GRILE: A Benchmark for Grammar Reasoning and Explanation in Romanian LLMs

Adrian-Marius Dumitran [*]   **Alexandra-Mihaela Dănilă** [*]   **Angela-Liliana Dumitran**

University of Bucharest   University of Bucharest   University of Bucharest

mihaelaalexandradanila@gmail.com

marius.dumitran@unibuc.ro   dumitranangela@gmail.com

Abstract

LLMs (Large language models) have revolutionized NLP (Natural Language Processing), yet their pedagogical value for low-resource languages remains unclear. We present GRILE (Grammar Romanian Inference and Language Explanations) , the first open benchmark of 1,151 multiple-choice questions harvested from Romanian high-stakes exams (National Evaluation, Baccalaureate, university admissions). GRILE enables us to probe two complementary abilities of seven state-of-the-art multilingual and Romanian-specific LLMs: (i) selecting the correct answer, and (ii) producing linguistically accurate explanations. While Gemini 2.5 Pro reaches 83% accuracy, most open-weight models stay below 65%, and 48% of their explanations contain factual or pedagogical flaws according to expert review. A detailed error analysis pinpoints systematic weaknesses in morphology and in applying the latest DOOM3 orthographic norms. All data, code and a public web demo are released to catalyze future research. Our findings expose open challenges for trustworthy educational NLP in low-resource settings and establish GRILE as a new test-bed for controllable explanation generation and evaluation.

1 Introduction

Recent advancements in NLP, particularly through LLMs, have unlocked new opportunities for intelligent computer-assisted language learning. Although extensive benchmarks exist for English and other high-resource languages (Hendrycks et al., 2021; OpenAI, 2023), their efficacy in low-resource contexts remains under-explored—especially for tasks that demand fine-grained grammatical reasoning. Romanian, spoken by roughly 24 million people worldwide, illustrates this gap: it is comparatively underserved by both annotated corpora and NLP research. (Păiş and Tufis, 2023) mention that the number of identified NLP tools for Romanian represent only 15% of the tools available for English.

This paper bridges that gap through an interdisciplinary collaboration between linguists and computer scientists. We introduce **GRILE**, a new benchmark of 1,151 single-answer multiple-choice questions (MCQs) extracted from high-stakes Romanian examinations, including the National Evaluation, the Baccalaureate, and several university entrance tests (e.g., Law School, Police Academy).

Our investigation pursues two principal goals. First, we quantitatively assess how accurately state-of-the-art multilingual and Romanian-specific LLMs answer the questions in our dataset. Second, we examine their pedagogical potential by requiring each model to justify its choice with a concise grammatical explanation. The evaluation covers commercial systems such as **Gemini 2.5 Pro Experimental** and open-weight models such as **DeepSeek V3-0324**.

Initial results reveal substantial variation, with answer accuracy spanning from **38%** to **85%**, and chain-of-thought prompting generally yielding higher scores. Beyond answer prediction, we instruct Gemini 2.5 Pro Experimental to categorise every question as *Lexical*, *Morphological*, *Syntactic*, or *Phonetic*. Manual verification confirms more than 99% agreement, offering an additional lens for analysing error patterns. A qualitative evaluation of the generated explanations—conducted by a specialist in Romanian grammar—highlights persistent issues in terminological precision and normative compliance, underscoring the challenges of deploying LLMs as trustworthy tutors for low-resource languages.

Our main contributions are:

- **GRILE**, the first public Romanian grammar

---

*Equal contribution.

benchmark paired with expert-validated explanations;

- a comprehensive evaluation of multilingual and Romanian-specific LLMs on both *answering* and *explaining* high-stakes grammar questions;

- a linguist-driven qualitative study that uncovers systematic weaknesses and normative misalignments;

- an open-access dataset and interactive web application, released under a permissive licence on Zenodo[1], to catalyse future research and educational use.

These resources and findings advance the study of educational NLP in low-resource settings and lay the groundwork for more reliable, Romanian-focused grammatical tutoring and assessment tools.

