Benchmarking Language Models on Romanian Eighth-Grade National Exams

Radu Ion
Research Institute for AI
"Mihai Drăgănescu",
Romania
radu@racai.ro

Verginica Barbu Mititelu
Research Institute for AI
"Mihai Drăgănescu",
Romania
vergi@racai.ro

Elena Irimia
Research Institute for AI
"Mihai Drăgănescu",
Romania
elena@racai.ro

Maria Mitrofan
Research Institute for AI
"Mihai Drăgănescu",
Romania
maria@racai.ro

Vasile Păiș
Research Institute for AI
"Mihai Drăgănescu",
Romania
vasile@racai.ro

Abstract

This paper evaluates a set of Large Language Models (LLMs) on the task of automatically solving national standardized Romanian language tests for eighth-grade students. The models are provided only with the question text, in Romanian, and are required to generate answers across various formats, such as multiple choice, fill-in-the-blank, true/false, short answer, and composition. Their responses are automatically evaluated against the official scoring rubrics, which detail how points are awarded for each question and specify the correct answer where applicable. This study's goal is to establish a benchmark for the performance of Romanian-aware and state-of-the-art LLMs on standardized eighth-grade Romanian exams, and to determine whether state-of-the-art LLMs can serve as accurate eighth-grade exam evaluators.

Keywords: Romanian LLM evaluation, automatic test scoring with LLMs, Romanian language standardized tests, LLM-as-a-judge

1 Introduction

Every year, eighth-grade schoolchildren[1] in Romania must take nationwide, standard exams[2] of Romanian language and mathematics, whose grades will determine their ranking in the high school admissions competition. Each exam is designed to evaluate their knowledge of Romanian and mathematics, acquired during the fifth to eighth grades. During simulation and exam days, and additionally, during supplementary sessions, teachers from all over Romania are tasked with the grading of these tests, according to the scoring rubric that contains detailed instructions on how to score each item in the test, awarding points for each sub-goal that has been achieved by the test taker, for each exam question.

Grading these tests by two independent graders is a tedious task, but one of tremendous importance for the future of eighth-grade graduates. Cononenco (2025) advances an amount of 70–80 tests per teacher to be graded, a task that collectively required "«an impressive deployment of forces» — Sorin Ion, Secretary of State in the Ministry of Education", according to the cited news article. In 2024, 152,903 schoolchildren took the Romanian language and mathematics exams in the National Evaluation, and 74.5% obtained more than 5 (on a scale from 1 to 10), as the average of the two exams.

In this context, leveraging LLMs to assist with test grading seems to be the obvious solution that can assist (not replace) teachers with test scoring. To assist with test grading, the LLMs should first be able to solve the tests themselves, demonstrating that "they know what they are doing" by obtaining high grades, and second, exhibit grading reasoning skills that human evaluators (i.e., Romanian

[1] Typically aged 14.

[2] Called the "National Evaluation of the Eighth Grade".

eighth-grade teachers) typically apply to these tests.

To assess the capabilities of Romanian and state-of-the-art LLMs in answering Romanian-language questions from the National Evaluation for eighth-grade students, we first parsed the publicly available PDF test documents and their corresponding answer keys into structured JSON format. Each test item is annotated with its point value, question type (e.g., multiple-choice, fill-in-the-blank, true/false, short answer, or composition), and the question text. From the associated answer key, we extracted either the exact correct answer or, for open-ended questions (specifically short-answer items), a set of example responses. Additionally, each question is accompanied by a detailed scoring rubric that outlines how points are allocated based on specific criteria or sub-goals the response meets. These rubrics, along with the reference answers, are used by a judge LLM (Gu et al., 2025) to automatically evaluate the quality of generated responses and assign points accordingly.

In what follows, we will review the relevant literature in Section 2, we will present the compilation of the JSON dataset from the Romanian language tests of the National Evaluation for the eighth grade in Section 3, and we will present the evaluation results of current Romanian-aware and state-of-the-art multilingual LLMs in the zero-shot scenario in Section 4. We will conclude the paper with Section 5.

