuwain* on the Arabic language leaderboard [41], comparing it against a diverse set of multilingual and Arabic-focused models. As shown in Table 6, *Kuwain* delivers competitive performance while being considerably smaller than many of its peers. This demonstrates the power of strategic data mixing and model tuning in low-resource scenarios, enabling high performance without inflating model size or training costs. Maintaining just 20% of the original English data was sufficient to achieve strong Arabic performance while preserving English capabilities.

7 Conclusion and Future Work

We presented a cost-efficient method for injecting a new language into an existing LLM via vocabulary expansion, layer extension, and selective training of only the newly added layers. Applied to *TinyLlama 1.1B*, the resulting model ***Kuwain*** achieves an 8% average improvement on Arabic benchmarks while fully preserving English proficiency—a result that conventional continual pre-training approaches consistently fail to achieve. To support the research community and advance Arabic NLP, we release Kuwain's model weights openly.

Future Work. We will focus on two directions:

Model	Size	ARC-c	ARC-e	Boolq	Copa	HSwag	Obqa	Piqa	Avg
gemma	2B	27.67	27.66	52.76	46.67	25.61	34.75	49.37	37.784286
falcon	11B	28.10	25.80	51.81	46.67	25.40	37.17	49.65	37.800000
falcon	7B	27.93	25.34	57.52	43.33	25.20	36.16	50.68	38.022857
ArabianGPT	1.5B	25.86	27.41	62.12	47.78	24.35	30.51	48.83	38.122857
falcon	40B	26.55	25.76	52.85	50.00	25.37	36.77	50.08	38.197143
bloom-1	7B	28.62	25.85	62.12	44.44	25.31	35.56	50.95	38.978571
jais	13B	28.53	28.43	62.12	48.89	25.67	35.35	54.56	40.507143
Qwen1.5	4B	29.31	28.09	62.76	52.22	25.07	36.16	50.85	40.637143
AceGPT	7B	29.66	28.64	62.36	48.89	25.89	38.38	52.59	40.915714
jais-adapted	7B	30.60	31.01	63.50	48.89	25.55	38.38	52.97	41.557143
Qwen1.5	7B	33.71	33.33	62.12	47.78	25.70	38.59	53.79	42.145714
jais-v1	30B	32.24	32.83	62.70	48.89	25.82	39.60	56.57	42.664286
AceGPT	13B	33.36	33.76	63.74	51.11	25.09	39.19	54.17	42.917143
jais-adapted	13B	33.62	34.90	65.03	47.78	26.41	39.39	54.99	43.160000
AceGPT-v1.5	13B	36.55	37.61	62.24	45.56	26.59	41.82	54.99	43.622857
Meta-Llama-3.1	8B	36.21	37.77	63.34	50.00	26.45	40.61	51.99	43.767143
Qwen1.5	14B	35.60	37.23	65.09	47.78	26.79	39.60	54.39	43.782857
Meta-Llama-3	8B	35.69	39.00	62.12	50.00	26.65	42.63	53.68	44.252857
Kuwain	$\downarrow$ 1.5B	28.15	40.10	62.04	58.38	37.14	29.20	56.42	44.490000
jais-v3	30B	36.98	41.46	75.25	52.22	26.66	42.02	56.63	47.317143
gemma	7B	42.16	45.60	67.94	47.78	26.90	46.46	58.16	47.857143
Qwen1.5	32B	43.19	43.99	64.08	51.11	29.19	46.87	59.36	48.255714
aya-23	8B	40.09	40.31	77.91	54.44	30.86	44.44	64.70	50.392857
jais-adapted	70B	47.67	51.73	74.14	44.44	28.44	47.47	63.07	50.994286
Qwen1.5	72B	49.05	46.40	65.61	51.11	34.02	47.27	64.76	51.174286
gemma-2	9B	53.28	55.80	64.33	51.11	29.11	51.52	60.23	52.197143
Qwen1.5	110B	51.81	51.18	62.12	50.00	36.65	50.91	63.72	52.341429
Meta-Llama-3.1	70B	52.93	52.28	62.15	56.67	35.38	50.10	63.72	53.318571
Meta-Llama-3	70B	54.91	55.80	62.18	56.67	35.03	51.72	62.79	54.157143
gemma-2	27B	56.55	59.09	65.58	55.56	28.43	52.93	62.96	54.442857
aya-23	35B	49.22	50.80	82.94	55.56	35.81	49.49	73.87	56.812857
Qwen2	72B	56.55	57.45	69.97	55.56	38.33	57.17	69.67	57.814286
c4ai-command-r-v01	35B	54.66	60.24	82.70	52.22	33.42	54.14	69.99	58.195714
c4ai-command-r-plus	104B	57.76	60.87	85.00	52.22	37.88	52.73	70.92	59.625714

Table 6: Arabic Leader board evaluation

(1) large-scale Arabic data collection to further stress-test and refine the approach, and (2) scaling the method to larger base models, where we hypothesize that Arabic performance will correlate with the base model's original capability.

References

[1] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[2] Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. Don't trust chatgpt when your question is not in english: A study of multilingual abilities and types of llms. *arXiv preprint arXiv:2305.16339*, 2023.

[3] Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*, 2023.

[4] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[5] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical re-

port: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

[6] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[7] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[8] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[9] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *https://inria.hal.science/hal-03850124*, 2023.

[10] L Xue. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020.

[11] Terra Blevins and Luke Zettlemoyer. Language contamination helps explain the cross-lingual capabilities of english pretrained models. *arXiv preprint arXiv:2204.08110*, 2022.

[12] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[13] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[14] Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. Multilingual instruction tuning with just a pinch of multilinguality. *arXiv preprint arXiv:2401.01854*, 2024.

[15] Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats L Richter, Quentin Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. Continual pre-training of large language models: How to (re) warm your model? *arXiv preprint arXiv:2308.04014*, 2023.

[16] Yiming Cui, Ziqing Yang, and Xin Yao. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*, 2023.

[17] Björn Plüster. Leolm: Igniting german-language llm research, 2023.

[18] Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Rio Yokota, and Sakae Mizuki. Building a large japanese web corpus for large language models. *arXiv preprint arXiv:2404.17733*, 2024.

[19] Leon Voukoutis, Dimitris Roussis, Georgios Paraskevopoulos, Sokratis Sofianopoulos, Prokopis Prokopidis, Vassilis Papavasileiou, Athanasios Katsamanis, Stelios Piperidis, and Vassilis Katsouros. Meltemi: The first open large language model for greek. *arXiv preprint arXiv:2407.20743*, 2024.

[20] Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A Smith, and Luke Zettlemoyer. Breaking the curse of multilinguality with cross-lingual expert language models. *arXiv preprint arXiv:2401.10440*, 2024.

[21] Hao Zhou, Zhijun Wang, Shujian Huang, Xin Huang, Xue Han, Junlan Feng, Chao Deng, Weihua Luo, and Jiajun Chen. Moe-lpr: Multilingual extension of large language models through mixture-of-experts with language priors routing. *arXiv preprint arXiv:2408.11396*, 2024.

[22] Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, et al. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*, 2023.

[23] Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Ziche Liu, et al. Acegpt, localizing large language models in arabic. *arXiv preprint arXiv:2309.12053*, 2023.

[24] Anis Koubaa, Adel Ammar, Lahouari Ghouti, Omar Najar, and Serry Sibaee. Arabiangpt: Native arabic gpt-based large language. *arXiv preprint arXiv:2402.15313*, 2024.

[25] M Saiful Bari, Yazeed Alnumay, Norah A Alzahrani, Nouf M Alotaibi, Hisham A Alyahya, Sultan AlRashed, Faisal A Mirza, Shaykhah Z Alsubaie, Hassan A Alahmed, Ghadah Alabduljabbar, et al. Allam: Large language models for arabic and english. *arXiv e-prints*, pages arXiv–2407, 2024.

[26] Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *arXiv preprint arXiv:2309.09400*, 2023.

[27] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*, 2019.

[28] Chengyue Wu, Yukang Gan, Yixiao Ge, Zeyu Lu, Jiahao Wang, Ye Feng, Ping Luo, and Ying Shan. Llama pro: Progressive llama with block expansion. *arXiv preprint arXiv:2401.02415*, 2024.

[29] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

[30] Zoltan Csaki, Pian Pawakapan, Urmish Thakker, and Qiantong Xu. Efficiently adapting pretrained language models to new languages. *arXiv preprint arXiv:2311.05741*, 2023.

[31] Aleksandr Nikolich, Konstantin Korolev, and Artem Shelmanov. Vikhr: The family of open-source instruction-tuned large language models for russian. *arXiv preprint arXiv:2405.13929*, 2024.

[32] Ali Farghaly and Khaled Shaalan. Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*, 8(4):1–22, 2009.

[33] Sabri Elkateb, William J Black, Piek Vossen, David Farwell, Horacio Rodriguez, Adam Pease, Musa Alkhalifa, and Christiane Fellbaum. Arabic wordnet and the challenges of arabic. In *Proceedings of the International Conference on the Challenge of Arabic for NLP/MT*, pages 15–24, 2006.

[34] Iyad Issa. Morphological complexity in arabic spelling and its implication for cognitive processing. *Journal of Psycholinguistic Research*, 52(1):331–357, 2023.

[35] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arxiv 2018. *arXiv preprint arXiv:1808.06226*, 1808.

[36] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024.

[37] Andrey Gromov, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Daniel A Roberts. The unreasonable ineffectiveness of the deeper layers. *arXiv preprint arXiv:2403.17887*, 2024.

[38] Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*, 2020.

[39] Kaj Bostrom and Greg Durrett. Byte pair encoding is suboptimal for language model pretraining. *arXiv preprint arXiv:2004.03720*, 2020.

[40] Chaofan Tao, Qian Liu, Longxu Dou, Niklas Muennighoff, Zhongwei Wan, Ping Luo, Min Lin, and Ngai Wong. Scaling laws with vocabulary: Larger models deserve larger vocabularies. *arXiv preprint arXiv:2407.13623*, 2024.

[41] Ali Elfilali, Hamza Alobeidli, Clémentine Fourrier, Basma El Amel Boussaha, Ruxandra Cojocaru, Nathan Habib, and Hakim Hacid. Open arabic llm leaderboard. https://huggingface.co/spaces/OALL/Open-Arabic-LLM-Leaderboard, 2024.

A Neuro-Symbolic RAG System for Marine Environmental Law: from Domain Ontology to Knowledge Graphs and Context-Enriched Legal Question Answering

Assoumana Souley Hadiza
Laboratory of Computer Science and Smart Systems (LISI)
Faculty of Sciences Semlalia, Cadi Ayyad University (UCA)
Marrakesh, Morocco
h.assoumanasouley0329@uca.ac.ma

Youssef Al Mouatamid
LISI Laboratory, Cadi Ayyad University, Marrakesh, Morocco
LEMAR, Univ Brest, Plouzané, France
youssef.almouatamid@gmail.com

Marie Bonnin
IRD, Univ Brest (UBO), CNRS, Ifremer, LEMAR
Faculté des Sciences Semlalia (FSSM)
Université Cadi Ayyad, Maroc
marie.bonnin@ird.fr

Jihad Zahir
Laboratory of Computer Science and Smart Systems (LISI)
Faculty of Sciences Semlalia, Cadi Ayyad University (UCA)
UMMISCO-IRD
j.zahir@uca.ac.ma

Abstract

Legal question answering in multi-jurisdictional normative domains demands factual precision that flat vector retrieval cannot guarantee. We present a neuro-symbolic RAG system for expert-level marine environmental law QA targeting national legislation across 15 West and North African coastal states. The system is grounded in a formally verified OWL 2.0 DL ontology of 120 classes, 8,844 RDF triples, and 1,217 individuals covering six prohibition categories from 12 international conventions, aligned with LKIF-Core, a reference ontology widely used in legal knowledge representation and reasoning. Its central contribution is a controlled ontology-to-graph transition: 68 T-Box axioms filter LLM-extracted A-Box assertions; HermiT pre-materialises 299 inferences; and the validated graph is exported to Neo4j (1,038 domain nodes, 7,714 edges, 400+ lexical nodes), decoupling formal reasoning from retrieval. A SKOS lexical bridge drives synonym integration at three pipeline stages; BGE-M3 and BM25 retrieval are fused via Reciprocal Rank Fusion, then re-ranked before a seven-step Ontological Reasoning Agent injects multi-hop legal facts, partial-match annotations, and normative gap signals. An ablation on 66 expert-annotated queries, drawn by stratified sampling from a 264-query pool and using Llama 3.2 3B as a deliberate stress test, shows the full pipeline achieves Context Precision 0.8929 (+23.1% over the agent-only baseline), Context F1 0.8242, and a +32% Faithfulness gain, establishing that context construction quality matters more than generator capacity in formally structured legal domains. On the hardest exception-discrimination queries, the full pipeline reaches 62.5% accuracy against expert ground truth, against 54.2% for dense-only retrieval (§7).

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 388–397
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_040

1 Introduction

Retrieval-Augmented Generation has established itself as the dominant paradigm for grounding large language model outputs in documentary evidence (Lewis et al., 2020). Yet standard RAG architectures remain fundamentally insufficient for normative domains: legal question answering requires deductive precision, not semantic plausibility, and the two do not coincide. Dense vector retrieval measures lexical and semantic proximity between text chunks but cannot represent normative dependencies, exclusion rules, or actor-zone obligations. An answer that is semantically adjacent to the truth may still be legally wrong.

Marine environmental law exposes these limitations with particular acuity. Its normative architecture spans international framework conventions (e.g., UNCLOS, CBD), sector-specific treaties (e.g., MARPOL, ICRW), and the 15 national legal regimes considered in this study across West and North African coastal states, generating both deep structural complexity and pervasive terminological heterogeneity. A query about whale hunting may be answered only by texts referring to *cetaceans* or *marine mammals*. Vocabulary divergence across jurisdictions causes silent recall failures that flat retrieval cannot detect.

Three failure modes are architectural rather than incidental. First, the relational structure of legal knowledge (prohibitions derived from conventions, actor-specific obligations, sanction mechanisms, and exception carve-outs) is invisible to chunk-level embeddings. Second, cross-jurisdictional terminological variation creates systematic recall gaps whenever user vocabulary diverges from the exact phrasing of a national statute. Third, legal QA demands that every claim be traceable to a verbatim textual source; approximation is not a graceful degradation but a correctness failure.

Our response is a neuro-symbolic architecture whose central design principle is the temporal separation of formal reasoning from inference time. A formally verified OWL 2.0 DL ontology encodes the normative structure of marine prohibition law; 68 strict axioms filter LLM-extracted knowledge before graph population; HermiT materialises all implicit inferences; and the validated property graph is deployed in Neo4j before any query is processed. At inference time, a seven-step Ontological Reasoning Agent queries this pre-secured graph to enrich retrieved contexts with multi-hop legal facts and normative gap signals, while a SKOS lexical bridge resolves terminological fragmentation at three pipeline stages without neural retraining.

Contributions.

1. **A reusable, formally verified OWL 2.0 DL ontology** for marine environmental law (120 classes, 72 properties, 8,844 RDF triples, 1,217 axiomatically filtered individuals), covering six prohibition categories from 12 international conventions, aligned with LKIF-Core, and equipped with a SKOS lexical bridge for cross-jurisdictional synonym integration.

2. **A controlled ontology-to-graph framework** that decouples formal reasoning from inference time: 68 T-Box axioms govern A-Box population, HermiT pre-materialises 299 inferences, and the validated graph of 1,038 domain nodes, 7,714 typed edges, and 400+ lexical nodes is queried at Cypher speed at runtime, making OWL guarantees available during retrieval without live reasoner overhead.

3. **An ablation-validated neuro-symbolic RAG pipeline** demonstrating, through systematic ablation using Llama 3.2 3B as a deliberate stress test on 66 expert-annotated queries, that context construction quality, not generator capacity, is the primary performance determinant in formally structured legal domains.

2 Flat Vectors versus Normative Knowledge: Related Work

Retrieval-augmented generation. Lewis et al. (2020) introduced RAG for knowledge-intensive NLP; Dense Passage Retrieval (Karpukhin et al., 2020) established bi-encoder architectures as the foundation for scalable retrieval. Hybrid dense-sparse pipelines have since become the dominant retrieval baseline (Lin et al., 2021). HyDE (Gao et al., 2023) improves retrieval by generating hypothetical answers before searching; RAPTOR (Sarthi et al., 2024) builds hierarchical summaries for multi-granularity coverage. These systems optimise semantic proximity but remain relationally flat: none addresses the normative dependencies and multi-jurisdictional terminological fragmentation that characterise legal corpora.

Graph-augmented retrieval. GraphRAG (Edge et al., 2024) uses knowledge graphs as a direct retrieval channel, substantially improving coverage over flat RAG. OG-RAG (Sharma et al., 2025) extends this to ontological hypergraphs; SymRAG (Hakim et al., 2025) introduces adaptive symbolic-neural routing. Our architecture departs from this line in a critical design choice: the knowledge graph is not a retrieval channel, which would conflate textual evidence with ontological inference and inflate recall with ontologically related but textually unattested nodes. Instead, it operates exclusively as a post-retrieval enrichment layer, grounded in a HermiT-verified OWL 2.0 ontology aligned with LKIF-Core. We have not yet run GraphRAG or OG-RAG on our corpus as a direct empirical baseline; the comparison above is architectural rather than measured, and a quantitative head-to-head is left for future work (§7).

Legal NLP and ontologies. LEGAL-BERT (Chalkidis et al., 2020) and SaulLM (Colombo et al., 2024) establish domain-adapted language models for legal text. ChatLaw (Cui et al., 2023) and DISC-LawLLM (Yue et al., 2023) augment LLMs with legal knowledge bases, but without formal ontological consistency guarantees. LegalBench (Guha et al., 2023) demonstrates that even frontier-scale models struggle with complex norm application. LKIF-Core (Hoekstra et al., 2007) provides foundational legal concepts underpinning our ontology alignment; Khorshidi et al. (2025) and Langer et al. (2024) explore LLM-guided ontology construction from legal corpora. No prior system combines HermiT-verified ontological consistency, axiomatic A-Box filtering, and a SKOS lexical bridge for multi-jurisdictional terminological coverage in a deployed RAG pipeline.

3 Architecture of the Neuro-Symbolic RAG Pipeline

3.1 End-to-End Pipeline Overview

At query time, a Query Analyser infers an intent class and applies a country filter. The SKOS bridge expands the query with synonym sets (Phase 1). Dense and sparse retrievers operate in parallel; their ranked lists are fused via synonym-aware weighted Reciprocal Rank Fusion (Cormack et al., 2009) (Phase 2) and re-ranked by a cross-encoder (Nogueira and Cho, 2019). The Ontological Reasoning Agent then queries the Neo4j property graph to inject post-retrieval synonyms (Phase 3), multi-hop legal facts, partial-match annotations, and normative gap signals. The resulting `EnrichedContext` block is passed to the generator alongside the re-ranked passages. Figure 1 illustrates the full pipeline.

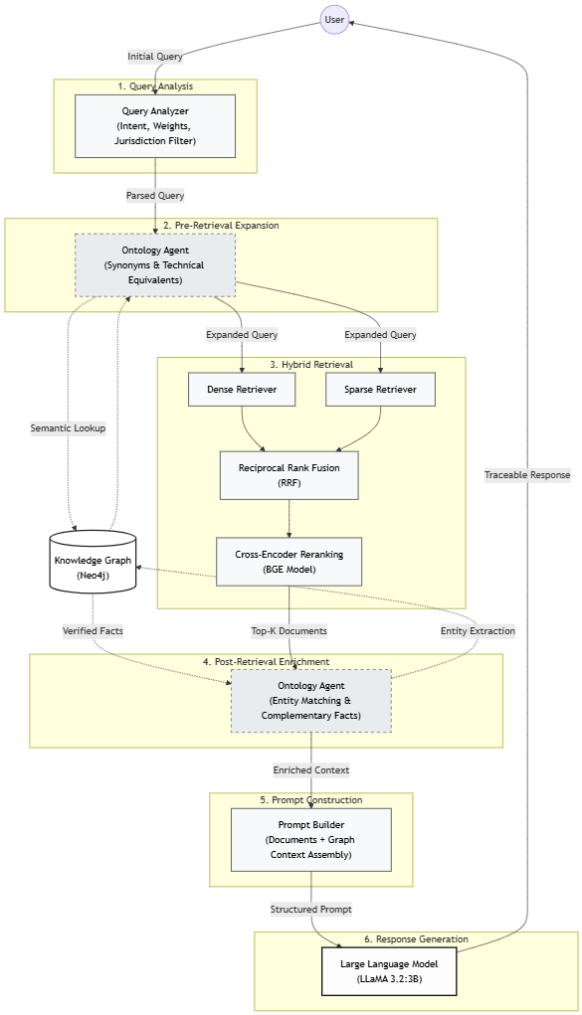


Figure 1: End-to-end pipeline. The Neo4j graph operates exclusively as a post-retrieval enrichment layer. Synonyms intervene at three stages.

3.2 Marine Environmental Law Ontology

T-Box coverage and LKIF-Core alignment. The T-Box is aligned with LKIF-Core: `Interdiction` is a subclass of `lkif:Prohibition`; `SourceJuridique` of `lkif:Legal_Source`; `Acteur` of `lkif:Agent`; `Zone` of `lkif:Place`. Six marine prohibition categories are formalised: cetacean hunting (ICRW 1946); coastal construction in sensitive zones (ICZM Protocol, Barcelona Convention 2008); marine bird protection (CMS/AEWA); marine sand extraction

(CBD 1992); hydrocarbon discharge (MARPOL 73/78 Annex I); and bottom trawling in Vulnerable Marine Ecosystems (A/RES/61/105, 64/72, 66/68, 71/123). The T-Box comprises 120 OWL 2.0 DL classes and 72 properties (45 object, 27 data), yielding 8,844 RDF triples and 1,217 individuals.

SKOS lexical bridge. The bridge attaches synonym sets via `skos:prefLabel` and `skos:altLabel`. For instance, `ChasseBaleine` links *Whale hunting* to *Capture of large cetaceans*, *whaling*, and *Hunting of rorquals*. This makes retrieval robust to cross-jurisdictional terminological variation and directly updatable when new legislation is added, without neural retraining.

Axiomatic safety net. The ontology adopts a lightweight axiomatisation paradigm: its 68 strict logical axioms act as a formal safety net rather than a classification driver, since LLMs already perform entity classification during ingestion. Table 1 groups these 68 axioms by OWL 2 construct, verified directly against the ontology-generation code. Among the 44 class-disjointness axioms, `SanctionPenale` is strictly disjoint from `SanctionAdministrative`; `EtatSouverain` from `NavirePeche`; `ChalutageFond` from `PecheINN`. `Cetace` is defined via `owl:disjointUnionOf` as the strict union of `BaleineMysticete` and `BaleineOdontocete`, enforcing a closed taxonomy that extraction models cannot deviate from. Among the 9 property-characteristic axioms, `couvreZone` is transitive (if zone A covers B and B covers C, then A covers C by inference); `entraineSanction` is asymmetric and irreflexive, preserving causal integrity. A further 10 axioms (6 equivalent-class definitions and 4 universal restrictions) drive automatic classification rather than safety filtering, and are reported separately from the 68 for that reason.

3.3 Controlled Ontology-to-Graph Transition

Axiomatic A-Box filtering. LLM-extracted entities and triples from the 12 international legal documents undergo T-Box validation before graph population. The disjointness constraint between `SanctionPenale` and `SanctionAdministrative` blocks dual classification; domain restrictions on `estActiviteDe` prevent a `Zone` from being asserted as the subject of an `Activite`. This yields a formally consistent A-Box of 1,217 individuals.

Axiom type (OWL 2 construct)	Count
Class disjointness (`disjointWith`)	44
Disjoint union (`disjointUnionOf`)	1
Existential restrictions (`someValuesFrom`)	10
Property chain (`propertyChainAxiom`)	1
Property characteristics (transitive, symmetric, asymmetric, irreflexive, functional)	9
Cardinality restrictions (min/maxQualifiedCard.)	3
Total (strict safety axioms)	**68**

Table 1: The 68 strict T-Box axioms by OWL 2 construct, counted directly from the ontology-generation code. An additional 6 equivalent-class and 4 universal-restriction axioms support automatic classification and are excluded from this count (see text).

Inference materialisation and graph export. HermiT 1.4 materialises 299 implicit inferences before export. An actor engaging in a prohibited activity within a protected zone receives the inferred label `ActeurEnInfraction`; the property chain `estActiviteDe` resolves into direct `[:EST_SOUMIS_A]` edges. Each `owl:NamedIndividual` becomes a labelled Neo4j node; each `owl:ObjectProperty` a typed relationship. Pre-materialising all inferences means that OWL guarantees are exploitable at Cypher traversal speed at inference time, with no live reasoner overhead. Crucially, the graph enriches context only after retrieval; it is not a retrieval channel, so ontologically related but textually unattested nodes cannot inflate recall.

3.4 Hybrid Retrieval with Three-Phase Synonym Integration

Dense and sparse retrievers. All 1,500+ chunks are encoded with BGE-M3 (Chen et al., 2024), producing 1,024-dimensional vectors stored in ChromaDB (cosine, HNSW). A contextual embedding strategy prepends each chunk's legal breadcrumb (country, document title, article hierarchy) before encoding, anchoring embeddings in their normative context. A BM25Okapi index uses a French-aware tokeniser preserving apostrophes and hyphens, reliably surfacing exact legal references (MARPOL Annex I, EEZ, PSSA) that dense embeddings may subsume into broader semantic clusters.

Query Analyser and synonym-aware weighted fusion. Nine intent classes are assigned by regex

and mapped to weight pairs $(\alpha_{\text{dense}}, \alpha_{\text{sparse}})$, biasing sparse retrieval for citation-exact queries and dense retrieval for conceptual and comparative ones. The synonym-aware RRF score is:

$$\text{Score}_{\text{RRF}}(d) = \sum_{r \in \mathcal{R}} \frac{\alpha_r}{k + \text{rank}_r(d)}, \quad k = 60.$$

with a synonym boost of $+0.05$ per `skos:altLabel` match (capped at $1.5\times$). The top-30 candidates are re-scored by `BAAI/bge-reranker-v2-m3`; the top-5 passages form the retrieved context passed to the agent.

3.5 Ontological Reasoning Agent and Constrained Generation

Seven-step reasoning agent. After retrieval, `Neo4jOntologyAgent` enriches the context through: (1) entity detection; (2) Phase 3 synonym injection into the prompt; (3) multi-hop fact retrieval (Species → Prohibition → Convention → Zone); (4) generic-to-subtype activity expansion; (5) cross-document enrichment via Neo4j lookup; (6) partial-match detection, annotating chunks on *protected species* as partial evidence; and (7) normative gap signalling, generating an explicit absence notice when no national text attests the queried prohibition.

Constrained answer generation. The prompt comprises three blocks: the ontological block (synonyms, multi-hop facts, annotations, gap signals); the documentary block (re-ranked passages with breadcrumb citations); and the user question with an intent-specific directive. A constraint-driven system prompt (Figure 2) governs output behaviour, enforcing binary format and verbatim citation. Generation is performed by **Llama 3.2 3B** via Ollama, functioning as an architectural stress test: performance gains attributable to a 3B generator can be assigned unambiguously to pipeline design rather than parametric capacity.

4 Building the Legal Knowledge Graph

Construction followed three phases. Phase 1 formalised twelve competency questions as SPARQL queries to delimit ontological scope and guide T-Box design. Phase 2 carried out multi-stage knowledge extraction from the source documents. Phase 3 assembled, formally verified, and exported the graph.

System prompt (translated from French).

You are a legal assistant. Answer based **solely** on the provided documents. If an [ONTOLOGICAL ENRICHMENT] block is present, use it to identify the relevant article. Each document begins with a bracketed header; reproduce it verbatim.

YES, there is an article on [brief answer]. According to [document header], "[exact citation]".
NO, there is no article on [subject] in the provided texts.

[Fundamental Rule] Answer YES only if a document addresses the question **exactly**. If the text covers a related topic but does not answer precisely, answer NO. **Do not make inferences.**

Figure 2: Constraint-driven system prompt. The binary format enforces verbatim citation; the Fundamental Rule is intended as the primary safeguard against fabricating content on trap queries, though §7 reports a more modest measured effect than originally claimed.

4.1 Complementary Multi-Model Triple Extraction

Transforming formal juridical text into RDF triples requires both exhaustivity and precision, objectives that no single extraction model reliably achieves simultaneously across heterogeneous document types. Two models with architecturally distinct designs were therefore run independently on each source document: Claude Sonnet 4.6, which produces compact, precision-oriented output, and LlamaParser, which achieves substantially higher recall at the cost of greater noise.

Their complementarity is quantitatively striking (Table 2). On the bottom-trawling prohibition, the two models share 29 of the 39 triples produced by Claude Sonnet 4.6, a 74% overlap. On the whale-hunt prohibition, however, Claude Sonnet 4.6 extracted 40 triples and LlamaParser extracted 329, with only 6 shared, an overlap below 2%.

Prohibition	Model	Ent.	Trip.	Shared
Bottom trawl	Sonnet 4.6	31	39	29
	LlamaParser	70	200	29
Whale hunt	Sonnet 4.6	38	40	6
	LlamaParser	331	329	6

Table 2: Extraction complementarity across two prohibitions. Overlap ranges from 2% to 74%; their union maximises coverage.

All merged triples were validated by maritime law specialists over one week. The review was conducted as a single joint reading session involving three of the co-authors – including a maritime and environmental law specialist – together with

several doctoral students in maritime law, rather than as independent parallel annotation. Reviewers flagged LLM-extracted triples judged incorrect or unsupported by the source text; flagged triples were discarded rather than retained, yielding a conservative A-Box. Because review was joint rather than independent, no formal inter-annotator disagreements arose and no inter-annotator agreement score was computed.

4.2 Legal Definitions and SKOS Synonym Seeding

Legal definition extraction. A pool of four models extracted legal definitions from the corpus in parallel: Mistral-8B, LlamaParser, Gemma-4-E4B-it, and Qwen. Each proved most reliable on a specific document class: LlamaParser achieved the most faithful reproduction of technical parameters in scientifically dense conventions (ICRW, MARPOL); Gemma-4-E4B-it produced the cleanest administrative definitions for coastal management instruments (ICZM Protocol, CMS/AEWA); Mistral-8B excelled on national texts with high concentrations of technical nautical terminology. Rather than imposing a single extractor on the full corpus, a domain-wise selection strategy retained the best-performing output per thematic cluster, achieving comprehensive and precisely worded coverage across all six prohibition categories.

SKOS synonym seeding. A dedicated synonym-generation pass addressed a fundamental retrieval risk: the same legal reality may be described as *whale*, *cetacean*, or *marine mammal* depending on the jurisdiction. Synonym sets were generated for all six prohibition categories across 15 national legal systems and encoded via `skos:altLabel`. This lexical layer makes retrieval robust to cross-jurisdictional terminological variation without requiring neural retraining when new legislation is added.

4.3 Formal Verification and Graph Export

The T-Box was generated programmatically via RDFLib for full reproducibility. After A-Box population under axiomatic filtering, HermiT 1.4 verified global consistency and pre-materialised 299 inferences. The validated graph was then exported to Neo4j.

The resulting property graph comprises 1,038 domain nodes representing OWL individuals, connected by 7,714 typed edges instantiating the full set of object properties and pre-materialised inferences. A dedicated lexical layer contributes 400+ `ConceptLexical` nodes encoding technical vocabulary and legal definitions extracted from the international corpus. The graph is structured around 10 node types and 15 relation types, enabling fine-grained Cypher traversals at inference time without live reasoner overhead. Population combines a top-down component, in which the six core prohibition categories were manually encoded as seed triples, with a bottom-up component, in which complementary triples automatically extracted by the LLM pipeline were filtered through the 68-axiom safety net before insertion.

5 Experimental Protocol

Corpus. The international corpus (12 documents, 364 pages) covers ICRW 1946, four UN GA resolutions on bottom trawling (A/RES/61/105, 64/72, 66/68, 71/123), CMS 1979/AEWA, MARPOL 73/78 Annex I, Barcelona Convention Protocol, UEMOA Regulation 06/2022, ICZM Protocol 2008, Loi Littoral 1986/2021, and CBD 1992. The national corpus (99 documents, 15 countries) covers all six prohibition categories. Chunks scoring ≥ 0.5 on a normative-term density metric are indexed, yielding 1,500+ chunks with hierarchical breadcrumb metadata.

Benchmark design. A pool of 264 specialist-authored queries was built over two prohibition categories selected for their deep textual coverage (cetacean hunting and hydrocarbon discharge), across 14 of the 15 national legal systems, along four analytical dimensions further split into 11 axes (existence, zoning, temporality, activity type, two exception axes, financial and penal sanctions, and three control axes). For the ablation reported here, a stratified subset of 66 queries was drawn from this pool: one query per (theme $\times$ axe $\times$ country-group) cell, with countries split into three balanced groups, using a fixed random seed for reproducibility and preferring non-trap queries within each cell. Trap queries – concentrated on one axis (exceptions outside health/order/research grounds) and designed as near-miss discrimination tests rather than queries with a guaranteed negative gold answer – are represented by 6 queries in the evaluation subset (one per theme $\times$ country-group). §7 reports a manually-verified re-analysis of this axis, since an earlier, retrieval-level automated metric on this axis proved uninformative (see below). Re-

trieval is evaluated by MRR, NDCG@5, Recall@5, P@5, and latency. End-to-end quality uses RAGAS (Es et al., 2024) with Llama 3.1 8B as judge (a model distinct from the 3B generator), reporting Faithfulness, Answer Relevancy, Context Precision, Context Recall, and Context F1 (harmonic mean of Ctx.P and Ctx.R). Of these, Context Precision and Context Recall are computed against the expert-authored ground-truth answer (RAGAS's reference-based implementations); Faithfulness is checked only against the retrieved context and Answer Relevancy only against the question itself, so the judge model is not asked to arbitrate legal correctness directly for either of those two metrics, and carries no ground-truth anchor for them. Four configurations are compared: (1) Dense+Ctx (BGE-M3 dense retrieval with contextual chunking, no BM25, no agent); (2) Dense+Agent (BGE-M3 with Ontological Reasoning Agent, no BM25); (3) Dense+Sparse (BGE-M3 and BM25 fused via RRF, no agent); and (4) Full (complete system: Dense+Sparse + agent).

6 Results and Ablation Analysis

6.1 Retrieval Performance

Method	MRR	NDCG@5	R@5	P@5	ms
Dense only	0.870	0.880	**0.771**	0.732	148
Sparse only	0.477	0.506	0.560	0.451	**82**
Dense+Sparse	**0.901**	**0.890**	0.744	**0.735**	245
+Rerank	**0.901**	**0.890**	0.744	**0.735**	251

Table 3: Retrieval performance. Bold: best per column.

Dense+Sparse raises MRR from 0.870 (dense alone) to 0.901, confirming the complementarity of the two modalities: dense retrieval handles paraphrastic queries while BM25 reliably surfaces exact legal references (MARPOL Annex I, EEZ, PSSA) that embeddings subsume into broader semantic clusters. The cross-encoder adds 6 ms with no rank-level change; its contribution to generation quality is captured in Table 4.

6.2 End-to-End Quality

The ontological agent as faithfulness driver. Comparing Full to the Hybrid configuration (identical retrieval stack, agent removed) isolates the agent's contribution: Faithfulness rises from 0.3436 to 0.4538 (+32.1%) and Context Precision from 0.7425 to 0.8929 (+20.3%). Removing the agent yields the lowest Faithfulness across all configurations, confirming that partial-match detection and normative gap signalling are the primary hallucination-prevention mechanisms. The Fundamental Rule of the system prompt reinforces this by prohibiting any inference beyond attested evidence.

Configuration	Faith.	Rel.	Ctx.P	Ctx.R	Ctx.F1
Dense+ctx	0.4053	0.2738	0.8385	0.7617	0.7983
Dense+Agent	0.4415	**0.6479**	0.7254	0.8373	0.7773
Dense+Sparse	0.3436	0.5636	0.7425	**0.8513**	0.7932
Full	**0.4538**	0.5061	**0.8929**	0.7653	**0.8242**

Table 4: RAGAS ablation results on the 66-query stratified evaluation subset. Dense+Ctx: BGE-M3 + contextual chunking shared by all configurations, no BM25, no agent. Dense+Sparse: BGE-M3 + BM25 via RRF, no agent. Generator: Llama 3.2 3B. Judge: Llama 3.1 8B (independent model). Ctx.F1: harmonic mean of Ctx.P and Ctx.R. Bold: best per column.

