Assessment of Human-in-the-Loop Multi-LLMs for Low-Resource Educational Data Expansion

Christophe Friezas Gonçalves[*†] **, Hedi Tebourbi**[†], **Christoph Schommer**, **Salima Lamsiyah**

Faculty of Science, Technology and Medicine (FSTM),
University of Luxembourg, Luxembourg

christophe.friezas@uni.lu, hedi.tebourbi.001@student.uni.lu,
christoph.schommer@uni.lu, salima.lamsiyah@uni.lu

[†] Equal contribution [*] Corresponding author

Abstract

Low-resource languages often lack the pedagogically grounded datasets needed for language learning and assessment. To address this gap, we assess the potential of large language models (LLMs) to synthetically expand educational datasets in a low-resource language setting. We evaluate LLM performance within a complex, multi-layer annotation framework. Our approach builds upon LuxDiagRC, a Luxembourgish reading comprehension dataset fully annotated by human experts, which serves both as a methodological foundation and as a gold-standard benchmark. The evaluation follows the two-layer design of LuxDiagRC. In the first layer, three LLMs annotate linguistic features in Luxembourgish source texts. In the second layer, the models generate multiple-choice reading comprehension questions grounded in these annotations. Our results indicate that a multi-LLM approach combined with simple human curation outperforms individual and combined model approaches. The framework produces diagnostically grounded reading comprehension questions that closely align with the structure and pedagogical objectives of the original dataset. Overall, the findings suggest that combining multiple LLMs with expert human oversight form a promising strategy for the scalable expansion of educational corpora in a low-resource language setting, while highlighting the continued necessity of human expertise in dataset quality and pedagogical validity.[1]

1 Introduction

Second language acquisition is a complex process shaped by multiple interacting factors, including learners initial abilities, instructional conditions, and the availability of appropriate learning and assessment resources (Duke and Cartwright, 2021; Sabatini et al., 2024). Reading research and diagnostic assessment studies have shown that effective comprehension support requires not only accurate evaluation of answers, but also pedagogically grounded materials that capture the linguistic and cognitive sources of learner difficulty (Duke and Cartwright, 2021; Hemati and Baghaei, 2020; Liao et al., 2024). When it comes to low-resource languages, the availability of such materials is particularly constrained.

For Luxembourgish, a language spoken by approximately 400,000 people, the label of low-resource language applies (Gilles, 2019). Despite this classification, the digital era has led to a gradual increase in available data, largely driven by the expansion of social media platforms. In addition, the scientific community, specifically Natural Language Processing (NLP), found an increased interest in the language due to its inherently multilingual nature and culture (Plum et al., 2025). Recent work has contributed important Luxembourgish NLP resources and models, including LuxemBERT, LuxBank, and new benchmarks for generation and proficiency evaluation (Lothritz et al., 2022; PLUM et al., 2024; Plum et al., 2025; Lothritz et al., 2026). Nevertheless, within the educational sector, resource scarcity remains especially pronounced, as language teaching requires highly specialized and pedagogically tailored materials.

Reading comprehension (RC) datasets have played a central role in both NLP and educational assessment, with benchmarks such as SQuAD and RACE supporting large-scale evaluation, and frameworks such as STARC highlighting the importance of structured, diagnostically meaningful question design (Rajpurkar et al., 2016; Lai et al., 2017; Berzak et al., 2020). For Luxembourgish, LuxDiagRC addressed this gap by introducing a diagnostically rich reading comprehension corpus annotated by human experts, combining linguistic

[1] GitHub Repository

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 172–180
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_019

annotations with cognitively grounded multiple-choice questions (Gonçalves et al., 2026). However, because such resources require substantial expert effort, their expansion remains difficult to scale.