2 Related work

Masala et al. (2024) describe most of the Romanian-aware LLMs that we evaluate here. Their paper is the most recent and comprehensive evaluation of Romanian LLMs, built upon established English-language models. Starting with a pre-trained English LLM, they first fine-tune it on Romanian corpora to produce a foundational Romanian-aware model. This base model is then further fine-tuned using Romanian-translated instruction datasets to yield an instruction-following Romanian-aware LLM. Using LLaMA-2-7B as the starting point, they develop RoLLaMA2-7B-Base (the foundational model) and RoLLaMA2-7B-Instruct (the instruction-tuned variant). Additional Romanian instruction-aware models are fine-tuned directly, without intermediate Romanian pre-training, from models such as Mistral-7B, LLaMA-3-8B, LLaMA-3.1-8B, and Gemma-7B. These models are evaluated on reference datasets that were machine-translated into Romanian without human post-editing. On average, across multiple benchmarks, the Romanian instruction-tuned models outperform their original English counterparts.

Păiș et al. (2025) describe two Romanian-aware language models that were fine-tuned starting from Llama3.2-1B base models. The first one[3] was further trained on a portion of the Representative Corpus of the Contemporary Romanian language (CoRoLa). Starting from it, the second one[4] was further fine-tuned on several Romanian instruction sets.

RoQLlama-7B (Dima et al., 2024) is another Romanian-aware LLM, built upon the Llama2-7B model. Its development follows the same methodology as RoLLaMA2-7B-Base (Masala et al., 2024) but incorporates QLoRA-based fine-tuning with 4-bit quantization over a 7.3-billion-token Romanian corpus. In addition to general NLP benchmarks, RoQLlama-7B is also evaluated on a Romanian medical question-answering dataset introduced in the same study. Under the zero-shot evaluation setting adopted in our work, the model achieves an accuracy of 21.57%. This model will also be evaluated on our dataset.

Dumitran et al. (2025) introduce GRILE, a dataset comprising 1,151 multiple-choice questions sourced from major Romanian high-stakes examinations, including the National Evaluation, Baccalaureate, and university entrance exams. GRILE serves as a benchmark for evaluating multilingual and Romanian-specific LLMs in two key capacities: (a) their ability to select correct answers, and (b) their capacity to generate coherent and informative explanations that justify their choices. While Gemini 2.5 Pro achieves 83% accuracy on the GRILE benchmark, most open-weight multilingual models remain below 65%, with Romanian-specific models not exceeding 41%. Furthermore, expert analysis reveals that 48% of their generated explanations contain factual inaccuracies or pedagogical flaws.

BacPrep (Dumitran et al., 2026) is an online platform that provides high-quality preparation and

[3] https://huggingface.co/racai/corola-llama-3.2-1b-e3

[4] https://huggingface.co/racai/corola-instruct-llama-3.2-1b-e2

feedback for the Romanian Baccalaureate exam, particularly targeting high school students from economically disadvantaged backgrounds who may lack access to private tutoring. The platform leverages Gemini 2.0 Flash, one of Google's latest LLMs, to deliver automated feedback on practice exam responses. Currently under development, BacPrep is actively collecting real-world student answer data. In future iterations, once student responses and LLM-generated feedback have been reviewed by human evaluators, the platform aims to provide reliable, pedagogically sound guidance to learners.

In line with current research on the automatic assessment of LLM-generated output (Dinh et al., 2024; Center for AI Safety et al., 2026; Shetty et al., 2025), we use the LLM-as-a-judge paradigm to automatically determine whether responses are correct. Dinh et al. (2024) observe that automatic evaluation of generated text is difficult and, even though machine translation metrics such as BLEU or ROUGE can be used, LLM-as-a-judge is preferable, as it is well correlated with human judgments, especially in the textual modality.

Center for AI Safety et al. (2026) conduct a study similar to ours on academic assessment, evaluating generative AI capabilities on a set of 2,500 challenging academic questions. The "Humanity's Last Exam" dataset features questions contributed by nearly 1,000 subject-matter experts from over 500 institutions across 50 countries, predominantly professors, researchers, and individuals holding graduate degrees. The o3-mini is the judge LLM used to verify the correctness of answers.