BM25 as the precision driver. Comparing Full to Dense+Agent isolates the contribution of sparse retrieval: Context Precision rises from 0.7254 to 0.8929 (+23.1%). BM25 surfaces exact legal references that dense embeddings conflate with semantically adjacent but legally distinct passages. The resulting Context F1 of 0.8242 is the highest across all configurations, a 6.1% gain over Dense+Agent.

Precision-recall trade-off. The Full pipeline does not lead on Answer Relevancy (0.5061 vs. 0.6479 for Dense+Agent) or Context Recall (0.7653 vs. 0.8513 for Hybrid). Both trade-offs are expected: BM25 raises precision but can demote passages that synonym expansion would rank first. For legal compliance auditing, where false positives are more consequential than false negatives, this precision-first profile is architecturally coherent. The very low Answer Relevancy of the dense-only configuration (0.2738) confirms that without ontological enrichment, the 3B generator lacks sufficient legal context to produce topically grounded responses.

6.3 Qualitative Illustration

The following example instantiates the Faithfulness and Context Precision gains reported in Table 4. Both pipelines receive the identical query; the difference in output traces directly to the agent's ontological concept expansion.

Query (Cameroon – coastal construction):(translated from French)
Is there an article on the prohibition of coastal construction in Cameroon?

RAG (Dense only). YES, there is an article on the prohibition of certain harmful activities on the Cameroonian coastline. According to `[Cmr] cmr19578 | Title III | Ch. III | § III | Art. 31`, the discharge, immersion, and incineration in maritime waters of substances harmful to human health or marine biological resources are prohibited, along with the development of certain fishing structures.

→ Semantically adjacent but legally incorrect: the retrieved passage covers marine pollution, not coastal construction.

Full Pipeline (Neuro-Symbolic RAG). YES, there is an article on the prohibition of coastal construction in Cameroon. According to `[Cmr] cmr19578 | Title d | Art. 35`: *"A non-aedificandi zone is delimited along maritime coasts, river and lake banks; its regime is governed by land law."*

→ Exact match: the agent resolved the query via the `ConstructionLittorale` concept and its `ZoneNonAedificandi` subtype, which dense retrieval conflated with semantically adjacent pollution passages.

Figure 3: Contrastive output on a coastal-construction query (Cameroon). Classic RAG retrieves a legally incorrect passage through semantic proximity; the neuro-symbolic pipeline identifies the exact prohibitive article via ontological concept expansion, illustrating the Context Precision gain (+20.3%) reported in Table 4.

7 Discussion

The ablation isolates two orthogonal failure modes, each addressed by a distinct architectural component. The ontological agent is not a marginal refinement: its removal drops Faithfulness to the lowest value across all configurations, below even the Dense+Agent configuration that lacks sparse retrieval. BM25 addresses a complementary failure: its removal reduces Context Precision by 18.8% relative even when the agent is present. These two components address different stages of the failure chain and must operate jointly.

The use of Llama 3.2 3B as the generator is a deliberate methodological choice. A larger model might compensate for pipeline weaknesses through parametric knowledge, confounding the attribution of gains. The 3B model's known limitations make every performance improvement attributable unambiguously to the pipeline. The absolute Faithfulness score (0.4538) reflects the known limitations of a 3B generator on dense legal text rather than a pipeline failure: the gain (+32.1% over the Dense+Sparse baseline) is attributable to the agent's gap signals, which suppress fabrication even when the generator lacks parametric capacity to complete missing legal content.

Retraction of the original trap-query claim. An earlier version of this paper reported a 100% true-negative rate on trap queries. On inspection, this figure came from a retrieval-level metric (`piege_tnr` in the released code) that scores 1.0 whenever the retrieval call completes without a runtime exception, regardless of whether the retrieved passages support the generated answer – it is a crash-rate indicator, not a discrimination measure, and we retract the claim built on it. We instead report a manually-verified re-analysis: for the 24 country×theme cells for which manual evaluation grids exist on the exception-discrimination axis, the generated answers were compared directly against the expert-authored gold answer. The full pipeline reaches 15/24 (62.5%) correct discriminations, against 13/24 (54.2%) for the dense-only baseline – a real but modest gain, far from the originally claimed 100%. A representative failure case illustrates why: for a Congo query on whale-hunting exceptions, the gold answer confirms an exception under Article 20; the dense-only baseline answers correctly, citing Article 20, while the full pipeline's ontological agent retrieves the same article yet concludes NO – a case where ontological enrichment surfaced the right source but did not improve the final inference. Conversely, for a Comoros query on the same axis, the dense-only baseline answers NO while the full pipeline correctly answers YES, citing a relevant (if not gold-identical) article. These mixed outcomes indicate that exception-discrimination is a harder task than the six core prohibition-existence categories, and that the agent's contribution there is real but partial, not the clean hallucination-blocking result originally claimed.

Limitations. Coverage extends to six prohibition categories; broader marine law requires sustained expert effort. Graph maintenance is an ongoing cost as legislation evolves. Restricting adaptive synonym expansion to conceptual queries only could further sharpen the precision-recall balance. The system is intended as a drafting assistant for compliance audits, not as a substitute for legal counsel.

With 66 queries and no confidence intervals reported in Table 4, the robustness of the percentage gains should be read with appropriate caution; a larger held-out evaluation is future work. RAGAS scores throughout depend on a single LLM

judge (Llama 3.1 8B). Context Precision and Context Recall are anchored to expert-authored ground truth, but Faithfulness and Answer Relevancy are reference-free: the judge alone decides whether the answer is consistent with the retrieved context and relevant to the question, with no expert answer to check it against. Residual judge-specific bias on legal text therefore cannot be ruled out for these two metrics without a human expert cross-check, which we have not yet conducted. Similarly, the architectural contrast with graph-as-retrieval-channel approaches (GraphRAG, OG-RAG) drawn in §2 is not yet backed by a quantitative comparison on our corpus; establishing one is a priority for future work.

8 Conclusion

We have presented a neuro-symbolic RAG system in which a formally verified OWL 2.0 DL ontology serves as the primary knowledge artifact. The controlled transition from T-Box axioms through axiomatically filtered A-Box instantiation to a pre-secured Neo4j property graph decouples formal reasoning from inference time, making description-logic guarantees available at Cypher traversal speed without live reasoner overhead. On a stratified sample of 66 expert queries drawn from a 264-query pool over 15 coastal states, the system achieves Context Precision 0.8929 (+23.1%), Context F1 0.8242, and a +32.1% Faithfulness gain, using a 3B generator as a deliberate stress test. On the harder exception-discrimination axis, a manually-verified re-analysis shows a modest but real accuracy gain over dense-only retrieval (62.5% vs. 54.2%); we explicitly retract an earlier, artefact-driven 100% true-negative claim on this axis. The results establish a clear principle: in specialised legal domains characterised by rich relational structure and multi-jurisdictional terminological heterogeneity, the quality of context construction matters more than the capacity of the generator.

Availability. Code, the OWL 2.0 ontology, the SKOS lexical bridge, the Neo4j export scripts, the full 264-query benchmark pool, the 66-query stratified evaluation subset, and the prompts are available at https://github.com/Assoumana-Souley-Hadiza/neurosymbolic-rag-maritime.

References

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2309.07597*.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre F. T. Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, and Michael Desa. 2024. SaulLM-7B: A pioneering large language model for law. *arXiv preprint arXiv:2403.03883*.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. ChatLaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2024. RAGAS: Automated evaluation of retrieval augmented generation. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1088–1099. Association for Computational Linguistics.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. LegalBench: A collaboratively built

benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Sumit Bhatia Hakim et al. 2025. Efficient neuro-symbolic retrieval-augmented generation through adaptive query routing. *arXiv preprint arXiv:2506.12981*.

Rinke Hoekstra, Joost Breuker, Marcello Di Bello, and Alexander Boer. 2007. The LKIF core ontology of basic legal concepts. In *Proceedings of the Workshop on Legal Ontologies and Artificial Intelligence Techniques (LOAIT 2007)*, pages 43–63.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Seyedeh Somayyeh Khorshidi et al. 2025. ODKE+: Ontology-guided open-domain knowledge extraction with LLMs. *arXiv preprint arXiv:2509.04696*.

Sebastian Langer, Frank Neuhaus, and Andreas Nürnberger. 2024. CEAR: Automatic construction of a knowledge graph of chemical entities and roles from scientific literature. In *Proceedings of the Joint Ontology Workshops (JOWO 2024), co-located with FOIS 2024*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2356–2362.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

Kartik Sharma, Peeyush Kumar, and Yunqing Li. 2025. OG-RAG: Ontology-grounded retrieval-augmented generation for large language models. *arXiv preprint arXiv:2412.15235*.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Huang, Xinyu Ye, Junjie Yao, Zechen Wang, Jia-Long Li, Ji-Rong Wen, and Rui Yan. 2023. DISC-LawLLM: Fine-tuning large language models for intelligent legal services. *arXiv preprint arXiv:2309.11325*.

Evaluating Model-Task Fit in Arabic Word Sense Disambiguation

Yousef Younes
Universität Ulm
Germany
yousef.younes@uni-ulm.de

Abdelhalim Dahou
GESIS - Leibniz Institute for the Social Sciences
Germany
Abdelhalim.dahou@gesis.org

Brigitte Mathiak
GESIS - Leibniz Institute for the Social Sciences
Germany
brigitte.mathiak@gesis.org

Abstract

Word Sense Disambiguation (WSD) is the computational task of identifying a word's intended meaning within a specific context when the word exhibits lexical ambiguity. It remains a core and challenging problem in Natural Language Processing (NLP). We formulate the problem in two ways: binary classification (context-gloss alignment) and multi-choice selection (identifying the correct sense from candidate glosses) using Large Language Models (LLMs) and Small Language Models (SLMs). Our results on the KSAA-CAD shared task dataset show a critical task-model dependency, where fine-tuned SLMs excel in binary classification (mBERT achieving 83.74% F1), while LLMs show strong performance on multi-choice (fine-tuned Gemma2 7B attaining 84.26% F1). Notably, Camel-BERT-MSA shows robust cross-task consistency, securing second place in both evaluations. This insight underscores that the effectiveness of a model in WSD depends heavily on the specific task. It also emphasizes the value of aligning model architecture with task requirements, offering useful insights for applying WSD models in real-world settings.

1 Introduction

The Word Sense Disambiguation (WSD) task aims to identify the intended meaning of a word based on its context. This task is crucial because many words in natural language are polysemous—having multiple related meanings (e.g., "paper" as a material vs. a publication)—or homonymous—having multiple unrelated meanings (e.g., "bark" as the sound a dog makes vs. the outer layer of a tree). Solving the polysemy and homonymy problems is important for language understanding tasks such as information retrieval, question answering, sentiment analysis, text summarization, and machine translation (Rahman and Borah, 2021; Pu et al., 2018), where accurately interpreting a word's intended meaning is crucial for overall system performance.

The WSD task helps resolve both polysemous and homonymous ambiguities by using context to select the correct meaning. Distributional word embeddings like Word2Vec (Church, 2017) and GloVe (Pennington et al., 2014) assign a single vector representation to each word, thereby failing to capture this semantic variation. To address this limitation, contextual word embeddings, such as those derived from transformer-based models like BERT (Devlin et al., 2019), were introduced. These embeddings dynamically generate different representation vectors for the same word depending on its context, enabling more effective handling of semantic variation (Devlin et al., 2019; Abdul-Mageed et al., 2020).

Although the WSD task has been extensively studied in English (Huang et al., 2019; Navigli, 2009), it remains particularly challenging for morphologically rich languages like Arabic. In Arabic, polysemy is highly prevalent: approximately 43% of words with diacritics have multiple meanings, and this figure rises sharply to 72% when diacritics are removed (Alshammari et al., 2024). Beyond its inherent difficulty, Arabic WSD is also significantly underexplored. This is illustrated by the recent KSAA-CAD shared task (Alshammari et al., 2024), where only one submission was made to the WSD subtask—and it was disregarded due to its performance being 64% below the baseline. Another shared task ArabicNLU (Khalilia et al., 2024) provide dataset of just 100 sentences and received just three submissions, without outperforming the baseline.

Existing approaches to the WSD problem adopt context-gloss pair classification (Huang et al., 2019) and relevance ranking (Yap et al., 2020). In this work, we frame the Arabic WSD problem as both a binary classification task and a multi-choice ((MC)) selection task. We run a

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 398–407
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_041

set of experiments using Small Language Models (SLMs), e.g., mBERT (Pires et al., 2019), and Large Language Models (LLMs), e.g., Mistral (Albert Q. Jiang, 2023), on the WSD dataset introduced in the KSAA-CAD shared task. The dataset is described in Section 4.1.

Our study is guided by the following research questions and corresponding contributions:

1. **RQ1. Do multilingual models outperform Arabic monolingual models, and how does task format impact cross-lingual transfer?** We compare multilingual models (e.g., mBERT, Gemma2, and others) against Arabic-specific monolingual models (e.g., AraBERT) under identical fine-tuning protocols. Performance is evaluated on both binary and multi-choice task formulations to isolate cross-lingual transfer effects.

2. **RQ2. How does task formulation affect model performance on WSD?** We conduct paired evaluations where each model executes both: (a) *binary classification*: context-gloss alignment; (b) *Multi-choice selection*: identifying the correct sense from candidate glosses. Task superiority is determined based on the F1-score metric on the dev set.

3. **RQ3. How do LLMs compare to SLMs on Arabic WSD under fine-tuning and zero-shot settings?** We benchmark open-source LLMs against SLMs under two conditions:
   - *Fine-tuned*: Models trained on our Arabic WSD training set.
   - *Zero-shot*: LLMs evaluated without task-specific training.

4. **RQ4. Can open-source LLMs match proprietary models in zero-shot Arabic WSD with identical prompts?** Top-performing open-source LLMs from RQ3 are compared against proprietary LLMs (e.g., GPT-4, Claude 3) using identical prompts and zero-shot settings. Closed-source models are accessed via API with standard hyperparameters.

Through our experiments, we found that fine-tuned multilingual models consistently outperform Arabic monolingual models. Additionally, task formulation significantly impacts performance, with some models showing clear preferences for binary classification or multi-choice formats. Finally, in the zero-shot setting, proprietary models (e.g., GPT-4, Claude 3) outperform open-source models, highlighting the performance gap between closed and open models in Arabic WSD.

The rest of this paper is organized as follows. Section 2 overviews some works that discuss the WSD problem. Then, we talk about the methods that we use to solve the problem in Section 3. After that, the data, models, hyperparameters, and metrics used in the experiments are described in Section 4. The results are reported in Section 5 and discussed in Section 6. Finally, the paper concludes in Section 7.

2 Related Work

The WSD task has received significant attention from researchers working on English and other Indo-European languages (Huang et al., 2019). However, handling this task in morphologically rich languages like Arabic is more challenging due to the ambiguity and rich morphology present in these languages.

Huang et al. (Huang et al., 2019) fine-tuned BERT-based models to give a *yes* or *no* label for each context-gloss pair. Next, the gloss from the pair that has the highest score for *yes* is selected as the target gloss. Another work (Yap et al., 2020) formulated the WSD task as a relevance ranking task where each context has a list of 4-6 glosses. They fine-tuned BERT to give scores to each gloss, then selected the gloss with the highest scores as the target sense of the target word. In this work, we use both binary classification and multi-choice tasks to tackle the WSD problem.

Researchers use different ways to emphasize the target word in the context. For example, Huang et al. (Huang et al., 2019) surrounded the target word by quotation marks and also added the target word followed by a colon at the start of each gloss in the training instances. On the other hand, some researchers proposed to use special tokens to mark the start and end of the target word. These tokens could be the same as in (Yap et al., 2020), which used " [TGT] " , or different like in (Botha et al., 2020; Lei et al., 2017). In this paper, we use different tokens to mark the start and end of the target word.

The authors in (El-Razzaz et al., 2021) fine-tuned AraBERTv2 (Antoun et al., 2020) and ARBERT (Abdul-Mageed et al., 2020) in binary classification on a small WSD dataset and claimed an

F1-score of 89%. However, the author of (Al-Hajj and Jarrar, 2021), investigated the results and found that the reason behind this high score, is the great intersection between the training and testing sets. To account for that, they introduced a WSD dataset and used it to fine-tune three Arabic BERT-based models, AraBERTv02, CAMeLBERT (Inoue et al., 2021), and QARiB (Abdelali et al., 2021), to tackle the problem as a binary classification one. The AraBERTv02 model performed the best and produced an accuracy of 84% on their dataset. A recent study (Wael et al., 2024) showed that using transformer-based models with siamese networks is an effective choice for the Arabic WSD task.

3 Methodology

This study focuses on Arabic WSD by tackling this problem as a binary classification task and also as a multi-choice task. We use small language models like mBERT (Pires et al., 2019) and Camel-BERT-MSA (Inoue et al., 2021) and also large language models like Mistral (Albert Q. Jiang, 2023), Llama3.1 (Grattafiori et al., 2024), and Gemma2 (Team et al., 2024). We test these models in both zero-shot and fine-tuned settings.

3.1 Binary Classification

This is the traditional way to handle the WSD task. Here every sample has a label, which is 1 to indicate that the given meaning is meant by the target word in the context and 0 otherwise. The top part of Figure 1 presents an example of a positive sample. During training, we inform the model about the target word using

```
``<token>TARGET_WORD</token>''
```

For LLMs, we use the following prompt:

```
< Given a context, a target word, and
a candidate sense (all provided
in Arabic), determine if the candidate
sense correctly
matches the meaning of the target
word in the given context.
The target word appears between <token>
and </token> tags in the context.
You MUST respond with exactly one number:
- Output "1" if the candidate sense
matches
the meaning of the word in the context.
- Otherwise, output "0".>
```

3.2 Multi-choice task

Here we have the context as the question and the different senses as possible choices of the target word. We instruct the model to focus on the target word using the same method used for the binary task and provide the possible meanings as choices for the model to select from. One possible problem here is how to specify the number of choices. We found that the best way is to use the highest number of possible choices for all the samples. Investigating the data, we found that in the worst case, a sample in the used dataset could have 43 choices, so we use that for all the samples in the training and development sets. Of course, a lot of samples do not have that number of choices, so we use "null" to fill in the missing choices as shown in the lower part of Figure 1. For LLMs, we use the following prompt:

```
<Your task is to choose **only one**
correct gloss that best matches the sense
of the word in the given context.
Please respond with **just the number**
of the correct answer (e.g., `0`, `1`,
`2`, ...). Do not include any text,
explanation, or justification.>
```

The high number of choices incurs a higher computational cost. However, we have tried to limit the number of choices using some similarity measure and found that doing so has a negative impact on the results. Because the system would then have two steps: one to select the most probable choices, say 4 or 5, and the other to make the final decision and choose the best choice. In other words, we would have two error sources instead of one.

4 Experimental Apparatus

4.1 Dataset

We use the data from the KSAA-CAD shared task (Alshammari et al., 2024). The dataset consists of two core components: the WSD context gloss mapping and dictionary data. The WSD context gloss mapping is divided into training, development, and testing sets containing 22,404, 2,801, and 2,801 samples, respectively. Each sample consists of a word, context, context ID, lemma ID, and corresponding gloss ID as depicted in part (a) of Figure 2. The dictionary contains 15,865 entries and is derived from the Contemporary Arabic Language Dictionary (Omar, 2008). Each dictionary entry contains lemma_id, gloss, and gloss_id as

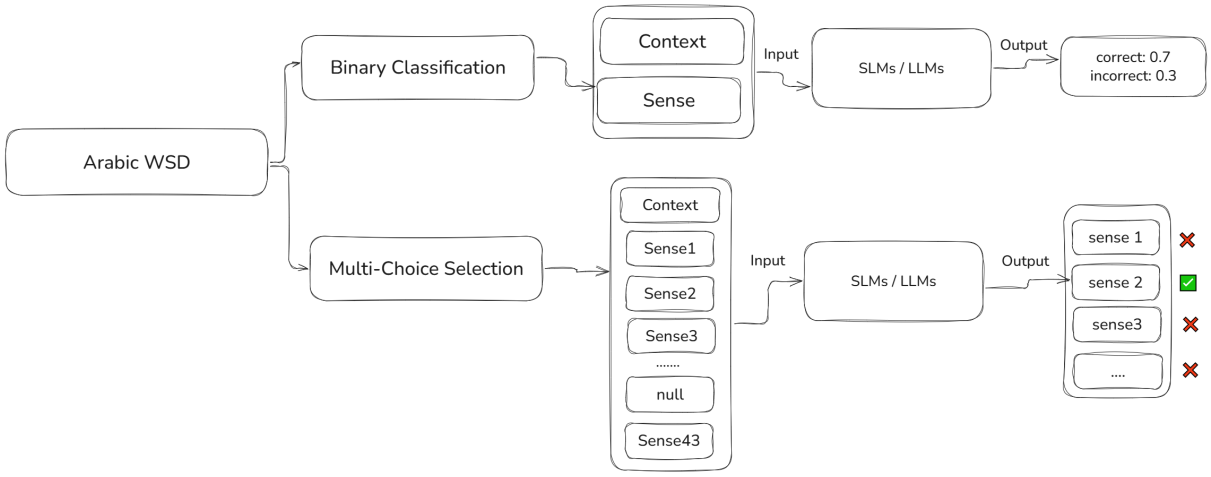


Figure 1: General workflow of the multi-choice and binary classification tasks for Arabic WSD.

shown in part (b) of Figure 2. The lemma_id is the link between the dictionary entry and the context gloss mapping. Unfortunately, the test set is not available, so we use the development set to compare the performance of the different models.

4.2 Models and Baselines

Baselines. We include two baselines from (Alshammari et al., 2024), which utilize transformer-based models in a binary classification setting, specifically CAMeLBERT-MSA and AraBERTv2.

To address our research questions, we experiment with a range of models representing diverse architectures. As an Arabic monolingual encoder-only models, we use AraBERT, CAMeLBERT-MSA, and ARBERT, bidirectional transformers pre-trained on MSA text and effective for Arabic contextual understanding. We also include multilingual encoder-only models such as mBERT (Pires et al., 2019), which is pre-trained on 104 languages, enabling cross-lingual transfer and multilingual understanding. This model serves as a comparative model to evaluate the effectiveness of multilingual versus monolingual representations in Arabic WSD. For decoder-only models, we employ instructed models including Llama3.1 (Grattafiori et al., 2024), Gemma2 (Team et al., 2024) with two variants in terms of size, Mistral (Albert Q. Jiang, 2023) version 0.3, and Qwen2.5 (Team, 2024). Additionally, we leverage advanced LLMs, including Claude 3.5 Sonnet (Anthropic, 2024)(version 2024-10-22), GPT-4o (Achiam et al., 2023) (version 2024-08-06), Gemini 1.5 Pro (Team et al., 2023), and DeepSeek V3 (Liu et al., 2024), which excel in text-related tasks due to extensive pre-training on natural language.

4.3 Hyperparameter

For multi-choice small models' fine-tuning, we trained for 3 epochs using a learning rate of 5×10^{-5}, a weight decay of 0.01, and a batch size of 2. In the other task, we set the batch size to 64, and used a learning rate of 3×10^{-5}, and models were trained for 3 epochs. All experiments were conducted using a fixed random seed of 42 to ensure reproducibility.

For instruction fine-tuning of Gemma-2B using QLoRA, we used a learning rate of 2×10^{-4}, a batch size of 4, weight decay of 0.001, and trained for 1 epoch with a maximum sequence length of 40. QLoRA parameters were set to $r = 4$, $\alpha = 16$, and dropout $= 0.1$. All training was performed with 8-bit quantization.

4.4 Metrics

We evaluate performance using accuracy, Macro-F1, and average precision/recall. Due to extreme class imbalance—particularly in the binary classification task (context-gloss alignment), where false labels substantially dominate true labels—we prioritize Macro-F1 across both tasks. Accuracy becomes particularly problematic here, as it fails to reflect model effectiveness. LLMs' responses that have no output or do not match any class label are regarded as misclassifications.

5 Results

5.1 Binary Configuration

Table 1 shows the evaluation results for the binary classification approach. The highest perfor-

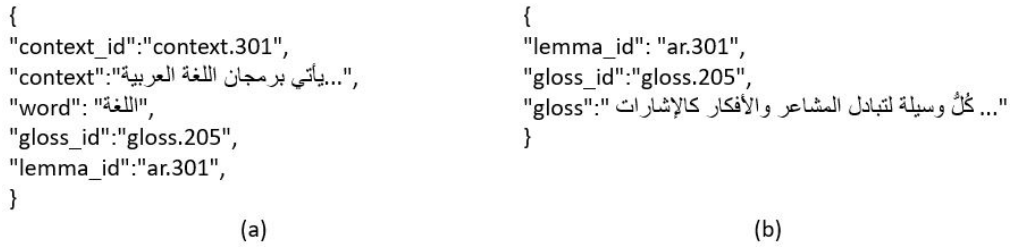


Figure 2: A training example of the KSAA-CAD dataset. Part (a) shows an entry from the context gloss mapping component. Part (b) displays a dictionary entry. This figure is reproduced from (Alshammari et al., 2024).

mance was achieved by fine-tuning mBERT model on Arabic data augmented with its English translation (En+Ar) reaching 83.74% Macro-F1 and the Arabic-only version (Ar) reaching 83.41%. Both significantly exceeded baseline models, including our experiment with the CAMeLBERT-MSA (80.04% Macro-F1). Our experiment with CAMeLBERT-MSA model also exceeded prior reported baselines achieving 94.17% accuracy.

Among zero-shot models, Qwen2.5 with 14B consistently achieved the best results, with Macro-F1 scores of 75.53% (fp16) and 75.57% (q8). It notably outperformed other zero-shot models, such as Gemma2 with 27B and Llama3.1. In contrast, Mistral-v0.3 stay far behind (13.63% fp16, 11.26% q8).

5.2 Multi-Choice Configuration

Results of the multi-choice approach are provided in Table 2. The multi-choice evaluation showed that the fine-tuned Gemma2 with 9B parameters achieved the highest Macro-F1 score (84.26%), surpassing both CAMeLBERT-MSA (81%) and AraBERT (77.08%).

Considering the models under zero-shot settings, proprietary models outperformed open-source models. GPT-4 led proprietary models with a Macro-F1 score of (55.36%), followed by Gemini-1.5-Pro (52.18%) and Claude3.5-Sonnet (40.61%). For open-source models in zero-shot settings, Gemma2 with 27B parameters achieved the best performance (53.94% Macro-F1), substantially outperforming Qwen2.5 with 14B parameters (45.14%) and Mistral-v0.3 with 7B parameters (14.45%). All fine-tuned models except ARBERT, consistently exceeded models under zero-shot settings.

6 Discussion

6.1 Key Scientific Insights

Our evaluation clearly demonstrates that fine-tuned multilingual models significantly outperform monolingual Arabic-specific models, and that task formulation and training paradigms critically influence performance outcomes in Arabic WSD task as shown in Table 1 and Table 2. In addition, fine-tuned multilingual models, notably Gemma2 with 9B parameters and mBERT (En+Ar), consistently achieved higher Macro-F1 scores compared to monolingual Arabic models such as AraBERT and ARBERT.

Insights by Research Questions:

1. **RQ1. Do multilingual models outperform Arabic monolingual models, and how does task format impact cross-lingual transfer?**

   Figure 3 confirms the results of Tables 1 and 2 that multilingual models significantly outperform Arabic-specific models under identical fine-tuning conditions. Gemma2 with 9B parameters and mBERT (En+Ar) consistently delivered higher Macro F1-scores compared to Arabic-specific models such as AraBERT and CAMeLBERT-MSA, indicating strong cross-lingual transfer capability. Multilingual models leverage semantic patterns from diverse languages (e.g., English and Arabic), enriching Arabic representations beyond monolingual training.

2. **RQ2. How does task formulation affect model performance on WSD?**

   The evaluations revealed that task formulation significantly impacts performance. Multi-choice selection task yielded substantially higher Macro-F1 scores across models compared to binary classification (context-gloss alignment). Gemma2:9B notably excelled in

Model	Accuracy	Macro F1	Precision	Recall	Provenance
Baselines					
CAMeLBERT	**94.17**	80.04	83.47	78.90	our experiment
CAMeLBERT	91.54	–	–	–	[x]
AraBERT	91.32	–	–	–	[x]
Zero-shot LLMs (fp16)					
Gemma-2:9B	*75.58*	*68.77*	*67.59*	*75.91*	
Gemma-2:27B	56.90	42.34	47.43	36.56	
Mistral-v0.3:7B	22.25	13.63	38.19	34.41	
Llama-3.1-8B	57.45	50.31	53.29	55.27	
Qwen2.5:7B	84.29	69.70	75.29	67.09	
Qwen2.5:14B	84.50	75.53	74.76	76.42	
Zero-shot LLMs (q8)					
Gemma-2:9B	73.14	67.66	67.59	77.54	
Gemma-2:27B	79.61	72.46	70.49	78.17	
Mistral-v0.3:7B	23.93	11.26	29.37	26.39	
Llama-3.1-8B	58.25	52.63	56.21	60.09	
Qwen2.5:7B	84.17	69.43	75.03	66.84	
Qwen2.5:14B	84.44	75.57	74.69	76.61	
Fine-tuned Models					
mBERT (En+Ar, B=64)	*94.14*	**83.74**	*86.60*	**81.73**	
mBERT (Ar, B=64)	94.02	83.41	**86.85**	80.90	

Table 1: Performance of language models on binary classification across different configurations

the multi-choice setup, achieving a Macro-F1 score of 84.26%, highlighting multi-choice as a superior task formulation for capturing contextual nuances in Arabic WSD. The multi-choice format provides a clear distinction among candidate senses, and modeling sense competition by ensuring rare senses contribute equally. Crucially, it supports words with diverse sense inventories (2-10 glosses), as shown in Table 4.

3. **RQ3. How do LLMs compare to SLMs on Arabic WSD under fine-tuned and zero-shot settings?**

   Fine-tuned models consistently outperformed zero-shot counterparts, demonstrating the substantial benefits of task-specific training. The performance gap was especially pronounced, with fine-tuned Gemma2 surpassing the top-performing models in zero-shot settings, GPT-4, by 28.9 Macro-F1 points. This emphasizes that despite the reasoning capabilities of models under zero-shot settings, fine-tuning remains critical for high-quality Arabic WSD.

4. **RQ4. Can open-source LLMs match proprietary models in zero-shot Arabic WSD with identical prompts?**

   Proprietary LLMs like GPT-4 outperformed the leading open-source model in zero-shot settings (Gemma2:27B), although the margin was relatively modest. GPT-4 (55.36% Macro-F1) surpassed Gemma2:27B (53.94%), suggesting that while open-source models are competitive, proprietary LLMs still maintain a slight performance advantage under identical zero-shot conditions and prompts.

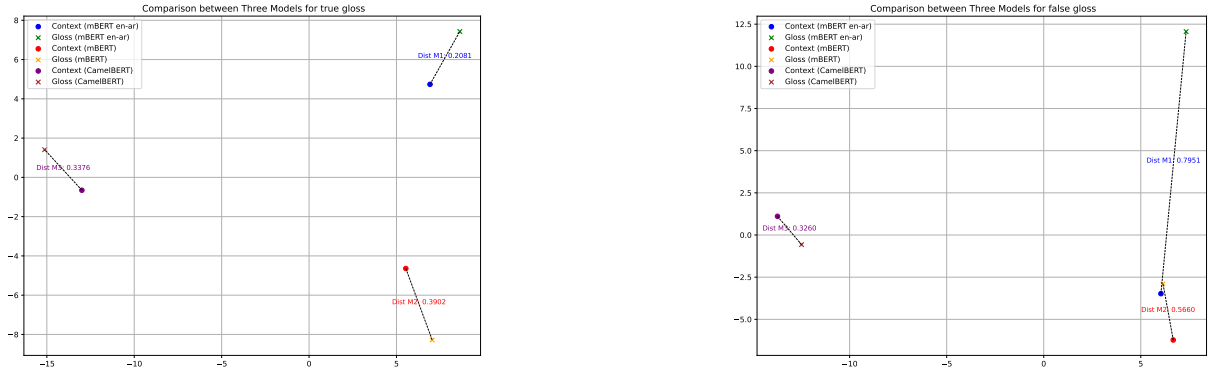
In summary, multilingual fine-tuned models substantially outperform monolingual counterparts and zero-shot models, with task formulation strongly influencing WSD performance outcomes. While proprietary models under zero-shot settings slightly outperform open-source models, fine-tuning significantly elevates overall performance.

6.2 Error Analysis

Analysis of the errors of the best performing model (Gemma2:9B) for multi-choice task in Table 3.

Model	Accuracy	Macro F1	Precision	Recall
Zero-shot Closed Models				
GPT-4	77.22	55.36	63.94	51.05
Claude3.5-Sonnet	77.12	40.61	44.80	37.94
Gemini-1.5-Pro	73.72	52.18	57.84	50.33
DeepSeek-V3	73.04	44.50	56.17	39.42
Zero-shot Open Models				
Gemma-2:9B	66.72	32.64	34.36	31.84
Gemma-2:27B	73.00	53.94	63.45	49.88
Mistral-v0.3:7B	44.55	14.45	18.07	13.07
Llama-3.1-8B	39.16	18.03	25.40	15.39
Qwen2.5:14B	68.26	45.14	50.86	44.29
Qwen2.5:7B	63.12	31.95	39.04	28.75
Fine-tuned Models				
Gemma-2:9B	**90.14**	**84.26**	**87.84**	**82.84**
CAMeLBERT	*87.00*	*81.00*	*86.84*	*78.63*
AraBERT	84.64	77.08	80.78	75.56
mBERT	67.08	56.59	62.24	59.17
ARBERT	50.55	21.90	23.35	56.59

Table 2: Performance results for MC WSD on dev set across different models.



(a) **True gloss** comparison.

(b) **False gloss** comparison.

Figure 3: Cosine distance comparison of context and gloss embeddings across mBERT (En–Ar), mBERT, and CamelBERT for both true and false gloss predictions.

We found the following errors categories: Semantic confusion between subtly distinct glosses, surface-level biases favoring lexical cues over contextual meaning, domain mismatches misinterpreting, literal-figurative misinterpretations, and gloss issues from ambiguous distractors.

These repeated errors reveal how complex Arabic WSD can be, especially when dealing with words that demand nuanced contextual and cultural interpretation.

7 Conclusion

This work provides an evaluation of different methods and models for Arabic WSD. Multilingual models like mBERT and Gemma2 consistently outperform Arabic-specific models by 3–7% Macro-F1, confirming strong cross-lingual transfer. The multi-choice task formulation proves more effective than binary classification. Fine-tuned models deliveres strong results overall, while proprietary LLMs maintained a 1.4–15% edge in zero-shot settings over open-source counterparts using identical prompts.

Error Category	Description	Example
Semantic Confusion	Model fails to distinguish between **subtly different glosses** due to close meaning or vague wording.	"تصيبهم" → gold: "afflicted and harmed" (نال منه وجرحه), predicted: "befell him" (نزلت به وحلَّت)
Surface-Level Bias	Model relies on **surface cues** like shared words or short/abstract glosses instead of deep semantic reasoning.	"أباطيل"/**Falsehoods** → distractor gloss included "خرافات"/**myths** from the context
Domain/Context Mismatch	Model fails to adapt to **domain-specific** usage (e.g., medical, religious, economic).	"الهياكل"/**Structures** in anatomy → predicted gloss was "huge from everything"
Literal vs. Figurative Misreading	Model interprets a **figurative or abstract concept** in a **literal** or common sense.	"الاستقبال"/**Welcoming** of Ramadan → predicted "greet", gold was "begin anew"
Gloss Design Quality Issue	The gloss set itself contains **poor quality distractors**, vague definitions, or domain-inappropriate candidates.	"الأعمال"/**business** in economy → predicted "intentional act", gold was "profession"

Table 3: General Error Categories in Arabic WSD Mutli-choice Model Predictions where the number of glosses is 7 and 9.