2 Related Work

Benchmarking LLMs on Multiple-Choice Grammar Tasks. Recent studies have shown that large language models (LLMs) can achieve impressive results on standardized multiple-choice tests across various domains (Hendrycks et al., 2021; OpenAI, 2023). However, fine-grained grammatical assessments remain challenging for LLMs. For example, multilingual benchmarks focusing on grammar, such as the TELEIA (Mayor-Rocher et al., 2025) dataset for Spanish, demonstrate that LLMs struggle to reach native-like accuracy in nuanced grammar tasks (Mayor-Rocher et al., 2024). This motivates specialized evaluations of grammatical knowledge, especially for lower-resource languages like Romanian.

Multilingual vs. Language-Specific LLMs. Most large language models are predominantly trained on English data, which results in significantly better performance for English compared to other languages (Hu et al., 2020). Consequently, this has encouraged the development of monolingual or language-specific NLP models that often outperform multilingual counterparts on local linguistic tasks. For instance, Masala et al. (2020) introduced RoBERT, a Romanian-specific BERT-based model, demonstrating superior results over multilingual BERT across various Romanian NLP

tasks such as sentiment analysis, dialect identification, and diacritic restoration.

Similar outcomes have been consistently reported for other languages, including French, Polish, and Japanese, where monolingual transformer models regularly surpass general multilingual baselines in language-specific evaluations (Martin et al., 2020; Kuratov and Arkhipov, 2019; Conneau et al., 2020). More recently, the NLP community has seen efforts toward developing large-scale LLMs tailored specifically for individual languages, such as Finnish GPT-style models (Luukkonen et al., 2023) and Chinese LLaMA2-based models (Cui et al., 2024), achieving state-of-the-art performance on localized benchmarks.

In the Romanian context, Masala et al. (2024) released RoLLaMA2, the first open-source Romanian-specific large language model based on LLaMA-2, trained on curated Romanian corpora and fine-tuned through instruction tuning. RoLLaMA2 achieved leading performance across multiple Romanian NLP benchmarks, underscoring the effectiveness of language-specific training. Our study contributes to this growing body of literature by directly comparing multilingual models against Romanian-specific models like RoLLaMA2, focusing particularly on grammatical tasks to evaluate the effectiveness and viability of smaller, targeted language models against more extensive multilingual counterparts.

LLM Explanation Generation and Educational Feedback. Generating explanations for educational purposes extends beyond correctness, encompassing the quality and pedagogical soundness of explanations. Prior works in educational NLP formalized feedback comment generation tasks, highlighting the need for meaningful automated feedback (Nagata, 2019). Recently, generative LLMs like GPT-4 demonstrated potential in automatically generating high-quality explanations in science education, closely matching human-generated feedback (López Cortez et al., 2024). Human-in-the-loop validation remains crucial to ensure explanation accuracy and educational value.

Our work builds upon these directions, benchmarking Romanian-specific and general LLMs against a Romanian grammar multiple-choice dataset, examining common errors, and assessing explanation quality through expert validation. This contributes toward understanding LLM capabilities and limitations within educational contexts for
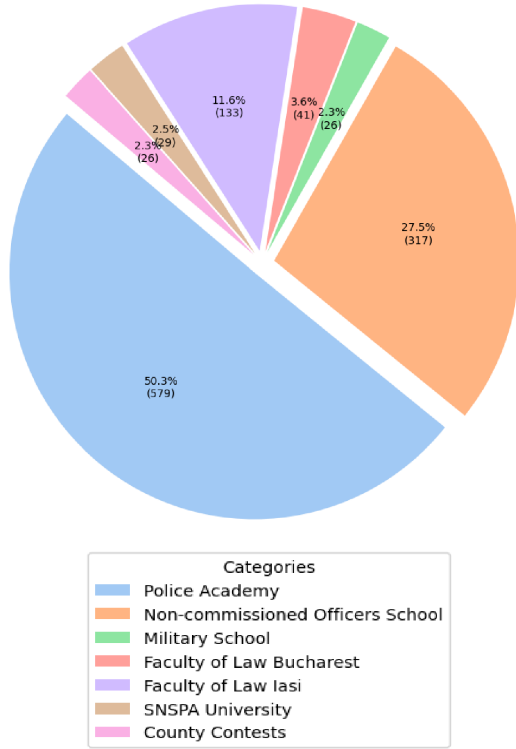
---

[1] https://zenodo.org/records/15504323

Figure 1: Dataset Sources

low-resource languages.