Recent studies show that LLMs can support annotation, synthetic data generation, and educational content creation, although human oversight remains important for quality control, especially in low-resource settings (Nasution and Onan, 2024; Pangakis and Wolken, 2024; Yu et al., 2024; Wang et al., 2024; Dutulescu et al., 2025; Hu et al., 2025; Säuberli and Clematide, 2024). In this paper, we assess a human-in-the-loop pipeline for synthetical expansion of a RC dataset for second language acquisition. Our assessment builds on LuxDiagRC, which serves both as the methodological foundation and as the gold-standard benchmark for evaluation (Gonçalves et al., 2026). Following its two-layer design, three LLMs first generate linguistic annotations for Luxembourgish short stories and then generate multiple-choice questions based on these annotations. We evaluate both individual and merged model outputs to assess the value of a multi-LLM strategy under human supervision.

The results show that the pipeline performs best when combining the outputs of multiple models in the annotation step, while also generating meaningful and textually grounded multiple-choice questions. These findings support the potential of using LLMs to expand diagnostically rich datasets, while also highlighting limitations in low-resource and instruction-sensitive settings.

2 Related Work

This section reviews the literature relevant to our study, focusing on three areas: Luxembourgish NLP, reading comprehension datasets, and LLM-based annotation and data synthesis.

2.1 Luxembourgish in NLP

Luxembourgish presents a complex multilingual context for NLP research, shaped by strong influences from German and French (Gilles, 2019). The language is further characterized by substantial orthographic and linguistic variation, partly due to limited standardization and formal teaching (Lutgen et al., 2025). These characteristics, combined with data scarcity, make Luxembourgish a challenging setting for both model development and evaluation. Recent work has led to the development of Luxembourgish-specific resources and models, including LuxemBERT (Lothritz et al., 2022) and LuxBank (PLUM et al., 2024). Despite these advances, such models remain constrained by limited training data and often struggle to capture the full variability of the language. In contrast, recent studies show that large commercial LLMs can achieve competitive performance on Luxembourgish tasks, leveraging cross-lingual transfer from high-resource languages (Lothritz et al., 2026). This makes them a promising tool for supporting resource development in low-resource settings.

2.2 Reading Comprehension Datasets

Reading comprehension datasets play a central role in both NLP and educational assessment. Large-scale benchmarks such as SQuAD (Rajpurkar et al., 2016) and RACE (Lai et al., 2017) have driven progress in extractive and multiple-choice question answering. More recent datasets, such as SciDQA (Singh et al., 2024) and MedMCQA (Pal et al., 2022), extend this paradigm to specialized domains. However, most existing datasets are designed for model evaluation rather than for pedagogical or diagnostic purposes, and often lack explicit links between textual evidence and answer choices. The STARC framework addresses this limitation by structuring answer options around text spans and supporting fine-grained diagnostic analysis of comprehension (Berzak et al., 2020). Building on this idea, LuxDiagRC (Gonçalves et al., 2026) introduces a diagnostically rich Luxembourgish RC dataset with a two-layer annotation scheme linking linguistic features to cognitively grounded questions. While this design enables detailed analysis of learner understanding, the reliance on expert human annotation limits scalability, motivating the need for automated expansion methods.

2.3 Automatic Annotation and Data Synthesis

Automatic annotation has been widely studied in NLP, particularly for tasks such as classification and linguistic tagging (Nasution and Onan, 2024; Pangakis and Wolken, 2024; Yu et al., 2024). Recent work has explored the use of LLMs as annotators, showing that they can achieve competitive performance while significantly reducing annotation cost (Yu et al., 2024; Nasution and Onan, 2024). However, their outputs can be inconsistent or biased, especially for complex or low-resource tasks, requiring human verification (Rønningstad et al., 2024). Hybrid human–LLM approaches have

therefore been proposed, where LLM-generated labels are filtered or corrected by human annotators to ensure quality (Wang et al., 2024). In parallel, synthetic data generation using LLMs has shown promise for improving downstream models and reducing reliance on manually labeled data (Pangakis and Wolken, 2024). In educational contexts, LLMs have also been used for generating assessment items, including multiple-choice questions, although ensuring pedagogical validity and high-quality distractors remains a challenge (Dutulescu et al., 2025).