#Gloss	mBERT	Gemma2:9B	Camel-Bert	AraBERT	ARBERT	#Samples
2	77.18	91.63	88.58	86.49	45.13	526
3	63.87	90.97	84.51	83.51	45.88	388
4	62.77	90.05	87.96	84.66	45.09	333
5	52.98	85.83	82.85	80.70	33.99	240
6	47.70	85.92	78.93	77.37	28.00	187
7	59.60	88.49	83.66	81.77	39.78	174
8	53.69	82.41	75.40	65.75	24.95	96
9	59.47	90.58	84.64	78.71	20.91	105
10	45.49	73.91	83.80	69.83	25.42	51

Table 4: Fine-tuned models performance on MC configuration by gloss count in terms of macro F1 score.

In the future work, we will address semantic granularity through contrastive gloss pairing and integrate Arabic knowledge resources to mitigate cultural-contextual gaps as highlighted by our error analysis. In addition, exploring more sophisticated prompting strategies such as few-shot and Chain-of-Thought (CoT) could better leverage LLMs' full capabilities.

Limitations

This study has three main limitations. First, the evaluation was limited to Arabic WSD using a single dataset, which may constrain generalizability to other languages or domains. Second, zero-shot evaluations utilized only one prompt engineering approach which may not fully exploit the potential of LLMs. Third, fine-tuning experiments were

conducted exclusively on open-source models (e.g., Gemma, mBERT). The high computational and financial costs associated with fine-tuning closed-source proprietary models hinder direct performance comparisons with such models.

References

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-training bert on arabic tweets: Practical considerations. *arXiv preprint arXiv:2102.10684*.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020. Arbert & marbert: Deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Moustafa Al-Hajj and Mustafa Jarrar. 2021. ArabGloss-BERT: Fine-tuning BERT on context-gloss pairs for WSD. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 35–43, Held Online. INCOMA Ltd.

Arthur Mensch Chris Bamford Devendra Singh Chaplot others Albert Q. Jiang, Alexandre Sablayrolles. 2023. Mistral 7b. *https://arxiv.org/abs/2310.06825*.

Waad Alshammari, Amal Almazrua, Asma Al Wazrah, Rawan Almatham, Muneera Alhoshan, and Abdulrahman Alosaimy. 2024. KSAA-CAD shared task: Contemporary Arabic dictionary for reverse dictionary and word sense disambiguation. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 677–685, Bangkok, Thailand. Association for Computational Linguistics.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Jan A. Botha, Zifei Shan, and Daniel Gillick. 2020. Entity Linking in 100 Languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7833–7845, Online. Association for Computational Linguistics.

Kenneth Ward Church. 2017. Word2vec. *Natural Language Engineering*, 23(1):155–162.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mohammed El-Razzaz, Mohamed Waleed Fakhr, and Fahima A Maghraby. 2021. Arabic gloss wsd using bert. *Applied Sciences*, 11(6):2567.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3509–3514, Hong Kong, China. Association for Computational Linguistics.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in arabic pre-trained language models. *arXiv preprint arXiv:2103.06678*.

Mohammed Khalilia, Sanad Malaysha, Reem Suwaileh, Mustafa Jarrar, Alaa Aljabari, Tamer Elsayed, and Imed Zitouni. 2024. Arabicnlu 2024: The first arabic natural language understanding shared task. *arXiv preprint arXiv:2407.20663*.

Wenqiang Lei, Xuancong Wang, Meichun Liu, Ilija Ilievski, Xiangnan He, and Min-Yen Kan. 2017. Swim: A simple word interaction model for implicit discourse relation recognition. In *IJCAI*, pages 4026–4032.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Ahmed Mukhtar Omar. 2008. Dictionary of contemporary arabic language.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Xiao Pu, Nikolaos Pappas, James Henderson, and Andrei Popescu-Belis. 2018. Integrating weakly supervised word sense disambiguation into neural machine translation. *Transactions of the Association for Computational Linguistics*, 6:635–649.

Nazreena Rahman and Bhogeswar Borah. 2021. A method to disambiguate a word by using restricted boltzmann machine. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 388–398, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Qwen Team. 2024. Qwen2 technical report. *arXiv preprint arXiv:2412.15115*.

Tasneem Wael, Eman Elrefai, Mohamed Makram, Sahar Selim, and Ghada Khoriba. 2024. Pirates at arabicnlu2024: Enhancing arabic word sense disambiguation using transformer-based approaches. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 372–376.

Boon Peng Yap, Andrew Koh, and Eng Siong Chng. 2020. Adapting BERT for word sense disambiguation with gloss selection objective and example sentences. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 41–46, Online. Association for Computational Linguistics.

Enhancing Urdu ASR with Whisper v3: Fine-Tuning on Latest Datasets and Realistic Multi-Speaker Evaluation with SLM Post-Processing

Zehra Ahmed Farah Inayat Zuha Aqib Sajjad Haider
Department of Computer Science
Institute of Business Administration (IBA), Karachi, Pakistan
{z.ahmed.26965, f.inayat.26912, z.aqib.26106}@khi.iba.edu.pk
sahaider@iba.edu.pk

Abstract

Automatic Speech Recognition (ASR) for low-resource languages like Urdu remains challenging due to limited training data, dialectal variation, code-switching, and orthographic inconsistencies. This work evaluates OpenAI Whisper large-v3-Turbo for Urdu ASR and performs Low Rank Adaptation (LoRA) fine-tuning on Common Voice v23, FLEURS and CSaLT datasets, drawn from a pooled corpus of 31 hours of Urdu speech. To gauge robustness in the real world, we also created an evaluation set from YouTube that includes news segments with two people, namely an anchor and an on-site reporter, with natural accents, interruptions, and background noise. Apart from this, the paper explored Urdu-capable Small Language Models (SLMs) ranging from 3.5B to 14B parameters for generative error correction (GEC). Whisper-Turbo achieved 25.33% WER on the evaluation set from YouTube, whereas Whisper-Turbo with 4-bit quantized Gemma3-12B as the SLM GEC attained 21.75% WER, which translates to a 14.14% relative reduction. The results indicate that there is a potential application of SLM post-processing for long-form Urdu transcription, at the cost of extra runtime.

1 Introduction

Although ASR has significantly advanced in recent years, progress on low-resource languages remains limited (Ahlawat et al., 2025). One such language is Urdu, which is widely used in Pakistan and across South Asia, but the problem of reliable Urdu ASR is challenging because of the lack of annotated speech data, dialectal differences, code switching, and orthographic irregularities. Urdu is also affected by regional languages like Punjabi, Pashto and Sindhi native speakers, giving rise to accented and variable pronunciations among many people who speak Urdu (Afsheen et al., 2014). While this cross-linguistic influence is not explicitly annotated in our training or evaluation data, it motivates the need for ASR systems that are robust to accent variation in real-world Urdu speech. Additionally, everyday Urdu speech is also often interspersed with Urdu and English, and written Urdu does not have standardized spelling and spacing conventions, which makes it harder to evaluate and transcribe.

The recent studies have provided important baselines for Urdu and other South Asian languages. (Arif et al., 2025b) evaluated Urdu ASR models like Whisper, MMS, and Seamless-M4T, with the issues of orthographic variation, code-switching, and the inadequacy of WER as the sole evaluation metric. (Sehar et al., 2025) investigated Whisper on three languages: Pashto, Punjabi and Urdu, demonstrating that fine-tuning can enhance the performance of low-resource ASR models, but the dialectal and script-related issues still pose a challenge. Existing evaluations, however, are mostly conducted on short and cleaner texts, single speaker or controlled speech, and not on real-world Urdu audio with interruptions, background noise, accented speech, and rapid alternation of speakers.

This distinction is particularly relevant for media and communication settings, where Urdu speech often appears in long-form, conversational, and multi-speaker formats. While post-processing methods have improved transcript readability in a few low-resource contexts (Znotins et al., 2025), their use for Urdu ASR remains underexplored, particularly for correcting spelling and orthographic variation in long-form conversational speech. Recently, Large Language Models (LLMs) have shown strong performance as ASR post-processors through Generative Error Correction (hereafter GEC) [(Ma et al., 2023) , (Yamashita et al., 2025)]. This approach suits Urdu well because its Perso-Arabic script and complex morphology often cause homophone errors and orthographic ambiguities. However, prior work mostly relies on proprietary models such as

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 408–417
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_042

ChatGPT (Ma et al., 2023), which introduce API costs, latency, and privacy concerns that hinder offline or budget-constrained deployment. The emergence of capable open-weight Small Language Models (SLMs) presents an opportunity to perform GEC with lower computational overhead and without dependence on external APIs, offering a practical, locally-deployable alternative. This motivates our focus on evaluating open-weight SLMs rather than proprietary large-scale models for Urdu GEC.

To address these gaps, we evaluate OpenAI Whisper large-v3-Turbo (hereafter, Whisper-Turbo) (Radford et al., 2023) for Urdu ASR, applying Low Rank Adaptation (LoRA) fine-tuning across Common Voice v23 (Ardila et al., 2020), FLEURS (Conneau et al., 2023), and CSaLT (Arif et al., 2025a). To assess whether this performance generalizes to unscripted speech, we introduce a YouTube-based multi-speaker evaluation dataset featuring unconstrained conversational news broadcasts, directly addressing the lack of real-world benchmarks for Urdu. Finally, we explore SLM-based Generative Error Correction (GEC) to reduce residual transcription errors while evaluating the latency trade-offs of post-processing.

2 Design of Experiments

2.1 Research Questions

The conducted experiments address two primary research questions. First, we investigate whether Low-Rank Adaptation (LoRA) fine-tuning on curated Urdu datasets improves transcription performance on real-world, multi-speaker conversational audio. Second, we examine whether SLM-based GEC further reduces errors in long-form Urdu transcripts.

2.2 Experimental Setup

All experiments were conducted on an NVIDIA RTX 4080 GPU with 16 GiB VRAM. Models were implemented using PyTorch and Hugging Face Transformers with mixed-precision training. Fine-tuning used LoRA with AdamW optimization. To ensure fair comparison, we fixed random seeds across Python, NumPy, and PyTorch, and used identical train/validation/test split logic, audio preprocessing, text normalization, evaluation datasets, and metric calculation scripts across experiments.

2.3 Models

This paper focuses on OpenAI Whisper large-v3-Turbo (Whisper-Turbo) (Radford et al., 2023) for Urdu ASR. Prior Urdu ASR benchmarking has shown Whisper-large variants to be competitive on conversational Urdu (Arif et al., 2025b). Whisper-Turbo offers this same competitive performance with greater efficiency than the full Whisper large-v3 model, which is why we selected it. We initially evaluated Whisper large-v3 alongside Whisper-Turbo to confirm this choice; however, Whisper large-v3's zero-shot and fine-tuned WER trends closely matched those already reported for Whisper-large variants in prior work (Arif et al., 2025b), offering no additional insight beyond what Whisper-Turbo captured. We therefore report results for Whisper-Turbo only throughout this paper. Whisper-Turbo was evaluated in both zero-shot and LoRA fine-tuned settings. We publicly release our fine-tuned LoRA adapter on Hugging Face to support reproducibility.[1]

On the YouTube evaluation set, we additionally evaluated SLM-based GEC. In this context, we evaluated four open-weight SLMs with multilingual or Urdu-specific capabilities, using a mix of quantized and full-precision models to enable fair comparison on the same machine: (a) Gemma3-12B (4-bit quantized) (Gemma Team, 2025), (b) Qwen3-14B (4-bit quantized) (Qwen Team, 2025), (c) Tiny-Aya-Fire (Cohere Labs, 2024), and (d) Qwen3-4B (Qwen Team, 2025). Due to the hardware constraints of the machine described in Section 2.2, quantized versions of the larger models were used to enable local inference. We additionally evaluated Gemma3-12B at 8-bit quantization; however, as its performance differed only marginally from the 4-bit version, we retain and report results for Gemma3-12B (4-bit) only, to allow for a more consistent comparison across models. These SLMs were employed strictly as text-in-text-out correctors; they receive ASR transcripts as input without access to the original audio or gold references. To rigorously isolate the contribution of post-processing, we applied SLM correction to the outputs of both Whisper-Turbo and Whisper-Turbo Fine-Tuned.

2.4 Datasets

We used three publicly available Urdu speech datasets for training and benchmark evaluation, as

[1] https://huggingface.co/awaaz-se-alfaaz/whisper-turbo-v3-urdu

Dataset	Source	Use	Duration (in mins.)	Speech Type
Common Voice v23	Public corpus	Train + Evaluation	1267	Short Read Speech
FLEURS	Public corpus	Train + Evaluation	515	Short Read Speech
CSaLT	Public corpus	Train + Evaluation	81	Urdu Speech Corpus
YouTube Evaluation Set	Public news clips	Evaluation only	40	Two-speaker conversation

Table 1: Datasets used in training and evaluation.

shown in Table 1 : Mozilla Common Voice v23 , FLEURS , and CSaLT. Together, these datasets provide diverse speakers and recording conditions, but most utterances are short, relatively clean, and single-speaker, making them different from real-world conversational Urdu.

To evaluate real-world performance under unconstrained conditions, we created a multi-speaker evaluation set comprising 30 independent YouTube news clips totaling approximately 40 minutes of audio. Prioritizing 30 distinct clips ensured coverage across diverse speaker pairs, topics, and acoustic environments. At an average duration of roughly 80 seconds, these clips are longer than standard benchmark utterances while remaining representative of short-form news broadcasts. Expanding the dataset with longer videos would have introduced different interaction patterns and altered the scope of the evaluation; we therefore treat this dataset as a controlled real-world stress test for multi-speaker Urdu ASR.

The clips feature interactive discussions between news anchors and on-site reporters, containing natural interruptions, regional accents, background noise, and extended contextual segments. This dataset was not used during fine-tuning, and no audio clips were directly redistributed due to copyright constraints. Instead, we release the corresponding YouTube Evaluation Set as video URLs, reference transcripts, and our best-performing pipeline outputs on Hugging Face to facilitate future research.[2]

Reference transcripts for the YouTube evaluation set were created through a human-in-the-loop annotation process. Audio was first extracted from each YouTube clip via an ffmpeg-based Python pipeline, and initial transcriptions were generated using Whisper large-v3. One author (a native Urdu speaker) then listened to the audio and corrected the machine-generated transcript. A second author (also a native Urdu speaker) independently reviewed each corrected transcript while listening to the same audio, verifying and further correcting any remaining errors. Both authors followed a shared correction guideline emphasizing verbatim transcription of spoken content over normalization. Spoken content (disfluencies, code-switched words, and speaker turns) was retained in the final references, and no forced alignment was performed.

2.5 Dataset Mixing and Splits

The three public datasets were standardized into a common metadata format, including audio paths, normalized transcripts, a speaker information field, and duration information. Speaker information was included in the metadata schema but was not systematically verified or used in this work. Each dataset was independently shuffled and split into train, validation, and test subsets using an 80/10/10 ratio with a fixed random seed. The resulting train, validation, and test subsets were then pooled across datasets to form the final training, validation, and test sets used for fine-tuning and evaluation. This produced a pooled training set of 16,491 samples, comprising 13,975 from Common Voice (84.8%), 2,140 from FLEURS (13.0%), and 376 from CSaLT (2.3%). No additional balancing or weighting was applied across datasets; the pooled training set therefore reflects each dataset's natural size, with Common Voice dominating in absolute sample count.

2.6 Text and Audio Pre-processing

Experiments were run using the same pipeline for all transcripts. This encompassed Unicode Urdu normalization, removal of punctuation/symbols, standardization of Arabic/Urdu characters into various forms, number normalization, and cleanup of whitespace. Lexical content was not normalized in any other way and stemming was not applied, since this would have been detrimental to the evaluation of the ASR. All audio files were converted to mono 16 kHz, 16-bit PCM format. Leading and trailing silence were removed, and natural pauses within speech were retained. The extraction of the log-Mel spectrograms was performed as per Whisper standard.

[2] https://huggingface.co/datasets/awaaz-se-alfaaz/YouTube-Evaluation-Set

Prompt
You are an Urdu ASR error correction expert. Your ONLY task is to replace incorrectly transcribed Urdu words with their correct forms. CRITICAL RULES: • ABSOLUTELY NO punctuation (no ! . , : ; ” ’ or any symbols) • NO new words or phrases • NO reordering • NO grammar changes • ONLY replace wrong words with correct ones • If unsure about a word, leave it unchanged Think of this as a word-by-word dictionary replacement, not a rewrite. Fix ONLY the incorrectly transcribed words in this Urdu text. Replace wrong words with correct ones based on context. Add NO punctuation. Urdu text: {text} Corrected text:

Table 2: SLM Prompt

2.7 Fine-Tuning Setup

Fine-tuning was done with AdamW with a learning rate in the range of 1e-4 to 1e-6, a batch size of 8, and epochs varied between 8 and 12. To improve the efficiency of training, the algorithm used mixed-precision training, the maximum label length was set to 128–256 tokens, and a learning-rate scheduler was employed to prevent overfitting. For LoRA, rank 32, alpha 64, and dropout 0.05 were applied to the query, key, value, and output projection layers. We did not do a full hyperparameter sweep because of compute limitations; instead, a few stability runs were performed with varying learning rate, number of epochs, and LoRA rank that were not overfitting and were not diverging on validation WER.

2.8 Benchmark Evaluation

The data used for benchmark evaluation was the pooled dataset described in 2.4: each of the three datasets (Common Voice v23, FLEURS, CSaLT) was independently split into train, validation, and test subsets using an 80/10/10 ratio, and the resulting subsets were pooled across datasets. The held-out test set was used for evaluation, with the same preprocessing, normalization, and scoring applied throughout. Results are reported for Turbo in both zero-shot and LoRA fine-tuned settings, measured using Word Error Rate (WER).

In general, we ran 30+ fine-tuning experiments with different learning rates, epoch sizes, and LoRA parameters, and selected the configuration with the best WER on the pooled validation set across all three datasets.

2.9 SLM-Based Generative Error Correction

When it comes to long-form YouTube transcriptions, we looked at Urdu-capable SLMs as text-to-text post-processing models. The SLMs were only provided with transcripts generated by the ASR and not the original audio or gold reference transcripts. This correction step was tested for both Whisper-Turbo and Whisper-Turbo Fine-Tuned, to study the effect of SLM post-processing.

All SLMs were prompted with the same conservative correction instruction: replace only incorrectly transcribed Urdu words, avoid punctuation restoration, do not add or reorder words, and preserve the original spoken style. This constrained the models to minimal ASR-error correction rather than fluent rewriting. Because the prompt restricts the SLM to replacing individual words without insertion, deletion, or reordering, this correction step is inherently limited to fixing substitution errors; insertion and deletion errors introduced by the ASR system fall outside its scope. The corrected outputs were then passed through the same normalization and jiwer-based[3] scoring pipeline used for all other experiments.

2.10 Evaluation Protocol and Metrics

All ASR and post-processing outputs were evaluated using the same normalization and scoring pipeline described in Section 2.6. The SLM correction stage was guided by a single fixed prompt, arrived at through iterative experimentation on the 30-clip YouTube evaluation set. Prompt variants were independently trialed by multiple team members across the different SLMs, with each variant assessed on its resulting WER; candidate prompts were then discussed collectively, and a single final prompt was adopted for use across all three SLMs. Each SLM was queried using its provider-recommended prompting/chat-template configura-

[3] https://github.com/jitsi/jiwer

Model / Dataset	CSaLT	Common Voice23	Fleurs
Whisper-Turbo	27.51	37.23	24.07
Whisper-Turbo Fine-Tuned	20.77	25.78	16.86

Table 3: Benchmark WER (%) for Whisper-Turbo using fine-tuning.

tion, ensuring the correction prompt was presented to every model in its intended input format.

The final prompt, shown in Table 2, was selected as it most reliably elicited minimal character- and word-level corrections without rephrasing, reordering, stylistic rewriting, or punctuation restoration. SLM correction was evaluated on both Whisper-Turbo and Whisper-Turbo Fine-Tuned outputs to separate the effect of post-processing from the effect of fine-tuning. Performance was measured using the jiwer library. WER was used as the primary metric across benchmark and YouTube evaluations.

3 Results and Findings

This section reports WER results for Whisper Turbo under zero-shot, LoRA fine-tuned settings, and SLM-based post-processing results.

3.1 Benchmark Evaluations

Table 3 summarizes WER on the three benchmark corpora for Whisper-Turbo, before and after LoRA fine-tuning. Fine-tuning yields a consistent reduction in WER across all three datasets: from 27.51% to 20.77% on CSaLT (a 24.5% relative reduction), from 37.23% to 25.78% on Common Voice v23 (a 30.8% relative reduction), and from 24.07% to 16.86% on FLEURS (a 29.9% relative reduction). In preliminary experiments using simple proportional mixing rather than stratified per-dataset splitting, we observed less consistent gains across the three corpora. Adopting stratified splitting, in which each dataset is independently divided into train, validation, and test subsets before pooling (Section 2.4), produced the uniform improvement reported above and was retained as our final experimental design.

We attribute this improvement in part to the acoustic diversity of Common Voice v23, which, despite being the noisiest of the three corpora, constitutes the majority of the pooled training set (84.8%, Section 2.4) under stratified sampling. Rather than degrading fine-tuning quality, this diversity appears to have improved the model's robustness to acoustic variation, contributing to the largest relative gain observed on Common Voice v23 itself (30.8%) while also supporting strong transfer to FLEURS (29.9%) and CSaLT (24.5%), which contributed far smaller shares of the training data (13.0% and 2.3%, respectively).

Taken together, these results indicate that LoRA fine-tuning meaningfully improves Whisper-Turbo's Urdu transcription accuracy across benchmark conditions of varying acoustic diversity and dataset scale, with exposure to Common Voice's noisier, more variable speech plausibly contributing to the model's improved generalization across all three corpora.

Model	WER (%)
Whisper-Turbo	25.33
Whisper-Turbo Fine-Tuned	24.20

Table 4: WER (%) on YouTube evaluation set.

3.2 Real-World Multi-Speaker Evaluation

To assess performance under noisy, unconstrained conditions, we evaluated our models on the curated YouTube news dataset (Section 2.4). As expected for real-world conversational audio, per-clip performance exhibits greater variability than on the standard benchmark corpora. Results fluctuate noticeably across clips, particularly those featuring strong regional accents, outdoor background noise, or colloquial speech.

Table 4 summarizes the evaluation results on this out-of-domain set. Whisper-Turbo achieves a baseline WER of 25.33%, whereas fine-tuning reduces WER to 24.20% (a relative reduction of 4.46%). However, as confirmed by our statistical significance analysis (Table 6), this 1.13% absolute gain is not statistically significant ($p = 0.8870$, 95% CI $[-4.38\%, +1.40\%]$). This confirms that acoustic fine-tuning alone fails to yield reliable improvements in unscripted, multi-speaker environments—directly motivating the need for higher-level semantic post-processing.

This relative improvement is considerably smaller than the 24.5–30.8% relative reductions observed on the benchmark corpora in Section 3.1.

SLM Model	Whisper-Turbo	Whisper-Turbo Fine-Tuned
4-bit Quantized Models		
Gemma3-12B	21.93	21.75
Qwen3-14B	23.58	22.39
Full-Precision Models		
Tiny-Aya-Fire	28.81	24.64
Qwen3-4B	30.57	29.41

Table 5: SLM post-processing Word Error Rates (WER %) across baseline and fine-tuned pipelines on the YouTube evaluation set.

	Wilcoxon Signed-Rank Test			Paired Bootstrap Resampling ($B = 10,000$)		
Comparison	**Stat (W)**	**p-value**	**Sig.**	**Mean ΔWER**	**95% CI**	**Sig.**
Turbo FT vs. Turbo	225.0	0.8870	n.s.	-1.00%	$[-4.38\%, +1.40\%]$	n.s.
Turbo+Gemma3-12B vs. Turbo	5.0	1.86×10^{-8}	***	-3.30%	$[-4.26\%, -2.41\%]$	***
Turbo FT+Gemma3-12B vs. Turbo FT	32.0	5.14×10^{-6}	***	-2.45%	$[-3.37\%, -1.44\%]$	***
Turbo FT+Qwen3-14B vs. Turbo FT	33.0	6.62×10^{-5}	***	-1.81%	$[-2.46\%, -1.12\%]$	***
Turbo FT+Qwen3-14B vs. Turbo FT+Gemma3-12B	105.0	0.0440	*	$+0.64\%$	$[+0.05\%, +1.25\%]$	*

Table 6: Statistical significance analysis across YouTube evaluation set ($N = 30$ clips). Significance levels: * $p < 0.05$, *** $p < 0.001$, n.s. = not significant.

Given the acoustic and structural differences between the fine-tuning data and the YouTube set outlined in Section 2.4, this gap suggests that fine-tuning primarily adapts the model to domain-specific training distribution rather than conferring general robustness to unscripted conversational audio. Consequently, benchmark gains do not transfer directly to noisy, real-world conditions, underscoring the necessity of evaluation sets like ours when assessing deployment readiness for Urdu ASR systems.

3.3 SLM-Based Post-Processing

To address these residual errors, we evaluate SLMs as text-to-text post-processors. Because long-form YouTube clips provide richer cross-sentence context than short benchmark utterances, they are well-suited for post-ASR error recovery while preserving the original spoken style. The SLMs were prompted conservatively to avoid adding punctuation, reordering words, rewriting sentences, or converting conversational Urdu into overly formal text.

The central challenge lies in balancing error correction with verbatim preservation. Overly aggressive correction risks normalizing colloquialisms, regional expressions, or code-switched terms into formal Urdu, ultimately worsening WER. This trade-off is particularly pronounced in clips featuring strong regional accents, informal vocabulary, or non-standard word spacing.

SLM post-processing yields mixed results depending on the corrector used, but significantly outperforms ASR fine-tuning alone (Table 6). Applying Gemma3-12B (4-bit) reduces WER to 21.93% on Whisper-Turbo (a 13.4% relative reduction; $p = 1.86 \times 10^{-8}$) and to 21.75% on Whisper-Turbo Fine-Tuned (a 10.1% relative reduction; $p = 5.14 \times 10^{-6}$), making it the strongest correction model overall. Qwen3-14B (4-bit) similarly yields statistically significant improvements across both pipelines, reducing WER to 23.58% on Whisper-Turbo (a 6.9% relative reduction) and to 22.39% on Whisper-Turbo Fine-Tuned (a 7.5% relative reduction; $p = 6.62 \times 10^{-5}$). Directly comparing the two top correctors confirms that Gemma3-12B provides a statistically significant advantage over Qwen3-14B ($p = 0.0440$).

In contrast, the two smaller full-precision models degraded performance. Tiny-Aya-Fire increases WER to 28.81% on Whisper-Turbo (a 13.7% relative increase) and to 24.64% on Whisper-Turbo Fine-Tuned (a 1.8% relative increase), while Qwen3-4B increases WER to 30.57% on Whisper-Turbo (20.7% relative increase) and to 29.41% on Whisper-Turbo Fine-Tuned (21.5% relative increase). Notably, both full-precision models degrade the baseline pipeline more severely than the fine-tuned pipeline, suggesting that a stronger starting transcript is somewhat more resilient to unhelpful corrections than a noisier one.

Model / Stage	Transcription Text (WER)
Raw ASR	سابق کپتان رمیز راجہ کراچی کنگز کے نوجوان کھلاڑیوں سے متاثر کہتے ہیں کہ کراچی کنگز کے ٹیم میں باسلاحیت نوجوان کھلاڑی شامل ہیں قاسم اکرم، ارشد اقبال، ابباس آفریڈی اور محمد حارس لمبے عرصے تک کراچی کنگز کے لیے کھیل سکتے ہیں کنگز کے پاس ینگ پلیز ہیں جو آپ کی توجہ کا مرکز بن سکتے ہیں آگے چل قاسم اکرم، گوڈ لوکنگ بیٹسمن ہے لیک سائٹ پر جو ان کی بیٹنگ ہے وہ ویرات کولی کی آپ کو یاد آجاتی ہے جیسے یہ فلک کرتے ہیں بال کو کنگز کے پاس دوینگ اچھے فاس بولرز بھی ہیں عباس آفریڈی اور عرشد اقبال لمبے کٹ کے ہیں سپیڈ بڑی اچھی یا ہارڈ لینڈ کو ہٹ کرتے ہیں کسی بھی موقع پہ آکے یہ آپ کو بولنگ کر سکتے ہیں اور عباس آفریڈی نئے بال پہ یورکر کروا سکتے ہیں بہت اچھے دو یینگ فاس بولرز ہیں جو کہ آگے کراچی کنگز کو کئی سال تک یہ سروس پروائیڈ کر سکتے ہیں کراچی کنگز کی یہ ایک اور اچھی ینگ پک ہے ان کا نام ہے محمد حارس بہت اچھی پوٹینچل ہے یہ بھی ٹیمپرومنٹلی سٹرونگ ہے اور کالٹیز بے انتہاں ہیں موسیقی
GEC: Gemma3-12B (4-bit)	سابق کپتان رمیز راجہ کراچی کنگز کے نوجوان کھلاڑیوں سے متاثر کہتے ہیں کہ کراچی کنگز کے ٹیم میں باصلاحیت نوجوان کھلاڑی شامل ہیں قاسم اکرم عرشد اقبال عباس آفریدی اور محمد حارث لمبے عرصے تک کراچی کنگز کے لئے کھیل سکتے ہیں کنگز کے پاس یہ ینگ پلیز ہیں جو آپ کی توجہوں کا مرکز بن سکتے ہیں آگئے چل قاسم اکرم گوڈ لکنگ بیٹسمن ہے لیک سائیڈ پہ جو ان کی بیٹنگ ہے وہ ویراٹ کولی کی آپ کو یاد آجاتی ہے جیسے یہ فلک کرتے ہیں بال کو کنگز کے پاس دوینگ اچھے فاس بولرز بھی ہیں عباس آفریدی اور عرشد اقبال لمبے کٹ کے ہیں سپیڈ بڑی اچھی یا ہارڈ لینکھ کو ہٹ کرتے ہیں کسی بھی موقع پہ آکے یہ آپ کو بولنگ کر سکتے ہیں اور عباس آفریدی نئے بال پہ یورکر کروا سکتے ہیں بہت اچھے دو یینگ فاس بولرز ہیں جو کہ آگے کراچی کنگز کو کئی سال تک یہ سروس پروائیڈ کر سکتے ہیں کراچی کنگز کی یہ ایک اور اچھی ینگ پک ہے ان کا نام ہے محمد حارث بہت اچھی پوٹینچل ہے یہ بھی ٹیمپرومنٹلی سٹرانگ ہے اور کالیٹیز بے انتہاں ہیں

Table 7: Qualitative comparison of Raw ASR vs. Post-SLM corrected Urdu transcripts. Highlighted text indicates residual recognition, orthographic, or spacing errors.

Comparing correction gains across the Whisper-Turbo baseline and fine-tuned pipelines also suggests differing relationships between SLM correction and fine-tuning. For Gemma3-12B, the relative gain from correction is larger on the Whisper-Turbo baseline (13.4%) than on the already fine-tuned pipeline (10.1%), consistent with SLM correction and fine-tuning addressing at least partially overlapping error patterns. For Qwen3-14B, the relative gain is similar across both pipelines (6.9% vs. 7.5%), suggesting a more complementary relationship between correction and fine-tuning for this model. The best overall WER, 21.75%, is achieved by combining fine-tuning with Gemma3-12B correction.

A qualitative examination of the post-processed outputs reveals distinct failure modes that explain the degradation observed in smaller models. Despite strict prompts prohibiting formatting changes, Tiny-Aya-Fire consistently inserted unprompted Urdu punctuation (full stops '۔' and commas), which artificially inflated edit distances against unpunctuated reference transcripts. Qwen3-4B exhibited even more severe instruction-following breakdowns: on certain clips, it entered repetitive generation loops (reproducing the transcript up to three times consecutively), while on another, it translated the Urdu transcript into English. Because these

Error Source	Observed Effect
Regional accents and colloquial pronunciation	Increased substitutions and non-standard lexical forms
Outdoor noise and rapid speaker alternation	Missed or confused words during turn changes
Urdu orthographic variation	Semantically correct outputs were penalized due to spacing or alternate spellings
SLM over-correction	Conversational Urdu sometimes normalized into overly formal wording

Table 8: Error Analysis.

catastrophic failure modes yielded clip-level WERs approaching 100%, they disproportionately inflated Qwen3-4B's average error rate. In contrast, Table 7 demonstrates a successful case where Gemma3-12B (4-bit) corrects several recognition and orthographic errors while preserving the underlying spoken content.

3.4 Error Analysis

A closer examination of the YouTube evaluation showed that a few common conditions repeatedly resulted in high WER.

Table 8 summarizes the main error sources observed in the YouTube evaluation set and their effect on transcription quality. WER increased with stronger regional accents, outdoor or noisy recording conditions, faster speaker alternation, and more colloquial speech. These conditions are not well represented in the benchmark datasets, which tend to be shorter, cleaner, and structurally simpler.

A further source of measured error was Urdu's orthographic variation itself, rather than genuine recognition failure: outputs that were semantically correct were still penalized by WER when they differed from the reference in spacing conventions, compound-word segmentation, or alternate valid spellings. SLM-based cleanup was more effective on clips that were more lexically dense and contextually continuous, and less effective on clips that were sparser or more fragmented.

GEC Model	Per Min (s)
Tiny-Aya-Fire	5.5
Qwen3-4B	10.5
Gemma3-12B (4-bit)	7.5
Qwen3-14B (4-bit)	12.5

Table 9: Inference latency for the SLM GEC stage.

3.5 Deployment Considerations and Runtime Trade-offs

Beyond transcription accuracy, practical deployment of Urdu ASR systems depends on processing speed. We therefore report latency, measured in seconds per minute of audio processed, for each pipeline configuration alongside the corresponding WER results from Section 3.3, to characterize the trade-off between correction quality and computational cost.

ASR inference alone is fast, with fine-tuning introducing no latency overhead while improving accuracy. However, adding SLM-based correction increases processing time in exchange for varying accuracy gains. Table 9 details the latency added by each GEC model in seconds per minute of audio processed.

Among the evaluated correctors, Gemma3-12B (4-bit) offers the strongest performance-to-latency trade-off: as shown in Table 9, it requires only 7.5 seconds per minute of audio to achieve the best overall WER of 21.75%. Notably, despite having three times as many parameters, Gemma3-12B (4-bit) processes audio faster than Qwen3-4B (7.5s vs. 10.5s per minute) while yielding significantly better accuracy gains. Qwen3-14B requires the highest runtime overhead at 12.5 seconds per minute of audio, whereas Tiny-Aya-Fire is the fastest corrector at 5.5 seconds per minute, though at the expense of severe accuracy degradation.

As noted in Section 2.3, we additionally tested Gemma3-12B at 8-bit quantization; while WER improved marginally (21.68% vs. 21.75% at 4-bit), inference latency increased substantially to nearly five times that of the 4-bit variant due to hardware limitations. This confirms 4-bit quantization as the more practical configuration for real-time and resource-constrained deployment.

4 Conclusion

This work establishes updated baselines for Urdu ASR using Whisper-Turbo, evaluated across curated benchmark corpora and a real-world YouTube

evaluation set. LoRA fine-tuning yields consistent WER reductions across all three benchmark datasets (24.5–30.8% relative), driven largely by Common Voice v23's acoustic diversity, but this gain transfers less strongly to real-world YouTube audio (4.46% relative), underscoring a persistent benchmark-to-deployment gap.

SLM-based GEC offers the largest further improvement, but its benefit is highly model-dependent: Gemma3-12B (4-bit) achieves the best overall result (21.75% WER, a 10.1% relative reduction over fine-tuned Whisper-Turbo), whereas Tiny-Aya-Fire and Qwen3-4B degrade performance. Overall, combining targeted fine-tuning with a carefully selected, lightweight SLM corrector offers a practical path toward more robust Urdu ASR for real-world, multi-speaker audio. To support reproducible research and benchmark reuse, we make our fine-tuned Whisper-Turbo checkpoint and the 30-clip YouTube evaluation corpus publicly available on Hugging Face.

In future work, we plan to explore decoder-level hyperparameter optimization (such as beam search, N-best reranking, and temperature scaling) alongside fine-tuning and SLM correction, as our preliminary tests showed promising WER reductions on FLEURS without modifying model weights. Additionally, future efforts will focus on disentangling model scale from quantization effects, developing domain-adapted correction models to mitigate over-formalization of colloquial speech, and expanding the YouTube evaluation set for community reuse.