3 National Evaluation Romanian language dataset

This dataset contains JSON-formatted versions of model, simulation, and official Romanian language tests, originally published as PDF files by the Romanian Ministry of Education for the National Evaluation at the eighth-grade level[5]. Each year, exam simulation sessions are held using tests that mirror the structure and content requirements – both literary and grammatical – of the official exams. Prior to these simulations, the Ministry also releases additional model tests that follow the same format and content expectations as the official ones.

Each test consists of a pair of files: the test variant ("variantă" in Romanian), which includes the exam questions, and the answer key ("barem" in Romanian). Multiple model and simulation pairs are available each year, while only two official test pairs are released – one for the main session and one as a backup. The Hungarian-speaking minority in Romania also takes this exam, and dedicated versions of the test are provided for them.

The Romanian language exam is divided into two units: Unit I and Unit II. Unit I typically comprises two parts (A and B) that assess students' reading comprehension and grammatical knowledge. Unit II requires students to write a short essay (150–300 words). The comprehension questions in Unit I refer to one or both of the two short literary texts included in the test (referred to as "Textul 1" and "Textul 2" in Romanian).

The test contains the following types of questions:

- **Short answer**: Students are required to produce one or more sentences that follow specific grammatical constraints or to analyze a phrase from a grammatical perspective. For example, "Rewrite the sentence « At the end of the year, before college admission, … » replacing each underlined word with a synonym."
- **Fill-in-the-blank**: Students must complete sentences by inserting the correct forms of words, based on definitions provided in parentheses following the blanks. For example, "Fill in the missing words in the sentence below according to the indications in parentheses: "____ (to say, imperative, plural) if you ____ (to like, conjunctive, present tense, 3rd person singular) …"
- **True/false**: Students evaluate the truth value of 4 to 6 statements using information from the accompanying literary texts, e.g., "From Text 1 above, determine if the following sentence is true or false: « Adina Nanu's desk mate looked like a porcelain doll. »"
- **Multiple choice**: Students select the correct answer by letter, based either on information from the literary texts or on their

[5] The National Evaluation tests can be found at http://subiecte202X.edu.ro, where "X" can be replaced with "3", "4" and "5".

grammatical knowledge, e.g. “Both words in the series contain a diphthong: a) „aproape”, „înviorat”; b) „fiindcă”, „voit”; c) „aceasta”, „coroană”; d) „mamei”, „deslușeam””

- **Composition**: Students write a short essay expressing their opinion in response to a prompt related to the included literary texts. For instance, “Do you think it is important to take an exam without fear? Justify your answer in 50–90 words, drawing on Text 2.”

Each question is assigned a number of points proportional to its difficulty. One full test is worth 90 points, with an additional 10 points awarded by default. The final grade is calculated by dividing the total number of points earned by 10, resulting in a grade on a scale from 1 to 10.

The answer key is always provided alongside the test. It includes exact answers when available or example responses when multiple valid answers are acceptable. Based on the question types defined above, the answer key contains the following information:

- **Short answer**: For questions requiring grammatical analysis, example responses are typically provided (e.g., grammatical breakdowns of word forms, phrases, or sentences that meet specific constraints).
- **Fill-in-the-blank**: The list of correct word forms.
- **True/false**: The correct truth values for each statement.
- **Multiple choice**: The letter corresponding to the correct answer.