2.4 Positioning of Our Work.

This work builds on these directions by combining LLM-based annotation, synthetic data generation, and human-in-the-loop validation in a unified assessment paradigm. We adopt a heterogeneous multi-LLM approach, employing three LLMs as annotators alongside a human overseer for quality control. Unlike prior work that focuses primarily on high-resource languages or single-model settings, we investigate a multi-LLM strategy for low-resource educational data expansion. Furthermore, our work explicitly targets the educational sector, leveraging the LuxDiagRC dataset as both a foundation and a benchmark. This allows us to evaluate not only annotation quality, but also the ability of LLMs to generate pedagogically grounded assessment items in a challenging low-resource setting.

3 Methodology

This section presents the human-in-the-loop multi-LLM pipeline for the assessment of the synthetical expansion of LuxDiagRC. We retain its two-layer design and extend it through LLM-based generation under human supervision. As shown in Figure 1, each model independently processes both layers. In Layer 1, models generate linguistic annotations reviewed at both individual and merged levels. In Layer 2, models generate diagnostically grounded questions, which are validated by the human annotator.

3.1 LuxDiagRC Overview

LuxDiagRC (Gonçalves et al., 2026) consists of 640 multiple-choice questions across 16 Luxembourgish short stories, fully annotated by human experts. The dataset follows a two-layer design. In the first layer, source texts are annotated with linguistic features that capture phenomena relevant to Luxembourgish reading difficulty. In the second layer, multiple-choice reading comprehension questions are constructed and linked to critical and distractor spans following the STARC framework (Berzak et al., 2020). Together, these layers support fine-grained analysis of learner comprehension by connecting answer choices to underlying linguistic phenomena.

In this work, we use LuxDiagRC as the reference dataset for evaluating LLM-generated annotations and questions. The linguistic feature set used in Layer 1 is summarized in Table 1. To align with token-level annotation constraints, we exclude tags requiring sentence-level analysis (e.g., LEX-IDIOM-ZLS, SYN-EMBEDDED). Experiments are conducted on 7 of the 16 texts in LuxDiagRC. We chose to use only texts with a word count higher than 1000 to provide sufficient context for LLM processing and evaluative analysis.

3.2 Automatic Annotation Pipeline

Rather than recreating the original dataset, we use LLMs to reproduce its two-layer annotation process and assess how well synthetic and human supervised outputs align with the fully human-created benchmark. The human-in-the-loop evaluation entails a shallow revision of the resulting two layer annotations.

3.2.1 Models

The pipeline employs three LLMs in a heterogeneous multi-LLM setup: GPT (5.4)[2], Claude Opus (4.6)[3], and DeepSeek (3.2)[4]. These models are selected based on prior evidence of complementary strengths in Luxembourgish tasks (Lothritz et al., 2026). The multi-LLM design aims to mitigate individual model weaknesses and improve coverage through aggregation.

For Layer 1, models are evaluated under zero-shot, few-shot, and chain-of-thought prompting. Zero-shot and few-shot serve as the primary evaluation settings, while chain-of-thought is used for qualitative analysis of model reasoning. In contrast, Layer 2 relies on zero-shot prompting, as few-shot settings led to imitation of the few-shot examples provided to the models.

3.2.2 Linguistic Annotation (Layer 1)

Layer 1 replaces human annotation with LLM-

[2]https://openai.com/index/introducing-gpt-5-4/
[3]https://www.anthropic.com/news/claude-opus-4-6
[4]https://api-docs.deepseek.com/news/news25120

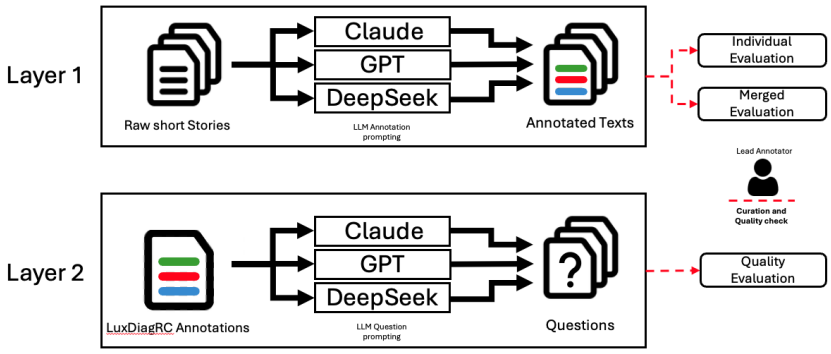

Figure 1: Overview of the automatic annotation pipeline and evaluation methodology. Layer 1 is evaluated in two steps: model outputs are assessed individually and after merging. All evaluations are supervised by a lead annotator for quality control.