5 Ethics

All the training sets used in this work were obtained from public resources: Mozilla Common Voice v23, FLEURS and CSaLT, all of which are publicly available, and used within their respective terms. YouTube Clips were not used to train, redistribute, or release, respecting copyright and platform policies. No personally identifiable or private data was collected.

The only post-processing done in the SLM system was minimal correction, but language models might have favored standard Urdu over other Urdu variants. We thus stress the need for careful deployment of Urdu ASR systems, taking into consideration privacy, fairness, and dialects.

Acknowledgments

The authors would like to thank the anonymous reviewers for their detailed feedback, which significantly improved the quality of this paper.

References

Afsheen, Saad Irtza, Mahwish Farooq, and Sarmad Hussain. 2014. Accent classification among punjabi, urdu, pashto, saraiki and sindhi accents of urdu language. In *Proceedings of the Conference on Language and Technology (CLT 2014)*. Center for Language Engineering (CLE), UET Lahore.

Harsh Ahlawat, Naveen Aggarwal, and Deepti Gupta. 2025. Automatic speech recognition: A survey of deep learning techniques and approaches. *International Journal of Cognitive Computing in Engineering*, 6:201–237.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 4211–4215, Marseille, France. European Language Resources Association (ELRA).

Samee Arif, Aamina Jamal Khan, Mustafa Abbas, Agha Ali Raza, and Awais Athar. 2025a. CSaLT-voice. Hugging Face.

Samee Arif, Aamina Jamal Khan, Mustafa Abbas, Agha Ali Raza, and Awais Athar. 2025b. WER we stand: Benchmarking Urdu ASR models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5952–5961, Abu Dhabi, UAE. Association for Computational Linguistics.

Cohere Labs. 2024. Tiny aya fire. Hugging Face.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. FLEURS: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Rao Ma, Mengjie Qian, Potsawee Manakul, Mark Gales, and Kate Knill. 2023. Can generative large language models perform ASR error correction? ArXiv preprint arXiv:2307.04172.

Qwen Team. 2025. Qwen3 technical report. ArXiv preprint arXiv:2505.09388.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International*

Conference on Machine Learning, volume 202, pages 28492–28518, Honolulu, Hawaii, USA. PMLR.

Najm Ul Sehar, Ayesha Khalid, Farah Adeeba, and Sarmad Hussain. 2025. Benchmarking Whisper for low-resource speech recognition: An n-shot evaluation on Pashto, Punjabi, and Urdu. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 202–207, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Natsuo Yamashita, Masaaki Yamamoto, Hiroaki Kokubo, and Yohei Kawaguchi. 2025. LLM-based generative error correction for rare words with synthetic data and phonetic context. In *Proceedings of Interspeech 2025*, pages 3653–3657, Rotterdam, The Netherlands. International Speech Communication Association (ISCA).

Arturs Znotins, Normunds Gruzitis, and Roberts Dargis. 2025. From conversational speech to readable text: Post-processing noisy transcripts in a low-resource setting. In *Proceedings of the Tenth Workshop on Noisy and User-generated Text*, pages 143–148, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Baseer: A Vision-Language Model for Arabic Document-to-Markdown OCR

Khalil Hennara
Misraj AI
Khobar, Saudi Arabia
hennara@misraj.ai

Muhammad Hreden
Misraj AI
Khobar, Saudi Arabia
hreden@misraj.ai

Mohamed Motasim Hamed
Misraj AI
Khobar, Saudi Arabia
hamed@misraj.ai

Ahmad Bastati
Misraj AI
Khobar, Saudi Arabia
bastati@misraj.ai

Zeina Aldallal
Misraj AI
Khobar, Saudi Arabia
aldallal@misraj.ai

Sara Chrouf
Misraj AI
Khobar, Saudi Arabia
chrouf@misraj.ai

Safwan AlModhayan
Misraj AI
Khobar, Saudi Arabia
safwan@misraj.ai

Abstract

Arabic Optical Character Recognition (OCR) remains a challenging task due to the language's cursive script, diverse fonts, diacritics, and right-to-left orientation. While modern Multimodal Large Language Models (MLLMs) have advanced document understanding for high-resource languages, their performance on Arabic remains limited. In this work, we introduce **Baseer**, a vision-language model fine-tuned specifically for Arabic document OCR. Leveraging a large-scale dataset combining synthetic and real-world documents, Baseer is trained using a decoder-only fine-tuning strategy to adapt a pre-trained MLLM while preserving general visual features. We also present Misraj-DocOCR, a high-quality, expert-verified benchmark designed for rigorous evaluation of Arabic OCR systems. Our experiments show that Baseer consistently outperforms existing open-source and commercial solutions on the benchmarks evaluated in this work, achieving state-of-the-art results for Arabic document OCR under these settings. Our results highlight the benefits of domain-specific adaptation of general-purpose MLLMs and establish a strong baseline for high-accuracy OCR on morphologically rich languages like Arabic.

1 Introduction

The recent and rapid advancements in Multimodal Large Language Models (MLLMs) have fundamentally reshaped the landscape of how machines perceive and process complex visual and textual data (Hurst et al., 2024; Comanici et al., 2025; Zhu et al., 2025; Bai et al., 2025). Among the myriad applications of these models, OCR and comprehensive document understanding continue to present significant challenges. This is particularly true for languages that are morphologically rich and structurally complex, such as Arabic. While contemporary OCR solutions have achieved remarkable performance for English and other high-resource languages (Comanici et al., 2025; Hurst et al., 2024), their efficacy does not readily generalize to Arabic documents. The inherent complexities of Arabic script, including its cursive nature, extensive ligature formation, the wide variety of fonts and styles, the critical role of diacritics, and the right-to-left text orientation, render Arabic OCR a task of considerable difficulty.

In parallel, progress in multimodal architectures has paved the way for unified vision-language reasoning, which enables models to concurrently extract both textual content and structural information from documents (Li et al., 2025; Mandal et al., 2025). Despite these technological strides, modern multimodal frameworks have seldom been specialized for the distinct demands of Arabic OCR and document parsing. This significant gap in research and development leaves academics, practitioners, and industries without robust, dedicated tools for processing real-world Arabic documents, which are prevalent across academic, commercial, and cultural heritage domains.

[*] **Baseer:** meaning *"one who sees clearly"* and *"insightful."* The name reflects the model's ability to "see" and interpret documents with clarity.

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 418–428
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_043

In this work, we introduce **Baseer**[†], a vision-language model fine-tuned for Arabic document OCR. Our approach adapts a powerful general-purpose MLLM to the unique challenges of Arabic document analysis. To facilitate this specialization, Baseer was trained on a large-scale, diverse dataset composed of both synthetically generated and authentic real-world Arabic documents. This dataset was curated to encompass the extensive variety of formats, fonts, and layouts encountered in practical applications. Furthermore, we present **Misraj-DocOCR**, a novel benchmark specifically engineered for the evaluation of Arabic OCR systems, featuring high-quality, expert-verified annotations to ensure reliability.

Our primary contributions are threefold:

1. We present the development and fine-tuning of **Baseer**, demonstrating that an efficient, decoder-only fine-tuning strategy can achieve state-of-the-art performance on the Arabic document OCR benchmarks evaluated in this work.

2. We introduce **Misraj-DocOCR**[*], a new, reliable, and openly available benchmark designed to provide a standardized and rigorous evaluation framework for Arabic OCR systems.

3. We conduct a thorough analysis of the **KITAB-pdf-to-markdown**[†] benchmark, providing a revised and improved version that addresses significant inaccuracies to enhance its accuracy and utility for the research community.

The remainder of this paper is organized as follows: Section 2 reviews related work, while Section 3 describes the dataset. Section 4 introduces our new benchmark, followed by Section 5, which details the proposed methodology. Finally, Section 6 outlines the experimental setup, and Section 7 presents the results and analysis.

2 Related Work

We situate our work at the intersection of two major research domains. First, we review the rapid advancements in Multimodal Large Language Models (MLLMs), which provide the architectural foundation for our approach. Second, we delve into the field of Optical Character Recognition (OCR) and Document Understanding, examining its evolution and highlighting the persistent challenges that motivate our research, particularly for morphologically complex languages like Arabic.

2.1 Multimodal Large Language Models

The paradigm of Large Language Models (LLMs) has recently been extended to handle multimodal input, leading to the development of powerful models capable of joint vision-language reasoning. Research in this area has generally progressed along two main architectural paths. One approach involves the modular integration of pre-trained components, where a specialized frozen vision encoder is connected to a large language decoder via a lightweight adapter. This design is seen in influential models like LLaVA (Liu et al., 2023, 2024; Li et al., 2024), Aya-Vision (Dash et al., 2025), Idefics (Laurençon et al., 2024b,a), and more compact architectures such as SmolVLM (Marafioti et al., 2025). These architectures achieve impressive zero-shot and few-shot performance on a diverse range of multimodal tasks with high parameter efficiency. The second approach focuses on training massive, end-to-end vision-language models. This category includes state-of-the-art systems such as InternVL (Chen et al., 2024; Zhu et al., 2025), Gemma (Team et al., 2025), PaliGemma (Steiner et al., 2024), and Qwen-VL (Bai et al., 2025). These models, with parameter counts scaling up to 70 billion in the case of Qwen2.5-VL, have demonstrated remarkable general-purpose capabilities. However, their broad, generalist training often leaves them unspecialized for precision-critical, niche domains. As we will discuss, high-fidelity document OCR represents one such domain where these powerful models still exhibit significant limitations.

2.2 OCR and Document Understanding

Leading efforts in this domain, such as Idefics3 (Laurençon et al., 2024a), Monkey-OCR (Li et al., 2025), SmolDocling (Nassar et al., 2025), and commercial systems such as Nanonets-OCR-s (Mandal et al., 2025), have established high benchmarks for performance on standard document types. More recent attempts, such as Qari (Wasfy et al., 2025), have tackled Arabic OCR directly, but their scope remains

[†] https://baseerocr.com/

[*] https://huggingface.co/datasets/Misraj/Misraj-DocOCR

[†] https://huggingface.co/datasets/Misraj/KITAB_pdf_to_markdown_reviewed

limited compared to comprehensive document understanding systems.

A small but growing body of work targets Arabic document OCR and the resources needed to train it. Arabic-Nougat (Rashad, 2024) adapts the Nougat architecture to convert Arabic book pages into Markdown and releases *arabic-img2md*, a synthetic image–Markdown corpus of roughly 13.7K pages; its focus, however, is on running text rather than the full range of structured elements (e.g., HTML tables) targeted here. SARD (RIOTU Lab, 2025) contributes a large-scale (∼843K pages) synthetic Arabic corpus that systematically varies fonts and book-style layouts, but by construction it is noise-free and omits real-world scans, embedded tables, and multi-column layouts. QARI-OCR (Wasfy et al., 2025) similarly releases synthetic Arabic OCR resources and a fine-tuned model. Baseer is complementary to and broader than these efforts along two axes: (i) its training mixture pairs synthetic data with a large set of *expert-verified real-world* pages and dedicated table-centric and non-text data, and (ii) it explicitly models structured output. We position *Misraj-DocOCR* against these resources accordingly: whereas KITAB-Bench (Heakl et al., 2025), arabic-img2md, and SARD are dominated by synthetic or automatically extracted ground truth, Misraj-DocOCR is a fully expert-verified benchmark that mixes real and synthetic pages and covers structured layouts, making it a stricter test of layout-aware transcription.

However, a critical challenge remains: the generalization of these models to languages with scripts that are fundamentally different from Latin-based languages. Arabic serves as a prime example of this challenge. Its inherent characteristics include cursive script, context-sensitive character shapes, optional but meaningful diacritics, right-to-left orientation, and a wide variety of fonts and styles. These characteristics often cause state-of-the-art document OCR systems to degrade sharply in performance when applied to Arabic texts.

To the best of our knowledge, the application of modern MLLM frameworks to the specific, challenging problem of Arabic document OCR remains a largely unexplored area. Although several multilingual and multimodal models include Arabic in their training, they are not optimized for the script-specific and structural challenges posed by Arabic documents. This work aims to bridge this critical gap by introducing *Baseer*, a system specifically engineered for the complexities of Arabic documents. Our results demonstrate that this specialized approach yields substantial improvements over general-purpose baselines on the benchmarks we evaluate.

3 Data

A primary motivation for our work is the severe scarcity of high-quality Arabic data for document OCR, a challenge that firmly positions Arabic as a low-resource language within this specific research domain. To overcome this critical bottleneck and address the shortcomings of existing resources, we dedicated substantial effort to sourcing, curating, and annotating an original, high-quality dataset tailored for complex document understanding.

To support effective document OCR, it is essential to represent textual content in a format that preserves both structure and semantics. In our dataset, the text corresponding to each image is formatted in Markdown, providing a clean and standardized representation of content. Tables are represented in HTML to accurately capture diverse table structures and complex layouts. Furthermore, specialized tags were introduced to mark specific elements within the text including watermarks, page numbers, and the presence of images enabling precise supervision for layout-aware OCR and document parsing tasks.

The dataset itself was constructed as a hybrid collection, combining a large corpus of synthetically generated documents with a carefully curated set of real-world publications, supplemented by targeted data covering tables, and non-text imagery. Each of these sources is described in detail below.

3.1 Synthetic Data

The first component of our dataset was generated synthetically using an in-house pipeline, designed to capture the diverse formatting and layout variations commonly found in word-processing documents. The foundation for this synthetic data is a corpus of markdown-formatted documents, which were downloaded and filtered from the Common Crawl archive. To ensure the quality and relevance of the source material, the raw data were subjected to the following preprocessing filters:

1. **Perplexity Filtering:** An in-house language model based on KenLM (Heafield, 2011) was employed to calculate perplexity scores, re-

taining only the most linguistically cohesive text samples.

2. **Table Sparsity Filtering:** To ensure structural integrity, documents containing markdown tables with more than 25% empty cells were identified and discarded.

The filtered markdown documents were then converted into image-text pairs. From the pool of generated images, 200,000 samples were randomly selected and underwent random transformations to consist an augmentation data. To prevent redundancy, the original, pre-augmentation versions of these images were discarded. In total, this synthetic pipeline produced **400,000** high-quality image-text pairs, comprising 200,000 clean rendered samples and 200,000 augmented variants designed to simulate diverse real-world document conditions.

3.2 Open-Source Books and Magazines

The second component of our dataset was sourced from real-world documents, including a diverse collection of books, magazines, educational documents, and academic papers. In contrast to synthetic data, these samples reflect authentic publishing environments, capturing genuine layout complexities and typographic conventions.

To ensure maximum diversity, the selected pages span a broad spectrum of layout complexities, identified using vision-based algorithms. Specifically, bounding boxes were first detected at the paragraph level, and their alignment and overlap were analyzed to capture challenging structures such as tables, figures, index pages, and skewed layouts. Page color distributions were also examined to include samples with embedded images, colorful backgrounds, and multi-colored text.

Ground-truth text for the real-world documents was obtained using a state-of-the-art vision–language model, *Gemini-2.5-pro*. To ensure the highest level of fidelity in our training data, we implemented a rigorous human-in-the-loop validation stage following the initial VLM extraction. For each sample, one of 20 trained annotators performed a side-by-side review of the original document image and the model-generated output. In this phase, the annotator's role was strictly binary: to either *accept* or *reject* the sample based on the presence of any errors in the extracted text or structural formatting.

Given our commitment to high-quality supervision, we maintained a strict threshold for accuracy; consequently, approximately 35% of the processed data was rejected due to inconsistencies or extraction errors. While these rejected samples were excluded from the current training cycle to maintain data purity, they represent a valuable resource for future work. In subsequent research, we plan to implement a corrective pipeline to refine these samples rather than discarding them, further expanding the dataset and iteratively improving the model's robustness. From this source, we curated **257,000** document images paired with their corresponding ground-truth text.

3.3 Table-Centric Data

To enhance tabular recognition and reconstruction, we developed a dedicated dataset of **111,000** samples using two complementary pipelines: First, *Template-Based Generation from Real Financial Documents*, Human-curated Jinja templates derived from real-world financial documents were populated with 20 unique LLM-generated instances each. A diversity-aware prompting strategy was employed to minimize redundancy. Second, *Translation and Augmentation of HTML Tables*, structured HTML tables were translated into Arabic using 12 stylistic prompts. These variations included full content translation, preservation of technical terms, and conversion between Western and Arabic numerals to reflect real-world document diversity.

3.4 Non-Text Images

To mitigate hallucination tendencies and reduce the model's bias toward assuming that every input image contains extractable text, we incorporated a set of $\sim$**1,500** images that contain no textual content. This subset includes blank pages, natural scenes, and other non-document imagery. Exposing the model to such samples during training encourages it to correctly produce an empty output when presented with text-free inputs, rather than fabricating non-existent content.

Collectively, the four data sources described above yield a total of approximately **769,000** image–text pairs used for training our model. A detailed breakdown of the dataset distribution across sources is shown in Figure 1.

4 Misraj-DocOCR: An Arabic Document OCR Benchmark

The evaluation of OCR models for Arabic text requires robust and accurate benchmarks. Our ini-

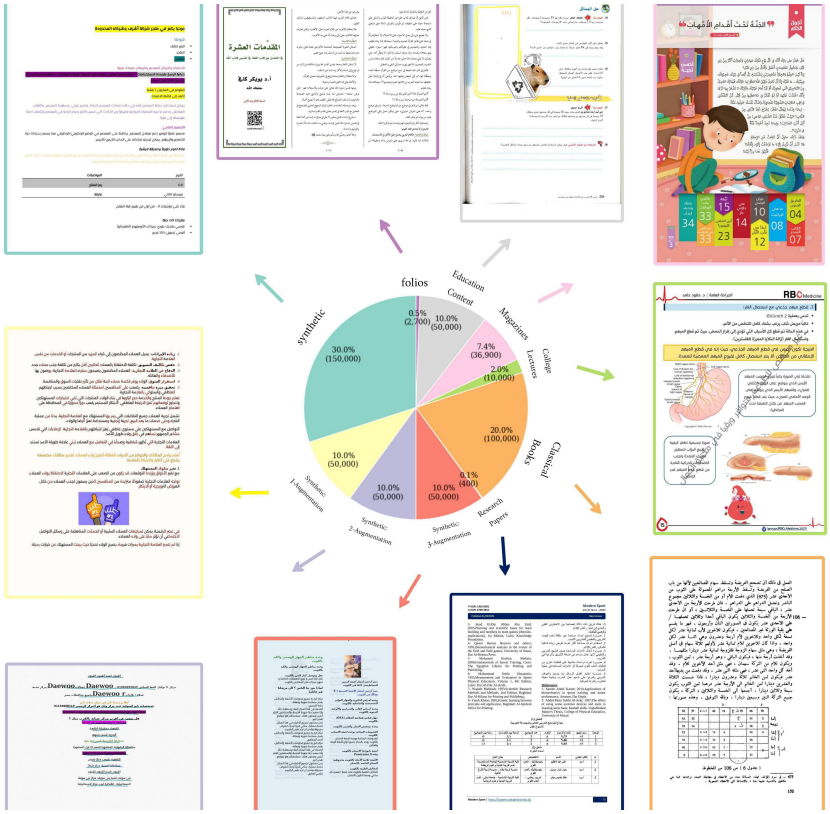


Figure 1: Distribution of training samples across the four data sources (synthetic, real-world books and magazines, table-centric, and non-text), totalling approximately 769K image–text pairs. The synthetic portion is split into clean rendered pages and their augmented variants.

tial investigation involved assessing existing benchmarks, such as the KITAB-bench (Heakl et al., 2025) pdf-to-markdown dataset. During this analysis, we identified significant shortcomings that compromise its reliability for model evaluation. A primary issue discovered was the presence of numerous errors in the ground truth data. We observed multiple instances of hallucinatory text, where the ground truth contained phrases not present in the source documents likely originating from a data creation or annotation tool rather than authentic content[‡]. Furthermore, our review revealed that many examples lacked corresponding page numbers, and small-font text was frequently omitted from the ground truth. These inaccuracies suggest that the dataset may not have undergone a thorough verification process after the initial data extraction. To address these deficiencies and provide a more reliable resource for the research community, we undertook a comprehensive correction of the KITAB-bench PDF-to-markdown dataset. This corrected version, with all identified errors rectified, has been made publicly available for academic use[§].

Beyond the inaccuracies, our examination of existing resources also indicated a lack of diversity in the style and type of documents. To foster more generalized and robust model development, a benchmark should encompass a wide variety of real-world scenarios. Therefore, we introduce *Misraj-DocOCR*, a new, comprehensive benchmark specifically designed for evaluating Arabic Document OCR models.

The primary contributions of this benchmark are:

- **Diverse and Comprehensive Content:** The benchmark consists of 400 high-quality images, curated to include a wide variation of document types, layouts, and fonts, and comprising both synthetic (150) and real-world (250) pages. The selection deliberately includes challenging conditions such as small-font and low-contrast text, dense tables, multi-column layouts, and diacritized passages, so

[‡]For example, one entry included the English sentence: "**You're right - let me write it exactly as it appears in the image, maintaining the right-to-left direction:**"

[§]https://huggingface.co/datasets/Misraj/KITAB_pdf_to_markdown_reviewed

that the benchmark stresses layout-aware transcription rather than clean single-column text alone.

- **Expert-Verified Ground Truth:** To ensure the highest level of accuracy, every image in the dataset has been meticulously reviewed by human experts. This verification process guarantees that both the transcribed text and the document structure are correct, eliminating the types of errors found in previous benchmarks.

- **Open Access:** Misraj-DocOCR is open-source and publicly available to all researchers. By providing this resource, we aim to facilitate further advancements and foster reproducible research in the field of Arabic OCR.

We evaluate many models on this benchmark and the corrected version of KITAB-Bench, all results on the section 7.

5 Methodology

The overall process for developing Baseer, involved a comprehensive data collection stage followed by a targeted fine-tuning stage.

The development of our model, Baseer, followed a two-stage methodology designed to tailor a powerful, pre-trained foundation model to our specific needs. The first stage involved the comprehensive collection and curation of a high-quality dataset, the details of which are described in Section 3. The subsequent stage, which is the focus of this section, consisted of fine-tuning the selected base model to align with our data and enhance its capabilities for Arabic document processing. For the base architecture of Baseer, we selected the *Qwen2.5-VL-3B-Instruct* model (Bai et al., 2025). This decision was predicated on its robust and state-of-the-art performance on multimodal tasks, particularly its demonstrated proficiency with the Arabic language compared to other open-source alternatives.

Despite its advanced capabilities, our preliminary analysis revealed that the base Qwen2.5-VL-3B-Instruct model exhibited certain limitations relevant to our use case. These included occasional reversions to left-to-right text generation, suboptimal handling of diacritized Arabic text, and other performance artifacts. Our fine-tuning strategy involved updating all model parameters, except for the vision encoder, which remained frozen. This approach allows the model to adapt its language and reasoning capabilities to our specialized dataset while preserving the powerful, generalized visual features learned during its original pre-training. The specific hyperparameters, hardware used, and other details of the training procedure are provided in Table 1.

Parameter	Value
Training Epochs	3
Learning Rate Schedule	Cosine decay
Learning Rate	1e-4
Batch Size	640
Weight Decay	0.01
Warm-up Steps	100
Optimizer	AdamW
Max Sequence Length	8192
GPU	8xH100

Table 1: Training Hyperparameters for *Baseer* Model

6 Experiments and Results

This section details the series of experiments conducted to systematically determine the optimal architecture and training configuration for Baseer. Our experimental process was designed to isolate variables and build upon the findings of each preceding stage.

6.1 Base Model Selection

The initial experiment was focused on selecting the most suitable base model for our task. To this end, we conducted a qualitative evaluation of several prominent open-source vision-language models. A curated set of representative examples, designed to test key capabilities in Arabic document understanding, was used as the input.

The outputs from each model were then subjected to a rigorous manual review by our evaluation team. The models were assessed based on criteria such as text recognition accuracy, preservation of right-to-left directionality, and overall coherence. This qualitative analysis concluded that Qwen2.5-VL-3B-Instruct demonstrated superior performance on Arabic-language tasks compared to the other candidates, making it the clear choice for our foundation.

6.2 Fine-Tuning Strategy Evaluation

After selecting the base model, our next objective was to identify the most effective fine-tuning strat-

egy. We designed a controlled experiment to compare three distinct approaches:

1. **Full Fine-Tuning (Baseer-Full):** All model parameters, including the vision encoder, were made trainable.

2. **Decoder-Only Fine-Tuning (Baseer-Decoder):** Only the parameters of the language decoder were updated, while the vision encoder remained frozen.

3. **Parameter-Efficient Fine-Tuning (Baseer-LoRA):** Low-Rank Adaptation (LoRA) was employed to update a small subset of parameters. LoRA adapters were applied to the linear projections of the language decoder (with the vision encoder frozen), using rank $r = 256$, scaling factor $\alpha = 128$.

To ensure a fair comparison, each of these strategies was tested on a 50,000-sample subset of our training data for two epochs, holding all other hyperparameters constant. We evaluated the models using ChrF, which measures OCR accuracy at the character level and captures text transcription quality. Table 2 summarizes the performance of the different fine-tuning strategies on the Baseer model.

Model	Trainable part	ChrF ↑
Baseer-Full	Full model	84.79
Baseer-Decoder	Language-decoder	**89.79**
Baseer-LoRA	LoRA weight	85.52

Table 2: Performance comparison of different fine-tuning strategies on Baseer model

As shown in Table 2, the results from our test set indicate that the decoder-only fine-tuning approach (Baseer-Decoder) significantly outperformed the other methods. This suggests that preserving the generalized features of the pre-trained vision encoder while adapting the language model to our specific data yields the best performance.

6.3 Impact of Sequence Length

Building on the previous finding, we adopted the decoder-only fine-tuning strategy and proceeded to investigate the effect of input sequence length on model performance. All training configurations were fixed while we experimented with three sequence length variants: 2048, 4096, and 8192 tokens.

Context Length	ChrF ↑
2048	82.69
4096	**89.79**
8192	88.52

Table 3: Performance comparison of different context lengths on Baseer model.

While the preliminary baseline metrics presented in Table 3 initially suggest an optimal performance at a sequence length of 4096, extensive real-world inference testing revealed a critical limitation. We observed that models constrained to 4096 tokens frequently failed to extract the complete context of highly dense visual data. This bottleneck was especially pronounced for images containing complex tables, which inherently necessitate a significantly larger token budget to accurately generate the exhaustive HTML table tags required for structural fidelity. To ensure comprehensive data extraction, we ultimately trained our final model using the 8192 context length, which substantially mitigates truncation issues on dense inputs and yields more robust results during inference.

7 Evaluation

We evaluate our model on our proposed Misraj-DocOCR benchmark and a corrected version of KITAB-Bench PDF-to-Markdown, alongside several open-source and commercial models. Text extraction performance is assessed using Word Error Rate (WER) and Character Error Rate (CER), which measure word and character-level transcription errors, BLEU for n-gram overlap, and ChrF, a character-level F-score suited for morphologically rich languages like Arabic. Structural and layout fidelity is measured with Tree Edit Distance Similarity (TEDS), capturing hierarchical document structures, and MARS (Heakl et al., 2025), which evaluates layout-aware alignment between predicted and reference renderings. Table 4 presents the results, showing that Baseer achieves state-of-the-art performance across both text and structural metrics, despite being smaller than competing models.

7.1 Evaluation Protocol

To ensure fair comparison across models, models designed for document understanding were evalu-

ated using their respective recommended system prompts, while general-purpose Multimodal Large Language Models (MLLMs) were provided with a single fixed instruction prompt requesting Markdown output with HTML tables. Because prompting can materially affect MLLM behaviour, we report the exact prompt templates used for every model in the Appendix, and all models were run with greedy decoding (temperature 0) to make outputs deterministic and reproducible. All outputs were then standardized using the following post-processing steps:

1. Remove HTML tags outside table structures
2. Convert Markdown tables to HTML format for consistency
3. Normalize horizontal line representations (—, ***, etc. → —)
4. Standardize header formatting
5. Unify formatting tags within HTML tables (, <b>→ <b>)

This standardization is critical because different models may produce semantically equivalent but syntactically different outputs, which would unfairly penalize models based on formatting choices rather than content accuracy.

Table 4 presents a comparative evaluation of different OCR and vision-language models using multiple metrics. The results indicate that **Baseer** achieves the best performance across all metrics, attaining the lowest WER and CER , as well as the highest TEDS and MARS scores. The **Gemini-2.5-pro** model follows closely, obtaining the highest BLEU and CHRF scores and ranking second overall. Notably, Baseer consistently outperforms large commercial systems such as GPT-based models and Azure AI Document Intelligence,[¶] underlining its robustness in both text and structure recognition. This is especially significant given that the evaluation dataset, Misraj-DocOCR, was deliberately designed to be highly diverse and challenging, with wide variation in layout and typography. The results also highlight a sharp performance gap between the top-performing systems and smaller or less specialized models (e.g., Gemma3, AIN, Aya-vision), underscoring the difficulty of this benchmark. Overall, Baseer and Gemini-2.5-pro emerge as the strongest systems in this comparison.

Table 5 reports the results on the KITAB-Bench PDF-to-Markdown dataset, which was carefully reviewed and corrected by domain experts to ensure high-quality ground truth annotations. For this corrected subset we restrict the comparison to open-source systems, so as to provide a fully reproducible, cost-free reference point on a small set of samples; commercial systems are instead evaluated on the larger and more diverse Misraj-DocOCR benchmark (Table 4). **Baseer** achieves the strongest performance across all metrics, attaining the best WER, CER, BLEU, CHRF, TEDS, and MARS scores, with **Dots.ocr** consistently ranking second across text-centric metrics. It is also worth noting that the KITAB-Bench subset is relatively small, consisting of only 30 samples, which makes every misprediction more impactful on the reported scores. In contrast, on the larger and more challenging **Misraj-DocOCR** benchmark with 400 diverse examples, Baseer's advantage over both open-source and commercial systems becomes more pronounced, highlighting its robustness across varied document types and layouts.

8 Conclusion

In this paper, we introduced Baseer, a vision-language model tailored for Arabic Document OCR, and presented Misraj-DocOCR, a high-quality benchmark designed for rigorous evaluation. By training on a diverse dataset of approximately 770K document-image pairs, we demonstrated that decoder-only fine-tuning is a powerful strategy that enables Baseer to achieve superior performance compared to a wide range of existing systems. Our detailed experimental analysis highlighted the importance of sequence length, fine-tuning scope, and dataset diversity in achieving robust performance. Notably, Baseer consistently achieved the best or near-best scores across Word Error Rate, Character Error Rate, and structure-aware metrics such as TEDS and MARS, surpassing both open-source and proprietary alternatives. These positive results underscore the value of domain-specific adaptation of general-purpose MLLMs, and provide new insights into how tailored data and efficient training strategies can push the boundaries of OCR for complex scripts. We

[¶] https://azure.microsoft.com/en-us/products/ai-services/ai-document-intelligence

Model	WER ↓	CER ↓	BLEU ↑	CHRF ↑	TEDS ↑	MARS ↑
Baseer (ours)	**0.17**	**0.1**	77.28	89.29	**61.27**	**75.28**
Gemini-2.5-pro	0.37	0.31	**77.92**	**89.55**	52	70.775
Gemini-3-pro	0.69	0.6	60.39	78.6	59	68.8
Azure AI	0.44	0.27	62.04	82.49	42	62.245
Dots.ocr	0.50	0.40	58.16	78.41	40	59.205
Nanonets	0.71	0.55	42.22	67.89	37	52.445
Qari	0.76	0.64	38.59	64.50	21	42.750
Qwen2.5-VL-32B	0.76	0.59	37.62	62.64	41	51.820
GPT-5	0.86	0.62	40.67	61.6	48	54.8
Qwen2.5-VL-3B-Instruct	0.87	0.71	25.39	53.42	27	40.210
Qwen2.5-VL-7B	0.92	0.77	31.57	54.70	27	40.850
Gemma3-12B	0.96	0.80	19.75	44.53	33	38.765
Gemma3-4B	1.01	0.85	9.57	31.39	28	29.695
GPT-4o-mini	1.36	1.1	22.63	47.04	26	36.52
AIN	1.23	1.11	1.25	2.24	21	11.620
Aya-vision	1.41	1.07	2.91	9.81	26	17.905

Table 4: Comparison of models across multiple evaluation metrics on Misraj-DocOCR. Best values are highlighted in **bold** and the second-best values are underlined.

Model	WER ↓	CER ↓	BLEU ↑	CHRF ↑	TEDS ↑	MARS ↑
Baseer (ours)	**0.22**	**0.17**	**77.24**	**89.39**	**63.79**	**76.59**
Dots.ocr	0.39	0.28	59.28	83.16	55	63.08
Nanonets	0.51	0.40	51.37	77.45	33	55.225
Qari	0.65	0.48	44.61	71.45	43	57.225
Qwen2.5-VL-3B-Instruct	0.70	0.57	40.44	66.78	31	48.89
Qwen2.5-VL-7B	0.76	0.63	36.76	62.45	24	43.225
Gemma3-12B	0.85	0.69	27.56	52.09	43	53.545
Gemma3-4B	0.95	0.82	12.94	31.72	27	29.36
Aya-vision	1.27	0.96	5.58	16.19	26	21.095
AIN	1.18	1.08	2.61	3.99	24	13.995

Table 5: Comparison of models across multiple evaluation metrics on KITAB-BenchPDF-to-Markdown dataset. Best values are highlighted in **bold** and the second-best values are underlined.

believe that this work establishes a strong baseline for future research and will accelerate the development of practical, high-accuracy OCR solutions for Arabic and other morphologically rich languages.

Limitations

While Baseer sets a strong baseline, several limitations should be noted. First, our evaluation relies on two benchmarks that we ourselves constructed or corrected: Misraj-DocOCR (400 expert-verified pages) and the corrected KITAB-Bench PDF-to-Markdown subset (60 pages). Both are modest in size, and the KITAB-Bench subset in particular is small enough that individual mispredictions can noticeably shift aggregate scores. We report single-run point estimates without statistical significance testing; the differences between the top systems on the small subset should therefore be read as indicative rather than definitive. Evaluation on independent, third-party Arabic OCR benchmarks would further strengthen these conclusions and is left to future work. To support scrutiny, we release both benchmarks, the evaluation scripts, and the exact prompts; we nonetheless caution that MLLM results are sensitive to prompt wording and post-processing, and that different choices may change the relative ordering of the general-purpose baselines.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, Shivalika Singh, Madeline Smith, Bharat Venkitesh, Vlad Shmyhlo, Viraat Aryabumi, Walter Beller-Morales, et al. 2025. Aya vision: Advancing the frontier of multilingual multimodality. *arXiv preprint arXiv:2505.08751*.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Ahmed Heakl, Abdullah Sohail, Mukul Ranjan, Rania Hossam, Ghazi Shazan Ahmad, Mohamed El-Geish, Omar Maher, Zhiqiang Shen, Fahad Khan, and Salman Khan. 2025. Kitab-bench: A comprehensive multi-domain benchmark for arabic ocr and document understanding. *arXiv preprint arXiv:2502.14949*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024a. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024b. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37:87874–87907.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Zhang Li, Yuliang Liu, Qiang Liu, Zhiyin Ma, Ziyang Zhang, Shuo Zhang, Zidun Guo, Jiarui Zhang, Xinyu Wang, and Xiang Bai. 2025. Monkeyocr: Document parsing with a structure-recognition-relation triplet paradigm. *arXiv preprint arXiv:2506.05218*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Souvik Mandal, Ashish Talewar, Paras Ahuja, and Prathamesh Juvatkar. 2025. Nanonets-ocr-s: A model for transforming documents into structured markdown with intelligent content recognition and semantic tagging.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. 2025. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*.

Ahmed Nassar, Andres Marafioti, Matteo Omenetti, Maksym Lysak, Nikolaos Livathinos, Christoph Auer, Lucas Morin, Rafael Teixeira de Lima, Yusik Kim, A Said Gurbuz, et al. 2025. Smoldocling: An ultra-compact vision-language model for end-to-end multi-modal document conversion. *arXiv preprint arXiv:2503.11576*.

Mohamed Rashad. 2024. Arabic-nougat: Fine-tuning vision transformers for arabic ocr and markdown extraction. *arXiv preprint arXiv:2411.17835*.