3 Dataset and Methodology

In this section we explain how we collect the data and the models we decided to benchmark in our test.

3.1 Data Collection and Structure

Figure 1 charts the distribution of our 1 151 single-answer multiple-choice questions (MCQs) were harvested from publicly available high-stakes Romanian examinations—the *National Evaluation*, *Baccalaureate*, and university entrance tests for Law School and the Police Academy (2010–2024). Source PDFs and scans were converted with Tesseract OCR, then parsed by rule-based scripts and manually spot-checked. Each record is stored as JSON with fields `question`, `options`, `answer`, `source`, and `year`; underlined focus words are preserved via surrounding underscores. Full OCR pipeline details and an example entry appear in Appendix A.[2]

---
[2]https://zenodo.org/records/15504323

3.2 LLM Selection

Our model selection aimed to evaluate both a state-of-the-art model known for strong reasoning capabilities and several other prominent models readily accessible for research purposes. Specifically, we included **Gemini 2.5 Pro Experimental** (Team et al., 2024) as it represented the leading edge of publicly available models at the time of our experiments. The other benchmarked models, including variants of **DeepSeek**, **Mistral**, **Llama 3.3**, and **Qwen**, were chosen primarily for their strong performance in various benchmarks and their convenient accessibility via the Together AI API[3], which facilitated systematic evaluation.

- **DeepSeek V3-0324** (DeepSeek-AI et al., 2025)

- **Mistral Small 24B**

- **Llama 3.3 70B Instruct** (Grattafiori et al., 2024)

- **Qwen 2.5 Coder 32B** (Yang et al., 2024)

- **Gemini 2.5 Pro Experimental**, **Gemini 2.0 Flash** (Team et al., 2024)

Furthermore, our few-shot prompting experiments (Section 4.2) involved evaluating **multiple fine-tuned Romanian variants** (e.g., RoLlama, RoMistral, RoGemma based on OpenLLM-Ro (Masala et al., 2024)) alongside their respective base models to analyze the interaction between fine-tuning and in-context learning. The full list of models and variants tested in the few-shot setting is provided in Appendix B.

3.3 Prompting Strategies

We evaluated LLM performance using two primary prompting approaches applied to each MCQ in the dataset:

3.3.1 Direct Questioning (Zero-Shot)

This baseline approach involved providing the model directly with the question text and multiple-choice options, instructing it to select the correct answer. The prompt template was:

```
Question: [Question text here]
Choices:
(A) [Choice A text]
(B) [Choice B text]
(C) [Choice C text]
```

---
[3]https://www.together.ai/

```
(D) [Choice D text, if applicable]
Select the correct answer. Ensure that
    the final answer is only the letter
    of the correct option, without any
    additional text or symbols.
```

The final instruction was added to minimize parsing errors during evaluation.

3.3.2   Chain-of-Thought (CoT) Prompting

To encourage step-by-step reasoning and generate explanations, we employed a CoT approach. The core CoT instruction added was:

```
Let's think step by step.
Provide your reasoning in
several steps, and then output
the final answer on a new line
starting with "Final Answer:".
Ensure that the final answer is
only the letter of the correct
option, without any additional
text or symbols.
```

This strategy aimed to improve answer accuracy by forcing a reasoning process and simultaneously generating explanatory text for qualitative analysis.

3.3.3   Few-Shot Prompting

To assess the impact of in-context examples, we also evaluated models using few-shot prompting. This involved prepending the Direct Questioning or CoT prompt (as applicable) with 1, 3, or 5 randomly selected question-answer pairs from the dataset (excluding the question being tested) formatted similarly to the main task.