For all question types except multiple-choice, the answer key includes detailed instructions specifying how points should be awarded for partial fulfillment of the question requirements. These guidelines are especially comprehensive for composition questions, outlining point allocations for aspects such as grammatical accuracy, spelling, clarity of expression, and other relevant criteria. For example, for the composition question “Do you think it is important to take an exam without fear? Justify your answer in 50–90 words, drawing on Text 2.”, the associated guidelines for the human evaluator are:

- stating the answer to the given question, **1 point**
- justification of the stated answer, **3 points**:
  - nuanced and appropriate, making effective use of the indicated text – 3 points
  - appropriate but schematic, making use of the indicated text – 2 points
  - tendency to generalize, attempt at justification, no use of the indicated text – 1 point
  - lack of justification – 0 points
- observance of spelling and punctuation rules (at most one error – 1 point; 2 or more errors – 0 points), **1 point**
- staying within the indicated word limit, **1 point**

To convert a PDF test file into its JSON equivalent, we followed a four-step process:

1. The PDF file was first transformed into a UTF-8-encoded text file, with careful attention to preserving formatting features such as bold, italic, and underline.
2. The resulting text file was segmented into units, parts, and individual questions using regular expressions.
3. Each question was parsed into the JSON format using the `gpt-oss-20b` LLM[6] (OpenAI, 2025), which was prompted to extract the question ID, type, point value, and text. Prompts were issued in English, as Romanian prompts yielded less reliable results, particularly regarding the JSON output, which was not supplied consistently with Romanian prompts.
4. Each parsed question was manually reviewed to ensure that the LLM did not introduce any unintended modifications. During prompting, we explicitly instructed the model to copy question and answer key details verbatim, without rephrasing or altering the original content.

Table 1 presents statistics of the newly developed `en_viii_rolang` dataset from the National Evaluation at the eighth-grade level.

[6] https://huggingface.co/openai/gpt-oss-20b

Year	SA	FI	TF	MC	CO	ALL
2021	34	4	9	31	28	**106**
2022	18	2	4	28	20	**72**
2023	5	1	1	7	4	**18**
2024	5	1	1	7	4	**18**
2025	48	11	10	65	51	**185**
ALL	**110**	**19**	**25**	**138**	**107**	**399**

Table 1: Statistics of the dataset `en_viii_rolang` per year and question type

For each year[7], we report the number of questions by type: SA (short answer), FI (fill-in-the-blank), TF (true/false), MC (multiple choice), and CO (composition). Composition questions account for 26% of the dataset and will be evaluated subjectively, taking into account potential biases introduced by the judge LLM. For all other question types, we can compute exact accuracy metrics, both in terms of accuracy on correct answers (percentage of correct answers) and the resulting grade, on a scale from 1 to 10.

4 Evaluation

4.1 Estimating judge LLM scoring bias

We chose[8] `gemma3-27b-it`[9] (Gemma Team and Google DeepMind, 2025) as the LLM to serve as the judge responsible for evaluating the responses generated by the "student" LLMs. To assess its potential bias – either in overestimating or underestimating student responses – we *manually reviewed its judgments* with the help of three Romanian language teachers in eighth grade.

Each of the Romanian language teachers had to review the judge LLM's scoring justifications on a set of three test files:

- ENVIII_2023_romana_var_model.json
- ENVIII_2024_romana_var_model.json
- ENVIII_2025_romana_var_model.json

The answers to the questions in these three files were supplied by the `gpt-oss-20b` "student" LLM, and *each teacher evaluated the judge LLM scoring justifications for these answers*, so that we could compare their evaluations.

When the judge LLM provided an incorrect evaluation for a given question, the teacher annotated the explanation, identified the issues, and adjusted the number of points awarded – either increasing or decreasing the score. As a result, for the three manually reviewed tests, we can compare the grades assigned by the judge LLM with those assigned by human evaluators.

Table 2 provides a breakdown of the *completion grades* assigned by the judge LLM (J) and human teachers (T1–T3) for each question type, across the three manually reviewed tests. The completion grade for a question type is a percentage of the points obtained by the "student" LLM out of all available points for that question type. Column OG presents the *official test grade* (on a scale from 1 to 10), computed across the three-file test set. Knowing that perfectly solving all three tests would yield 90 × 3 = 270 points plus 10 × 3 = 30 points offered by default, for a total of 300 points, the official test grade OG is computed as follows (P is the number of points obtained by the "student" LLM over all three tests):

$$OG = \frac{P + 30}{300} \times 10$$

	SA	FI	TF	MC	CO	OG
J	66%	71%	89%	62%	98%	8.3
T1	62%	79%	89%	62%	83%	7.63
T2	57%	79%	89%	62%	94%	7.97
T3	66%	79%	83%	62%	89%	7.97

Table 2: Completion and official grades offered by the judge LLM and three teachers, on the three-file test set

If we look at the official test grade (OG), we see that all human evaluators are more conservative than the LLM, providing evidence of the LLM judge's overestimation of grades.