Prince Sultan University RIOTU Lab. 2025. Sard: A large-scale synthetic arabic ocr dataset for book-style text recognition. *arXiv preprint arXiv:2505.24600*.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, et al. 2024. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Ahmed Wasfy, Omer Nacar, Abdelakreem Elkhateb, Mahmoud Reda, Omar Elshehy, Adel Ammar, and Wadii Boulila. 2025. Qari-ocr: High-fidelity arabic text recognition through multimodal large language model adaptation. *arXiv preprint arXiv:2506.02295*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Worldview Annotation for Low-Resource Languages

Tetiana Ilman
Independent Researcher
worldviewannotationschema@gmail.com

Abstract

Low-resource language technology has advanced significantly in preserving linguistic form, yet a critical gap remains: the conceptual systems that languages encode — how communities structure personhood, agency, time, evidence, relationality, and moral responsibility — are systematically lost in current documentation pipelines. We argue that the problem is not only the annotation scale but annotation level. We introduce the Worldview Annotation Schema (WAS), presented here as a three-layer framework that begins with community documentators in the field and ends with machine-readable structured data, catching culturally grounded meaning before downstream AI analysis can homogenize it. We track where worldview meaning dissolves in existing pipelines, propose the WAS architecture with a worked example from Maa (Eastern Africa), and outline a governance model that grounds curator interpretation in community-provided evidence while producing graph-compatible structured output. In the age of AI, language preservation without worldview preservation risks producing languages that survive archivally while disappearing cognitively.

1 Introduction

More than 7,000 languages are spoken worldwide (Hammarström et al., 2020), yet global NLP resources remain concentrated in a small number of high-resource languages (Joshi et al., 2020), while many languages face ongoing endangerment—concerns reflected in international initiatives such as the International Decade of Indigenous Languages (UNESCO, 2022). Language loss is not merely the loss of another dictionary: languages carry oral traditions, ecological knowledge, social memory, and specific ontologies. When a language disappears, humanity loses a knowledge operating system — an epistemic and ontological loss that may also contribute to future AI bias.

This paper introduces a conceptual annotation framework designed to preserve worldview information in low-resource language documentation. Our goal is to define the architecture, annotation workflow, and governance principles that can support future empirical evaluation.

Traditional language preservation focused on documentation, literacy, and heritage transmission. Although essential, these approaches are no longer sufficient in the age of artificial intelligence. A language archived but not computationally integrated risks symbolic survival alongside functional disappearance. Preservation must now include the structured representation of the worldview systems encoded in language.

Advances in machine translation, speech recognition, corpus construction, and automated annotation have created new possibilities for low-resource language inclusion. Nevertheless, global NLP remains heavily concentrated in a small number of high-resource languages (Joshi et al., 2020). However, there is a foundational asymmetry. AI systems learn from what becomes machine-readable. Deeper conceptual distinctions encoded in the language often are not.

This study was motivated by an exploratory experiment that examined this asymmetry directly. In Ukrainian affective language contexts, human speakers consistently used diminutive morphology to express relational closeness and care, whereas the outputs of several AI systems omitted these forms almost entirely under neutral prompting. When explicitly prompted to use diminutives, outputs shifted to an exaggerated, saturated form that read as caricatured or performative rather than conveying genuine affective warmth. In our exploratory Ukrainian data, relational warmth was frequently expressed through layered diminutive morphology. This use is consistent with cross-linguistically documented affective functions of diminutives (Jurafsky, 1996). The exploratory com-

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 429–437
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_044

parison therefore suggests that one of Ukrainian's most productive worldview-bearing features is consistently misrepresented during current AI language generation. If AI systems fail to preserve this distinction even in a well-documented language, the same compression is likely to have greater consequences in languages where the documentation infrastructure itself is limited. The complete experimental methodology and quantitative analysis are beyond the scope of this paper and will be reported separately. Here, the observation serves only to motivate the annotation framework introduced below.

Current annotation systems scale the analysis of observable content: topics, sentiment, named entities, discourse functions. However, many culturally meaningful distinctions exist at another level — how speakers encode evidential source, distribute agency, conceptualize time, or assign moral responsibility. These distinctions are not reliably recoverable through surface semantic labeling alone. The problem is not only the annotation scale but annotation level.

A language that survives in archives but loses its conceptual architecture has not fully survived. Preserving a language's forms while losing its ontological structure is just a slower kind of disappearance. We frame WAS around low-resource languages specifically because these are the contexts where a lost worldview distinction has no fallback: no large parallel corpus, no established reference grammar, no standing community of professional linguists to reconstruct it later. Better-documented languages such as German or French retain enough alternative documentation to recover such distinctions after the fact even when an annotation pipeline flattens them once; thinly documented languages typically do not, so losses at the annotation stage tend to be final.

Contributions. This paper makes five contributions: (1) it repositions annotation as a site of epistemic decision-making where worldview distinctions are preserved or erased, introducing the Worldview Annotation Schema (WAS) to operationalize worldview through dimensions such as time, agency, and emotion; (2) it defines the current three-layer annotation architecture and a two-tier community/curator workflow, designed for non-technical first-tier language specialists (translators, educators, field linguists) and separating linguistic interpretation from technical curation; (3) it treats non-alignment and untranslatability as valid outcomes, addressing the upstream data formation stage that benchmarks implicitly assume already been handled; (4) it illustrates the framework through a fully worked Maa proverb annotation, including two distinct documented usage events and a complete literal gloss; and (5) it reports a pilot field application with a native Maa-speaking community documentator, illustrating that the workflow can be carried out by a non-technical community member in an authentic setting.

2 The Dissolution of Worldview in Current Pipelines

The dissolution of worldview-level meaning happens cumulatively, across three transitions, each introducing a compression that the next stage cannot recover.

From interaction to transcription: spoken language carries context, prosody, timing, gesture, silence, and relational positioning. When interaction becomes text, these channels are largely discarded. Culturally meaningful pause lengths, intonation contours, evidentiality, and bodily conduct either disappear or require specialist annotation that most low-resource pipelines cannot sustain.

From transcription to translation: the target language (most often English) lacks direct equivalents for many source-language distinctions. Evidential morphology may be rendered as a neutral assertion, echoing Jakobson's (1959) observation that translation entails loss of grammatically obligatory source-language distinctions. A Mayan greeting encoding a path-based model of life (*Bix a beel*: how-your-path) may be glossed simply as "Hello." Conceptual architecture is not preserved.

From translation to annotation: standard frameworks capture structural linguistic information or task labels, not worldview-level distinctions. An additional epistemic filter operates — for example, Maya field linguists have reported that greeting expressions encoding life as trajectory are often dismissed as romanticization, reducing systematic worldview patterns to noise.

The cumulative result is ontological compression: languages are digitized in ways that preserve surface structure but not the conceptual architectures generating meaning. A language may achieve formal digital survival while the conceptual systems it encodes remain inaccessible to computational tools.

3 Background and Related Work

3.1 Low-Resource Language Technology

Low-resource NLP has focused primarily on overcoming data scarcity for tasks such as machine translation, speech recognition, and corpus construction (Joshi et al., 2020), including documented challenges specific to indigenous languages of the Americas (Mager et al., 2018). Prior surveys demonstrate that global NLP remains concentrated on a small number of high-resource languages. Pava et al. (2025) document persistent asymmetries in evaluation and model coverage in low-resource language contexts.

van Rijn et al. (2026) demonstrate the scalability of automatically generated vocabulary-based proficiency assessments across a large number of languages. However, scalable language-level tools do not by themselves represent culturally specific conceptual distinctions: a model may process a language fluently while systematically underrepresenting the morphological, evidential, or relational dimensions through which it encodes meaning.

LLMs trained primarily on dominant linguistic and cultural contexts reproduce those structures when interpreting new material — reshaping it to fit familiar patterns. For under-documented cultures, this is particularly problematic: invented outputs can circulate as authentic representations in the absence of accessible reference corpora.

This work assumes languages encode distinctions not fully recoverable through translation alone, consistent with work on linguistic relativity, culturally structured meaning, and pragmatic interpretation (Lucy, 1992; Levinson, 1983; Wierzbicka, 1997), without strong claims about cognitive determinism. The key point is that these distinctions risk being lost in current documentation pipelines.

Our approach asks how meaning is structured before becoming data, capturing worldview patterns explicitly so they are not lost prior to annotation — shifting the task from labeling content to modeling how different systems organize meaning.

3.2 Annotation Frameworks and Related Work

Established annotation frameworks represent phonetics, morphology, syntax, discourse structure, and semantic roles. Architectures such as annotation graphs (Bird and Liberman, 2001) and the ISO Linguistic Annotation Framework (Ide and Romary, 2004) were developed to support interoperable representation across linguistic annotations and datasets. The issue is not that annotation is absent, but that most existing frameworks were designed to capture structural linguistic information rather than worldview-level distinctions. Table 1 summarizes how the proposed framework differs from other annotation approaches.

Existing semantic representation frameworks and lexical-semantic models—including FrameNet (Baker et al., 1998), AMR (Banarescu et al., 2013), UCCA (Abend and Rappoport, 2013), and OntoLex-Lemon (W3C Ontology-Lexica Community Group, 2016)—capture complementary aspects of lexical and utterance meaning, such as frames, concepts, participants, semantic relations, and links between lexical entries and ontological entities. WAS operates at a different dimension.Rather than modeling only the meaning expressed by an utterance, WAS represents evidence-grounded interpretations of the worldview patterns made salient by the utterance and its documented contexts of use. A FrameNet annotation would represent the event structure and participant roles evoked by the proverb. A WAS annotation instead represents its epistemic structure (signal $\neq$ outcome), ecological grounding (predator behavior), social function (calming fear in a hierarchical context), and boundary condition (inappropriate when harm has already occurred). These dimensions are not an extension of FrameNet. They operate at a different level of description, representing the worldview that semantic roles presuppose rather than the semantic roles themselves.

3.3 LLM-Based Cultural Annotation

Recent LLMs have enabled scalable annotation of narratives, values, and cultural discourse patterns. However, most such systems operate on observable outputs rather than the conceptual structures generating them. The standard pipeline is: Text $\rightarrow$ LLM $\rightarrow$ Annotation. This treats text and its meaning as already transparent. If culturally distinct meaning systems are interpreted only through dominant-language categories, the result may be efficient but reductive. WAS is designed to address this risk.

AlKhamissi et al. (2025) find that current cultural benchmarks reduce culture to static facts or homogeneous values, arguing that anthropological methods are largely absent from benchmark design. Khudayberganova et al. (2025) demon-

Approach	Preserves raw form	Captures context of use	Controlled worldview tags	Structured output for downstream use	Annotator role in worldview capture
Linguistic glossing	Yes	Limited	No	Variable	No structured worldview role
Metadata archiving	Partial	Partial	Limited	Yes — metadata	Consultative; not a structured annotator
FrameNet	No	Limited — frame-evoking context, not community usage	No	Yes	Expert lexicographer/annotator; no community role
AMR	No	No — sentence-level meaning graph	No	Yes	Trained annotator; no community role
UCCA	No	Limited	No	Yes	Trained annotator; no community role
OntoLex	No	Limited	Limited — lexical-semantic relations, not worldview	Yes	Lexicographer/ontology engineer; no community role
This framework	Yes	Yes — usage event required	Yes — Tier-1 tagset	Yes — graph-compatible	First annotator: evidence precedes and constrains curator interpretation

Table 1: Comparison of annotation approaches.

strate community-grounded ontology annotation for Uzbek folklore preservation; their system begins with transcribed text, while WAS targets the pre-textual documentation stage. Martinez Pandiani and Presutti (2024) formalize annotation context through SituAnnotate, explaining how labels are produced but not how culturally specific conceptual systems shape what can be expressed. WAS targets this internal layer, encoding schema-level distinctions — personhood, time, agency, morality — as structured dimensions of meaning.

4 The Worldview Annotation Schema

In the version presented here, the Worldview Annotation Schema is a three-layer framework for capturing culturally grounded meaning before downstream AI analysis normalizes it. Its central claim is that language is a constrained evidence layer for worldview, not the worldview structure itself. The framework is intended as a practical first step that extends existing documentation practice. This approach is grounded in the sociology of culture and cognition: Cerulo et al. (2021) review how culture operates through cognitive processes, including structured patterns of acquisition, storage, and retrieval. Building on this perspective, we argue that annotation limited to propositional content may miss aspects of the schema-level organization through which culturally grounded meaning is interpreted. Alwajih et al. (2025) demonstrate the value of culturally specific datasets and evaluation benchmarks, including attention to nuanced cultural variation that broad labels may obscure. We believe that non-equivalence should be treated as informative rather than as annotation failure — directly supporting the non-equivalence tolerance requirement built into the framework's design. The Ukrainian diminutive asymmetry introduced in Section 1 illustrates why this matters even in a well-resourced language; the Maa case study below shows the same compression mechanism operating in a language with substantially fewer digital resources than Ukrainian, making such losses more difficult to detect and recover after the fact.

WAS is implemented through a two-tier workflow: a community-facing documentator layer and a technical curator layer, detailed below.

4.1 The Initial Three-Layer Architecture

Layer 1 — Raw Input: the documentator records what speakers say, exactly as spoken. No interpretation, no simplification, audio recording preferred, variants captured. The raw layer is immutable: nothing recorded here is discarded or corrected at later stages. This layer ensures linguistic survival regardless of how tags evolve.

Layer 2 — Analysis: divided into three sequential components. Layer 2A (Conceptual Extraction) identifies the domain, conceptual axes, and candidate tags — detecting where conceptual systems diverge from dominant-language defaults. Layer 2B (Language/Form Analysis) uses lexical, metaphorical, grammatical, and discourse-level signals as evidence for the worldview reading. Layer 2C (Evidence Integration) combines 2A and 2B outputs, checks for recurrence, validates tags against recorded evidence, and assigns a confidence level. Without 2C, annotation collapses into subjective labeling.

Layer 3 — Structured Output: clean, standardized representation encoding the final cluster, axes, tags, and confidence level. No interpretation happens here — only encoding. Output is graph-compatible, machine-readable.

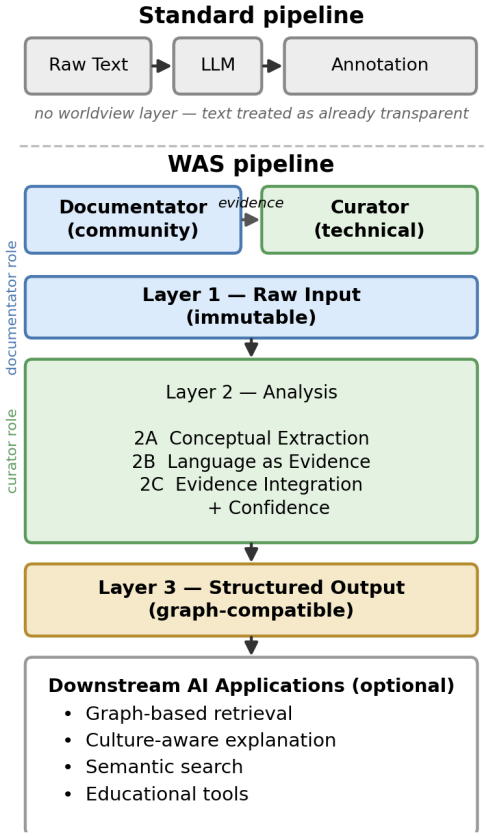


Figure 1: Standard pipeline versus the WAS pipeline.The graph-compatible structured output is designed to support future AI retrieval and comparison of worldview concepts alongside lexical and propositional content

4.2 The Two-Tier Workflow

The documentator is a community member or language worker with no required technical background. Using a structured field sheet, the documentator: (1) records the exact form with audio and variants; (2) collects one concrete remembered story of real use; (3) identifies 1–2 primary functions from a fixed list; (4) identifies the emotion targeted; (5) completes a metaphor map; (6) records boundary conditions. The documentator does not invent tag names.

The curator is a trained linguist or senior community worker who converts documentary evidence into structured annotation: normalizing form, creating usage event objects, extracting metaphor frames, encoding worldview assertions, and recording boundary conditions as explicit graph relations. The key constraint: all worldview assertions must be traceable to recorded documentary evidence. This traceability requirement ensures that curator interpretations remain constrained by community-provided evidence and open to community review.

4.3 Tag Governance

The governance model addresses the central tension: tags may not grow freely without losing interoperability, but cannot be locked too tightly without erasing worldview nuance. The solution is layered governance: field freedom in description, central discipline in vocabulary.

The Tier-1 tagset covers function (CalmFear, Warn, Correct, Encourage, AssertAuthority, ReduceConflict, TeachRespect, InduceShame, MotivatePatience, Humor), emotion (Fear, Anger, Shame, Pride, Grief, Envy, Impatience), domain (Conflict, Family, Leadership, Livestock, Marriage, Ritual, Community, Survival, YouthTraining), and epistemic hint (AppearancesMislead, WaitForEvidence, ActionsMatterMoreThanWords, TimeRevealsTruth, RespectHierarchy, CommunityBeforeSelf). New tags enter only after appearing in three independent cases, with a formal definition and two usage examples — mirroring controlled vocabulary evolution in FrameNet and WordNet.

Field	Content
Proverb ID	MAA-PROV-001
Language / Dialect	Maa (Eastern Africa)
Exact text	Me metan'ga olowuaru engutuk nijo kitoijoyie
Literal gloss	Me = not that | Metan'ga = opens | Olowuaru = a lion | Engutuk = the mouth | Nijo = you say | Kitoijoyie = it has swallowed you
Free translation	It does not mean that when a lion opens its mouth that it has eaten or swallowed you
Variants	Single-speaker record (Alois Thomas Porokwa); cross-speaker variants not yet collected
Audio	Recorded with speaker consent; available upon request, subject to community approval.

Table 2: Layer 1 raw input record for MAA-PROV-001.

The system distinguishes Closed Mode (highly under-resourced languages: restricted to usage events and Tier-1 tags) and Open Mode (once recurring patterns are available: full Layer 2 pipeline with deeper conceptual extraction and vocabulary expansion).

5 Worked Example: A Maa Proverb

We illustrate WAS through a fully worked annotation of a proverb from Maa, a Nilotic language spoken in Eastern Africa: *Me metan'ga olowuaru engutuk nijo kitoijoyie* — "It does not mean that when a lion opens its mouth that it has swallowed you." This example illustrates ecological grounding in local predator behavior, an epistemic structure distinguishing signal from outcome, and a socially situated function often omitted in summary translation.

5.1 Layer 1 — Raw Input

Table 2 presents the Layer 1 raw input record for MAA-PROV-001, documenting the proverb's exact text, literal gloss, free translation, variants, and audio provenance exactly as recorded by the documentator, prior to any interpretation or analysis.

5.2 Layer 2A — Conceptual Extraction

Two independent usage events (Table 3) — a medical crisis and a livelihood crisis — show a speaker with cultural authority reorienting a younger person from despair toward action, suggesting the proverb encodes a portable cognitive structure for managing fear amid unresolved threat. This structure rests on three dimensions: an epistemic heuristic (wait for evidence rather than inferring outcome from signal), a signal-outcome distinction (threat display vs. realized harm), and ecological grounding in predator behavior.

5.3 Layer 2B — Language as Evidence

Lexical and grammatical evidence: *olowuaru* ("lion") evokes a culturally important predator; *engutuk* ("mouth") foregrounds the display organ; *kitoijoyie* ("swallowed/eaten") encodes irreversible harm as the feared outcome. The negation structure (*me metan'ga*, "it does not mean that…") explicitly separates a perceived signal from its inferred consequence, a distinction confirmed by a substitution test: replacing "lion" with a generic threat weakens the display-versus-harm contrast, suggesting the proverb relies on shared knowledge of predator behavior rather than on danger alone.

5.4 Layer 2C — Evidence Integration

The lexical, grammatical, and contextual evidence converges on the same interpretation. Grammatically, the proverb distinguishes a threat signal from realized harm through explicit negation, while lexically it contrasts opening the mouth with swallowing, reinforcing the distinction between warning and outcome. Shared knowledge of predator behavior provides the ecological grounding, and recurrent usage suggests a social function of regulating fear under uncertainty. Confidence in the item-level interpretation is assessed as high based on convergence across evidence types and recurrence across the two documented usage events; however, data sufficiency for broader generalization remains limited.

5.5 Layer 3 — Structured Output

Example RDF graph relations:

MAA-PROV-001 **hasFunction** CalmFear;
MAA-PROV-001 **hasEpistemicHeuristic** WaitForEvidence;
MAA-PROV-001 **hasMetaphorSource** PredatorThreatDisplay;
MAA-PROV-001 **hasMetaphorTarget** ThreatSituation;
MAA-PROV-001 **inappropriateWhen** ActualHarmOccurred.

The Layer 3 representation supports graph storage and queries such as retrieving expressions that distinguish a threat signal from its outcome.

5.6 Pilot Field Application

To illustrate the practical application of the proposed workflow in an authentic documentation set-

Usage Event Field	UE-001	UE-002
Usage Event ID	MAA-PROV-001-UE-001	MAA-PROV-001-UE-002
Speaker role	Maa community member (adult, cultural authority)	Elder (family leader, cultural authority)
Addressee role	Non-Maa interlocutor (younger, outside community)	Younger community member (adult male)
Context type	Medical crisis — serious illness, acute uncertainty	Livelihood crisis — near-total livestock loss due to drought
Trigger	Addressee expressing fear and distress	Addressee in hopelessness; considering abandoning pastoral livelihood
Speech act	Direct comfort; reassurance through ecological-epistemic reframing	Encouragement; reframing of crisis as non-terminal
Emotional climate	High anxiety; fear of irreversible harm	Despair, grief, fear of permanent loss
Outcome	Addressee reported proverb as meaningfully calming	Addressee reported renewed courage; rebuilt livestock holdings

Table 3: Usage Event records for MAA-PROV-001. Two independent usage contexts illustrate portability across threat types. UE-002 reported by Alois Thomas Porokwa, 28 April 2026.

ting, we conducted a pilot field application with Alois Thomas Porokwa, a native Maa speaker and community documentator. The pilot demonstrates the workflow end-to-end—from documentator-captured raw input through curator-produced Layer 3 structured output—for a single illustrative item.

The documentator independently completed the Layer 1 record for MAA-PROV-001 and assigned the Tier-1 tags Encourage and MotivatePatience (function), Fear (emotion), Survival (domain), and AppearancesMislead (epistemic hint), without reference to the curator's analysis. These annotations were then compared with the curator's independently produced annotation of the same raw evidence: CalmFear (function), Fear (emotion), Conflict/Leadership (domain), and WaitForEvidence (epistemic hint).

The two selections fully agree on emotion, partially agree on epistemic hint—representing adjacent framings of the same underlying logic—and diverge on function and domain in ways that appear context-sensitive rather than erroneous. The documentator's tags foreground the encouragement given to a specific addressee, while the curator's tags foreground the social-hierarchical frame of the exchange.

We report this comparison to illustrate feasibility in practice, not as an inter-annotator agreement study. A single comparison point cannot support a reliability statistic such as Cohen's κ; a formal agreement study involving multiple documentators remains future work. The comparison nonetheless provides a preliminary signal that a non-technical documentator using only the field cheat sheet can independently produce tags that meaningfully align with expert curation, which is the minimum bar required by the proposed two-tier workflow.

6 Implications for AI and Low-Resource NLP

If culturally meaningful distinctions are systematically absent from training data, AI systems may perform well on surface tasks while remaining shallow in cross-cultural settings. Structured Layer 3 outputs are designed to support the downstream applications described below, complementing the scalable vocabulary-based language assessment demonstrated by van Rijn et al. (2026) by extending infrastructure from language coverage toward worldview coverage.

AlKhamissi et al. (2025) show that current cultural benchmarks often reduce culture to static facts or survey-derived values. WAS-based evaluation sets could instead test whether systems retain features such as evidential marking, relational morphology, or ecologically grounded metaphor — dimensions not explicitly evaluated in existing benchmarks.

Operationally, a Layer 3 record can serve as structured contextual knowledge for downstream AI applications. Rather than relying only on the literal wording of an expression, an AI system can use a Layer 3 record to explain the worldview it encodes. In the single worked example presented here, an assistant asked to explain the Maa proverb could present not only the literal translation ("a lion opening its mouth does not mean it has caught you") but also an explanation of its documented epistemic stance (WaitForEvidence), the observation that appearances may be misleading

Field	Value
CLUSTER	Knowledge & Truth
AXES	Signal vs. Outcome; Appearance vs. Reality
FUNCTION (Tier-1)	CalmFear
EMOTION TARGET (Tier-1)	Fear → reduced / redirected
DOMAIN (Tier-1)	Survival, Conflict / Leadership
EPISTEMIC HINT (Tier-1)	WaitForEvidence; AppearancesMislead
METAPHOR MAPPING: SourceFrame	PredatorThreatDisplay (Agent: Lion; Action: OpenMouth; PotentialOutcome: Swallow)
METAPHOR MAPPING: TargetFrame	ThreatSituation (Agent: PowerfulPerson or ExternalThreat; Action: ThreatDisplay; PotentialOutcome: Harm)
Derived Assertion: EncodesNorm	DoNotInferHarmFromThreatSignal
EncodesEpistemicHeuristic	WaitForEvidence
EcologicalAnchor	Savannah predator behavior (lion threat display)
ReinforcesHierarchy	Contextual — acknowledges power differential while moderating fear response
BoundaryCondition	Inappropriate when actual harm has already occurred; may be used sarcastically to minimize real danger (documentator flag required)
CONFIDENCE	High (based on convergence across independent evidence types; see §5.6 documentator–curator comparison)

Table 4: Layer 3 structured output for MAA-PROV-001. All fields are traceable to documentary evidence recorded at Layers 1 and 2.

(AppearancesMislead), its ecological grounding, and its conditions of use. These concepts are explicitly represented in the Layer 3 record rather than remaining implicit in free-text explanation. The same structured representation also supports graph-based retrieval and provides a reference for detecting cases where distinct worldview concepts have been collapsed into an apparent equivalent during translation or AI processing. For example, a query such as "find expressions that distinguish a threat signal from its outcome" would return MAA-PROV-001 because these worldview concepts are encoded as searchable fields rather than embedded only in prose. This example illustrates the type of AI-supported explanation and retrieval enabled by WAS on a single annotated item. As larger collections of WAS-annotated records become available, the same representation could support retrieval and comparison of analogous worldview structures across languages. Building and evaluating such a multilingual corpus remains future work.

Regarding scalability: a single proverb annotation took about 15 minutes with no technical training, and the framework fits existing documentation practice, with trained curators able to support multiple community projects remotely. Deployment must respect restricted or ceremonial knowledge through informed consent.

7 Limitations

The framework is illustrated through one worked example; broader validation across typologically diverse languages remains future work. The documentator–curator comparison reported in the pilot field application (§5.6) involves a single community documentator rather than multiple independent annotators; a full inter-annotator agreement study with several documentators per language remains future work. The workflow assumes access to a trained curator, which may limit applicability in severely under-resourced settings. The Tier-1 tagset is based on a limited empirical base and may require expansion. The governance model has not yet been tested in situations where it could be misused. The architecture presented here is an initial conceptual model and may be refined as its provenance relations, and governance procedures are tested across additional languages and documentation settings.

8 Conclusion

Low-resource language technology increasingly preserves speech, text, and computational access, but linguistic form alone does not capture how speakers represent evidence, personhood, agency, time, and moral responsibility. These distinctions can be reduced across transcription, translation, and annotation.

This paper introduced an initial three-layer annotation framework with a two-tier community/curator workflow, illustrated through a worked Maa proverb. Non-technical community members act as first annotators, providing the evidence from which curators build controlled, structured representations.

As AI systems increasingly learn from recorded language data, the question is not only which languages are present, but which meaning-encoding features remain representable. WAS supports collaborative annotation of alternative interpretations without forced equivalence or cultural collapse.

References

Omri Abend and Ari Rappoport. 2013. Universal Conceptual Cognitive Annotation (UCCA). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 228–238, Sofia, Bulgaria. Association for Computational Linguistics.

Mai AlKhamissi, Yunze Xiao, Badr AlKhamissi, and Mona Diab. 2025. Hire your anthropologist! rethinking culture benchmarks through an anthropological lens.

Fakhraddin Alwajih et al. 2025. PEARL: A multimodal culturally-aware Arabic instruction dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23048–23079, Suzhou, China. Association for Computational Linguistics.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Steven Bird and Mark Liberman. 2001. A formal framework for linguistic annotation. *Speech Communication*, 33(1–2):23–60.

Karen A. Cerulo, Vanina Leschziner, and Hana Shepherd. 2021. Rethinking culture and cognition. *Annual Review of Sociology*, 47:63–85.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2020. Glottolog Database 4.3 as CLDF.

Nancy Ide and Laurent Romary. 2004. International standard for a linguistic annotation framework. *Natural Language Engineering*, 10(3–4):211–225.

Roman Jakobson. 1959. On linguistic aspects of translation. In Reuben A. Brower, editor, *On Translation*, pages 232–239. Harvard University Press, Cambridge, Massachusetts.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Daniel Jurafsky. 1996. Universal tendencies in the semantics of the diminutive. *Language*, 72(3):533–578.

Durdona Khudayberganova, Nilufar Tosheva, Hafiza Aslanova, Olim Kayumov, Gulnoz Khallieva, Otabek Mukhitdinov, I. B. Sapaev, and Dilfuza Tursunova. 2025. Hybrid ontology and NLP-based annotation model for indigenous lexicon and folklore preservation. *Indian Journal of Information Sources and Services*, 15(2):333–339.

Stephen C. Levinson. 1983. *Pragmatics*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.

John A. Lucy. 1992. *Language Diversity and Thought: A Reformulation of the Linguistic Relativity Hypothesis*. Number 12 in Studies in the Social and Cultural Foundations of Language. Cambridge University Press, Cambridge.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Delfina Sol Martinez Pandiani and Valentina Presutti. 2024. Situated ground truths: Enhancing bias-aware AI by situating data labels with SituAnnotate.

Juan N. Pava, Caroline Meinhardt, Haifa Badi Uz Zaman, Toni Friedman, Sang T. Truong, Daniel Zhang, Elena Cryst, Vukosi Marivate, and Sanmi Koyejo. 2025. Mind the (language) gap: Mapping the challenges of LLM development in low-resource language contexts. White paper, The Asia Foundation and Stanford Institute for Human-Centered Artificial Intelligence.

Pol van Rijn, Yue Sun, Harin Lee, Raja Marjieh, Ilia Sucholutsky, Francesca Lanzarini, Elisabeth André, and Nori Jacoby. 2026. One test, many tongues: Surveying language proficiency across the globe. *Proceedings of the National Academy of Sciences*, 123(13):e2420179123.

UNESCO. 2022. International decade of indigenous languages 2022–2032.

W3C Ontology-Lexica Community Group. 2016. Lexicon model for ontologies. Final community group report, World Wide Web Consortium.

Anna Wierzbicka. 1997. *Understanding Cultures through Their Key Words: English, Russian, Polish, German, and Japanese*, volume 8 of *Oxford Studies in Anthropological Linguistics*. Oxford University Press, New York.

A Multilingual Sentence-Transformer Baseline for Serbian News Topic Classification

Saša Petalinkar*, Milica Ikonić Nešić†, Ranka Stanković*, Jelena Graovac‡
*University of Belgrade, Faculty of Mining and Geology, Serbia
{sasa.petalinkar, ranka.stankovic}@rgf.bg.ac.rs
†University of Belgrade, Faculty of Philology, Serbia
milica.ikonic.nesic@fil.bg.ac.rs
‡University of Belgrade, Faculty of Mathematics, Serbia
jgraovac@matf.bg.ac.rs

Abstract

We present a baseline for topic classification of Serbian newspaper articles. Starting from a corpus of roughly 123,000 newspaper articles annotated with section labels, we apply a deterministic label-normalization procedure that corrects encoding-corrupted spellings, merges truncated labels, and removes a non-coherent catch-all category, resulting in 35 normalized classes. The multilingual MiniLM sentence-transformer is fine-tuned with a multiple-negatives ranking objective and, at test time, all 35 section labels are ranked by cosine similarity, with the top-ranked label treated as a single direct classification hit. On a held-out test split of 18,487 articles, the model reaches a top-1 accuracy of 0.784 (Wilson 95% interval 0.778–0.790), with a weighted F1 of 0.790 and a macro F1 of 0.653. Per-section analysis shows strong performance on high-frequency sections and substantial over-prediction of small, semantically adjacent classes. The fine-tuned model, per-section metrics, a confusion analysis, and documented label-cleaning corrections are released to support further work on Serbian news classification.

1 Introduction

Serbian is a less-resourced South Slavic language for which large, cleanly labeled text-classification resources remain scarce. However, newspaper archives are routinely tagged with editorial section (*rubrika*) labels that describe the topical section to which an article belongs. These labels offer a natural supervisory signal for topical classification, which is useful for archival indexing, retrieval, and media monitoring.

In this paper, a baseline section classifier is built and evaluated over a large Serbian news corpus. The central modeling choice is to treat section assignment as a ranking problem: a multilingual sentence-transformer encodes both the article text and each candidate section label, and sections are ranked by cosine similarity. At evaluation time, the first ranked label is taken as the single direct classification hit, which makes the model directly comparable to a standard top-1 classifier.

The main contributions of this work are:

- a normalized 35-section Serbian news dataset with a fully documented, deterministic label-cleaning procedure;
- a multilingual MiniLM sentence-transformer baseline that reaches a top-1 accuracy of 0.784 (Wilson 95% interval 0.778–0.790) and a weighted F1 of 0.790;
- per-section metrics and a confusion analysis that characterize head- and tail-section behavior and expose the dominant inter-section confusions.

2 Related Work

Early approaches contrasted document clustering and latent semantic indexing with probabilistic topic models such as Latent Dirichlet Allocation (Blei et al., 2003; Blei, 2012), which represent each article as a mixture over latent topics rather than assigning it to a single cluster. Newman et al. (2006) built on this lineage by coupling statistical topic models with named-entity recognition to jointly model the topics discussed and the persons, organizations, and places mentioned across 330,000 New York Times articles, extracting topics, topic trends, and entity-linking topics. More recent embedding-based methods reframe topic discovery as clustering over contextual document vectors (Grootendorst, 2022). These methods are unsupervised; in contrast, editorial section labels provide direct supervision against a fixed inventory.

Using supervised classification, Daud et al. (2023) categorize real-world Reuters news with

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 438–448
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_045

classical models over TF-IDF features, reporting that a hyperparameter-optimized SVM with a linear kernel outperforms stochastic gradient descent, logistic regression, k-nearest neighbors, random forests, and naive Bayes. Their survey of the area surfaces several recurring findings: support vector machines are strong text categorizers, as in the Reuters-21578 work of Debole and Sebastiani on supervised term weighting; SVM and naive Bayes have been compared for news classification on social-media streams; and frequency-based descriptors such as TF*PDF have been used for topic extraction and tracking over news archives. The review also notes a move toward deep models for some languages (e.g. CNN/LSTM hybrids for Chinese news), feature selection over "bags of topics" rather than words, and neural topic models. Two cross-cutting observations are directly relevant here: much work labelled news "classification" in fact targets fake-news detection rather than genuine topical categorization, and most studies rely on pre-organized benchmark datasets whose clean structure does not reflect noisy real-world news.

Hu (2025) survey small-sample text classification and group existing methods into data augmentation (synonym replacement, back-translation), meta-learning (MAML, prototypical networks), and transfer learning by fine-tuning pre-trained encoders (Devlin et al., 2019), noting that augmentation can inject noise, meta-learning is unstable on text, and fine-tuning needs substantial data and is sensitive to distribution shift. They instead adopt prompt engineering, casting classification as masked language generation, and combine dynamic prompt generation with knowledge distillation from a large pre-trained teacher over unlabeled news, reporting gains in 5/10/20-shot settings and identifying class imbalance as a key difficulty.

