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**

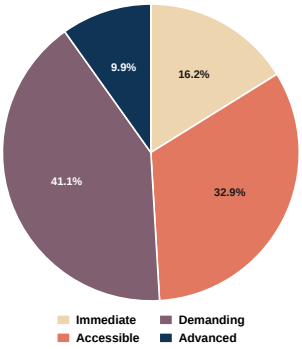


Figure 1: Distribution of the four readability levels in BAYAN.

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

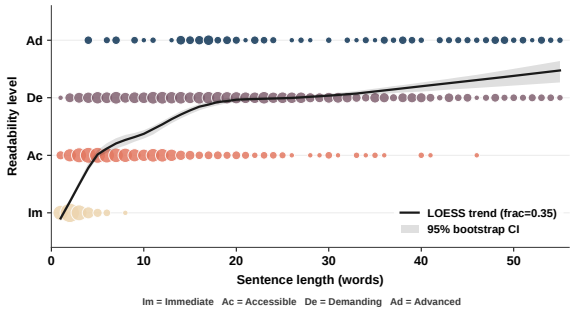


Figure 2: Association between sentence length and annotated readability level. The line shows the LOESS trend, and the shaded area represents the 95% bootstrap confidence interval.

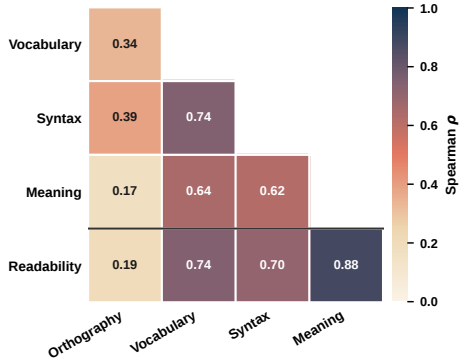


Figure 3: Spearman correlations between the diagnostic dimensions and the overall readability label in BAYANEX.

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.

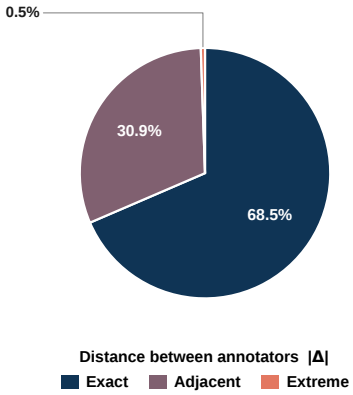


Figure 4: Distribution of ordinal distances between the two readability labels for sentences retained by both annotators.

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.

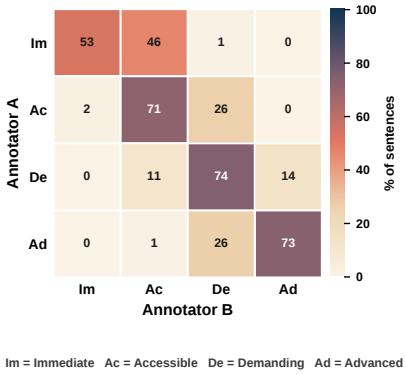
C.6 Sentence Length and Readability

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.

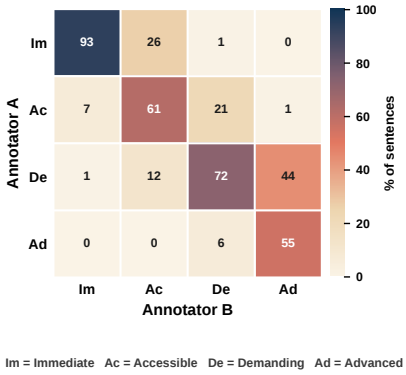
To further examine this relationship, we report the correlations between sentence length and the four diagnostic dimensions in BAYANEX. Sentence length is positively associated with all four dimensions, with the strongest association observed for syntactic difficulty ($\rho = 0.866$), followed by vocabulary ($\rho = 0.683$), 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

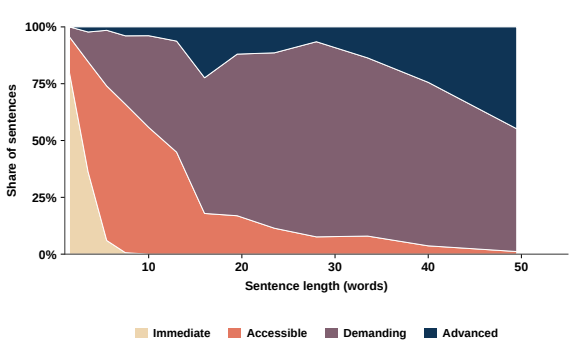


(a) Agreement normalised by Annotator A's labels.