3.4   Explanation Generation and Validation Setup

To facilitate efficient **linguistic expert validation** of both the answers and reasoning of the best-performing model identified in preliminary tests (Gemini 2.5 Pro Experimental), we used a simple, direct prompt asking it to provide its chosen answer, the correct answer (based on its knowledge), its reasoning, and a category classification. This generated a structured output specifically formatted for review, exemplified below:

```
Question 10: Din campul lexical al cuvantului "
    electorat" fac parte:
Options:
  a) electron, neutron, proton
  b) electric, energetic, curent
  c) senator, primar, magistrat
  d) alegeri, vot, candidat
 Gemini's Answer: d
 Correct Answer: d
Explanation: Campul lexical include cuvinte
    inrudite ca sens. "Electorat" se refera la
    totalitatea alegatorilor, fiind direct legat de
    "alegeri", "vot", "candidat".
```

```
Category: Lexical
```

Listing 1: Example structured output from Gemini 2.5 Pro for validation.

This structured format, including the category classification (Lexical, Morphological, Syntactic, Phonetic) assigned by Gemini 2.5 Pro Experimental, was used by the linguist expert to validate the model's answer, reasoning quality, and classification accuracy (which exceeded 99% for the categorization task itself).

3.5   Evaluation Metrics

Performance was primarily measured by **answer accuracy**. Additionally, a **qualitative linguistic analysis** was conducted on the explanations generated via CoT (facilitated by the structured output from Gemini 2.5 Pro for validation) to assess their correctness, precision, and pedagogical suitability.

4   Quantitative Results

This section presents quantitative findings from our benchmarking evaluations. We analyze the impact of different prompting strategies (Chain-of-Thought vs. Direct, Few-Shot vs. Zero-Shot) and compare the performance of Romanian-specific fine-tuned models against their multilingual base versions.

4.1   Impact of Chain-of-Thought Prompting

We first compared model accuracy using direct zero-shot questioning versus Chain-of-Thought (CoT) prompting, which instructs the model to provide reasoning steps. Table 1 summarizes these results.

Model	Accuracy without CoT	Accuracy with CoT
DeepSeek V3.0324	**53.26%**	64.62%
Mistral small 24B	39.10%	38.31%
Llama 3.3 70B	49.70%	55.60%
Qwen 2.5 Coder 32B	48.22%	51.61%
Gemini 2.5 Pro	—	**82.8%**

Table 1: Comparison of model accuracy (%) with and without Chain-of-Thought (CoT) prompting. Top score in each column is bolded. Gemini 2.5 Pro was evaluated only with CoT.

CoT prompting generally led to notable accuracy improvements, particularly for DeepSeek

`V3` (+11.36 pp) and `Llama 3.3 70B` (+5.9 pp), confirming its benefit for reasoning on this task. An exception was `Mistral Small 24B`, which saw a slight decrease, possibly due to difficulties following complex instructions. `Gemini 2.5 Pro`, tested only with CoT, achieved the highest score (82.8%), significantly outperforming others. However, accuracies for most models (38-65% range with CoT) remain substantially below proficient human levels (>80%) on comparable exams[4], highlighting a remaining gap in nuanced Romanian grammar understanding for most LLMs.

4.2 Impact of Few-Shot Prompting

We also investigated whether providing in-context examples (1, 3, or 5 shots) could improve performance compared to zero-shot prompting. The results, detailed fully in Appendix B , show **mixed and generally modest benefits**. While accuracy often increased slightly, gains were typically small (2-5 pp) and inconsistent across models and shot counts. Table 2 illustrates this variability for the Llama 3.1 8B family.

Llama 3.1 8B Variant	FS=0	FS=1	FS=3	FS=5
Base Instruct	35.62	38.84	**40.75**	39.70
Ro-Instruct (25-04)	36.06	37.71	**40.49**	39.36
Ro-Instruct-DPO (25-04)	38.49	37.53	**39.53**	39.27

Table 2: Illustrative accuracy (%) for Llama 3.1 8B variants with Few-Shot (FS) prompting. Best few-shot score per model in bold. Full results in Appendix B .

