PHEB:
An European Portuguese High School-Level LLM Benchmark

Diogo Tavares, Rafael Ferreira, Afonso Simplício, Gonçalo Vinagre
Ana Carolina Condez, Inês Calvo, Inês Vieira, David Semedo, João Magalhães
NOVA School of Science and Technology
{dc.tavares, rah.ferreira, am.simplicio, gv.martins, a.condez}@campus.fct.unl.pt
{jmag, df.semedo}@fct.unl.pt

Abstract

We present PHEB (**P**ortuguese **H**ighschool **E**xams **B**enchmark), a comprehensive benchmark designed to evaluate Large Language Models (LLMs) on real high school level national exams in European Portuguese. The goal is to promote the development of NLP tools and provide a reliable resource for evaluating LLM performance in European Portuguese across educational domains. Covering over 3,500 questions spanning 18 years (2006–2023) across six core subjects, the benchmark compiles high-quality questions from Portuguese National Exams, written and thoroughly curated by professors to ensure topic diversity, linguistic accuracy, and alignment with national curricula. PHEB spans a wide range of subjects, including Mathematics, Portuguese Language and Literature, History, Geography, Biology/Geology, and Philosophy. Questions incorporate both multiple-choice and long-form answers to assess factual knowledge, reasoning capabilities, and language understanding. We comprehensively benchmark state-of-the-art LLMs, shedding light on key challenges such as models' knowledge, language coverage, answer format biases and robustness to machine translation.

Keywords: European Portuguese, High-school exams, Benchmark, LLMs

1. Introduction

Despite the remarkable progress in Large Language Models (LLMs), the development and evaluation of these models, especially in educational capabilities, remain heavily skewed toward high-resource languages. As a result, European Portuguese (PT-PT) lags behind in both model performance and assessment, with the existence of high-quality benchmarks tailored to its linguistic and cultural context being severely limited. Although LLMs have shown impressive capabilities in English at language comprehension, understanding and reasoning, their performance in PT-PT remains weakly assessed and often unreliable.

Existing benchmarks for Portuguese are predominantly machine translated datasets—typically from English (Thellmann et al., 2024). However, translation-based resources suffer from two critical issues: 1) from a broad perspective, they overlook PT-PT context (linguistic, historical, social), and 2) from a linguistic perspective, they suffer from overly literal translations, loss of context, and misalignment with the educational or cultural realities of Portugal. Even benchmarks designed for Brazilian Portuguese (PT-BR) cannot be a substitute for PT-PT evaluation, as, despite sharing the same base language, the variants differ significantly in vocabulary, grammar, orthography and usage in a way that significantly impacts both model evaluation and real-world application.

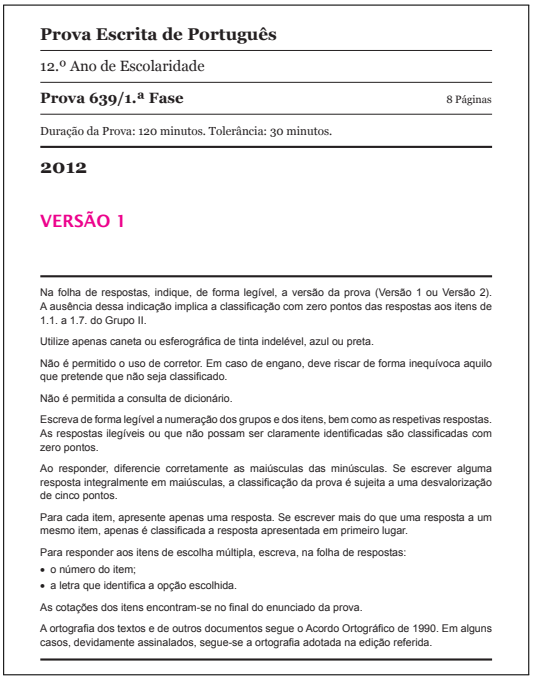
At the same time, there is a growing need to rigorously evaluate LLMs' reasoning capabilities across different domains of knowledge, especially in non-English languages. In educational settings, this means going beyond surface-level language understanding to test how well models can reason through problems, infer meaning, and generate coherent long-form answers in the target language. Few existing benchmarks attempt this in a structured or scalable way for PT-PT.

To address these gaps, we introduce PHEB[1], a benchmark designed to evaluate LLMs on European Portuguese high school level tasks. This resource provides a structured and linguistically authentic way to assess both language understanding and reasoning in PT-PT, across a wide variety of subjects. Our main contributions are as follows:

- **Official Portuguese High School Exams written by native speakers:** The dataset includes over 3,500 questions across 6 academic subjects, including Mathematics, Portuguese, Geography, Biology/Geology, Philosophy, and History. All questions are written natively in PT-PT by fluent speakers, ensuring fidelity to the language and educational context.
- **Benchmarks structured to elicit distinct LLM capabilities:** The benchmark includes both multiple choice questions (to probe factual comprehension and discriminative reasoning), and long-form open-ended questions

[1] https://github.com/AMALIA-LLM/pheb

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4673–4683
11–16 May 2026.