based tagging of linguistic features. All three models receive identical prompts applied to the same seven raw Luxembourgish texts, including detailed instructions and linguistic feature definitions (Table 1). Each text is processed independently by each model, resulting in 21 prompt executions per prompting paradigm. To ensure comparability across models, strict formatting constraints are enforced, including token-level tagging, preservation of the original text, and exhaustive annotation of relevant features.

The prompting strategies differ only in the level of guidance: zero-shot uses base instructions, few-shot includes annotated examples, and chain-of-thought adds brief reasoning for ambiguous cases. This design allows us to assess both performance and the interpretability of model outputs.

3.2.3 Question Generation (Layer 2)

Layer 2 focuses on generating diagnostically grounded multiple-choice reading comprehension questions using the three selected LLMs. The generation process is guided by the STARC framework (Berzak et al., 2020) and the PIRLS cognitive taxonomy (Sabatini et al., 2024), ensuring both structural alignment and pedagogical validity. Each model is prompted to generate ten questions per short story, following the predefined distribution of four cognitive types: retrieval, interpretation, inferential, and evaluative (Gonçalves et al., 2026).

Prompts enforce strict constraints on both question structure and span selection. Each question must include four answer options: a correct answer, a misunderstanding-based option, a distractor-based option, and a textually unsupported option. All options are derived from critical and distractor spans explicitly grounded in the source text, in accordance with the STARC paradigm.

To isolate the evaluation of question generation, models are provided with gold-standard annotations from LuxDiagRC as input, ensuring a stable and controlled basis for generation independent of Layer 1 variability.

Table 1: Linguistic feature annotation tags with short descriptors used in layer 1 of the LuxDiagRC dataset creation (Gonçalves et al., 2026).

Tag ID	Description
LEX-LOAN-FR	Word of clear French origin.
LEX-LOAN-DE	Word of clear German origin.
LEX-FALSE-FRIEND	Word prone to cross-lingual confusion.
LEX-IDIOM-ZLS	Idiomatic expression.
LEX-RARE-ZLS	Rare or low-frequency word.
SYN-VERB-FINAL	Verb-final word order.
SYN-VERB-SEP	Split verbal construction.
SYN-EMBEDDED	Embedded or nested relative clause.
MORPH-N-RULE	Instance of the *Eifeler Regel*.
DISC-COREF-AMBIG	Pronoun with ambiguous referent.
ORTHO-PHONO-DIVERGE	Word exhibiting divergence between written and spoken forms.

3.3 Multi-LLM Framework with Human-in-the-Loop Verification

In Layer 1, the LLMs generate individual annotation outputs in parallel, which are subsequently combined using two distinct merging strategies. The first is a simple aggregation approach, in which all annotations are merged via union over token-level tags, resulting in a comprehensive but poten-

tially noisy annotation set. The second approach incorporates human-in-the-loop curation. A human annotator reviews the aggregated output and removes incorrect tags to produce a shallow yet refined annotation. This strategy mitigates common annotation errors while substantially reducing the required annotation time, requiring approximately one day compared with the considerably more time-intensive process used to construct the human-annotated dataset (Gonçalves et al., 2026). In Layer 2, the human annotator assumes an evaluative role, assessing both the difficulty level of the generated questions and the coherence between selected spans and corresponding answer options.

4 Results & Discussion

This section presents and discusses the results of the two annotation layers, with the goal of evaluating both model performance and the effectiveness. Beyond reporting quantitative scores, we analyze how different prompting strategies and aggregation mechanisms influence the quality of linguistic annotations and question generation. The results highlight both the limitations of current LLMs in low-resource language settings and the potential of multi-LLM and human-in-the-loop approaches.