Few-shot prompting did not consistently favor base models over fine-tuned ones, nor did a clear optimal number of shots emerge. Crucially, even with examples, peak accuracies for the models tested in this setting (detailed in Appendix B) remained significantly below top performers like Gemini 2.5 Pro and human levels. This suggests limited in-context examples are insufficient to overcome the core challenges of this task for these models.

4.3 Performance of Romanian Fine-tuned Models vs. Base Models

A key goal was to assess the impact of Romanian-specific fine-tuning. We compared the performance of various RoLLM variants (developed by Masala

et al. (2024)) against their respective base multilingual models using data from our few-shot experiments (details in Appendix B).

The results indicate that Romanian-specific fine-tuning **does not consistently yield significant improvements** on this specific grammatical MCQ task, especially compared to strong base models.

- For the **Llama** family (Llama-2, Llama-3, Llama-3.1), the RoLLM variants generally performed similarly to or sometimes slightly worse than their corresponding base Instruct models in zero-shot settings. Few-shot prompting occasionally led to RoLLM variants slightly outperforming their base, but the difference was often marginal. DPO fine-tuning on RoLLM variants did not produce consistent gains over standard RoLLM instruction tuning for this task.

- For **Mistral-7B**, the RoMistral variants showed slightly better average performance than the base Mistral-7B Instruct v0.3, but the difference was small (within 1 pp on average across FS settings).

- For **Gemma** models, the results were particularly striking. The base `google/gemma-2-9b-it` significantly outperformed all its RoGemma2-9B fine-tuned variants (including DPO versions) across all few-shot settings, achieving an average accuracy of 47.13% compared to averages around 39-42% for the RoGemma2 variants. A similar, though less pronounced, trend was observed for Gemma-1.1-7b.

These findings suggest that for complex grammatical reasoning tasks like those in our dataset, the benefits of general multilingual pre-training in strong base models (like Llama 3.1 Instruct or Gemma-2-9B-it) might outweigh the current Romanian-specific fine-tuning approaches applied in the RoLLM project, at least for zero-shot and few-shot settings. The fine-tuning might be more beneficial for other NLP tasks or require different strategies (e.g., incorporating grammatical knowledge more explicitly) to show significant gains on this benchmark. The relatively low overall scores for both base and fine-tuned models (excluding Gemini) further emphasize the difficulty of this task.

---

[4]Example results: https://cdn.edupedu.ro/wp-content/uploads/2023/01/lista_anonimizata_rezultate_finale_LC.pdf

5 Qualitative Analysis: Explanations, Categorization, and Dataset Challenges

Beyond quantitative accuracy, understanding LLM capabilities for educational use requires qualitative assessment. This section presents findings from an in-depth analysis focusing primarily on explanations and category classifications generated by **Gemini 2.5 Pro Experimental** (using CoT prompting). The analysis, conducted by a linguistic expert, evaluated 200 explanations in detail and reviewed LLM category assignments across the full dataset (1151 items). This revealed insights into LLM reasoning, limitations in handling linguistic nuances, and characteristics of the benchmark dataset itself.

The involvement of a linguistic expert was essential to the present study due to the complex and nuanced nature of the materials under analysis. The apparent uniformity of item structure did not preclude the presence of subtle or highly context-dependent linguistic phenomena. Certain errors identified in the dataset were marked by fine semantic or syntactic distinctions that would likely elude detection without expert philological judgment, such as putting the error into the correct language category (see questions 51, 77, 211, 228, etc. where the lexical errors are considered morphological and the opposite) or presenting the explanation in an inadequate style or erroneously.

5.1 Quality of LLM-Generated Explanations

While CoT prompting improved answer accuracy (Table 1), the quality of the accompanying explanations varied significantly. A detailed review of 200 explanations revealed that **48% (96 explanations) were problematic** (incomplete, incorrect, misleading, or containing flaws). Key issues included:

- **Incorrect Justifications (67/96):** Explanations frequently provided factually incorrect grammatical reasoning, even when the selected MCQ answer was correct.