Prova Escrita de Português

12.º Ano de Escolaridade

Prova 639/1.ª Fase 8 Páginas

Duração da Prova: 120 minutos. Tolerância: 30 minutos.

2012

VERSÃO 1

Na folha de respostas, indique, de forma legível, a versão da prova (Versão 1 ou Versão 2). A ausência dessa indicação implica a classificação com zero pontos das respostas aos itens de 1.1. a 1.7. do Grupo II.

Utilize apenas caneta ou esferográfica de tinta indelével, azul ou preta.

Não é permitido o uso de corretor. Em caso de engano, deve riscar de forma inequívoca aquilo que pretende que não seja classificado.

Não é permitida a consulta de dicionário.

Escreva de forma legível a numeração dos grupos e dos itens, bem como as respetivas respostas. As respostas ilegíveis ou que não possam ser claramente identificadas são classificadas com zero pontos.

Ao responder, diferencie corretamente as maiúsculas das minúsculas. Se escrever alguma resposta integralmente em maiúsculas, a classificação da prova é sujeita a uma desvalorização de cinco pontos.

Para cada item, apresente apenas uma resposta. Se escrever mais do que uma resposta a um mesmo item, apenas é classificada a resposta apresentada em primeiro lugar.

Para responder aos itens de escolha múltipla, escreva, na folha de respostas:

- o número do item;
- a letra que identifica a opção escolhida.

As cotações dos itens encontram-se no final do enunciado da prova.

A ortografia dos textos e de outros documentos segue o Acordo Ortográfico de 1990. Em alguns casos, devidamente assinalados, segue-se a ortografia adotada na edição referida.

Figure 1: PHEB is a collection of 279 real exams across different disciplines of knowledge from 2006 to 2023, written in European Portuguese.

(to assess coherent language generation and structured reasoning). This design enables a comprehensive evaluation of model understanding, inference, and generation capabilities in PT-PT beyond surface-level fluency.

- **Evaluation of General-Purpose, Multilingual, and PT-Specific LLMs:** We evaluate a set of leading general-purpose, open LLMs, alongside multilingual and Portuguese-specific models, offering insight into how well models trained or adapted for PT-PT perform across domains. This comparison helps quantify the gap between multilingual and native-oriented performance.

2. Related Work

Evaluating LLMs across languages and domains has driven the creation of numerous benchmarks. However, most focus on English or high-resource languages, offering limited support for European Portuguese (PT-PT) and even fewer resources for educational and reasoning tasks across multiple school subjects.

Multilingual and Domain-General Benchmarks. MMLU (Hendrycks et al., 2021) is a standard benchmark covering over 50 subjects via multiple-choice questions (MCQ) in English. Translations of this dataset to PT-PT (Thellmann et al., 2024) allow multilingual testing but often degrade performance and may introduce errors, particularly for lower-resource languages. Global-MMLU (Singh et al., 2025) extends MMLU to 42 languages with professional translations and cultural sensitivity annotations. IberBench (González et al., 2026) supports Iberian languages for general NLP tasks (e.g., sentiment, classification, translation), yet excludes academic content or reasoning evaluation. MALAMUTE (Shaier et al., 2025) offers a textbook-based multilingual probing dataset relying on cloze-style questions, limiting its coverage of complex reasoning or generation tasks. AI4Math (Perez et al., 2025) evaluates mathematical reasoning in Spanish, offering depth within one domain but no coverage of broader subjects or grade levels.

For Brazilian Portuguese, the Sabiá project (Almeida et al., 2024; Abonizio et al., 2024) has produced evaluation datasets drawn from university exams and professional certifications across domains including law, economics, medicine, and public administration. While valuable for assessing PT-BR capabilities, these resources focus on MCQ, do not address the PT-PT variant, and do not target long-form answers.

PT-PT Specific Benchmarks. Portuguese-language benchmarks are fewer and mostly focus on linguistic competence rather than curriculum-based reasoning. CALAME-PT (Lopes et al., 2024) evaluates fluency and context sensitivity in PT-PT short texts. PORTULAN ExtraGLUE (Osório et al., 2024), adapted from GLUE (Wang et al., 2019b) and SuperGLUE (Wang et al., 2019a), includes entailment, paraphrase, and classification tasks across PT-BR and PT-PT variants but remains sentence-level. BATS-PT (Oliveira et al., 2024) measures lexical relations such as analogies. However, none of these benchmarks cover educational reasoning or support long-form response generation.