Teacher 1 (T1) stands out in relation to the composition question type (CO), having assigned a significantly lower completion grade (83%) than the LLM judge (98%). This discrepancy can be attributed to T1 being by far the most experienced of the three teachers, with many years of experience evaluating national eighth-grade assessment tests. Her extensive experience enables her to accurately assess the quality of the LLM judge's evaluations. Below are several of her

[7] Regarding 2023 and 2024, there were no model or simulation tests available for download, just the official tests.

[8] This LLM has been favored by all eighth-grade Romanian language teachers who participated in our study, following trial-and-error tests conducted independently by each teacher.

[9] https://huggingface.co/google/gemma-3-27b-it

insightful comments regarding the LLM's assessment of compositions:

- The dialogue is not a characterization method; thus, a loss of 1 point.
- Two words are inappropriately used: "poet" and "Arghezi"; "poetic voice" would be the right word; thus, a loss of 1 point. The speech figure "personificarea" is not exemplified; thus, a loss of 1 point. The interpretation of the cited metaphor is incorrect, a loss of 1 point.
- The theme and the message are not correctly identified, thus a loss of 2 points.
- The answer does not provide a motivation at all.
- The two texts should not be associated with each other, so no new text is mentioned; thus, there is a loss of 1 point. No value is named, thus a loss of 1 point. The reference to the suggested text is missing, resulting in a 1-point deduction.

The composition question is the most challenging, as it demands creativity from students, as well as clarity and logical cohesion in writing. It is therefore particularly noteworthy that an LLM such as `gpt-oss-20b` received a completion grade of 83% from an experienced Romanian language teacher. This sets a high benchmark that will be difficult to surpass.

Table 3 reports the percentage of questions correctly evaluated by the judge LLM (i.e., the judge LLM's accuracy) and the judge LLM's overall judgment accuracy rate, as evaluated by each teacher.

	SA	FI	TF	MC	CO	ALL
T1	64%	67%	100%	100%	39%	74%
T2	79%	67%	100%	100%	85%	89%
T3	64%	67%	67%	100%	54%	76%

Table 3: Accuracy of per-question judgment decisions by the LLM, according to each teacher

The LLM judge achieved a perfect success rate on multiple-choice (MC) and true/false (TF) questions. The overall accuracy rate of the LLM judge is around 75%, except for T2, who reported a higher rate. However, as shown in Table 2, the LLM judge tends to overestimate the final grade by 0.33 to 0.67 points. Given that T1 is the most experienced evaluator of such assessments, we consider her final grade of 7.63 as the reference point. Accordingly, the LLM judge's 0.67-point inflation will be treated as bias and subtracted from the "student" LLM's overall grade to yield a better estimate of the true grade.

The 0.67-point inflation bias is specific to our selected pair of judge and student LLMs. Its determination was possible through the manual evaluation performed by our teacher team, as previously shown, to obtain a more accurate estimate of the official grades. We hypothesize[10] that its value is determined primarily by the judge LLM rather than the student LLM. This is why we chose a strong student model rather than a weak one for computing the bias, since bad answers are easier to justify (more obvious) than good or partially good ones.

4.2 Student LLM performance

We ran all selected Romanian models on the `en_viii_rolang` dataset in a zero-shot setting – that is, by prompting them only with the information available to an eighth-grade student – and used the judge LLM to score each model's outputs. For comparison with the state-of-the-art LLMs, we also ran `gemma3-27b-it` (G3) and `gpt-oss-20b` (OS) on the same dataset.