- **Imprecision and Irrelevance (29/96):** Explanations often suffered from redundancy, imprecise terminology, irrelevant points, or incoherent concluding statements that undermined the reasoning (e.g., the concluding phrase for Question 26 regarding pronoun agreement).

This high error rate in explanations limits the reliability of current LLMs as standalone grammar tutors without expert validation.

5.2 LLM Performance on Linguistic Categorization

We also evaluated Gemini 2.5 Pro Experimental's ability to classify each of the 1151 questions into four linguistic domains: Lexical, Morphological, Syntactic, or Phonetic. While the overall accuracy on this task was very high (over 99% agreement with expert classification), the **five instances of misclassification** are revealing:

- *Morphology vs. Lexicon:* Question 51 (abbreviation of *dumneavoastră*) was misclassified as "Lexical" instead of "Morphological". Question 77 (use of *optim* creating a pleonasm) was misclassified as "Morphological" instead of "Lexical-Semantic".

- *Syntax vs. Morphology:* Question 211 (morphological value of pronoun *ce*) and Question 228 (identifying an interjection) were misclassified as "Syntactic". Question 252 (focusing on the preposition *după*) was also misclassified as "Syntactic" instead of "Morphological".

These errors typically occurred in borderline cases involving multiple linguistic layers (form, function, meaning), suggesting LLMs may struggle to pinpoint the primary linguistic concept being tested beyond surface cues.

5.3 Dataset Characteristics and Normative Alignment

The qualitative analysis also highlighted characteristics and potential issues within the source materials:

- **Unclear/Flawed Prompts:** Seven questions out of the 200 deeply analyzed contained ambiguities or errors affecting comprehensibility and reliable evaluation.

- **Incorrect/Incomplete Official Answers:** Four analyzed items had designated correct answers conflicting with current standards.

- **DOOM 3 Misalignment:** [5]This was a notable issue. Analysis of the 19 DOOM-related questions in the dataset revealed:

  - *LLM Errors:* In six explanations for DOOM-related questions, Gemini 2.5 Pro provided false or incorrect information regarding DOOM 3 (Institutul de Lingvistică "Iorgu Iordan – Alexandru Rosetti", 2021) norms. Examples include incorrectly stating DOOM 3 "accepts" both forms for *cofeina/cafeina* (Q626) or *biscuit/biscuite* (Q1128), misstating the preferred form for *barocă* (Q733), inaccurately describing the abbreviation rule for *dumneavoastră* (Q51), or misrepresenting the recommendation for *cuvânt-înainte* (Q896).

  - *Outdated Source Questions:* Some source questions themselves appeared outdated relative to DOOM 3 (2021 edition). For example, Question 144 (plural of *proroc*) lacked the standard DOOM 3 form among its options, suggesting the question predates the latest normative update.

These findings underscore the need for careful curation and normative alignment of benchmark datasets, especially when evaluating performance on rule-based linguistic tasks.

5.4 Implications

The qualitative results emphasize the need for evaluation metrics beyond accuracy, particularly when assessing LLMs for educational roles requiring explanations. The inconsistencies found highlight the importance of dataset quality control and alignment with current linguistic standards (like DOOM 3) for reliable benchmarking in high-stakes domains like language assessment.

---

[5]DOOM3, published in 2021, serves as the current standard reference for Romanian spelling (ortografic), pronunciation (ortoepic), and morphology (morfologic). While the core orthographic, ortoepic, and morphological norms largely remain unchanged, DOOM3 introduces over 3,000 discrete normative updates. These include the addition or removal of definite/article forms, select flexional variants, and adjustments to accentuation in compound words. It intentionally curtails the proliferation of orthographic variants, helping users to clearly identify the recommended form. It represents the official norms, reason why explanations such as "DOOM 3 prefers" or "recommends" from questions 733 or 896 are not adequate.