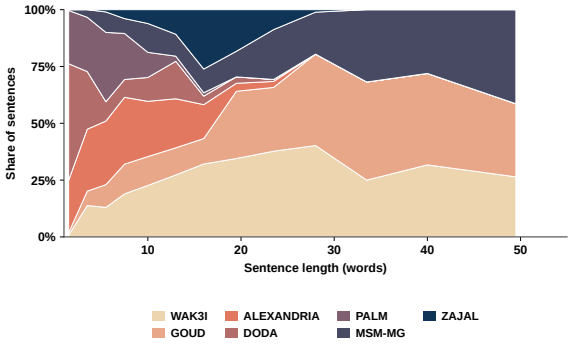


(b) Agreement normalised by Annotator B's labels.

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.

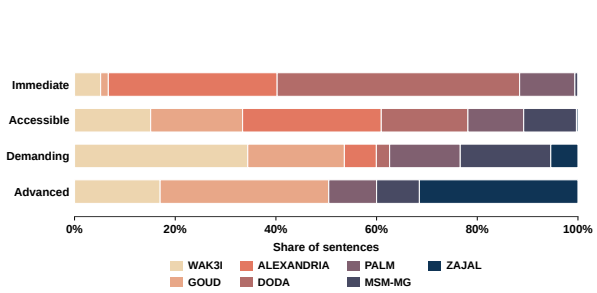


(a) By readability level.

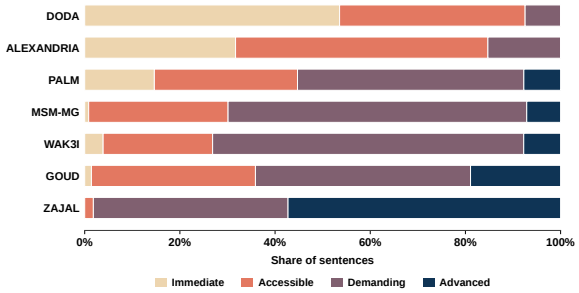


(b) By source subset.

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.

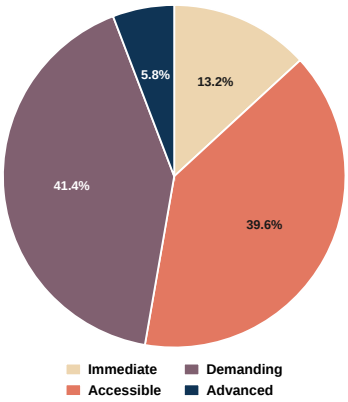


Figure 8: Readability distribution of BAYANEX.