The present work sits between these lines. As in the supervised classifiers above, articles are mapped to a fixed label set, but rather than TF-IDF features or a softmax head, a multilingual sentence-transformer (Reimers and Gurevych, 2019, 2020; Wang et al., 2020) is used and classification is cast as embedding-based ranking over section labels. Unlike the balanced, accuracy-only setup of Daud et al. (2023) and the genuinely few-shot regime of Hu (2025), the supervision here is abundant but highly skewed over 35 Serbian sections, so the central difficulty is tail-class precision. We therefore report class-balanced and distribution-weighted metrics and compare the sentence-transformer models with conventional TF–IDF classifiers. Large-scale section classification for Serbian remains comparatively underexplored, although a few related studies address smaller samples. Živanović and Stefanović (2025) investigate topic modeling in newspaper headlines from Serbia, Bosnia and Herzegovina, and Croatia using LDA and NMF models. The study identifies dominant topics and compares thematic similarities across the three countries. Model performance is evaluated using coherence, perplexity, Hellinger distance, and Silhouette score. Batanović et al. (2018) presents the Serbian Semantic Textual Similarity News Corpus (STS.news.sr), a dataset of 1,192 Serbian sentence pairs annotated with fine-grained semantic similarity scores where several STS models are evaluated on the corpus. A WordNet-based approach for Serbian text categorization was proposed (Graovac, 2013), leveraging morphological, syntactic, and semantic information from lexical resources and representative synsets to improve classification performance on Ebart corpus, classifying on 3 classes: politics, economy and sport. Graovac (2014) uses the same corpus for classification with different, n-gram-based methods. This line of work also includes language-independent n-gram-based categorizers with weighting factors (Graovac et al., 2015), a structural SVM over n-gram features for Serbian (Kovačević and Graovac, 2015), and a hierarchical formulation (Graovac et al., 2017) of the kind this work proposes as a coarse-to-fine extension, all on small, balanced datasets with only a few classes. Đukić (2024) explore topic classification of economic newspaper articles for Serbian.

Annotated news corpora are a common way to build foundational resources for lower-resourced languages, most often for named-entity recognition. Ruokolainen et al. (2020) release a Finnish news corpus of 953 articles (193,742 tokens) with six entity classes (organization, location, person, product, event, and date) drawn from a single online outlet, with rule-based and neural baselines on in- and out-of-domain test sets. Yusufu et al. (2025) similarly present an Uzbek news corpus of 500 manually annotated articles (222,536 tokens) with three entity classes (person, location, organization) from a single source, reporting that the pre-trained CINO model outperforms other multilingual encoders. At the opposite scale, Mackenzie et al. (2020) derive CC-News-En, a 44-million-

document English news corpus from Common Crawl collected between 2016 and 2018, primarily for information-retrieval experiments. For Serbian, newspaper text features in the TESLA-NER-NEL corpus (Ikonić Nešić and Utvić, 2024), which provides more than 150,000 annotated sentences with four named-entity classes (person, organisation, location, and role) over 12,216 publicly available news articles drawn from the SrpKor reference corpus (Vitas et al., 2025). In addition, newspaper articles annotated with seven entity classes have been used to train a CNN model for named-entity recognition and linking (NER and NEL) (Ikonić Nešić et al., 2026). The resource introduced here differs from all of these along two axes: it is labeled with editorial sections (topical sections) rather than named entities or IR topics, and it is large (123,000 articles) for a single lower-resourced language, which allows topical section classification and its class-imbalance behavior to be studied directly.

3 Corpus Preparation and Label Normalization

The corpus is a slice of the Ebart media archive for 2023. Digitization, OCR correction, and the assignment of editorial sections (topics) were performed manually by archive staff.

The training dataset is prepared in several steps. The raw corpus index contains one row per newspaper article with the fields `unid`, `file_name`, `naslov` (title), and `rubrika` (section). Raw metadata is read with fallback decoding, and each row is linked to a plain-text article file. The dominant raw encoding is `cp1250`.

Label normalization is deterministic and is applied before any split generation, so that all train, validation, and test statistics are computed over the same normalized label inventory. The pass (i) standardizes known encoding-corrupted (mojibake[1]) spellings of canonical labels, (ii) merges truncated fragment labels into their parent section, (iii) drops a generic non-coherent catch-all label, and (iv) excludes one malformed row with an empty article identifier. Table 1 summarizes the normalized and raw corpus accounting; Tables 2–4 document the individual corrections.

The spelling correction pass standardized 10,316 rows whose canonical labels were corrupted by encoding. Table 2 lists the affected labels, where "[?]" marks a byte that was corrupted (mojibake) in the raw label.

Quantity	Value
Raw metadata rows	123,196
Raw article files on disk	123,635
Linked metadata rows	123,196
Metadata rows missing files	0
Raw orphan article files	439
Raw topics before norm.	43
Normalized rows	123,179
Dropped rows in norm.	17
Canonical spelling rewrites	10,316
Fragment-label merges	28
Final normalized sections	35
Raw encodings	cp1250=122,926, utf-8=270
Raw words/file	min 7, median 298, mean 402.35, p75 517, max 23,005

Table 1: Raw and normalized corpus accounting.

Twenty-eight rows carried truncated fragment labels, isolated content words left over from clipped or corrupted section strings that were deterministically merged into their parent sections. The largest group comprised 11 rows of the fragment *nepogode* ("disasters"), folded into *Nesreće i elementarne nepogode* (Accidents and natural disasters) and raising that section from 4,288 to 4,299 articles. Seven rows of *finansije* ("finance") were merged into *Ekonomija i finansije* (Economy and finance; 3,692 to 3,699), three rows of *informatika* ("IT") into *Telekomunikacije i informatika* (Telecommunications and IT; 600 to 603), and three rows of *potroša[?]a*, a mojibake-truncated form of *potrošača* ("consumers") into *Zaštita prava potrošača* (Consumer rights protection; 174 to 177). The remaining merges were single- or double-row corrections: *rudarstvo* ("mining") into *Energetika i rudarstvo* (Energy and mining; 1,563 to 1,565), *društvene* ("social") into *Ostale društvene teme* (Other social topics; 4,911 to 4,912), and *prava* ("rights") into *Ljudska prava* (Human rights; 440 to 441).

After that, seventeen rows were removed during normalization. Sixteen of them belonged to the generic label *teme* ("topics"), a catch-all bucket whose mixed semantics and low internal coherence made it unsuitable as a research-grade class. The remaining row was a malformed metadata entry, a blank article identifier paired with a placeholder `.txt` file, that could not be traced through preprocessing, splitting, or evaluation. Both were ex-

[1]Mojibake is the corruption of readable text into garbled characters caused by interpreting encoded text with the wrong character encoding.

Raw label	Normalized label	English gloss	Rows
Nesre[?]e i elementarne nepogode	Nesreće i elementarne nepogode	Accidents and natural disasters	4,288
Pravosu[?]e	Pravosuđe	Judiciary	3,264
Saobra[?]aj	Saobraćaj	Traffic / transport	1,839
Ratni zlo[?]ini/Haški tribunal	Ratni zločini/Haški tribunal	War crimes / Hague Tribunal	751
Zaštita prava potroša[?]a	Zaštita prava potrošača	Consumer rights protection	174

Table 2: Canonical spelling corrections for encoding-corrupted section labels.

Category	Item	Rows	Rationale
Dropped section	teme	16	Generic catch-all bucket with mixed semantics and too little internal coherence for a research-grade class.
Malformed row	blank article id, placeholder `.txt`	1	Could not be traced through preprocessing, splitting, or evaluation.

Table 3: Noisy or malformed items removed during normalization.

cluded before any split generation, so they affect neither the normalized label inventory nor the reported statistics.

Several small but semantically coherent sections were deliberately kept as distinct classes rather than merged for balance (Table 4).

Section	English gloss	Rows	Rationale
Civilno društvo	Civil society	23	Small but semantically coherent and distinct from the generic social-topics bucket.
Korupcija	Corruption	312	Corruption coverage is substantively distinct and analytically useful.
EU	European Union	422	EU coverage is narrower than general foreign policy and remains traceable on its own.
Nauka	Science	427	Science content is thematically coherent even if relatively small.
Ljudska prava	Human rights	441	Human-rights coverage is analytically meaningful as its own class.
Religija	Religion	973	Religion is a clear semantic class, not merged solely for balance.
Istorijske teme	Historical topics	1,518	Historical topics form a coherent genre that would be diluted by broader cultural categories.

Table 4: Small coherent sections preserved as distinct classes after normalization.

4 Model and Training

The baseline uses the multilingual MiniLM sentence-transformer[2] (Wang et al., 2020; Reimers and Gurevych, 2019, 2020). It is fine-tuned with a multiple-negatives ranking loss over positive article-text/section pairs, relying on in-batch negatives (Henderson et al., 2017). Validation during training uses an explicit validation pair file with a binary-classification evaluator. At test time, the model ranks all 35 section labels for each held-out article by cosine similarity between ℓ_2-normalized article and section embeddings; the direct-hit classification prediction is the top-ranked section.

We compared the sentence-transformer models with two conventional classifiers trained on the same training split and evaluated on the same held-out test set. Articles were represented by unigram TF–IDF vectors with at most 100,000 features, a minimum document frequency of two, a maximum document frequency of 0.95, and sublinear term-frequency scaling. We trained a Linear SVM and multinomial logistic regression on these representations. Both conventional models perform direct single-label classification over the 35 sections, without ranking section labels.

Because the top-1 prediction of the ranking model corresponds to a single-label decision, micro precision, micro recall, micro F1, and accuracy are numerically identical: each article contributes exactly one gold section and one top-1 predicted section. Macro- and weighted-averaged precision, recall, and F1 are additionally reported.

We also report balanced accuracy, Matthews correlation coefficient (MCC), and Cohen's κ. In this single-label multiclass setting, balanced accuracy equals macro recall: it assigns equal weight to the recall of each section, irrespective of its frequency. This measure complements overall accuracy, which is dominated by the most frequent sections in this highly imbalanced dataset. Equal treatment of sections is relevant to archival indexing and media monitoring because the observed frequency of a section reflects the news cycle and editorial organization rather than its analytical importance. Low-frequency sections such as Corruption, Human rights, Science, Civil society, and Consumer rights may therefore remain important despite their limited test support. We interpret balanced accuracy as a measure of class-balanced thematic coverage and consider it together with accuracy, macro

[2] `paraphrase-multilingual-MiniLM-L12-v2`

Split	Articles	Ratio	Pos. pairs	Neg. pairs	Total pairs	Median words	Mean words
train	86,224	0.6999	86,224	172,448	258,672	298.0	402.2
val	18,468	0.1499	18,468	36,936	55,404	300.0	405.0
test	18,487	0.1501	18,487	36,974	55,461	295.0	400.3

Table 5: Article split sizes, pair counts, and article-length statistics.

F1, weighted F1, MCC, and per-class precision and recall.

For the direct-hit rate, Wilson score intervals (Wilson, 1927) are reported, which are appropriate for binomial proportions.

The normalized index is split at the article level with iterative multilabel stratification (Sechidis et al., 2011; Szymański and Kajdanowicz, 2017), using a 70/15/15 train/validation/test ratio and seed 42. The resulting artifacts are effectively single-label: the average number of sections per article is 1.0 in all three splits. Training data are prepared as article-text/section pairs, with two sampled negative labels per positive article-label pair. Table 5 reports split sizes, pair counts, and article-length statistics; Table 6 reports the section distribution across splits.

The modeling index does not contain newspaper or publication-date fields. Newspaper and date analyses are therefore joined from an auxiliary export and reported with explicit coverage. The join matched 11,782 of 123,179 normalized articles (Table 7); all matched rows with year metadata are from 2023, with 8,264 articles in train, 1,776 in validation, and 1,742 in the test set. The auxiliary join covers twelve distinct national newspapers, each present in all three splits. Because the auxiliary source was sampled at roughly one thousand articles per outlet, the per-newspaper totals are near-uniform (about 980 to 1,000 articles each), and the counts within each split simply track the 70/15/15 ratio. The five most frequent outlets in the training split are `Blic`, `Politika`, `NIN`, `Novi Magazin` and `Vreme`. The set mixes daily newspapers (`Blic`, `Politika`, `Večernje novosti`, `Kurir`, `Alo!`, `Informer`, `Danas`) with weekly news magazines (`NIN`, `Vreme`, `Novi Magazin`, `Nedeljnik`, `Pečat`); this editorial mix matters because section conventions differ across outlets, which plausibly contributes to some of the inter-section overlap observed in the confusion analysis. A sample of the training-data export is shown in Listing 1, which includes the article text together with its newspaper, date, and topic labels.

Listing 1: A sample record from the training-data export.

```
<record n="9" articleId="
    C1258927004CD0B8C125896C0044DDC6" split="train"
     sourceWordCount="111" sourceCharCount="792"
    sampleWordCount="80" sampleCharCount="557"
    sampleSentenceCount="4" truncated="true"
    candidateCountForNewspaper="709">
<newspaper>NIN</newspaper>
<date year="2023" month="2023-02">2023-02-16</date>
<sourceDatabase>Develop\Arhiva2023.nsf</
    sourceDatabase>
<title>FILMOVI DINKA TUCAKOVI?A</title>
<topicMatch>
<goldTopic>Kultura</goldTopic>
<sourceTopic>Kultura</sourceTopic>
<canonicalTopic>Kultura</canonicalTopic>
</topicMatch>
<textPath>articles/C1258927004CD0B8C125896C0044DDC6.
    txt</textPath>
<text encoding="cp1250" role="shortCorpusSample">
    FILMOVI DINKA TUCAKOVI?A KINOTEKA Nakon što je
    ovu nedelju i ciklus „Velikani svetskog filma“
    posvetila Đini Lolobriđidi, prikazavši njene
    filmove Nevernice (1953), Hotel Raj (1966),
    Dobro ve?e, gospo?o Kembel (1968), Zavodnice
    (1965), Žena od slame (1964), Hleb, ljubav i ma
    šta (1953), Carska venera (1962), Hleb, ljubav
    i ljubomora (1954) - Jugoslovenska kinoteka
    obeleži?e desetogodišnjicu smrti Dinka Tucakovi
    ?a. Tako su u nedelju i ponedeljak, 19. i 20.
    februara, na programu ostvarenja The Rubber
    Soul Project (2004), Šest dana juna (1985),
    Zabranjeni</text>
</record>
<record n="10" articleId="
    C1258927004CD0B8C125899B003CF0D0" split="train"
     sourceWordCount="36" sourceCharCount="259"
    sampleWordCount="36" sampleCharCount="259"
    sampleSentenceCount="3" truncated="false"
    candidateCountForNewspaper="706">
<newspaper>Novi Magazin</newspaper>
<date year="2023" month="2023-04">2023-04-20</date>
<sourceDatabase>Develop\Arhiva2023.nsf</
    sourceDatabase>
<title>MOTOCIKLISTI - HUMANITARCI</title>
<topicMatch>
<goldTopic>Religija</goldTopic>
<sourceTopic>Religija</sourceTopic>
<canonicalTopic>Religija</canonicalTopic>
</topicMatch>
<textPath>articles/C1258927004CD0B8C125899B003CF0D0.
    txt</textPath>
<text encoding="cp1250" role="shortCorpusSample">
    MOTOCIKLISTI - HUMANITARCI Mališani iz Niša
    takmi?ila su se u kucanju uskršnjim jajima sa
    motociklistima - učesnicima humanitarne akcije
    „Odaberi svog bajkera i odmeri snage“. Cilj je
    bio prikupljanje novca za lečenje bolesne dece.
     (BETAPHOTO/SASA DJORDJEVIC)</text>
</record>
```

5 Results and Discussion

On the held-out test split of 18,487 articles the model produces 14,502 correct top-1 predictions and 3,985 errors, giving a top-1 accuracy (equivalently micro F1) of 0.784. Table 9 reports the full

Section	English gloss	Full	Train	Val	Test
Civilno društvo	Civil society	23	16	3	4
EU	European Union	422	295	63	64
Ekologija	Environment	1,108	776	166	166
Ekonomija i finansije	Economy and finance	3,699	2,589	555	555
Energetika i rudarstvo	Energy and mining	1,565	1,095	235	235
Estrada/Zabava	Showbiz/entertainment	6,980	4,886	1,047	1,047
Infrastruktura	Infrastructure	1,252	876	188	188
Istorijske teme	Historical topics	1,518	1,063	227	228
Korupcija	Corruption	312	218	47	47
Kosovo	Kosovo	5,877	4,114	881	882
Kriminal	Crime	4,663	3,264	699	700
Kultura	Culture	14,764	10,335	2,214	2,215
Ljudska prava	Human rights	441	309	66	66
Mediji	Media	1,681	1,177	252	252
Nauka	Science	427	299	64	64
Nesreće i elem. nepogode	Accidents and natural disasters	4,299	3,009	645	645
Obrazovanje	Education	2,743	1,920	411	412
Ostale društvene teme	Other social topics	4,912	3,438	737	737
Policija	Police	3,331	2,332	499	500
Politika	Politics	8,489	5,942	1,273	1,274
Poljoprivreda	Agriculture	1,629	1,140	244	245
Pravosuđe	Judiciary	3,264	2,285	489	490
Privreda	Industry and commerce	2,048	1,434	307	307
Ratni zločini/Haški trib.	War crimes / Hague Tribunal	751	526	112	113
Region	Regional (Balkans)	3,831	2,682	574	575
Religija	Religion	973	681	146	146
Saobraćaj	Traffic / transport	1,839	1,287	276	276
Socijalna politika	Social policy	725	507	109	109
Spoljna politika	Foreign policy	2,830	1,981	424	425
Sport	Sports	16,175	11,322	2,426	2,427
Svet	World	15,761	11,033	2,364	2,364
Telekom. i informatika	Telecommunications and IT	603	422	90	91
Vojska	Military	751	526	112	113
Zaštita prava potrošača	Consumer rights protection	177	124	26	27
Zdravstvo	Healthcare	3,316	2,321	497	498

Table 6: Section distribution in the full normalized corpus and the train/validation/test splits, with an English gloss for each section.

Split	Matched	Unmatched	Coverage
train	8,264	77,960	0.0958
val	1,776	16,692	0.0962
test	1,742	16,745	0.0942

Table 7: Coverage of auxiliary newspaper/date metadata joined by article identifier.

set of aggregate metrics. The weighted F1 (0.790) exceeds the macro F1 (0.653), reflecting strong performance on frequent sections and weaker performance on small ones.

Table 8 compares the sentence-transformer models with the conventional classifiers. The TF–IDF Linear SVM achieves the highest accuracy (0.815), macro F1 (0.713), weighted F1 (0.812), and MCC (0.801). MiniLM achieves the highest balanced accuracy (0.734), compared with 0.694 for the Linear SVM. The Linear SVM therefore performs better under metrics influenced by the observed class distribution, whereas MiniLM has the highest mean recall across the 35 sections.

Table 10 reports per-section support, predictions, true/false positives and negatives, precision, recall, F1, and a Wilson 95% interval on recall. The macro recall (0.734) exceeds the macro precision (0.618), indicating that the model tends to over-predict minority sections: for many small classes the number of predictions far exceeds the support (e.g. EU, with support 64 but 156 predictions), which inflates recall while depressing precision.

Table 11 lists the largest top-1 confusions. The dominant errors form interpretable clusters: a law-enforcement cluster in which *Kriminal* (Crime), *Policija* (Police), and *Pravosuđe* (Judiciary) are mutually confused; a cultural overlap in which *Kultura* (Culture) is drawn toward *Estrada/Zabava* (Showbiz/entertainment) and *Istorijske teme* (Historical topics); and a political adjacency in which *Politika* (Politics) is drawn toward *Kosovo*, *Spoljna politika*

Model	Accuracy	Bal. acc.	Macro P	Macro R	Macro F1	Weighted F1	MCC
TF–IDF Linear SVM	**0.815**	0.694	0.741	0.694	**0.713**	**0.812**	**0.801**
TF–IDF logistic regression	0.802	0.616	**0.751**	0.616	0.657	0.797	0.786
MiniLM	0.784	**0.734**	0.618	**0.734**	0.653	0.790	0.770
GTE-512	0.722	0.639	0.534	0.639	0.555	0.736	0.705

Table 8: Aggregate single-label classification results on the common held-out test set. Balanced accuracy equals macro recall in this setting. Bold indicates the highest value in each column.

Metric	Value
Evaluated articles	18,487
Correct top-1 predictions	14,502
Top-1 errors	3,985
Micro precision / recall / F1	0.784
Wilson 95% interval	0.778–0.790
Macro precision	0.618
Macro recall	0.734
Macro F1	0.653
Weighted precision	0.805
Weighted recall	0.784
Weighted F1	0.790
Ranking evaluator Hit@1	0.784

Table 9: Held-out test results when the ranking model is evaluated as a direct top-1 section classifier.

(Foreign policy), and *Mediji* (Media). Several confusions reflect genuine semantic proximity, such as the mutual confusion between *Saobraćaj* (Traffic/-transport) and *Infrastruktura* (Infrastructure), and the tendency for *Spoljna politika* (Foreign policy) to be predicted as *EU* (European Union).

The Linear SVM's overall accuracy advantage is concentrated in several frequent sections. Differences in *Culture*, *World*, *Politics*, and *Crime* account for approximately 97% of its net accuracy advantage over MiniLM. By contrast, MiniLM has higher recall in 19 of the 35 sections. Its largest recall advantages occur for *Corruption* (+0.298), *EU* (+0.281), *Infrastructure* (+0.202), *Human rights* (+0.121), *Consumer rights protection* (+0.111), and *Science* (+0.109). This pattern is consistent with MiniLM's higher balanced accuracy.

Higher minority-class recall is generally accompanied by lower precision. Relative to the Linear SVM, MiniLM increases recall for *Corruption* by 0.298 but decreases precision by 0.318. For *EU*, recall increases by 0.281 and precision decreases by 0.309. The Linear SVM has higher per-class F1 in 31 of the 35 sections; MiniLM has higher F1 for *Infrastructure*, *Region*, *Judiciary*, and *Economy and finance*. Thus, the Linear SVM has stronger overall and class-level F1 performance, while MiniLM has higher mean class recall.

The baseline is strong on high-frequency sections: *Sport* (Sports; F1 0.965), *Svet* (World; 0.905), and *Kultura* (Culture; 0.865) are all reliably separated. The weakest classes are small and semantically adjacent to larger ones; for example, *Civilno društvo* (Civil society; support 4, precision 0.094), *Zaštita prava potrošača* (Consumer rights protection; precision 0.228), and *EU* (European Union; precision 0.282) are heavily over-predicted. The gap between macro recall and macro precision quantifies this tendency: the ranking objective favors recall on rare labels at the cost of precision, since small sections receive far more top-1 predictions than their true support.

Two regularities make this concrete. First, top-1 quality is strongly governed by support: every section with more than 2,000 test instances scores F1 ≥ 0.865 (Sport, Svet, Kultura), whereas nine of the ten smallest sections (support below 120) fall below 0.60.Second, the over-prediction is systematic rather than incidental: for almost every minority section the model assigns the label far more often than it actually occurs. *Civilno društvo* (Civil society) is predicted eight times as often as its true support, *Zaštita prava potrošača* (Consumer rights protection) nearly three times, and EU about 2.4 times, whereas the head sections are predicted slightly less often than their support (Sport and Kultura at 0.97 and 0.88 of their support, respectively). The ranking objective therefore redistributes decision mass toward rare labels, which is precisely what pushes macro recall (0.734) above macro precision (0.618) and opens the 0.137-point gap between weighted F1 (0.790) and macro F1 (0.653).

The confusion structure is largely interpretable and points to editorial overlap rather than arbitrary error. Law-enforcement sections, cultural subgenres, and political topics are the principal sources of confusion, all of which are areas where newspaper section boundaries are themselves fuzzy.

Two practical consequences follow from this structure. First, the largest confusions are bidi-

Section	English gloss	Support	Pred.	TP	FP	FN	Prec.	Rec.	F1	Rec. 95% CI
Sport	Sports	2,427	2,358	2,309	49	118	0.979	0.951	0.965	0.942–0.959
Svet	World	2,364	2,181	2,056	125	308	0.943	0.870	0.905	0.856–0.883
Kultura	Culture	2,215	1,954	1,804	150	411	0.923	0.814	0.865	0.798–0.830
Politika	Politics	1,274	1,049	893	156	381	0.851	0.701	0.769	0.675–0.725
Estrada/Zabava	Showbiz/entertainment	1,047	1,115	944	171	103	0.847	0.902	0.873	0.882–0.918
Kosovo	Kosovo	882	859	736	123	146	0.857	0.834	0.845	0.808–0.858
Ostale društvene teme	Other social topics	737	574	331	243	406	0.577	0.449	0.505	0.414–0.485
Kriminal	Crime	700	607	410	197	290	0.675	0.586	0.627	0.549–0.622
Nesreće i elem. nepogode	Accidents and natural disasters	645	641	514	127	131	0.802	0.797	0.799	0.764–0.826
Region	Regional (Balkans)	575	634	487	147	88	0.768	0.847	0.806	0.815–0.874
Ekonomija i finansije	Economy and finance	555	446	360	86	195	0.807	0.649	0.719	0.608–0.687
Policija	Police	500	490	301	189	199	0.614	0.602	0.608	0.558–0.644
Zdravstvo	Healthcare	498	504	413	91	85	0.819	0.829	0.824	0.794–0.860
Pravosuđe	Judiciary	490	491	367	124	123	0.747	0.749	0.748	0.709–0.785
Spoljna politika	Foreign policy	425	413	263	150	162	0.637	0.619	0.628	0.572–0.664
Obrazovanje	Education	412	475	338	137	74	0.712	0.820	0.762	0.780–0.854
Privreda	Industry and commerce	307	335	196	139	111	0.585	0.638	0.611	0.583–0.690
Saobraćaj	Traffic / transport	276	304	194	110	82	0.638	0.703	0.669	0.646–0.754
Mediji	Media	252	305	183	122	69	0.600	0.726	0.657	0.668–0.778
Poljoprivreda	Agriculture	245	246	189	57	56	0.768	0.771	0.770	0.715–0.820
Energetika i rudarstvo	Energy and mining	235	269	194	75	41	0.721	0.826	0.770	0.772–0.869
Istorijske teme	Historical topics	228	302	163	139	65	0.540	0.715	0.615	0.653–0.770
Infrastruktura	Infrastructure	188	286	135	151	53	0.472	0.718	0.570	0.650–0.778
Ekologija	Environment	166	210	123	87	43	0.586	0.741	0.654	0.669–0.802
Religija	Religion	146	203	113	90	33	0.557	0.774	0.648	0.700–0.834
Ratni zločini/Haški trib.	War crimes / Hague Tribunal	113	151	89	62	24	0.589	0.788	0.674	0.703–0.853
Vojska	Military	113	169	84	85	29	0.497	0.743	0.596	0.656–0.815
Socijalna politika	Social policy	109	169	56	113	53	0.331	0.514	0.403	0.421–0.606
Telekom. i informatika	Telecommunications and IT	91	135	63	72	28	0.467	0.692	0.558	0.591–0.778
Ljudska prava	Human rights	66	120	40	80	26	0.333	0.606	0.430	0.485–0.715
EU	European Union	64	156	44	112	20	0.282	0.688	0.400	0.566–0.788
Nauka	Science	64	127	51	76	13	0.402	0.797	0.534	0.683–0.877
Korupcija	Corruption	47	98	38	60	9	0.388	0.809	0.524	0.675–0.896
Zaštita prava potrošača	Consumer rights protection	27	79	18	61	9	0.228	0.667	0.340	0.478–0.814
Civilno društvo	Civil society	4	32	3	29	1	0.094	0.750	0.167	0.301–0.954

Table 10: Per-section direct-hit classification metrics on the held-out test split.

Gold section	Predicted section	Count	Gold support	Rate within gold
Kriminal (Crime)	Policija (Police)	100	700	0.1429
Kultura (Culture)	Estrada/Zabava (Showbiz/entertainment)	96	2,215	0.0433
Policija (Police)	Kriminal (Crime)	82	500	0.1640
Kriminal (Crime)	Pravosuđe (Judiciary)	57	700	0.0814
Kultura (Culture)	Istorijske teme (Historical topics)	56	2,215	0.0253
Politika (Politics)	Kosovo	47	1,274	0.0369
Ekonomija i finansije (Economy & finance)	Privreda (Industry & commerce)	44	555	0.0793
Kultura (Culture)	Ostale društvene teme (Other social topics)	44	2,215	0.0199
Ostale društvene teme (Other social topics)	Socijalna politika (Social policy)	43	737	0.0583
Politika (Politics)	Spoljna politika (Foreign policy)	43	1,274	0.0338
Estrada/Zabava (Showbiz/entertainment)	Kultura (Culture)	34	1,047	0.0325
Politika (Politics)	Mediji (Media)	34	1,274	0.0267
Kosovo	Spoljna politika (Foreign policy)	33	882	0.0374
Svet (World)	Regional (Balkans)	31	2,364	0.0131
Ostale društvene teme (Other social topics)	Infrastruktura (Infrastructure)	30	737	0.0407
Spoljna politika (Foreign policy)	EU (European Union)	30	425	0.0706
Saobraćaj (Transport)	Infrastruktura (Infrastructure)	29	276	0.1051
Kriminal (Crime)	Nesreće i elem. nepogode (Accidents & disasters)	28	700	0.0400
Politika (Politics)	Ostale društvene teme (Other social topics)	28	1,274	0.0220
Pravosuđe (Judiciary)	Kriminal (Crime)	28	490	0.0571
Ekonomija i finansije (Economy & finance)	Energetika i rudarstvo (Energy & mining)	25	555	0.0450
Nesreće i elem. nepogode (Accidents & disasters)	Kriminal (Crime)	25	645	0.0388
Ostale društvene teme (Other social topics)	Privreda (Industry & commerce)	25	737	0.0339
Spoljna politika (Foreign policy)	Kosovo	25	425	0.0588
Svet (World)	EU (European Union)	24	2,364	0.0102
Policija (Police)	Pravosuđe (Judiciary)	23	500	0.0460
Svet (World)	Spoljna politika (Foreign policy)	23	2,364	0.0097
Infrastruktura (Infrastructure)	Saobraćaj (Transport)	22	188	0.1170
Kultura (Culture)	Mediji (Media)	22	2,215	0.0099
Spoljna politika (Foreign policy)	Politika (Politics)	22	425	0.0518

Table 11: Largest top-1 confusions in the held-out test split.

rectional and occur between editorially adjacent sections: Kriminal (Crime) and Policija (Police) are confused in both directions, as are Saobraćaj (Traffic/transport) and Infrastruktura (Infrastructure), and Kultura (Culture) and Estrada/Zabava (Showbiz/entertainment). The embedding space therefore already places semantically related sections close together, and the model fails mainly where human section boundaries are themselves porous; the residual error is dominated by genuine label ambiguity rather than representational weakness. Second, because the head sections are already precision-strong, per-class decision thresholds or calibration targeted at the over-predicted tail (EU, Civilno društvo (Civil society), Zaštita prava potrošača (Consumer rights protection), and Socijalna politika (Social policy)) should raise macro precision and macro F1 at little cost to the head. The consistently symmetric confusions are, in turn, natural candidates for a coarse-to-fine label hierarchy or selective merging that would cut avoidable inter-section leakage while preserving the analytically useful small classes.

The final training-time validation row corresponds to epoch 5.0, with pairwise validation accuracy 0.95 and pairwise validation F1 0.92; the reported average precision and Matthews correlation were 0.0. These are pairwise validation diagnostics on the binary-classification evaluator, not the final held-out top-1 classification metrics, and should not be interpreted as the model's classification performance.

6 Conclusion

A documented label-normalization pipeline and a multilingual sentence-transformer baseline for section classification of Serbian newspaper articles are introduced, addressing a task that has so far been largely unaddressed at this scale for the language. By casting section assignment as embedding-based ranking and evaluating the top-ranked label as a direct classification decision, the model reaches a top-1 accuracy of 0.784 (Wilson 95% interval 0.778–0.790) and a weighted F1 of 0.790 over 35 sections and 18,487 held-out articles, with reliable separation of the high-frequency sections (F1 up to 0.965).

The comparison with conventional classifiers shows that unigram TF–IDF with a Linear SVM is a strong baseline for this task. It outperforms MiniLM in accuracy, macro F1, weighted F1, and MCC. MiniLM does not outperform the Linear SVM overall, but it achieves higher balanced accuracy and higher recall in 19 of the 35 sections, including several low-frequency sections. The two models therefore differ in their error distribution: the Linear SVM yields better overall and per-class F1 performance, whereas MiniLM yields higher mean recall across sections.

Beyond these headline numbers, the analysis isolates a single, actionable failure mode: performance is governed almost entirely by class support, and the ranking objective systematically over-predicts the long tail, which is what separates the weighted F1 (0.790) from the macro F1 (0.653). Crucially, the residual errors are not random; rather, they are concentrated in bidirectional confusions between editorially adjacent sections. This pattern suggests that the embedding space already captures the underlying topical structure, and that the remaining scope for improvement is constrained primarily by genuine label ambiguity rather than by limitations in the learned representations. This diagnosis directly motivates the most promising next steps: per-class thresholding or calibration on the over-predicted tail, class re-balancing, and a coarse-to-fine or explicitly multilabel formulation for articles that legitimately span several sections. The fine-tuned model, together with the per-section metrics and the documented cleaning corrections, will be released with the final version of this paper, the underlying newspaper corpus itself cannot be made publicly available under the terms of the source archive.

Acknowledgment

This research was supported by the Science Fund of the Republic of Serbia, #7276, *Text Embeddings - Serbian Language Applications - TESLA*. The authors gratefully acknowledge Media Archive Ebart for providing access to valuable resources that were essential to the completion of this research.

References

Vuk Batanović, Miloš Cvetanović, and Boško Nikolić. 2018. Fine-grained semantic textual similarity for Serbian. In *Proc. of 11th Conference on Language Resources and Evaluation (LREC 2018)*, pages 1370–1378.

David M. Blei. 2012. Probabilistic Topic Models. *Communications of the ACM*, 55(4):77–84.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Shahzada Daud, Muti Ullah, Amjad Rehman, Tanzila Saba, Robertas Damaševičius, and Abdul Sattar. 2023. Topic Classification of Online News Articles Using Optimized Machine Learning Models. *Computers*, 12(1):16.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *2019 Conf. of the North American Chapter of ACL: HLT (NAACL-HLT)*, pages 4171–4186.

Jelena Graovac. 2013. Wordnet-Based Serbian Text Categorization. *Infotheca - Journal for Digital Humanities*, 14(2):2a–17a.

Jelena Graovac. 2014. Text categorization using n-gram based language independent technique. *Intelligent Data Analysis*, 18(4):677–695.

Jelena Graovac, Jovana Kovačević, and Gordana Pavlović-Lažetić. 2015. Language Independent n-Gram-Based Text Categorization with Weighting Factors: A Case Study. *Journal of Information and Data Management (JIDM)*, 6(1):4–17.

Jelena Graovac, Jovana Kovačević, and Gordana Pavlović-Lažetić. 2017. Hierarchical vs. flat n-gram-based text categorization: can we do better? *Computer Science and Information Systems (ComSIS)*, 14(1):103–121.

Maarten Grootendorst. 2022. BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure. *arXiv preprint arXiv:2203.05794*.

Matthew Henderson, Rami Al-Rfou, and Brian Strope et. all. 2017. Efficient Natural Language Response Suggestion for Smart Reply. *arXiv:1705.00652*.

Liao Hu. 2025. Topic Classification of Small Sample News Based on Prompt Engineering. In *Proceedings of the 7th International Conference on Computing and Data Science*, Applied and Computational Engineering, pages 101–107.

Milica Ikonić Nešić, Saša Petalinkar, Olivera Kitanović, Ranka Stanković, and Miloš Utvić. 2026. CNN-based named entity linking: Serbian use case. *Poznan Studies in Contemporary Linguistics*, 62(2):249–280.

Milica Ikonić Nešić and Miloš Utvić. 2024. Overview of the Tesla-Ner-Nel-Gold Dataset: Showcase on Serbian-English Parallel Corpus. In *South Slavic Languages in the Digital Environment JuDig: Book of Abstracts*, page 57.

Jovana Kovačević and Jelena Graovac. 2015. Application of a Structural Support Vector Machine method to N-gram based text classification in Serbian. *INFOtheca - Journal of Information and Library Science*, 16(1).

Joel Mackenzie, Rodger Benham, and Matthias Petri et al. 2020. CC-News-En: A Large English News Corpus. In *Proc.s of 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 3077–3084. ACM.

David Newman, Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. 2006. Analyzing Entities and Topics in News Articles Using Statistical Topic Models. In *Intelligence and Security Informatics (ISI 2006)*, volume 3975 of *LNCS*, pages 93–104. Springer.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proc. of 2019 Conference on Empirical Methods in NLP and the 9th International Joint Conference on NLP (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. ACL.

Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proc. of 2020 Conference on Empirical Methods in NLP (EMNLP)*, pages 4512–4525. ACL.

Teemu Ruokolainen, Pekka Kauppinen, Miikka Silfverberg, and Krister Lindén. 2020. A Finnish news corpus for named entity recognition. *Language Resources and Evaluation*, 54(1):247–272.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the Stratification of Multi-Label Data. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2011)*, volume 6913 of *LNCS*, pages 145–158. Springer.

Piotr Szymański and Tomasz Kajdanowicz. 2017. A Network Perspective on Stratification of Multi-Label Data. In *Proc. of 1st International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA)*, volume 74, pages 22–35.

Mirko Đukić. 2024. Topic classification of economic newspaper articles in a highly inflectional language–the case of Serbia. Technical report, Working Papers Bulletin of the National Bank of Serbia, VI.

Duško Vitas, Ranka Stanković, and Cvetana Krstev. 2025. The Many Faces of SrpKor [O familiji korpusa savremenog srpskog jezika SrpKor]. In *South Slavic Languages in the Digital Environment JuDig Book of Abstracts*, pages 3–27.

Wenhui Wang, Furu Wei, and Li Dong et al. 2020. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In *Also in Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

Edwin B. Wilson. 1927. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, 22(158):209–212.

Aizihaierjiang Yusufu, Kamran Aziz, Aizierguli Yusufu, Abidan Ainiwaer, Fei Li, and Donghong Ji. 2025. Uzbek news corpus for named entity recognition: A. Yusufu et al. *Language resources and evaluation*, 59(3):3139–3152.

Marko M Živanović and Nenad Stefanović. 2025. Processing Balkan News Articles Using Selected Methods of Natural Language Processing. In *24th Int. Symposium INFOTEH-JAHORINA*, pages 11–16.

Wasm: A Pipeline for Constructing Structured Arabic Interleaved Multimodal Corpora

Khalil Hennara
Misraj AI
Khobar, Saudi Arabia
hennara@misraj.ai

Ahmad Bastati
Misraj AI
Khobar, Saudi Arabia
bastati@misraj.ai

Muhammad Hreden
Misraj AI
Khobar, Saudi Arabia
hreden@misraj.ai

Mohamed Motasim Hamed
Misraj AI
Khobar, Saudi Arabia
hamed@misraj.ai

Zeina Aldallal
Misraj AI
Khobar, Saudi Arabia
aldallal@misraj.ai

Sara Chrouf
Misraj AI
Khobar, Saudi Arabia
chrouf@misraj.ai

Safwan AlModhayan
Misraj AI
Khobar, Saudi Arabia
safwan@misraj.ai

Abstract

The performance of large language models (LLMs) and large multimodal models (LMMs) depends heavily on the quality and scale of their pre-training datasets. Recent research shows that LMMs trained on natural documents where images and text are interleaved outperform those trained only on image–text pairs across a wide range of benchmarks, leveraging advanced pre-trained models to enforce semantic alignment, image-sequence consistency, and textual coherence. For Arabic, however, the lack of high-quality multimodal datasets that preserve document structure has limited progress. In this paper, we present our pipeline ***Wasm***[†] for processing the Common Crawl[*] dataset to create a new Arabic multimodal dataset that uniquely provides markdown output. Unlike existing Arabic corpora that focus solely on text extraction, our approach preserves the structural integrity of web content while maintaining flexibility for both text-only and multimodal pre-training scenarios. We provide a comprehensive comparative analysis of our data processing pipeline against those used for major existing datasets, highlighting the convergences in filtering strategies and justifying our specific design choices. To support future research, we publicly release a representative dataset dump along with the full pipeline implementation.

1 Introduction

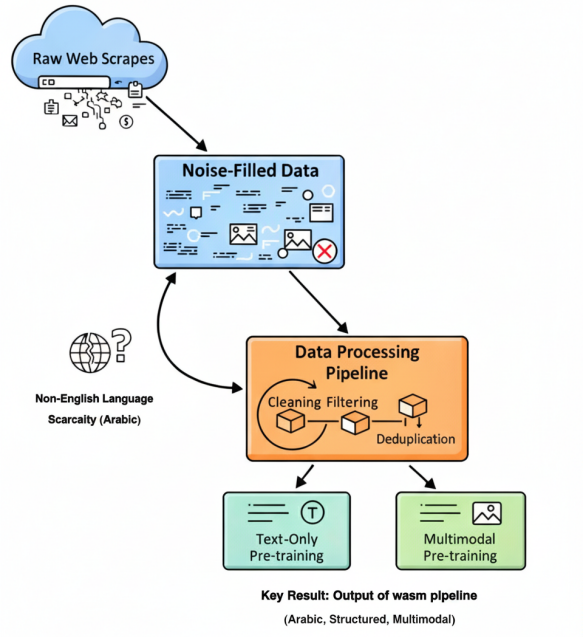


Figure 1: High-level overview of the *Wasm* pipeline, from Common Crawl WARC extraction to the final structured, Markdown-formatted output.

Arabic presents a compelling case study in low-resource Natural Language Processing (NLP). Despite its status as a major global language, high-

[†]**Wasm**: meaning 'tag' or'mark,' which reflects the unique preservation of web markup structures in our data set.
[*]https://commoncrawl.org

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 449–458
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_046

quality structured web corpora, particularly multimodal datasets, remain scarce. While data processing pipelines for English have matured considerably, Arabic-specific infrastructure and curation efforts have lagged behind, creating a significant bottleneck for regional AI development.

Recent studies emphasize that interleaved image–text data, which preserves the natural sequence of textual and visual elements within documents, is essential for training advanced multimodal models (Zhu et al., 2023; Laurençon et al., 2023; Li et al., 2024; Awadalla et al., 2024; Chen et al., 2025). Unlike isolated caption–image pairs, interleaved corpora capture document-level structure, allowing models to learn long-range dependencies, maintain narrative coherence, and align images with text across multiple segments or pages. This structure enables richer multimodal reasoning, such as following temporal or story-driven progressions, comparing various images within a context, and establishing complex instructions in visual evidence.

In contrast, existing Arabic resources (Aloui et al., 2024; Farhat et al., 2024; Nguyen et al., 2023) emphasize plain text extraction, discarding valuable structural cues such as document layout, formatting, and image associations, precisely the elements that interleaved corpora preserve and leverage for training advanced multimodal models. A recent effort, AMCrawl (Aboukozzana et al., 2025), addresses this gap by constructing a large-scale Arabic interleaved image-text corpus together with image-text pairs; however, it flattens documents into linear image-text sequences and does not preserve the underlying document structure (e.g., headings, lists, tables) or provide a structured, Markdown-based representation of the source pages.

In this paper, we present **Wasm**, an Arabic pipeline that produces both interleaved multimodal datasets and text-only corpora while fully preserving document structure Figure 1. To our knowledge, *Wasm* is the first Arabic pipeline to jointly preserve document-level structure (headings, lists, tables) and represent it as Markdown, offering a single output format that can be flexibly degraded into text-only or interleaved multimodal training data. *Wasm* retains document-level structure and the natural interleaving of text and images as they appear on their respective webpage, preserving layout cues and image–text associations needed for multimodal training. Figure 2 illustrates the structured output produced by our pipeline. Our work builds upon the OBELICS framework (Laurençon et al., 2023), adapting and extending it for Arabic to form an optimized pre-processing pipeline tailored to Arabic web data and multimodal use cases. Through systematic analysis of filtering techniques, corpus origins, and computational requirements across these datasets, our objective was to establish best practices for Arabic corpus construction and identify the most effective pre-processing strategies for multilingual and multimodal applications.

Our primary contributions are as follows.

1. We introduce *Wasm*, a new framework that processes and produces Arabic datasets. To our knowledge, it is the only Arabic pipeline that preserves structural information as Markdown while maintaining flexibility between text-only and interleaved multimodal outputs.

2. We release a part of the Wasm pipeline as open source[1], providing the Arabic-adapted module for interleaved multimodal unformatted data extraction. This component enables reproducible construction of Arabic web corpora with preserved text–image sequences and can serve as a foundation for further dataset development.

3. We provide a comprehensive comparative analysis of our data processing pipeline against those used for major existing corpora, highlighting the convergences in filtering strategies and justifying our specific design choices.

4. We release a Misraj Structured Data Dump (MSDD) [2] processed by *Wasm*.

2 Related Work

The creation of large-scale, high-quality datasets has been fundamental to recent breakthroughs in LLMs and LMMs. Dataset curation methodologies have evolved from simple crawling approaches to sophisticated multi-stage pipelines that balance scale with quality. This section reviews key developments in the construction of text-only and multimodal datasets, with particular attention to Arabic resources.

[1] https://github.com/misraj-ai/wasm
[2] https://huggingface.co/datasets/Misraj/msdd

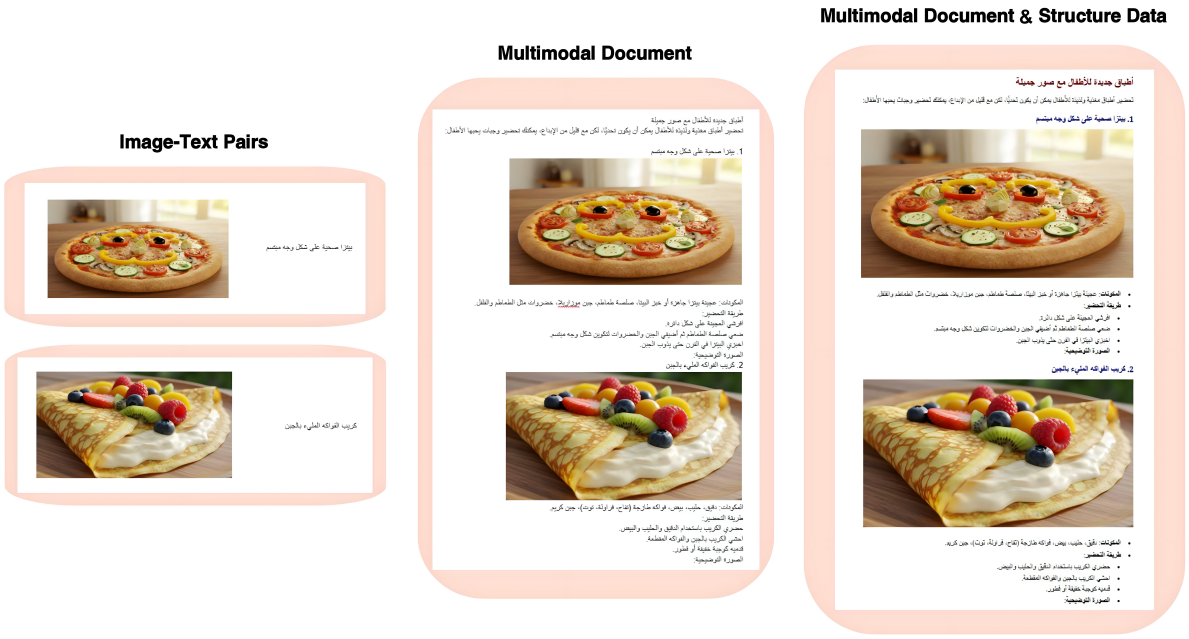


Figure 2: Example of the structured Markdown output produced by *Wasm* for a single web page, showing preserved headings, paragraphs, lists, and interleaved image placement.

2.1 Text-Only Corpora

Early large-scale text datasets established foundational principles for web data curation that remain influential today. The ROOTS corpus (Laurençon et al., 2022), developed for training BLOOM (Le Scao et al., 2023), exemplifies the construction of multilingual datasets by combining diverse sources such as Oscar (Ortiz Suárez et al., 2019), Common crawls, GitHub [3] and more, with extensive source-specific filtering and deduplication.

Subsequent work has focused on scaling multilingual coverage while refining quality metrics. For example, CulturaX (Nguyen et al., 2023) aggregates the dumps of mC4 (Xue et al., 2020) and mOSCAR (Futeral et al., 2024) while incorporating perplexity-based scoring, repetition ratios, and confidence thresholds for language identification. These techniques have since become standard in large-scale corpus construction.

FineWeb (Penedo et al., 2024) represents a recent state-of-the-art corpus. Building on the Massive Text pipeline, it combines neural quality classifiers with custom heuristics and introduces per-dump MinHash deduplication, shown to outperform global deduplication in preserving quality at scale. FineWeb2 (Penedo et al., 2025) generalizes this pipeline to more than 1000 languages, introducing a scalable multilingual dataset construction framework with language-adaptive filtering and deduplication. In addition, it proposes a principled rebalancing strategy based on duplication and quality metrics, yielding improved downstream model performance across diverse languages.

Arabic-specific curation efforts highlight how pre-processing pipelines adapt to the challenges of non-English web data. The 101 Billion Arabic Words corpus (Aloui et al., 2024) applied URL filtering, normalization (e.g., Unicode standardization), and document-level deduplication to Common Crawl WET [4] files, while ArabicWeb24 (Farhat et al., 2024) incorporated more advanced methods such as Gopher quality filtering (Rae et al., 2021) and MinHash-based deduplication. However, like all existing Arabic text-only resources, they remain restricted to plain-text extraction, discarding structural and multimodal information present in native web documents.

2.2 Multimodal Corpora

The evolution toward Vision Language Models (VLMs) has necessitated datasets that preserve the rich contextual relationships between textual and visual content as they naturally occur on the web.

[3] https://github.com/

[4] These files are a subset of the WARC (Web ARChive) format, specifically holding "extracted plaintext".

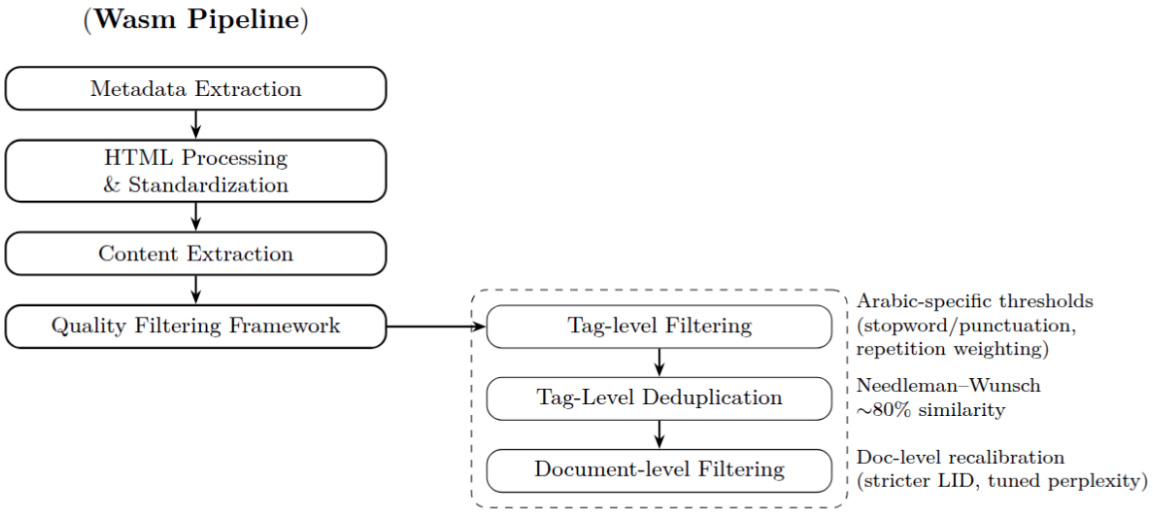


Figure 3: Detailed view of the *Wasm* pipeline stages: metadata extraction, HTML processing and standardization, content structuring into Markdown, and multilevel (tag- and document-level) quality filtering.

This requirement has driven methodological innovations beyond the simple extraction of image-text pairs. Early large-scale multimodal datasets, such as LAION-400M (Schuhmann et al., 2021) and LAION-5B (Schuhmann et al., 2022), were mainly based on large-scale capture of image capture pairs from Common Crawl, filtered using CLIP model (Radford et al., 2021). Although these resources proved invaluable for scaling multimodal training, their pair-based design eliminates broader document context and structural information.

In contrast, more recent work has focused on interleaved multimodal datasets that retain the sequential and structural interplay of text and images within documents. MMC4 (Zhu et al., 2023) advanced beyond basic co-occurrence by incorporating image-text alignment scores and document-level quality metrics into filtering pipelines. This work established the importance of semantic coherence in the construction of multimodal datasets. OBELICS (Laurençon et al., 2023) marked a further paradigm shift by prioritizing structural preservation: Rather than isolating image-text pairs, OBELICS maintains the interleaved document structure found in web content, creating a corpus of 141 million documents where images and text retain their natural sequential relationships. The processing pipeline operates at both document and HTML node levels, with final output representing documents as coherent sequences of text tokens and contextually-positioned images. Methodologically, OBELICS and MMC4 represent complementary approaches to HTML processing: OBELICS leverages the Document Object Model (DOM) tree structure for comprehensive content filtering, while MMC4 uses HTML primarily for image location and integration with existing text corpora like C4, highlighting the trade-off between structural fidelity and processing efficiency.

More recently, OmniCorpus (Li et al., 2024) has pushed this paradigm further by introducing a 10-billion-level interleaved dataset, incorporating more diverse sources, including English and non-English web domains, as well as video-centric sites, and offers flexible formatting that can be degraded into pure text corpora or image-text pairs.

MNiT-1T (Awadalla et al., 2024) extends this trajectory by introducing a trillion-token multilingual multimodal corpus constructed from Common Crawl, including new source types such as PDFs and arXiv papers. It combines large-scale image-text pairing with document-level filtering and quality heuristics for robust VLM pre-training.

Complementing these scale-focused efforts, CoMM (Chen et al., 2025) addresses the qualitative limitations of existing interleaved datasets. It introduces a coherent interleaved multimodal corpus, applying multiperspective filters (text coherence, image sequence consistency, image-text alignment) that improve the quality of interleaved training data and enhance the in-context learning capabilities of multimodal LLMs.

For Arabic specifically, early multimodal resources were mainly based on pairwise, often translated, image-text datasets (Alwajih et al., 2024). More recently, AMCrawl (Aboukozzana et al.,

2025) introduced a web-scale Arabic interleaved image-text corpus of 2.8 million documents together with 5 million QA pairs, and used it to train a native Arabic vision-language model. AMCrawl demonstrates the feasibility of large-scale Arabic interleaved corpus construction, but, like MMC4, OBELICS, OmniCorpus, MINT-1T, and CoMM, it represents documents as flattened sequences of text and images rather than preserving the original structural hierarchy (headings, lists, tables) of the source page. This gap motivates *Wasm*, which, to our knowledge, is the first Arabic framework to jointly preserve full document structure and provide a Markdown-based representation that supports both text-only and interleaved multimodal pre-training.

3 Wasm Pipeline

This section details the comprehensive data processing pipeline developed to create a structured Arabic dataset. Our pipeline closely follows the methodology established by OBELICS (Laurençon et al., 2023), adapted specifically for Arabic language processing requirements. Figure 3 provides an overview of the pipeline architecture. All steps in the pipeline were iteratively tuned and refined through careful observation of the outputs on large samples of the data, ensuring that filtering thresholds, normalization rules, and structural preservation techniques were well-suited to the characteristics of Arabic web content.

3.1 Metadata Extraction

The initial phase involves extracting metadata by filtering web pages containing Arabic language content (not necessarily exclusively Arabic) from selected Common Crawl dumps. Each dump was processed separately to facilitate efficient separation and deduplication. We perform a language-based filtering before loading the Web Archive (WARC) files. This early intervention not only ensures that irrelevant data is excluded at the source but also saves substantial computational resources, including time, memory, and storage. Each scraped article is then associated with a set of metadata that facilitates storage, retrieval, and analysis. This metadata includes the URL of the article, the storage location within the corresponding Common Crawl snapshot as WARC files, and the byte position indicating where the webpage begins within that file. In addition, it records the size of the webpage content in bytes, the detected language(s) of the webpage, and the domain name of the source website.

3.2 HTML Processing and Standardization

Using the extracted metadata (filename, offset, and length), the corresponding WARC files were accessed to retrieve the raw webpage content, which was subsequently converted to HTML format. To reduce noise and improve data quality, several preprocessing filters were applied. First, repeated line breaks and whitespace were normalized into single instances to ensure consistency in text formatting. The HTML comments, which do not have semantic value, were then eliminated. Finally, all CSS-related content was removed to eliminate styling artifacts that do not contribute to the linguistic or semantic properties of the data.

3.3 Simplifying and Structuring Web Content

Following HTML simplification, the content was converted into *Markdown* format to facilitate downstream text processing, preserve the document's basic structural hierarchy, and allow precise extraction of textual and visual elements. Subsequently, the extraction process categorized the webpage content into two primary types: textual and visual. Our framework transforms texts into structured text; this step encompasses elements such as headers, paragraphs, ordered and unordered lists, and tables. The visual category comprises figures (`fig`) and image (`img`) tags. To maintain semantic coherence, text elements sharing identical tags are concatenated, thereby preserving both structural integrity and the contextual flow of the content.

3.4 Quality Filtering Steps

The pipeline employs a multilevel filtering system at both the tag and document levels to ensure data quality and relevance. A *tag* refers to any connected textual unit (e.g., paragraph, list, or section), while a *document* corresponds to the entire web page.

3.4.1 Arabic Perplexity Model

To provide a robust filtering strategy similar to the one introduced in (Laurençon et al., 2023), we trained a custom perplexity model specialized for Arabic text based on the KenLM (Heafield, 2011), which we called *WasmKenLM*. The training corpus for this 5-gram model comprises 37,258,520 samples across three domains: news, scientific, and

dialectal Arabic. This totals approximately 2.98 billion tokens, as measured by the Kuwain tokenizer (Hennara et al., 2025).

Table 1 presents the token distribution across the corpus categories.

Domain	Tokens
News	2,129,569,014
Scientific	515,017,029
Dialectal Arabic	337,945,783
Total	**2,982,531,826**

Table 1: Token distribution of the perplexity model training corpus by domain.

The KenLM model was trained with a 5-gram order using the `lmplz` toolkit and subsequently compiled into a binary format via `build_binary` for efficient inference.

3.4.2 Tag-level Filtering

Compared to the baseline pipeline OBELICS, we introduced several manual modifications to adapt the filtering process to the specific characteristics of the Arabic language. These adjustments included relaxing or removing certain thresholds that were originally optimized for English but do not generalize well to Arabic.

For example, we reduce the weight of the *Word Repetition Ratio*, which measures the proportion of repeated words in a text. In Arabic, word repetition often carries stylistic and rhetorical significance rather than being indicative of low-quality content. Similarly, we removed the *Stopword Ratio* filter, as Arabic exhibits a rich vocabulary and flexible syntactic structures that allow grammatically correct sentences with relatively few function words. In the same spirit, the *Punctuation Ratio* was also discarded, since Arabic web content frequently lacks punctuation, and using this metric would disproportionately eliminate valid Arabic text. We also disabled the *Common Word Ratio* filter, which penalizes texts containing many high-frequency words. In Arabic, this would be biased against authentic content, given that the distribution of common words differs significantly from English. Likewise, the *Special Character Ratio* (e.g., emojis, abbreviations) was adjusted more lightly, since contemporary Arabic text often incorporates such elements without necessarily being low-quality. We applied a stricter *Language Identification* process to guarantee that the text is predominantly Arabic, while still allowing for the natural occurrence of foreign terms (for example, English terminology).

3.4.3 Visual Data Filtering

Our image data filtering strategy was tailored to the characteristics of Arabic web content, with an emphasis on maximizing data retention while upholding quality standards. Given the relative scarcity of Arabic multimodal resources, we adopted a conservative approach that avoids unnecessary exclusions.

Instead of downloading images directly, we collected their URLs to reduce storage costs, accelerate the acquisition process, and enable scalable filtering. This design naturally shifted the focus of filtering from individual images to the URL level. To ensure safety and appropriateness, we maintain a blacklist of websites that host explicit or unsuitable material and exclude all associated image URLs. In particular, Arabic web content is rarely hosted on mainstream platforms that contain prohibited content, which further justifies this site-level filtering strategy.

The resulting set of URLs forms a flexible foundation for subsequent stages of image processing and task-specific filtering, allowing later adjustments to be aligned with the requirements of different models and training objectives.

3.4.4 Tag-Level Deduplication

We have worked to remove implicit duplicates within documents. For example, some sites contain duplicate ads. To address substantial repetition at the tag level observed in the dataset, we implemented the Needleman–Wunsch algorithm (Needleman and Wunsch, 1970) with a similarity threshold of 80% to efficiently identify and remove nearly duplicate content.

3.4.5 Document-level Filtering

At the document level, we applied the same set of filtering criteria as in the tag-level filtering 3.4.2, but with different parameter values. These values were recalibrated to reflect document-wide characteristics rather than paragraph-level ones. In particular, thresholds were tuned to balance the need for higher-quality long-form content with the goal of retaining as much Arabic data as possible.

3.5 Dataset Statistics

Table 3 summarizes the overall scale of the pipeline. Starting from approximately 2.5 million Arabic web documents identified in Common Crawl, the

Dataset	Excluded Example (excerpt)
fine_web2	حظر تدريس المعلبين لأبنائهم في الفصول - yB الأخبار, الإدارات التعليمية, التعليم الإعدادي, التعليم الابتدائي, التعليم الثانوي, التعليم الخاص, التعليم الفني, التوجيه الفني, الديوان العام, الشئون القانونية, الطالب, المعلم, شئون الطلبة والامتحانات, عام ni - 01/60/9102 - وجهت وزارة التربية و التعليم ،والتعليم الفنى، خطاب إلى جميع مديريات الجمهورية، ينص على حظر قيام الآباء و الأمهات من المعلبين بالتدريس في الفصول التي يوجد بها أبنائهم. وأشارت إلي أن هذا القرار يأتي لصالح العملية التعليمية ، ودرءًا للمفاسد مع الالتزام بالحيادية، وعدم التحيز لطالب أو لفئة من الطلاب دون غيرهم. وشددت على أنه في حالة عدم الالتزام بما سبق ذكره سيتم اتخاذ الإجراءات القانونية الجزائية الفورية حيال الشخص أو الجهة المخالفة. siht ot knil tnenamreP gro.udeaibulaq//:ptth :elcitra
ara24	« -NEIEVLOS « IEVLOS « EVLOS « VLOS « LOS « OS « S « moc.ONseinapmoC ROTNOK NEIEVLOS ,ĀBERV¯Ā ZIL-YAM PAKSNGER GO ROTNOK NEIEVLOS آراء حول noinipOdiloS في هذه الصفحة سوف تجد تعاليق Āberv¯Ā ziL-yaM PAKSNGER GO حول ملاحظات يقدمها زوار Āberv¯Ā ziL-yaM PAKSNGER GO ROTNOK NEIEVLOS. الموقع moc.ONseinapmoC. وتظهر سجلات قاعدة البيانات لدينا أن هذه الشركة من الترويج، منصة التعليق: noinipOdiloS
cultura_x	الفك كسارات السوق في الهند. الفك محطم مصنع سعر الهند. حجر الجلخ telryot حجر صغير محطم سعر الجهاز cj حجر الفك محطم آلة حجر محطم سعر الجهاز في المكسيك حجر محطم ما هو مطحنة الكرة تأذن تاجر من الحجر محطم فيمصر. كسارات الحجر المستعملة للبيع في الهند. اسعار كسارات فى مصر 03 ebuTuoY أيلول (سبتمبر) 3102 وNOT هو الصانع المهنية اسعار كسارات صغيرة للبيع فى مصر في العالم، وتقع في . الحصول على السعر
dataset_101	شركة الزامل العقارية المحتوى ج دة الإم ارات دبي الم ارينا م ع ارض نا الع ق اري ة ج دة جميع الحقوق محفوظة لشركة الزامل العقارية 6102

Table 2: Examples of excluded data based on high perplexity values.

combined quality filtering and tag-level deduplication steps retained 1,758,151 documents, corresponding to an overall removal rate of approximately 29.7%. This relatively moderate reduction reflects our deliberate high-precision, low-recall design philosophy (Section 4.2.1): we prefer to err on the side of retention for a low-resource language such as Arabic, removing only content that the pipeline is highly confident is low quality or duplicated.

4 Discussion & Analysis

This section situates our methodology within the broader landscape of dataset construction. By addressing the specific constraints of previous models, we demonstrate how our design choices refine and enhance the process of Arabic multimodal curation.

4.1 Structured Data Preservation.

Unlike approaches that flatten web content into sequential text-image pairs (e.g., MMC4 (Zhu et al., 2023) and AMCrawl (Aboukozzana et al., 2025)) or transform documents into linear token sequences (e.g., OBELICS (Laurençon et al., 2023)), our methodology preserves the hierarchical structure inherent in web documents in Markdown format. This preservation maintains semantic relationships between content elements such as image-caption associations, section hierarchies, and contextual dependencies that are crucial for training models capable of understanding document-level coherence. Our approach also further preserving the underlying DOM hierarchy, enabling more sophisticated downstream applications that require an understanding of document organization and content relationships.

Table 4 summarizes how *Wasm* relates to prior Arabic and interleaved multimodal corpora along the dimensions most relevant to our contribution: whether the corpus targets Arabic, whether it preserves interleaved image-text structure, whether it retains document-level structural elements (headings, lists, tables), whether it offers a Markdown (or equivalent structured) output, and the granularity of its deduplication.

4.2 Perplexity-Based Quality Assessment

We evaluate the perplexity model's effectiveness against existing baselines and demonstrate the practical impact of this filtering approach by assessing exclusion rates across multiple Arabic datasets.

4.2.1 Perplexity Model and Baseline Comparison

To evaluate the effectiveness of *WasmKenLM* in identifying and removing incoherent or machine-generated content, we compared it against the widely used WikiKenLM baseline [5]. While WikiKenLM is trained exclusively on Arabic Wikipedia, our model benefits from the diverse, multi-domain corpus described in section 3.4.1, providing a broader representation of Arabic dialects and topical diversity while preserving the

[5] https://huggingface.co/edugp/kenlm

Stage	Documents	Images
Raw extraction	2.5M (100%)	30.3M (100%)
Filtering & deduplication	1.76M (70.3%)	21.8M (71.9%)

Table 3: Document and image counts before and after the *Wasm* filtering and deduplication stages.

Dataset	Arabic	Interleaved	Struct. preserved	Dedup. granularity
MMC4 (Zhu et al., 2023)	✗	✓	✗	Document
OBELICS (Laurençon et al., 2023)	✗	✓	✗	Document (MinHash)
OmniCorpus (Li et al., 2024)	✗	✓	✗	Document
MINT-1T (Awadalla et al., 2024)	✗	✓	✗	Document
101 Billion Arabic (Aloui et al., 2024)	✓	✗	✗	Document
ArabicWeb24 (Farhat et al., 2024)	✓	✗	✗	Document (MinHash)
AMCrawl (Aboukozzana et al., 2025)	✓	✓	✗	Document
Wasm (ours)	✓	✓	✓	Tag / node-level

Table 4: Comparison of *Wasm* with existing Arabic and interleaved multimodal corpora. "Struct. preserved" refers to retention of document-level structural elements such as headings, lists, and tables, rather than flattening documents into linear text-image sequences.

authenticity of human-authored text.

For a rigorous assessment, we constructed an evaluation suite comprising four distinct datasets: FineWeb-2, Ara24, CulturaX, and Dataset 101. We randomly sampled 100 examples from each dataset, yielding 400 evaluation samples in total. Human annotators then reviewed whether each example was correctly classified by each model.

Filtering philosophy: high precision over high recall. Because the goal of the perplexity filter is to *remove* bad text without discarding good text, we frame the human evaluation as a binary detection task. For each example flagged as low quality, the annotator judges whether the flag is correct (a *True Positive*, TP) or whether the example is in fact acceptable text that was incorrectly removed (a *False Positive*, FP). The failure mode we most want to avoid is the latter: good Arabic text being silently discarded.

Accordingly, our filtering strategy is designed to **minimize False Positives** at the cost of tolerating more False Negatives (borderline low-quality content that slips through the filter). This is achieved by setting a conservative, high perplexity threshold: *WasmKenLM* only flags an example as low quality when it is highly confident, prioritising **Precision** over Recall:

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

A high-precision filter guarantees that when *Wasm* removes an example, that removal can be trusted. We consider this trade-off appropriate for a low-resource language setting: for Arabic, which constitutes only ∼0.6% of Common Crawl content, over-filtering is considerably more costly than under-filtering.

Human review of the 400 sampled examples confirmed that *WasmKenLM* consistently outperforms the WikiKenLM baseline. WikiKenLM, trained solely on Arabic Wikipedia, showed limited sensitivity to the range of quality issues present in Arabic web text, frequently failing to flag incoherent or machine-generated content while in other cases incorrectly penalising authentic dialectal or informal text. *WasmKenLM*, by contrast, benefited from its multi-domain training corpus (news, scientific, and dialectal Arabic) and produced notably fewer False Positives across all four evaluation datasets, confirming the suitability of its thresholds for Arabic web curation.

4.2.2 Dataset Assessment and Exclusion Rates

Using the 4,000 manually reviewed samples described in Section 4.2.1, we analyzed the practical impact of our filtering thresholds. For each dataset, we calculated the *exclusion rate* the proportion of examples rejected for exceeding acceptable perplexity scores.

Table 5 reports the exclusion rates across the four datasets. The human review conducted on these samples confirmed the high accuracy of our model in successfully targeting low-quality content. Specific examples of text excluded by these thresholds are presented in Table 2. Ultimately, applying a specialized filtering technique like *WasmKenLM* yields significantly higher-quality data, which di-

rectly and positively affects the performance of downstream applications relying on these datasets.

Dataset	Exclusion Rate (%)
FineWeb-2	1.766
Ara24	7.820
CulturaX	8.605
Dataset 101	19.757

Table 5: Exclusion rates across datasets based on perplexity thresholds.

4.3 Granular Node-Level Deduplication.

While existing approaches typically perform deduplication at the document level (e.g., ROOTS (Laurençon et al., 2022), 101 Billion Arabic Words (Aloui et al., 2024), AMCrawl (Aboukozzana et al., 2025)) or apply MinHash deduplication globally or per-dump (e.g., FineWeb (Penedo et al., 2024)), our methodology implements deduplication at the HTML node level. This granular approach enables the preservation of documents that contain unique content alongside duplicated elements (such as navigation menus or boilerplate text), significantly improving content diversity while maintaining processing efficiency. Node-level deduplication is particularly valuable for web documents where substantial unique content may coexist with repeated structural elements, allowing for more nuanced quality preservation than binary document-level decisions.

5 Conclusion

This study introduces *Wasm*, an Arabic multimodal processing framework that processes Common Crawl data, designed to preserve the structural and semantic integrity of web documents. To our knowledge, *Wasm* is the first Arabic pipeline to jointly preserve document-level structure and offer a Markdown-based output that flexibly supports both text-only and interleaved multimodal pre-training. Wasm provides a flexible foundation for training both LMMs and LLMs by maintaining document-level coherence, cross-modal alignments, and hierarchical structures such as captions, sections, and contextual dependencies. By releasing both the dataset and pipeline code, Wasm not only democratizes access to advanced multimodal Arabic resources but also pushes the boundaries of Arabic NLP development, enabling reproducible research and laying the groundwork for future large-scale corpus construction. Wasm demonstrates that even for a language as syntactically rich and web underrepresented as Arabic, constituting only 0.6% of Common Crawl content careful pipeline design can yield a high-quality, structurally faithful multimodal corpus, offering a replicable template for other low-resource languages.

As an important direction for future work, we plan to use the Misraj Structured Data Dump (MSDD) to pre-train or continually fine-tune Arabic LLMs and LMMs, and to evaluate the resulting models on standard Arabic and multimodal benchmarks. Such experiments will provide direct empirical evidence of the downstream utility of structure-preserving, Markdown-based pre-training data compared to corpora that flatten document structure into linear text-image sequences.

References

Shahad Aboukozzana, Muhammad Kamran J Khan, and Ahmed Ali. 2025. AMCrawl: An Arabic web-scale dataset of interleaved image-text documents and image-text pairs. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 448–465, Suzhou, China. Association for Computational Linguistics.

Manel Aloui, Hasna Chouikhi, Ghaith Chaabane, Haithem Kchaou, and Chehir Dhaouadi. 2024. 101 billion arabic words dataset.