4.1 Evaluation Methods

The evaluation protocol is structured along two complementary dimensions for each layer: quantitative and qualitative analysis. Quantitative evaluation provides performance measures based on predefined metrics, while qualitative analysis enables a deeper examination of model behavior, focusing on error patterns, reasoning strategies, and alignment with task requirements. The evaluation covers 7 Luxembourgish short stories, each annotated by 3 models under 3 prompting paradigms, yielding 56 valid annotation runs per experimental condition and 504 annotation-category pairs scored against the gold standard (GPT zero-shot excluded). For Layer 2, 208 questions are generated across 3 models and 7 texts.

For Layer 1, we employ the Jaccard similarity (Jaccard, 1901) coefficient to measure the overlap between predicted and gold annotations. The metric is computed over sets of unique word forms, ensuring that repeated occurrences do not bias the evaluation. This formulation emphasizes correct identification of linguistic phenomena rather than frequency. Qualitative analysis examines representative annotation cases per category, chain-of-thought reasoning traces, and systematic tagging error patterns to reinforce the quantitative findings.

For Layer 2, the quantitative evaluation assesses three properties: (i) distinctiveness of answer options, (ii) grounding of critical and distractor spans in the source text via exact string matching, and (iii) adherence to the predefined PIRLS distribution of cognitive question types. In addition, inter-model Jaccard similarity is computed to assess whether different models focus on similar textual regions. Qualitative evaluation examines the semantic validity and pedagogical coherence of each item, including alignment between cognitive type, question formulation, and supporting spans.

4.2 Linguistic Annotation Results

The results demonstrate that all models show limited capability in capturing fine-grained linguistic phenomena in Luxembourgish, as reflected by consistently low Jaccard scores across most categories, as seen in Table 2. Notably, the zero-shot configuration of GPT fails entirely to follow the annotation format, producing structured JSON outputs instead of inline annotations. This indicates a lack of controllability under constrained output specifications. Thus achieving no comparable results.

Overall performance remains relatively similar across models, with Claude achieving slightly higher scores and GPT consistently underperforming. The highest performance is observed for LEX-LOAN-FR (0.43), while several categories, such as DISC-COREF-AMBIG, show near-zero overlap with the gold standard. These results suggest that current LLMs struggle to reliably capture complex linguistic structures in low-resource settings.

Lexical categories achieve comparatively higher performance than syntactic and morphological ones. This can be attributed to the presence of loanwords from high-resource languages such as German and French, which are well represented in LLM training data. In contrast, categories requiring deeper structural or contextual reasoning, such as false friends or morphological rules, remain challenging. This pattern indicates that model performance is strongly influenced by training data distribution and lexical familiarity.

The LEX-FALSE-FRIEND category yields particularly low scores, reflecting the difficulty of capturing subtle cross-linguistic semantic and phonetic distinctions. These phenomena require contextual

and multilingual reasoning beyond surface-level lexical matching, which current models fail to consistently exhibit.

Among syntactic features, SYN-VERB-FINAL achieves relatively stronger performance, suggesting that models can capture local and consistent grammatical patterns. However, SYN-VERB-SEP remains challenging, as models often identify only one component of separable constructions. This indicates partial structural understanding without full adherence to the annotation scheme.

Qualitative analysis further shows systematic overgeneralization in the MORPH-N-RULE category. Models tend to associate the rule with specific lexical items rather than correctly identifying its contextual application. This behavior extends to the ORTHO-PHONO-DIVERGE category, suggesting that models rely primarily on surface-level cues rather than rule-based generalization. Taken together, these findings highlight a fundamental limitation of current LLMs in modeling linguistically grounded transformations, which is particularly critical for educational applications requiring precise and explainable annotations.

4.2.1 Attention and Context Window

To assess the impact of context length on annotation quality, we introduce a chunking strategy in which texts are split into segments of approximately 200 words. This modification eliminates output truncation observed in the full-text condition, reducing unscored outputs to zero. As a result, the average Jaccard score increases from 0.095 to 0.113. Performance gains are most noticeable for syntactic categories, particularly SYN-VERB-FINAL (+0.074), indicating that local grammatical patterns are easier to capture within shorter contexts. Improvements are also observed for word-level features such as MORPH-N-RULE (+0.025) and LEX-FALSE-FRIEND (+0.027). However, performance decreases for LEX-LOAN-FR (-0.061), suggesting that certain lexical judgments require a broader discourse context. These results demonstrate that context management plays a critical role in LLM-based annotation tasks. While reducing context improves local pattern detection and output reliability, it may remove essential global information. This trade-off highlights the importance of adaptive context strategies in multi-LLM systems.