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

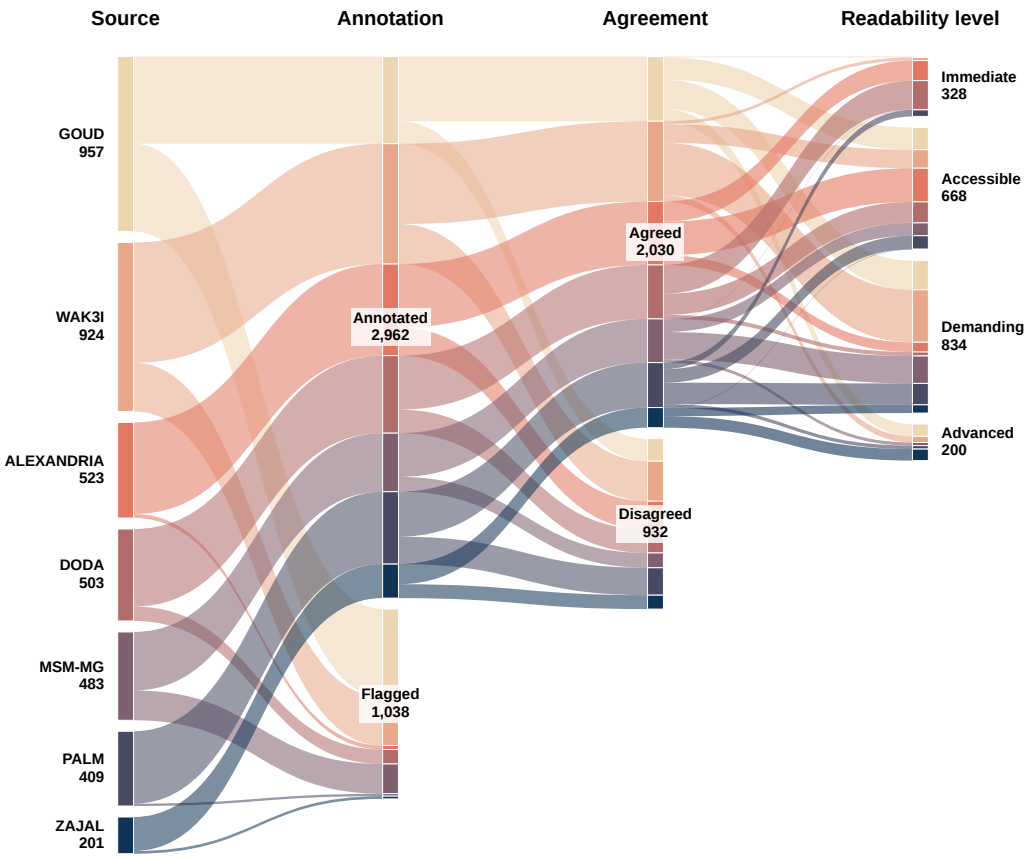
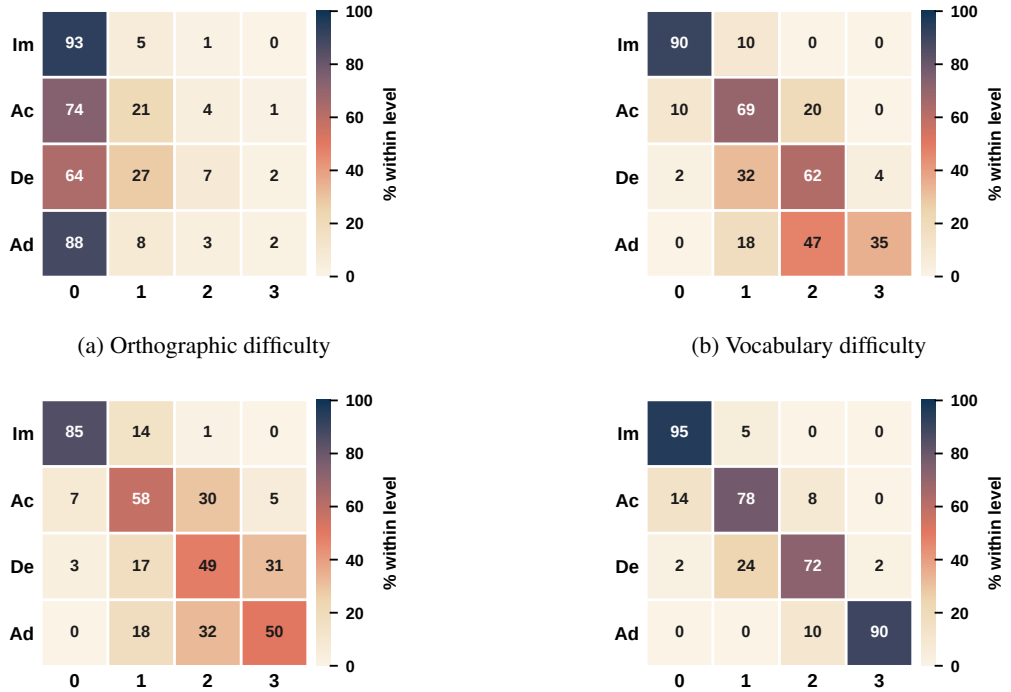


Figure 9: Flow from the 4,000 sampled candidates through suitability review and agreement filtering to BAYAN and BAYANEX.

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.