6 Application

In order to explore the applicability of the created dataset and the generated explanations, we developed a publicly accessible web application designed as a learning tool for Romanian language proficiency.
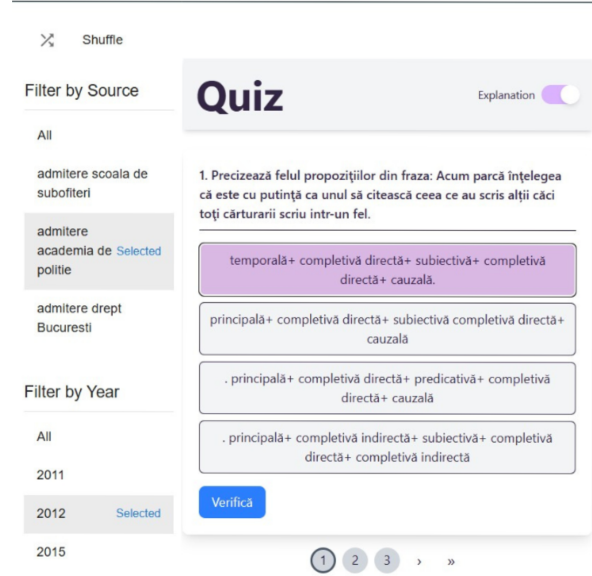

Figure 2: User interface of the web application, allowing practice with dataset questions.

The application utilizes the benchmark dataset as its question bank. Users can navigate questions filtered by source (allowing targeted practice based on exam type/difficulty proxy) and submit their answers within a timed or practice mode. Upon submission, the application reveals the correct answer and, importantly, displays the corresponding **LLM-generated grammatical explanation** (specifically, those generated by Gemini 2.5 Pro and validated for quality by our linguistic expert, as discussed in Section 5).
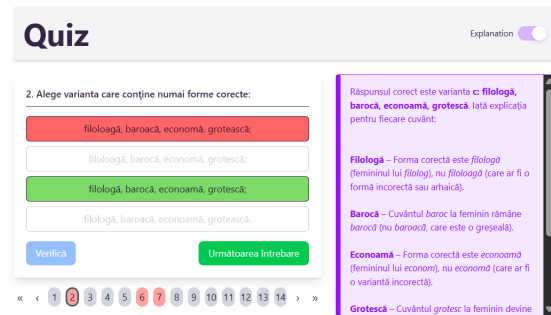

Figure 3: User interface of the web application, displaying explanations for the correct answers.

This feature allows learners to not only test their knowledge but also receive immediate feedback and reasoning, leveraging the LLM's explanatory capabilities in a practical educational context. Although explanation quality still requires ongoing validation, providing expert-vetted explanations offers significantly more pedagogical value than simple answer keys. The application is implemented in React and hosted statically, ensuring easy access and deployment.

7   Conclusion and Future Work

This study evaluated state-of-the-art LLMs on a novel dataset of Romanian language MCQs sourced from high-stakes exams, focusing on answer accuracy and explanation quality. Our benchmarking, utilizing direct and Chain-of-Thought (CoT) prompting, revealed significant performance variations and a considerable gap between most LLMs and proficient human performance, although `Gemini 2.5 Pro Experimental` achieved high accuracy (83%) using CoT.

Qualitative analysis of Gemini 2.5 Pro's explanations, while revealing flaws in nearly half, also found that the majority were linguistically valuable and useful as a starting point for pedagogical feedback, confirmed by expert assessment. This potential was demonstrated through their integration into our publicly available educational web application, providing learners with practice opportunities augmented by AI-generated, expert-validated grammatical reasoning. However, the analysis also highlighted challenges related to dataset quality and normative alignment (e.g., with DOOM 3).

Our findings underscore the need for both quantitative and qualitative metrics when evaluating LLMs for educational roles, particularly concerning explanation generation in low-resource languages. While current LLMs show promise as assistive tools, significant work remains to ensure consistent accuracy and pedagogical soundness.