4.2.2 Merged Annotation

The merged condition is based on a simple aggregation of the annotations produced by the individual models and yields mixed results. As shown in Tables 2, merging can improve weaker categories by increasing annotation coverage across models. However, it also reduces performance in categories where individual models already performed better. This decline is mainly caused by over-tagging, since each model introduces its own false positives and their combination expands the total number of errors, which is penalized by the Jaccard similarity metric. These findings indicate that naive ensemble strategies are insufficient for structured annotation tasks.

4.2.3 Human Guided Merging

To address this limitation, we evaluate a human-in-the-loop filtering step that removes over-tagged annotations from the aggregated outputs. This approach preserves the recall benefits of model aggregation while restoring precision through manual validation. As shown in Table 2, this hybrid strategy substantially outperforms both individual models and simple aggregation.

These findings provide strong evidence for the effectiveness of human-in-the-loop verification in multi-LLM settings. Rather than replacing human expertise, multiple imperfect models can be leveraged to maximize coverage, while human validation ensures accuracy. This hybrid paradigm is particularly well suited for complex linguistic annotation tasks in low-resource educational contexts.

Table 3: Quantitative results for the Layer 2 question generation task.

Metric	GPT 5.4	Claude 4.6	DeepSeek 3.2
PIRLS Distribution	70/70	68/70	70/70
Distinct Answers	70/70	68/68	70/70
Critical Span Found	61/70	60/68	58/70
Distractor Span Found	62/70	63/68	58/70

4.3 Question Generation Results

In contrast to the annotation task, models demonstrate good performance in question generation (Table 3). All models satisfy the distinctiveness constraint, and adherence to the PIRLS distribution is nearly perfect. A noteworthy observation regarding the Claude results is reported in the table. Specifically, the Claude output exhibited parsing errors for two questions (Text 2 and Text 7), which per-

Table 2: Jaccard similarity scores for Layer 1 annotations under the full-text condition and 200-word chunking. Each model was evaluated in three prompting settings: zero-shot (ZS), few-shot (FS), and chain-of-thought (CoT). The merged result is based on a simple aggregation of individual tags (Aggr) and a human-guided elimination of over-tagging (Hum).

Tag ID	Full Text										Chunking										
	GPT 5.4			Claude 4.6			DeepSeek 3.2			Merged	GPT 5.4			Claude 4.6			DeepSeek 3.2			Merged	
	ZS	FS	CoT	ZS	FS	CoT	ZS	FS	CoT	Aggr	ZS	FS	CoT	ZS	FS	CoT	ZS	FS	CoT	Aggr	Hum
LEX-LOAN-FR	/	0.39	0.42	0.27	0.27	0.28	0.29	0.32	**0.43**	0.18	/	0.35	0.36	0.24	0.20	0.23	0.24	0.25	0.31	0.12	**0.74**
LEX-LOAN-DE	/	0.16	0.15	0.26	**0.27**	0.23	0.17	0.25	0.25	0.22	/	0.17	0.23	0.21	0.21	0.23	0.18	0.22	0.21	0.17	**0.56**
LEX-FALSE-FRIEND	/	0.01	0.02	0.01	0.00	0.01	0.00	0.01	0.02	**0.06**	/	0.05	0.04	0.02	0.05	0.03	0.01	0.07	0.03	0.13	**0.17**
LEX-RARE-ZLS	/	0.13	0.12	**0.15**	0.13	0.02	0.13	0.13	0.14	0.14	/	0.12	0.11	0.12	0.14	0.13	0.13	0.18	0.28	0.12	**0.46**
SYN-VERB-FINAL	/	0.04	0.13	0.21	0.20	0.19	0.08	0.04	0.12	**0.22**	/	0.19	0.18	0.23	0.21	0.21	0.20	0.23	0.15	0.24	**0.50**
SYN-VERB-SEP	/	0.04	0.05	0.09	**0.11**	0.07	0.01	0.00	0.00	0.07	/	0.10	0.08	0.09	0.11	0.09	0.04	0.08	0.02	0.07	**0.22**
MORPH-N-RULE	/	0.01	0.01	0.06	0.06	0.06	0.00	0.01	0.00	**0.08**	/	0.02	0.02	0.09	0.12	0.09	0.01	0.03	0.02	0.14	**0.27**
DISC-COREF-AMBIG	/	0.00	0.01	**0.02**	**0.02**	0.00	0.00	0.00	0.00	0.01	/	0.01	0.01	0.02	0.01	0.01	0.01	0.08	0.05	0.09	**0.19**
ORTHO-PHONO-DIVERGE	/	0.01	0.01	0.01	0.02	0.02	0.00	0.00	0.01	**0.04**	/	0.02	0.02	0.02	0.03	0.03	0.00	0.03	0.02	0.08	**0.11**