The Romanian "student" LLMs evaluated in this study are listed below (all identifiers are valid on HuggingFace):

- `OpenLLM-Ro/RoLlama2-7b-Instruct-DPO` (R2)
- `OpenLLM-Ro/RoLlama3-8b-Instruct-DPO` (R3)
- `OpenLLM-Ro/RoLlama3.1-8b-Instruct-DPO` (R31)
- `OpenLLM-Ro/RoMistral-7b-Instruct-DPO` (RM)
- `OpenLLM-Ro/RoGemma2-9b-Instruct-DPO` (RG)
- `racai/corola-instruct-llama-3.2-1b-e2` (crl)
- `andreidima/Llama-2-7b-Romanian-qlora` (L2)

[10] Experimentally proving this hypothesis requires a lot of manual evaluation of lengthy LLM output, and it is thus left for future work.

	SA	FI	TF	MC	CO	OG
R2	22%	26%	19%	45%	60%	4.27
R3	24%	27%	27%	36%	60%	4.26
R31	28%	30%	52%	55%	72%	5.42
RM	34%	39%	69%	50%	79%	5.95
RG	30%	38%	58%	61%	55%	4.81
crl	.4%	0%	0%	3%	1%	0.11
L2	11%	24%	30%	17%	13%	1.47
G3	**76%**	**78%**	**93%**	74%	90%	8.30
OS	**76%**	75%	90%	**76%**	**91%**	**8.37**

Table 4: "Student" LLM completion grades and official test grades

	SA	FI	TF	MC	CO	ALL
R2	8%	5%	0%	45%	24%	25%
R3	15%	11%	0%	37%	23%	24%
R31	14%	5%	8%	57%	24%	31%
RM	16%	11%	24%	50%	32%	32%
RG	12%	16%	12%	62%	8%	29%
crl	0%	0%	0%	3%	0%	1%
L2	6%	5%	0%	16%	3%	8%
G3	52%	47%	**68%**	74%	68%	65%
OS	**54%**	**53%**	56%	**76%**	**72%**	**66%**

Table 5: Percents of correctly answered questions

	SA	FI	TF	MC	CO	ALL
R2	33%	53%	32%	0%	65%	31%
R3	26%	47%	56%	0%	65%	30%
R31	37%	63%	84%	0%	71%	38%
RM	44%	68%	72%	0%	65%	37%
RG	42%	58%	72%	0%	81%	41%
crl	1%	0%	0%	0%	4%	1%
L2	16%	32%	56%	0%	36%	19%
G3	45%	53%	32%	0%	30%	25%
OS	38%	42%	44%	0%	28%	23%

Table 6: Percents of partially correct answers

	SA	FI	TF	MC	CO	ALL
R2	59%	42%	68%	55%	11%	44%
R3	59%	42%	44%	63%	12%	46%
R31	49%	32%	8%	43%	5%	31%
RM	40%	21%	4%	50%	3%	31%
RG	46%	26%	16%	38%	11%	30%
crl	99%	100%	100%	97%	96%	98%
L2	78%	63%	44%	84%	61%	73%
G3	**3%**	**0%**	**0%**	26%	2%	**10%**
OS	8%	5%	**0%**	**24%**	**0%**	11%

Table 7: Percents of incorrectly answered questions

Tables 4 through 7 present the results for each question type (SA, FI, TF, MC, and CO) and for the entire dataset, as evaluated by our judge LLM. An underlined value indicates the highest performance among Romanian-aware LLMs, while a bolded value represents the overall best performer in each question category. For each model and question type, the sum of the percentages for correct, partially correct, and incorrect answers equals 100%. The definitions of these types of answers are as follows:

- A *correct answer* scores all points out of the total points assigned to the question.
- A *partially correct answer* is one that earns some of the points assigned to a question, but not all. It follows that multiple-choice questions, which can only be evaluated as correct or incorrect, cannot receive partial credit (column MC in Table 6 is 0%).
- An *incorrect answer* receives zero points.