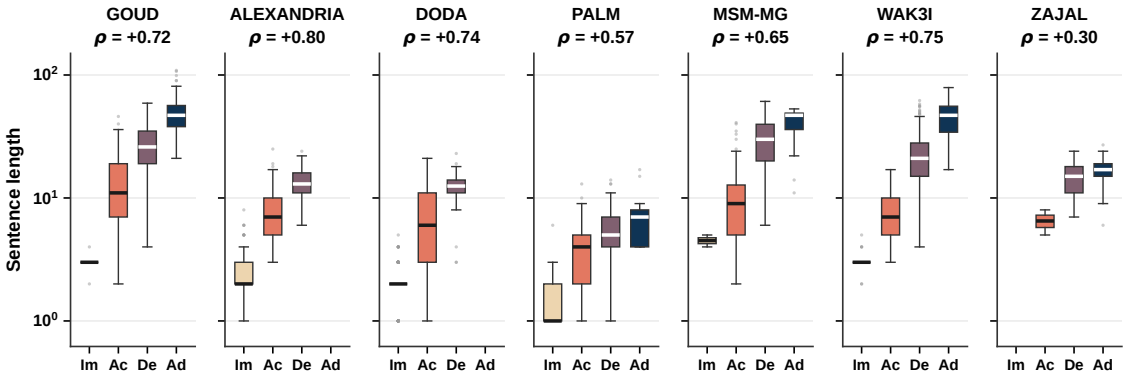


Figure 12: Sentence-length distributions across readability levels within each source.

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.

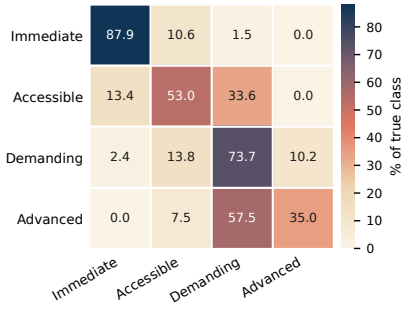
Features	Accuracy	Macro-F1	QWK
SVR → ordinal			
formulas	**68.6**	63.7	**0.756**
surface+formulas	64.4	58.3	0.714
surface	64.6	58.6	0.712
char TF-IDF	57.2	38.0	0.580
char+word TF-IDF	55.5	34.8	0.554
word TF-IDF	52.6	37.3	0.511
Logistic Regression			
surface+formulas	68.1	64.0	0.754
surface	68.1	63.9	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.

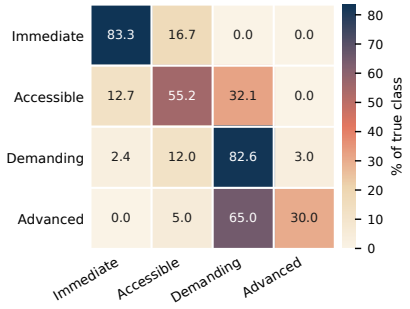
Model	Size	Accuracy ↑	Macro-F1 ↑	QWK ↑
Darija-specific				
Atlas-Chat-9B	9B	54.8	**51.5**	0.577
GemMaroc-Qwen2.5-7B-darija	7B	42.3	35.5	0.418
NileChat-3B	3B	53.6	45.2	0.560
Arabic				
Fanar-1-9B-Instruct	9B	41.0	32.6	0.480
Jais-2-8B-Chat	8B	37.3	32.9	0.406
ALLaM-7B-Instruct	7B	32.2	28.6	0.235
General-purpose				
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.

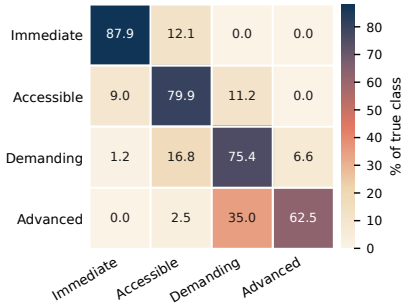
Model	Size	Accuracy ↑	Macro-F1 ↑	QWK ↑
Darija-specific				
Atlas-Chat-9B	9B	**62.4**	**61.9**	**0.723**
GemMaroc-Qwen2.5-7B-darija	7B	46.4	41.5	0.482
NileChat-3B	3B	51.4	42.0	0.522
Arabic				
Fanar-1-9B-Instruct	9B	56.5	49.3	0.600
Jais-2-8B-Chat	8B	45.9	44.7	0.581
ALLaM-7B-Instruct	7B	43.0	39.0	0.471
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



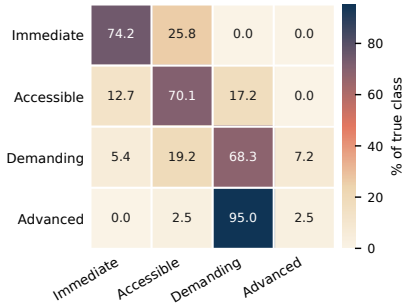
(a) AARI.



(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.