Fakhraddin Alwajih, El Moatez Billah Nagoudi, Gagan Bhatia, Abdelrahman Mohamed, and Muhammad Abdul-Mageed. 2024. Peacock: A family of Arabic multimodal large language models and benchmarks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand. Association for Computational Linguistics.

Anas Awadalla, Le Xue, Oscar Lo, Manli Shu, Hannah Lee, Etash Guha, Sheng Shen, Mohamed Awadalla, Silvio Savarese, Caiming Xiong, et al. 2024. Mint-1t: Scaling open-source multimodal data by 10x: A multimodal dataset with one trillion tokens. *Advances in Neural Information Processing Systems*, 37:36805–36828.

Wei Chen, Lin Li, Yongqi Yang, Bin Wen, Fan Yang, Tingting Gao, Yu Wu, and Long Chen. 2025. Comm: A coherent interleaved image-text dataset for multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8073–8082.

May Farhat, Said Taghadouini, Oskar Hallström, and Sonja Hajri-Gabouj. 2024. Arabicweb24: Creating a high quality arabic web-only pre-training dataset.

Matthieu Futeral, Armel Zebaze, Pedro Ortiz Suarez, Julien Abadji, Rémi Lacroix, Cordelia Schmid, Rachel Bawden, and Benoît Sagot. 2024. moscar: A large-scale multilingual and multimodal document-level corpus. *arXiv preprint arXiv:2406.08707*.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Khalil Hennara, Sara Chrouf, Mohamed Motaism Hamed, Zeina Aldallal, Omar Hadid, and Safwan AlModhayan. 2025. Kuwain 1.5 b: An arabic slm via language injection. *arXiv preprint arXiv:2504.15120*.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. Obelics: An open web-scale filtered dataset of interleaved image-text documents.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Qingyun Li, Zhe Chen, Weiyun Wang, Wenhai Wang, Shenglong Ye, Zhenjiang Jin, Guanzhou Chen, Yinan He, Zhangwei Gao, Erfei Cui, et al. 2024. Omnicorpus: A unified multimodal corpus of 10 billion-level images interleaved with text. *arXiv preprint arXiv:2406.08418*.

Saul B. Needleman and Christian D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all–adapting pre-training data processing to every language. *arXiv preprint arXiv:2506.20920*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. In *North American Chapter of the Association for Computational Linguistics*.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2023. Multimodal C4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*.

Linguistic Specialization of Arabic in Text Embedding Models via Multi-Teacher Knowledge Distillation

Ahmad Abdelfattah
Misraj.AI
Syria
ahmad.abdalfattah@misraj.ai

Zeina Aldallal
Misraj.AI
Syria
aldallal@misraj.ai

Sara Chrouf
Misraj.AI
Saudi Arabia
chrouf@misraj.ai

Safwan AlModhayan
Misraj.AI
Saudi Arabia
safwan@misraj.ai

Abstract

Knowledge distillation has proven to be a powerful paradigm for compressing large neural networks into compact yet capable models, embedding models being one such example. In this work, we investigate the use of ensemble-based teacher distillation to specialize a multilingual model for the Arabic language. We demonstrate that aggregating signals from multiple teacher models is a more effective strategy than the conventional single-teacher approach that is predominantly followed in current practice. Our findings further confirm that knowledge distillation serves as a general and reliable mechanism for enhancing the capabilities of embedding models for specific downstream tasks for the target language, with the resulting model's performance being directly shaped by the collective of the employed teachers. Building on these insights, we introduce *kawn-embed-light*, a lightweight Arabic embedding model that achieves state-of-the-art performance among models under 1 billion parameters on the Arabic information retrieval tasks of the Multilingual Massive Text Embedding Benchmark (MMTEB), reaching an average NDCG@10 (Normalized Discounted Cumulative Gain at rank 10) score of **0.7711**. Notably, *kawn-embed-light* remains competitive with significantly larger models and surpasses the majority of closed-source commercial embedding APIs. Finally, our work demonstrates that ensemble distillation provides a computationally efficient and high-performance pathway for model localization, offering a scalable solution to enhance Natural Language Processing (NLP) capabilities in low-resource languages.

1 Introduction

In Natural Language Processing (NLP), sentence embedding refers to a numeric representation of a sentence in the form of a fixed-length vector of real numbers, that encode meaningful semantic information. These embeddings can then be used in various downstream tasks, such as information retrieval, semantic text similarity, classification, clustering, and Retrieval-Augmented Generation (RAG) systems (Lewis et al., 2020).

Recently, there has been a growing focus on utilizing knowledge distillation in the context of embedding models (Zhao et al., 2025; Akram et al., 2026; Zhang et al., 2025a, 2024; Kim et al., 2023; Vera et al., 2025). Knowledge distillation (Buciluǎ et al., 2006; Hinton et al., 2015) provides an efficient mechanism for training smaller models to mimic the behavior of one or more teacher models (Akram et al., 2026).

Nevertheless, the application of knowledge distillation for acquiring or improving language-specific capabilities in embedding models remains largely unexplored. Moreover, Arabic remains underrepresented in embedding research; dedicated Arabic embedding models remain scarce, and the best-performing Arabic models on Multilingual Massive Text Embedding Benchmark (MMTEB) (Enevoldsen et al., 2025) are multilingual with parameter counts exceeding 4B (Zhang et al., 2025b; Babakhin et al., 2025) rendering them impractical for resource-constrained and production deployment scenarios due to their prohibitive computational cost.

In this work, we present a novel approach for dis-

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 459–467
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_047

tilling Arabic embedding knowledge from multiple large multilingual embedding models (i.e., teachers) into a smaller multilingual embedding model (i.e., student). We demonstrate that multi-teacher distillation (MTD) outperforms the commonly used single-teacher approach.

Our contributions are summarized as follows:

- We leverage Knowledge Distillation to enhance the Arabic language capabilities of embedding models, eliminating the need for expensive from-scratch training.
- We introduce a novel linguistic specialization through MTD approach for embedding models and empirically show that distilling from multiple teachers outperforms relying on the single best teacher alone.
- We introduce *kawn-embed-light*, a 300M-parameter Arabic embedding model that achieves state-of-the-art retrieval performance among sub-1B-parameter models on the Multilingual Massive Text Embedding Benchmark (MMTEB), with an average NDCG@10 (Normalized Discounted Cumulative Gain at rank 10) of 0.7711. This result validates our proposed multi-teacher distillation (MTD) approach for adapting embedding models to new languages.

2 Related Work

2.1 Text Embedding Models

Reimers and Gurevych (2019) proposed an early approach that leverages Siamese networks without contrastive learning. Subsequently, researchers introduced two-stage contrastive learning frameworks built on encoder-only transformers, where the first stage pre-trains on large-scale weakly supervised data and the second stage fine-tunes on smaller, higher-quality supervised datasets (Wang et al., 2022; Li et al., 2023; Sturua et al., 2024; Xiao et al., 2024).

Another line of research explores decoder-only transformers for text embedding. Some works incorporate a pre-training phase on synthetic data (Lee et al., 2025; Zhang et al., 2025b; Babakhin et al., 2025; Zhao et al., 2025), while others skip this phase entirely and apply direct fine-tuning (Hu et al., 2025; Wang et al., 2024a). The latter demonstrate that pre-training provides no significant benefit for decoder-only models, attributing this to their strong autoregressive capabilities.

Despite their differences, both approaches may rely on large scale weakly supervised data and always require high-quality supervised datasets for fine-tuning. For Arabic, such resources are extremely limited, making this a significant challenge.

2.2 Embedding Model Distillation

Knowledge distillation has been applied to embedding models through several complementary strategies. Yang et al. (2024) explore distillation for dense retrieval in a cross-lingual setting, using machine translation to transfer retrieval knowledge across languages.

(Kim et al., 2023) introduce a foundational approach for single-teacher embedding distillation, in which the student is trained to match the teacher's embedding space via an alignment objective, with a projection layer added on top of the student to handle any dimensional mismatch between the two models. Building on a similar single-teacher paradigm, Akram et al. (2026) apply distillation as a pre-training phase using a standard cosine-similarity objective, demonstrating that embedding supervision can be incorporated even before task-specific fine-tuning. They also follow this distillation stage with task-adapter fine-tuning.

A complementary line of work extends distillation to multiple teachers. Zhang et al. (2024) propose concatenating the embedding vectors of several teachers into a single target vector, training the student to reproduce this combined representation through both score-based distillation and embedding alignment, again relying on a projection layer to accommodate the enlarged target dimensionality. Zhang et al. (2025a) follow a similar approach, concatenating two teacher embeddings and optimizing the student with a cosine-similarity objective, demonstrating that ensemble distillation scales effectively to larger models.

On the other hand, Zhao et al. (2025) utilize distillation as a post-training phase, applied after an initial contrastive learning stage. Rather than minimizing cosine distance directly, they distill the normalized distribution of temperature-scaled cosine similarity scores via KL-divergence, transferring richer relational information encoded in the teacher's score distribution.

However, embedding space distillation is not well-suited for multi-teacher distillation because the differing embedding dimensions across multi-

ple teachers make projection-based alignment suboptimal, as no single projection can faithfully map the student to all teacher spaces. Furthermore, there is limited research on the use of distillation techniques to improve the capabilities of the embedding model in a specific language.

3 Dataset

This section describes the dataset used to train our Arabic embedding model. We detail how the data was collected from multiple complementary sources, how hard negatives were mined to construct effective training triplets, and the preprocessing steps applied to ensure data quality. The final dataset comprises approximately 2M training examples drawn from more than 20 distinct sources, spanning general, Islamic, and medical domains.

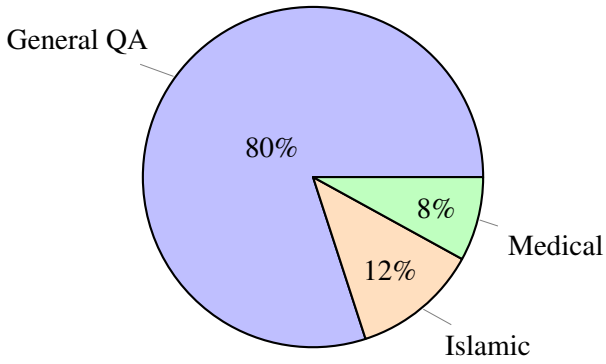


Figure 1: Distribution of data domains across the dataset. General QA accounts for the majority (80%), with Islamic (12%) and Medical (8%) content comprising the remainder.

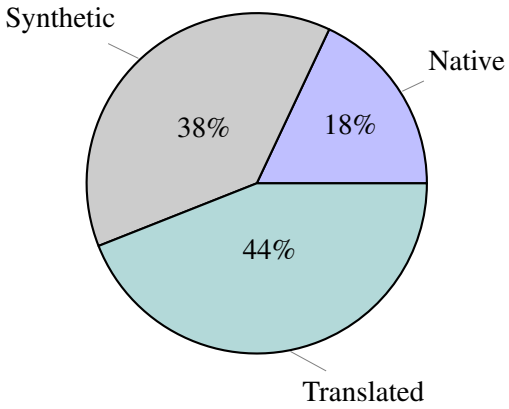


Figure 2: Distribution of data sources across the dataset. Translated data constitutes the largest share (44%), followed by Synthetic (38%) and Native (18%).

3.1 Data Collection

The training data used to train the embedding model is formed as triplets $\{q_i, d_i^+, \{d_{i,k}^-\}_{k=1}^n\}$, where d_i^+ denotes the positive document associated with query q_i, and $\{d_{i,1}^-, \ldots d_{i,n}^-\}$ represents a set of n hard negative (irrelevant) documents of q_i.

However, Arabic remains a low-resource language in this context, with very limited publicly available datasets conforming to this training scheme. To address this limitation, prior work has proposed several practical solutions, including translating existing datasets from high-resource languages into the target low-resource language (Babakhin et al., 2025; Sturua et al., 2024), or generating synthetic datasets by prompting large language models (LLMs) to produce queries for existing documents an approach that has gained considerable adoption in recent years (Babakhin et al., 2025; Hu et al., 2025; Zhang et al., 2025b; Wang et al., 2024a).

Guided by these strategies, our dataset construction follows four complementary directions: (1) leveraging existing Arabic datasets, (2) translating widely adopted information retrieval benchmarks including MS-MARCO (Bajaj et al., 2016), Natural Questions (Kwiatkowski et al., 2019), and SQuAD (Rajpurkar et al., 2016) into Arabic, (3) scraping a large volume of Arabic Islamic community question-answering content from online resources, and (4) generating synthetic data following the approach proposed by (Zhang et al., 2025b). Figure 1 illustrates the domain-based distribution of the compiled dataset. Approximately 80% of the data corresponds to general question-answering, encompassing translated, natively Arabic, and synthetically generated samples. Islamic-domain content accounts for 12% of the dataset, sourced entirely from scraped online resources, while the remaining 8% pertains to the medical domain.

From the perspective of data sourcing, Figure 2 reveals that 44% of the dataset is translated from well-established English information retrieval benchmarks, 38% is synthetically generated using Gemini 2.5 Flash and Gemini 2.0 Flash, following the approach introduced by Zhang et al. (2025b) by generating synthetic queries for documents using a two-stage prompting technique (in our case, using an initial corpus scraped from native Arabic websites covering diverse domains). Concretely, the first stage is a *configuration* stage, where the model is prompted to select a query type (e.g., key-

word, factual, summary), a difficulty level, and a persona likely to be interested in the given document; the second stage is a *query generation* stage, where these configuration choices condition the generation of a realistic query targeting that document. This persona-driven, multi-dimensional conditioning encourages diversity and realism in the resulting query distribution. The remaining 18% consists of natively Arabic content that is either sourced from open datasets or scraped from Islamic QA communities.

3.2 Hard Negatives Mining

To mine hard negatives, we utilize the positive-aware approach proposed by (Moreira et al., 2024); specifically, the `TopK-PercPos` strategy. This strategy retrieves top-k most similar documents for each query, then excludes passages whose similarity score exceeds a given percentage of the query-positive score, in order to eliminate false-negative cases. For the positive-aware mining algorithm, the similarity threshold was set to 95% of the query-positive score.

3.3 Data Pre-processing

Utilizing the embedding that we calculated in the hard negative mining step we filter out pairs whose query-positive similarity falls below 0.35, as well as pairs containing documents below a minimum length threshold, as these are considered noisy or uninformative training examples. After this cleaning process, we retain approximately 2M diverse training examples spanning more than 20 distinct sources.

4 Methodology

In this section, we present our methodology, which is based on knowledge distillation. Formally, let S denote a multilingual sentence embedding model (the *student*), and let $\mathcal{T} = \{t_1, t_2, \ldots, t_K\}$ denote a set of K stronger multilingual embedding models (the *teachers*). The method uses knowledge distillation to transfer Arabic language understanding from multiple teacher models into a single student model by minimizing the KL divergence between the soft probability distributions of the ensemble teachers and the student model.

Inspired by (Zhao et al., 2025), we distill normalized temperature-scaled cosine similarity scores into the student model. Our method is built upon two key design choices: (1) we employ a single unified training phase, eliminating the need for a separate post-contrastive fine-tuning stage in which the model would first be trained with a standard contrastive loss and subsequently fine-tuned; and (2) we leverage an ensemble of multiple teacher models rather than relying on a single best-performing teacher.

Formally, Training data consists of triplets $\{q_i, d_i^+, \{d_{i,j}^-\}_{j=1}^n\}$ (Section 3). For each sample, a model M scores every candidate document $d_i \in \{d^+\} \cup \{d_1^-, \ldots, d_n^-\}$ as a normalized cosine similarity: $z_i^M = \phi(M(q_i), M(d_i))$, where ϕ denotes the cosine similarity function. These scores form a logit vector $\mathbf{z}^M = (z_1^M, \ldots, z_N^M) \in \mathbb{R}^N$, with $N = 1 + n$, over which a temperature-τ softmax is applied to produce candidate probabilities, as defined in Equations 1 and 2 below.

The individual teacher probabilities $P_{t,k}(i)$ and the student probabilities $P_S(i)$ are obtained by applying a temperature-scaled softmax transformation over their respective similarity logits, where $\tau > 0$ is a temperature hyperparameter controlling the smoothness of the resulting distributions. For the student model with similarity logits $z_{s,i}$:

$$P_S(i) = \frac{\exp(z_{s,i}/\tau)}{\sum_{j=1}^{N} \exp(z_{s,j}/\tau)}, \tag{1}$$

and for the k-th teacher with similarity logits $z_{t,k,i}$:

$$P_{t,k}(i) = \frac{\exp(z_{t,k,i}/\tau)}{\sum_{j=1}^{N} \exp(z_{t,k,j}/\tau)}. \tag{2}$$

Higher values of τ produce softer distributions that emphasize relative ranking across all candidates, while lower values produce sharper distributions that concentrate probability mass on the highest-scoring candidate.

The learning objective is the Kullback–Leibler (KL) divergence between the aggregated teacher distribution $P_{\mathcal{T}}$ and the student distribution P_S. We adopt the direction $D_{KL}(P_{\mathcal{T}} \parallel P_S)$, treating the teacher ensemble as the reference distribution and penalizing the student for deviating from it:

$$\begin{aligned} \mathcal{L}_{KL} &= D_{KL}(P_{\mathcal{T}} \parallel P_S) \\ &= \sum_{i=1}^{N} P_{\mathcal{T}}(i) \log\left(\frac{P_{\mathcal{T}}(i)}{P_S(i)}\right). \end{aligned} \tag{3}$$

The aggregated target distribution $P_{\mathcal{T}}$ is constructed as a convex combination of the K individual teacher distributions, with each teacher's

contribution is modulated by a scalar weight w_k:

$$P_{\mathcal{T}}(i) = \sum_{k=1}^{K} w_k P_{t,k}(i) \quad \text{where} \sum_{k=1}^{K} w_k = 1, \quad w_k \geq 0. \tag{4}$$

This formulation guarantees that $P_{\mathcal{T}}$ constitutes a valid probability distribution. The weights $\{w_k\}$ thus provide fine-grained control over each teacher's relative contribution to the distillation signal.

5 Experiments and Results

5.1 Setup

The student model is initialized from *EmbeddingGemma-300M* (Vera et al., 2025), selected for its optimal trade-off between computational efficiency and performance on Arabic information retrieval tasks on MMTEB. Knowledge distillation is performed using two high-capacity multilingual teacher models: *Qwen3-Embedding-8B* (Zhang et al., 2025b) and *Llama-Embed-Nemotron-8B* (Babakhin et al., 2025), both of which rank among the strongest publicly available models for Arabic information retrieval tasks on the benchmark. Training proceeds for approximately 15,000 optimization steps over a corpus of 2 million high-quality, filtered instances with 4 hard negatives per query-document pair, with context length 1024 token. All experiments run on 2 × NVIDIA H100 (80GB) GPUs with a global batch size of 128. We use BFloat16 (bf16) mixed-precision to ensure numerical stability while reducing memory overhead, and optimize with a learning rate of 1×10^{-5} and a linear warmup schedule over the first 100 steps. For softmax temperature scaling we use $\tau = 1$. Following the standard protocol established for *EmbeddingGemma-300M*, we apply the prescribed prompt templates to both queries and documents during training. For knowledge distillation, we use a uniform ensemble weighting strategy, assigning equal weight to each teacher signal ($w_k = 0.5$ per teacher), and treating both teachers as equally reliable.

5.2 Single-Teacher vs. Multi-Teacher Distillation

To determine the optimal distillation strategy, we compare three configurations: distilling from *Qwen3-Embedding-8B* alone, from *LLaMA-Embed-Nemotron-8B* alone, and from the ensemble of both teachers (proposed approach). As shown in Table 1, ensemble distillation outperforms both single-teacher baselines, including the stronger *LLaMA-Embed-Nemotron-8B* teacher, demonstrating that combining multiple teachers yields richer training signal than any individual teacher alone. Furthermore, the single-teacher results exhibit a consistent ordering: distilling from *LLaMA-Embed-Nemotron-8B* yields superior performance compared to distilling from *Qwen3-Embedding-8B*, which aligns with the relative performance gap between the two teachers observed in Table 2. All results are reported on the MMTEB benchmark using the evaluation tasks described in Section 6.1.

Configuration	Avg
Teacher Models (Upper Bound)	
LLaMA-Embed-Nemotron-8B	0.7980
Qwen3-Embedding-8B	0.7883
Student Model (Lower Bound)	
EmbeddingGemma-300M	0.7259
Single-Teacher Distillation	
w/ LLaMA-Embed-Nemotron-8B	0.7626
w/ Qwen3-Embedding-8B	0.7565
Multi-Teacher Distillation	
kawn-embed-light (proposed)	**0.7711**

Table 1: Comparison of distillation configurations. Our proposed multi-teacher distillation (MTD) outperforms all single-teacher baselines despite using a significantly smaller student model (*kawn-embed-light*, 300M parameters).

6 Evaluation

6.1 Evaluation Settings

For evaluation we utilize Massive Multilingual Text Embedding Benchmark (MMTEB) (Enevoldsen et al., 2025), encompasses a comprehensive collection of evaluation tasks spanning over 250 languages and ten task types: Bitext Mining, Classification, Clustering, Instruction Retrieval, Multilabel Classification, Pair Classification, Reranking, Retrieval, Semantic Textual Similarity (STS), and Summarization.

As this work focuses on distilling information retrieval capabilities for an Arabic language model,

Model	Number Of Parameters(M)	Avg	MRHN	MLQA
Commercial APIs				
google/gemini-embedding-001	-	**0.7934**	**0.7877**	**0.7990**
Cohere/Cohere-embed-multilingual-v3.0	-	0.7445	0.7679	0.7210
voyageai/voyage-3	-	0.7201	0.7012	0.7390
openai/text-embedding-3-large	-	0.6812	0.6912	0.6712
Large Models (More than 4B Params)				
nvidia/llama-embed-nemotron-8b	8000	**0.7980**	**0.8100**	**0.7859**
Qwen/Qwen3-Embedding-8B	8000	0.7883	0.7921	0.7845
Small Models (Less than 1B Params)				
kawn-embed-light	300	**0.7711**	**0.7896**	**0.7527**
intfloat/multilingual-e5-large	596	0.7497	0.7658	0.7336
BAAI/bge-m3	568	0.7458	0.7890	0.7026
jinaai/jina-embeddings-v5-text-small	596	0.7404	0.7480	0.7328
google/embeddinggemma-300m	300	0.7259	0.7278	0.7239
Qwen/Qwen3-Embedding-0.6B	595	0.6829	0.7050	0.6608

Table 2: Arabic MMTEB Retrieval Tasks Evaluation Results. MRHN: MIRACL Retrieval Hard Negatives, MLQA: MultiLingual Question Answering

we leverage two prominent retrieval tasks from the benchmark, selected based on their prevalence among submissions by compared model authors. MIRACL Retrieval Hard Negative which is modified more compatitive version of MIRACL (Zhang et al., 2023), and MLQA. For both tasks, we adopt **NDCG@10** (Normalized Discounted Cumulative Gain at rank 10) as the evaluation metric, consistent with MMTEB's default scoring protocol for information retrieval tasks.

6.2 Main Results

We compare our model with various open-source text embedding models and commercial API services. The open-source models include two of the best Multilingual embedding models larger than 4B parameters: Llama-Embed-Nemotron-8B (Babakhin et al., 2025) and Qwen3-Embedding-8B (Zhang et al., 2025b), which also serve as our teacher models in the knowledge distillation process. For small models less than 1B parameters, we include Multilingual-E5-Large (Wang et al., 2024b), BGE-M3 (Chen et al., 2024), Jina-Embeddings-v5-Text-Small (Akram et al., 2026), EmbeddingGemma-300M (Vera et al., 2025), and Qwen3-Embedding-0.6B (Zhang et al., 2025b). The commercial APIs evaluated are Gemini-Embedding-001 from Google (Lee et al., 2025), Voyage3 from Voyage AI, Cohere-Embed-Multilingual-v3.0 from Cohere, and Text-Embedding-3-Large from OpenAI. For all models, we use results submitted by their authors to the MMTEB results repository[1].

Table 2 presents the evaluation results on Arabic MMTEB retrieval tasks, covering two tasks: MIRACL Retrieval Hard Negatives (MRHN) and MultiLingual Question Answering (MLQA). We compare *kawn-embed-light* against commercial APIs, large-scale open models (>4B parameters), and small open models (<1B parameters).

Among the commercial APIs, Gemini-Embedding-001 achieves the highest average score of 0.7934, setting a strong competitive ceiling. In the large model category, Llama-Embed-Nemotron-8B and Qwen3-Embedding-8B which serve as the teacher models in our ensemble distillation framework achieve average scores of 0.7980 and 0.7883, respectively, yielding a mean teacher performance of 0.7932.

Our proposed model, *kawn-embed-light*, is a compact 300M parameter model trained via ensemble knowledge distillation from both teacher models, using EmbeddingGemma-300M as the student backbone. Despite its significantly reduced size more than 26× smaller than each teacher *kawn-embed-light* achieves an average score of 0.7711, corresponding to 97.2% of the mean teacher performance. Compared to the base student model before distillation, EmbeddingGemma-300M, which scores 0.7259, *kawn-embed-light* yields an absolute improvement of +4.52%, demonstrating the substantial gain introduced by the ensemble dis-

[1]See the MMTEB results repository at https://github.com/embeddings-benchmark/results.

Model	Avg. Retr.
kawn-embed-light	**0.7711**
Merge-Distill (Qwen × LLaMA)	0.7634

Table 3: Model merging vs. proposed approach on average retrieval.

tillation process. Notably, *kawn-embed-light* outperforms all other small models (less than 1B params) by a substantial margin, surpassing the second-best small model, Multilingual-E5-Large (596M), by +2.14% in average score, while also exceeding several closed-source baselines including Cohere-Embed-Multilingual-v3.0 from Cohere, Voyage3 from Voyage AI, and Text-Embedding-3-Large from OpenAI.

7 Analysis and Discussion

In this section, we analyze two key aspects of our proposed approach. First, we investigate whether model merging can serve as a viable alternative to ensemble distillation. Second, we assess whether the distillation process causes catastrophic forgetting by evaluating the model on semantic textual similarity (STS) as a proxy for general language understanding.

7.1 Model Merging

Inspired by recent works applying model merging as a post-training technique (Zhang et al., 2025b; Lee et al., 2025; Vera et al., 2025), we investigate whether merging separately distilled models can yield competitive or superior performance compared to our proposed ensemble distillation approach. To this end, we apply Spherical Linear Interpolation (SLERP) to merge the two individually distilled teacher-based models described previously. As reported in Table 3, the ensemble distillation approach still yields substantial improvements over the merged model, demonstrating that joint optimization from multiple teachers during training captures richer complementary signals than post-hoc weight interpolation.

7.2 Catastrophic Forgetting and General Knowledge Improvement

To assess whether the distillation process degrades the student model's pre-existing knowledge given that training was conducted exclusively on information retrieval data we evaluate the resulting model on a semantically distinct task type: Semantic Textual Similarity (STS). Specifically, we benchmark on two widely adopted Arabic STS datasets, STS17 and STS22, both of which are available within the MMTEB benchmark suite. As shown in Table 4, despite the complete absence of STS data during training, the distilled model achieves an improvement of +1.2% over the original EmbeddingGemma-300M baseline. This result provides two key insights: first, it confirms that the distillation procedure does not induce catastrophic forgetting of previously acquired capabilities; second, it suggests that the process yields a broader improvement in the model's general understanding of Arabic, extending beyond the information retrieval domain.

Model	Avg. STS
kawn-embed-light	**0.7327**
embeddinggemma-300m	0.7204

Table 4: Semantic Textual Similarity (STS) benchmark results on the Arabic splits of STS-17 and STS-22. Spearman rank correlation is computed between model-predicted and ground-truth similarity scores. *embeddinggemma-300m* denotes the pre-fine-tuning baseline.

8 Conclusion

In this work, we presented *kawn-embed-light*, a compact 300M-parameter Arabic embedding model trained for information retrieval via multi-teacher knowledge distillation from *Qwen3-Embedding-8B* and *Llama-Embed-Nemotron-8B*. Despite its small size, *kawn-embed-light* outperforms all evaluated small models and several commercial APIs on Arabic MMTEB retrieval benchmarks. Our experiments demonstrate that ensemble distillation consistently surpasses single-teacher distillation, and STS evaluation confirms that the distillation process enhances general Arabic language understanding without causing catastrophic forgetting o overcome the scarcity of Arabic IR training data, we constructed a diverse corpus by combining LLM-based translation, synthetic data generation, and web scraping. These results highlight the potential of knowledge distillation as an effective strategy for building efficient, high-quality Arabic embedding models..

References

Mohammad Kalim Akram, Saba Sturua, Nastia Havriushenko, Quentin Herreros, Michael Günther, Maximilian Werk, and Han Xiao. 2026. jina-embeddings-v5-text: Task-targeted embedding distillation. *arXiv preprint arXiv:2602.15547*.

Yauhen Babakhin, Radek Osmulski, Ronay Ak, Gabriel Moreira, Mengyao Xu, Benedikt Schifferer, Bo Liu, and Even Oldridge. 2025. Llama-embed-nemotron-8b: A universal text embedding model for multilingual and cross-lingual tasks. *arXiv preprint arXiv:2511.07025*.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 4(5).

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, et al. 2025. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Xinshuo Hu, Zifei Shan, Xinping Zhao, Zetian Sun, Zhenyu Liu, Dongfang Li, Shaolin Ye, Xinyuan Wei, Qian Chen, Baotian Hu, et al. 2025. Kalm-embedding: Superior training data brings a stronger embedding model. *arXiv preprint arXiv:2501.01028*.

Seungyeon Kim, Ankit Singh Rawat, Manzil Zaheer, Sadeep Jayasumana, Veeranjaneyulu Sadhanala, Wittawat Jitkrittum, Aditya Krishna Menon, Rob Fergus, and Sanjiv Kumar. 2023. Embeddistill: A geometric knowledge distillation for information retrieval. *arXiv preprint arXiv:2301.12005*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, et al. 2025. Gemini embedding: Generalizable embeddings from gemini. *arXiv preprint arXiv:2503.07891*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Gabriel de Souza P Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. 2024. Nv-retriever: Improving text embedding models with effective hard-negative mining. *arXiv preprint arXiv:2407.15831*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 2383–2392.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, et al. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *arXiv preprint arXiv:2409.10173*.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, et al. 2025. Embeddinggemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

Eugene Yang, Dawn Lawrie, James Mayfield, Douglas W Oard, and Scott Miller. 2024. Translate-distill: Learning cross-language dense retrieval by translation and distillation. In *European Conference on Information Retrieval*, pages 50–65. Springer.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2024. Jasper and stella: distillation of sota embedding models. *arXiv preprint arXiv:2412.19048*.

Dun Zhang, Ziyang Zeng, Yudong Zhou, and Shuyang Lu. 2025a. Jasper-token-compression-600m technical report. *arXiv preprint arXiv:2511.14405*.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. 2025b. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Xinping Zhao, Xinshuo Hu, Zifei Shan, Shouzheng Huang, Yao Zhou, Xin Zhang, Zetian Sun, Zhenyu Liu, Dongfang Li, Xinyuan Wei, et al. 2025. Kalm-embedding-v2: Superior training techniques and data inspire a versatile embedding model. *arXiv preprint arXiv:2506.20923*.

Author Index

Abarnous, Jamal, 24
Abdelfattah, Ahmad, 459
Afrouni, Azzeddine, 24
Ahidar-Coutrix, Adil, 131
Ahmed, Zehra, 408
Al Ghussin, Yusser, 264
Al Mouatamid, Youssef, 318, 388
Alahapperuma, Malithi P., 304
Alberski, Bartłomiej, 18
Aldallal, Zeina, 378, 418, 449, 459
Alharbi, Maram I., 30, 237
AlModhayan, Safwan, 378, 418, 449, 459
Almutairi, Ali, 326
Alsuhaibani, Abdullah, 326
Aqib, Zuha, 408
Ataa Allah, Fadoua, 24
Atou, Houdaifa, 199

Bastati, Ahmad, 418, 449
Baumann, Timo, 264
Benouini, Rachid, 131
Bernardy, Laura Maria, 284
Bikakis, Antonis, 304
Bissyandé, Tegawendé, 284
Bissyandé, Tegawendé F., 272
Bonnin, Marie, 318, 388
Bouamir, Assia, 318
Boytcheva, Svetla, 309
Buzaaba, Happy, 68

Cabot Sagrera, Jordi, 102
Chaker, Ilham, 121
Chaouri, Yassine, 264
Chersoni, Emmanuele, 150
Chesi, Cristiano, 1
Chrouf, Sara, 378, 418, 449, 459
Collins, Kris, 309
Consuegra-Ayala, Juan Pablo, 96

Dahou, Abdelhalim Hafedh, 398
Daignault-Pichette, Loïc, 80
Dembowski, Paweł, 18
Derych, Alicja Helena, 18
Dev, Vishnu, 370
Diallo, El-Hacen, 272

El Bachyr, Omar, 284
El Houbri, Fatima Ezzahra, 246
El Moussa, Noura, 181
Estevanell-Valladares, Ernesto Luis, 350
Ezzini, Saad, 284

Fellbaum, Christiane, 68
Fernandes, Rafael Macario, 40
François, Alexandre, 59
Friezas Gonçalves, Christophe, 161, 172, 191
Fusco, Achille, 1

Garg, Meenakshi, 370
Ghezaiel Hammouda, Nadia, 237
Gowrishankar, Sudeep, 370
Graovac, Jelena, 438
Guan, Kevin, 68
Guo, Siwen, 102
Gupta, Divya, 361
Gutiérrez, Yoan, 96

Hadiza, Assoumana Souley, 388
Haider, Sajjad, 408
Hamed, Mohamed Motasim, 378, 418, 449
Hammale, Mourad, 131
Hennara, Khalil, 378, 418, 449
Hreden, Muhammad, 418, 449
Hübner, David, 309

Idrissi, Najlae, 246
Ignatev, Daniil, 257
Ikonić Nešić, Milica, 438
Ilman, Tetiana, 429
Inayat, Farah, 408
Ingason, Anton Karl, 140

Jameel, Shoaib, 326
Jankowski, Hubert, 18
Jha, Aryan Kumar, 339
Joshi, Aditya, 326

Kaboré, Abdoul Kader, 272
Kanel, Avery Cole, 264
Khallaf, Nouran, 199
Khan, Atai Waris, 339

Kirouchenassamy, Badmavasan, 370
Klein, Jacques, 272, 284
Kumar, Sandeep, 339

Lamsiyah, Salima, 150, 161, 172, 191, 199, 226
Laracha, Bouazza, 264
Lareau, François, 80
Lasmanis, Viesturs Jūlijs, 295
Lbouhli, Imrane, 264
Lothritz, Cedric, 102
Lotman, Eno-Martin, 309

Martín Arista, Javier, 8
Mathiak, Brigitte, 398
Mechler, Johanna, 140
Mekaoui, Salma, 121
Meroni, Fabio, 59
Mitkov, Ruslan, 30, 199, 226, 237, 350, 361
Miyagawa, So, 50
Mohammadi, Gelareh, 326
Mukusheva, Albina, 1
Muñoz Guillena, Rafael, 96
Murugaraj, Keerthana, 161

Nikolov, Nikola S., 121
Nikolova-Koleva, Ivelina, 309

Ouattara, Maimouna, 272

Petalinkar, Saša, 438
Philippy, Fred, 272, 284
Protonina, Kseniia, 257

Rabiei, Marzieh, 89
Razzak, Imran, 326
R'baiti, Jihad, 30
Roche, Mathieu, 246

Sawhney, Sukhna, 370
Schommer, Christoph, 172, 191
Schuler, Christian, 264
Sepúlveda-Torres, Robiert, 96
Shahid, M Kaab Bin, 339
Silberztein, Max, 59, 89
Singh, Pooja, 339
Singh, Rahul, 370
Soufiane, Mehdi, 131
Stanković, Ranka, 438
Stefánsdóttir, Lilja Björk, 140

Tebourbi, Hedi, 161, 172

Valentin, Sarah, 246
Valline, Julian, 102
Vlachidis, Andreas, 304

Xu, Han, 150

Younes, Yousef, 398
Yousefi, Shahin, 350

Zahir, Jihad, 318, 388
Zarghili, Arsalane, 121
Zerhouni, Kawtar, 131
Zhao, Xiaojing, 150