The top performer among the Romanian-aware LLMs is the `OpenLLM-Ro/RoMistral-7b-Instruct-DPO`, which achieved an official test grade of 5.95 on our dataset. After subtracting the 0.67-point bias introduced by the judge LLM's tendency to overestimate grades, the adjusted score becomes 5.28 – much lower than the adjusted scores of `gemma3-27b-it` (7.63) and `gpt-oss-20b` (7.7).

One obvious reason multilingual LLMs are significantly better than Romanian-aware LLMs is that they are roughly 3 times larger in terms of parameter count. On the other hand, and this is the main reason they evidently outperform the Romanian models, they were trained on much larger datasets (both during pre-training and fine-tuning) and used state-of-the-art fine-tuning techniques such as knowledge distillation and reinforcement learning. For instance, `gemma3-27b-it` was fine-tuned using a large instruction-tuned teacher model (Gemma Team and Google DeepMind, 2025: 4). This means it learned not just from conventional human-written prompt-response pairs, but also from richer probability distributions or generated behavior supplied by a more capable internal teacher. Pre-trained on 14 trillion tokens, the model was fine-tuned on data targeting conversational ability, instruction following, multilingual responses, and long-context behavior, among other learning objectives.

A note regarding `racai/corola-instruct-llama-3.2-1b-e2`: this model is the smallest among those evaluated, with only 1 billion parameters. Although it was fine-tuned on several instruction datasets, this training – combined with its initial pre-training on the CoRoLa corpus –

appears to have been insufficient for achieving better performance. To get a sense of how off the `racai/corola-instruct-llama-3.2-1b-e2`'s answers were, here is a representative sample comparison with `gpt-oss-20b` on a short-answer question asking the student to name two cities that they read about in the first supporting text ("Textul 1"):

	Answer	Judge LLM decision
crl	Assistant: Iată cele două nume menționate în textul 1. Numele orașelor sunt Berlin (Berlin) și Constantinopol (Constantinopol). (...) Analiza mecanismelor politice din România a furnizat dovezi că populația orașului Romanovski a crescut în ultimele decenii. Potrivit acestui raport, a crescut cu aproximativ 18 milioane în 2008. (...)	The student did not provide any of the place names that appear in "Textul 1" (such as Câmpina, Sinaia, București). None of the names mentioned (Berlin, Constantinopol, Romanovski, etc.) are present in the text, so no points are awarded.
OS	Localități menționate în Textul 1: „Câmpina", „Sinaia"	Student listed "Câmpina" and "Sinaia", both of which appear in "Textul 1". Each correct city earns 1 point, for a total of 2 points out of 2.

Table 8: Sample output of student LLM `racai/corola-instruct-llama-3.2-1b-e2`

We found that the judge LLM's evaluation accuracy is approximately 75% (see the "ALL" column in Table 3). Consequently, the percentages of correctly answered questions reported in Table 5 should be adjusted to reflect this finding. For instance, `OpenLLM-Ro/RoMistral-7b-Instruct-DPO`'s 32% correct response rate should be recalculated as 32% × 75% = 24%, with the remaining 8% *most likely reclassified as partially correct*. As a result, the percentage of incorrect answers could be a more reliable metric for ranking student LLMs, since it remains unaffected.

5 Conclusions

We evaluated seven Romanian-aware LLMs on the task of solving tests from the National Evaluation administered at the eighth-grade level in Romania. The results are particularly impressive for the multilingual models `gemma3-27b-it` and `gpt-oss-20b`, while the Romanian-aware models performed significantly lower. For this evaluation, we developed the `en_viii_rolang` dataset based on the official, publicly available National Evaluation tests for Romanian. The dataset contains 399 questions of various types, formatted in JSON. It is continuously being expanded, as tests older than three years are no longer accessible on official websites, though they remain available through other Romanian language learning resources online.

Correctly and partially answered questions together provide a valuable foundation for automatically generating an instruction dataset for fine-tuning student LLMs and improving their performance. Future research will focus on this direction, as well as on training Romanian-aware LLMs to generate reasoning paths that enable them to evaluate real-world tests in a manner similar to Romanian language teachers. To support this goal, we plan to involve additional Romanian teachers in the project, asking them to verbalize their evaluation thought processes – following the approach taken by the experienced Romanian language teacher featured in this study.