sisted despite multiple prompt adaptations. Thus, reducing the total from 70 to 68.

Over 80% of the generated questions correctly ground both critical and distractor spans in the source text. A key finding is the high diversity of span selection across models, with only 8 overlapping cases out of 208 questions. This indicates that independently operating models capture complementary aspects of the input, reinforcing the value of a multi-LLM approach for generating diverse and pedagogically rich question sets. Qualitative analysis shows that models successfully construct structurally valid questions aligned with their intended cognitive type. However, limitations emerge for evaluative and inferential questions, where complexity often exceeds the intended learner level. This suggests that controlling cognitive difficulty remains an open challenge. Overall, these results demonstrate that LLMs are effective for structured question generation. Their outputs benefit from multi-LLM diversity. They do require additional control mechanisms to ensure pedagogical appropriateness.

5 Conclusion

This paper examined a the use of LLMs for linguistic annotation and question generation in Luxembourgish reading materials. The results show that the models performed better on question generation than on fine-grained linguistic annotation. For Layer 1, performance was generally low, especially for categories requiring structural or contextual understanding. Chunking improved annotation stability, while simple merging increased false positives. Human-guided filtering improved the merged results. For Layer 2, the models generated well-formed questions with strong span grounding and high adherence to the required distribution. The low overlap in selected spans also suggests complementary behavior across models. However, some evaluative and inferential questions remained too difficult or ambiguous. Overall, the findings suggest that current LLMs are more reliable for question generation than for detailed linguistic annotation in this setting, and that combining multiple LLMs is most useful when supported by human verification.

Limitations & Future Work Several limitations should be noted. First, the study uses general-purpose LLMs that are not specifically optimized for Luxembourgish. Low performance was therefore expected, especially for fine-grained linguistic phenomena. The focus of the study is not model specialization, but the effect of a unified curation strategy within a multi-LLM framework. Second, the human-in-the-loop component is limited to a simplified elimination-based curation process. Although this improves scores with limited additional effort, it does not allow full correction or refinement of the aggregated outputs. In addition, using a single annotator may introduce individual bias. Thirdly, the study evaluates Layer 2 under an optimal setting rather than in an end-to-end configuration, as the performance observed in Layer 1 was insufficient to support a reliable transition to Layer 2. Future work will evaluate Luxembourgish-specific models and involve multiple annotators to compare alternative human-in-the-loop strategies and better balance annotation quality with time efficiency.

References

Yevgeni Berzak, Jonathan Malmaud, and Roger Levy. 2020. STARC: Structured annotations for reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Lin-*

guistics, pages 5726–5735, Online. Association for Computational Linguistics.

Nell K. Duke and Kelly B. Cartwright. 2021. The science of reading progresses: Communicating advances beyond the simple view of reading. *Reading Research Quarterly*, 56(S1):S25–S44.

Andreea Dutulescu, Stefan Ruseti, Denis Iorga, Mihai Dascalu, and Danielle S. McNamara. 2025. Ymcq: Reasoning-enhanced mcq generation. In *Artificial Intelligence in Education*, pages 308–315, Cham. Springer Nature Switzerland.