Future work should proceed along several key directions. Firstly, **expanding the benchmark scope** is essential: incorporating more questions, annotations (like difficulty levels or specific grammar subsections), evaluating a broader range of SOTA models (including newer Romanian-specific ones), and performing detailed cross-lingual analyses. Secondly, significant potential lies in **enhancing model performance and explanation quality** through advanced techniques. This includes ex-

ploring **Retrieval-Augmented Generation (RAG)** leveraging Romanian language manuals or grammar resources as external knowledge, targeted **fine-tuning** using a subset of the dataset questions, and experimenting with more sophisticated prompting strategies such as **few-shot learning** to potentially create more effective AI tutors. Finally, addressing the identified dataset limitations through continued **curation and normative alignment** remains vital for building truly robust evaluation benchmarks for Romanian educational NLP.

8   Limitations

Our dataset of 1 151 Romanian multiple-choice questions (MCQs) is smaller than the large benchmarks available for high-resource languages and represents mainly high-stakes exam material. Items were extracted with Tesseract OCR and then spot-checked, yet some annotation noise may remain. Question categorisation relied on an LLM-assisted workflow that, despite manual review, can still inherit model errors.

While we have categorized the questions when answering them and obtained the following split syntactic 382 questions, morphological 364, lexical 319 and phonetic 86, we have not annotated the initial data with this information.

Methodologically, the study is restricted to MCQs; open-ended generation and deeper discourse skills are outside our present scope.

9   Ethical Considerations

This research uses publicly available Romanian multiple-choice exam questions; all sources are free of licensing or privacy constraints. We release the dataset on Hugging Face to foster transparency, reproducibility, and further work on low-resource educational NLP. The free web application stores no PII. Dataset content mirrors the focus of the original exams. LLM-generated explanations were validated by a linguist to prevent misinformation. We acknowledge possible training-data biases and welcome community scrutiny; users should respect original copyrights.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised

cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Yiming Cui, Ziqing Yang, and Xin Yao. 2024. Efficient and effective text encoding for chinese llama and alpaca.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, et al. 2025. Deepseek-v3 technical report.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, et al. 2024. The llama 3 herd of models.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the 37th International Conference on Machine Learning*.

Institutul de Lingvistică "Iorgu Iordan – Alexandru Rosetti". 2021. *Dicționarul Ortografic, Ortoepic și Morfologic al Limbii Române (DOOM)*, 3 edition. Editura Univers Enciclopedic, București. Coordonatoare: Ioana Vintilă-Rădulescu.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language.

S. Magalí López Cortez, Mark Josef Norris, and Steve Duman. 2024. GMEG-EXP: A dataset of human- and LLM-generated explanations of grammatical and fluency edits. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7785–7800, Torino, Italia. ELRA and ICCL.

Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, Thomas Wang, Nouamane Tazi, Teven Scao, Thomas Wolf, Osma Suominen, Samuli Sairanen, Mikko Merioksa, Jyrki Heinonen, Aija Vahtola, Samuel Antao, and Sampo Pyysalo. 2023. FinGPT: Large generative models for a small language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2710–2726, Singapore. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Mihai Masala, Denis C. Ilie-Ablachim, Dragos Corlatescu, Miruna Zavelca, Marius Leordeanu, Horia Velicu, Marius Popescu, Mihai Dascalu, and Traian Rebedea. 2024. Openllm-ro – technical report on open-source romanian llms.

Mihai Masala, Stefan Ruseti, and Mihai Dascalu. 2020. RoBERT – a Romanian BERT model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6626–6637, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Marina Mayor-Rocher, Nina Melero, Elena Merino-Gómez, Miguel González, Raquel Ferrando, Javier Conde, and Pedro Reviriego. 2025. Teleia: A spanish language dataset for evaluating artificial intelligence models. *Data in Brief*, 59:111437.

Marina Mayor-Rocher, Nina Melero, Elena Merino-Gómez, María Grandury, Javier Conde, and Pedro Reviriego. 2024. Evaluating large language models with tests of spanish as a foreign language: Pass or fail?

Ryo Nagata. 2019. Toward a task of feedback comment generation for writing learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3206–3215, Hong Kong, China. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Vasile Păiş and Dan Tufis. 2023. *Language Report Romanian*, pages 199–202. Springer International Publishing.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, et al. 2024. Gemini: A family of highly capable multimodal models.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.