Acknowledgments

This work was carried out within the "Large Language Models for the European Union (LLMs4EU)", project no. 101198470, call DIGITAL-2024-AI-B-06-LANGUAGE, funded by the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

This work was also supported by a grant from the Ministry of Research, Innovation and Digitalization – UEFISCDI, project number PN-IV-P8-8.2-EUD-2025-0061, within PNCDI IV.

References

Center for AI Safety, Scale AI & HLE Contributors Consortium. 2026. A benchmark of expert-level academic questions to assess AI capabilities. *Nature* 649, 1139–1146. https://doi.org/10.1038/s41586-025-09962-4

Laura Cononenco. 2025. Profesorii au terminat corectarea lucrărilor la simularea Evaluării

Naționale mai devreme. News article, online at https://www.edupedu.ro/profesorii-au-terminat-corectarea-lucrarilor-la-simularea-evaluarii-nationale-mai-devreme-sorin-ion-a-fost-o-desfasurare-de-forte-impresionanta-vom-incepe-sa-trimitem-de-saptamana-aceasta-borderou/

George-Andrei Dima, Andrei-Marius Avram, Cristian-George Crăciun, Dumitru-Clementin Cercel. 2024. RoQLlama: A lightweight Romanian adapted language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4531–4541, Miami, Florida, USA. Association for Computational Linguistics.

Tu Anh Dinh, Carlos Mullov, Leonard Bärmann, Zhaolin Li, Danni Liu, Simon Reiß, Jueun Lee, Nathan Lerzer, Jianfeng Gao, Fabian Peller-Konrad, Tobias Röddiger, Alexander Waibel, Tamim Asfour, Michael Beigl, Rainer Stiefelhagen, Carsten Dachsbacher, Klemens Böhm, and Jan Niehues. 2024. SciEx: Benchmarking Large Language Models on Scientific Exams with Human Expert Grading and Automatic Grading. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11592–11610, Miami, Florida, USA. Association for Computational Linguistics.

Marius Dumitran, Angela Dumitran, Alexandra Mihaela Danila. 2025. GRILE: A Benchmark for Grammar Reasoning and Explanation in Romanian LLMs. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 316–324, Varna, Bulgaria.

Adrian-Marius Dumitran, Radu Dita, and Angela-Liliana Dumitran. 2026. BacPrep: Lessons from Deploying an LLM-Based Bacalaureat Assessment Platform. arXiv:2506.04989v2 [cs.SE]

Gemma Team, Google DeepMind. 2025. Gemma 3 Technical Report. arXiv:2503.19786v1 [cs.CL]

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A Survey on LLM-as-a-Judge. arXiv:2411.15594v5 [cs.CL]

Mihai Masala, Denis Ilie-Ablachim, Alexandru Dima, Dragos Georgian Corlatescu, Miruna-Andreea Zavelca, Ovio Olaru, Simina-Maria Terian, Andrei Terian, Marius Leordeanu, Horia Velicu, Marius Popescu, Mihai Dascalu, and Traian Rebedea. 2024. "Vorbești românește?" A Recipe to Train Powerful Romanian LLMs with English Instructions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11632–11647, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b Model Card. arXiv:2508.10925v1 [cs.CL]

Vasile Păiș, Dan Tufiș, Verginica Barbu Mititelu, Radu Ion, Maria Mitrofan, Elena Irimia, Eric Curea, and Andrei Păun. 2025. Towards Romanian Large Language Models Using the CoRoLa Corpus. In *Proceedings of the 20th International Conference on Linguistic Resources and Tools for Natural Language Processing – ConsILR 2025*, Bucharest, Romania.

Pranam Shetty, Abhisek Upadhayaya, Parth Mitesh Shah, Srikanth Jagabathula, Shilpi Nayak, and Anna Joo Fee. 2025. Advanced Financial Reasoning at Scale: A Comprehensive Evaluation of Large Language Models on CFA Level III. arXiv:2507.02954v2 [cs.CL]