Peter Gilles. 2019. *39. Komplexe Überdachung II: Luxemburg. Die Genese Einer Neuen Nationalsprache*, pages 1039–1060. De Gruyter Mouton, Berlin, Boston.

Christophe Friezas Gonçalves, Salima Lamsiyah, and Christoph Schommer. 2026. LuxDiagRC: A diagnostic reading comprehension corpus for Luxembourgish with linguistic and cognitive annotation layers. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 532–541, Rabat, Morocco. Association for Computational Linguistics.

Seyed Jamal Hemati and Purya Baghaei. 2020. A cognitive diagnostic modeling analysis of the english reading comprehension section of the iranian national university entrance examination. *International Journal of Language Testing*, 10(1):11–32.

Bihao Hu, Jiayi Zhu, Yiying Pei, and Xiaoqing Gu. 2025. Exploring the potential of llm to enhance teaching plans through teaching simulation. *npj Science of Learning*, 10(1):7.

Paul Jaccard. 1901. Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bull Soc Vaudoise Sci Nat*, 37:241–272.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Manqian Liao, Hong Jiao, and Qiwei He. 2024. Explanatory cognitive diagnosis models incorporating item features. *Journal of Intelligence*, 12(3):32.

Cedric Lothritz, Jordi Cabot, and Laura Bernardy. 2026. Testing low-resource language support in LLMs using language proficiency exams: the case of Luxembourgish. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 2453–2476, Rabat, Morocco. Association for Computational Linguistics.

Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Lisa Veiber, Tegawende Bissyande, Jacques Klein, Andrey Boytsov, Clément Lefebvre, and Anne Goujon. 2022. LuxemBERT: Simple and practical data augmentation in language model pre-training for Luxembourgish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5080–5089, Marseille, France. European Language Resources Association.

Anne-Marie Lutgen, Alistair Plum, Christoph Purschke, and Barbara Plank. 2025. Neural text normalization for Luxembourgish using real-life variation data. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 115–127, Abu Dhabi, UAE. Association for Computational Linguistics.

Arbi Haza Nasution and Aytuğ Onan. 2024. Chatgpt label: Comparing the quality of human-generated and llm-generated annotations in low-resource language nlp tasks. *IEEE Access*, 12:71876–71900.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Nicholas Pangakis and Sam Wolken. 2024. Knowledge distillation in automated annotation: Supervised text classification with LLM-generated training labels. In *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS 2024)*, pages 113–131, Mexico City, Mexico. Association for Computational Linguistics.

Alistair PLUM, Caroline DÖHMER, Emilia MILANO, Anne-Marie LUTGEN, and Christoph PURSCHKE. 2024. Luxbank: The first universal dependency treebank for luxembourgish. Association for Computational Linguistics. Accepted at 22nd Workshop on Treebanks and Linguistic Theories (TLT 2024).

Alistair Plum, Tharindu Ranasinghe, and Christoph Purschke. 2025. Text generation models for Luxembourgish with limited data: A balanced multilingual strategy. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 93–104, Abu Dhabi, UAE. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Egil Rønningstad, Erik Velldal, and Lilja Øvrelid. 2024. A GPT among annotators: LLM-based entity-level sentiment annotation. In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 133–139, St. Julians, Malta. Association for Computational Linguistics.

John Sabatini, Ann Kennedy, Erin Wry, and Matthias von Davier. 2024. PIRLS 2026 reading assessment framework. Technical report.

Andreas Säuberli and Simon Clematide. 2024. Automatic generation and evaluation of reading comprehension test items with large language models. In *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI) @ LREC-COLING 2024*, pages 22–37, Torino, Italia. ELRA and ICCL.

Shruti Singh, Nandan Sarkar, and Arman Cohan. 2024. SciDQA: A deep reading comprehension dataset over scientific papers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20908–20923, Miami, Florida, USA. Association for Computational Linguistics.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Danni Yu, Luyang Li, Hang Su, and Matteo Fuoli. 2024. Assessing the potential of llm-assisted annotation for corpus-based pragmatics and discourse analysis. *International Journal of Corpus Linguistics*, 29(4):534–561.