Education-Specific Benchmarks. Several English-language benchmarks target educational reasoning. RACE (Lai et al., 2017) and ARC (Clark et al., 2018) evaluate comprehension and science reasoning through standardized exam questions but lack language diversity or curriculum alignment. M3Exam (Zhang et al., 2023) introduces a multimodal, multilingual exam dataset with minimal PT-PT coverage, emphasizing image-based rather than textual reasoning. EXAMS (Hardalov et al., 2020) provides multilingual multiple-choice questions (including 920 in PT-PT across 4 subjects), though limited in scale and with no open-ended responses, while GPQA (Lewkowycz et al., 2022) and GSM8K (Cobbe et al., 2021) benchmark advanced and grade-school reasoning respectively, both in English.

[...] já que não podemos falar-lhes das vidas, por tantas serem, ao menos deixemos os nomes escritos, é essa a nossa obrigação, só para isso escrevemos, torná-los imortais, pois aí ficam, se de nós depende, Alcino, Brás, Cristóvão, Daniel, Egas, Firmino, Geraldo, Horácio, Isidro, Juvino, Luís, Marcolino, Nicanor, Onofre, Paulo, Quitério, Rufino, Sebastião, Tadeu, Ubaldo, Valério, Xavier, Zacarias, uma letra de cada um para ficarem todos representados [...].

José Saramago, *Memorial do Convento*, 27.ª ed., Lisboa, Caminho, 1998

Os trabalhadores da construção do Convento assumem o estatuto de heróis no romance *Memorial do Convento*.

Explique, fazendo apelo à sua experiência de leitura da obra, o modo como esses trabalhadores conquistam este estatuto, fundamentando a sua exposição em dois exemplos significativos.

Escreva um texto de oitenta a cento e trinta palavras.

(a) Example of a Portuguese language long-answer question.

4. Seja f uma função de domínio $\mathbb{R}$, definida por $f(x) = e^x - 3$

Em qual dos intervalos seguintes o teorema de Bolzano permite afirmar que a equação $f(x) = -x - \frac{3}{2}$ tem, pelo menos, uma solução?

(A) $]0, \frac{1}{5}[$ **(B)** $]\frac{1}{5}, \frac{1}{4}[$ **(C)** $]\frac{1}{4}, \frac{1}{3}[$ **(D)** $]\frac{1}{3}, 1[$

(b) Example of a mathematics multiple choice question.

9. Considere as condicionais seguintes.

1. Adília Lopes é poetisa se escreve rimas e quadras.
2. Escrever rimas e quadras é condição suficiente para Adília Lopes ser poetisa.

A proposição de que Adília Lopes escreve rimas e quadras

(A) é a consequente nas duas condicionais apresentadas.

(B) é a antecedente nas duas condicionais apresentadas.

(C) é a antecedente na condicional 1 e é a consequente na condicional 2.

(D) é a consequente na condicional 1 e é a antecedente na condicional 2.

(c) Example of a philosophy multiple choice question.

Os peixes-balão pertencem a um grupo com mais de 100 espécies que inclui a espécie marinha *Arothron hispidus*. Estes peixes, quando em perigo, expandem o corpo através da ingestão de água, que bombeiam para o estômago. Esta capacidade está dependente da compressão da cavidade bucal, da abertura e do fecho das fendas operculares e do funcionamento de uma válvula que está localizada imediatamente atrás das mandíbulas e que impede a saída de água.

Diversas espécies de peixes-balão apresentam em alguns órgãos, como a pele, o fígado e as glândulas sexuais, uma das toxinas mais letais que se conhecem – a tetrodotoxina (TTX) –, que, mesmo em pequenas doses, é capaz de matar um ser humano adulto. Esta substância é produzida por bactérias endossimbióticas que se alimentam dos nutrientes existentes no peixe. No ser humano, a TTX condiciona o movimento de iões sódio através da membrana celular, dificultando a propagação elétrica do impulso nervoso. Este processo começa por causar dormência da região oral, podendo conduzir a paralisia total e a insuficiência respiratória, por deficiente ventilação pulmonar, capaz de provocar a morte.

Baseado em A. Gomes, *et al.*, «Emprego terapêutico da tetrodotoxina em organismos animais», *Acta Veterinaria Brasilica*, Vol. 5, n.º 4, 2011 e em http://querosaber.com.pt (consultado em setembro de 2017).

1. O modo de atuação da TTX pode estar relacionado com

(A) o bloqueio de proteínas membranares intrínsecas.

(B) a degradação de alguns neurotransmissores.

(C) o bloqueio de proteínas membranares extrínsecas.

(D) a produção de neurotransmissores disfuncionais.

(d) Example of a biology multiple choice question.

Figure 2: A selection of examples of the PHEB benchmark.

Subject	MCQ	Open-Ended	Total
Portuguese Language	306	385	691
Mathematics	264	436	700
History	120	276	396
Geography	424	169	593
Philosophy	373	331	704
Biology/Geology	332	159	491
Total	**1,819**	**1,756**	**3,575**

Table 1: Distribution of questions across subjects and question types in PHEB.

PHEB Existing benchmarks demonstrate notable progress in multilingual and educational evaluation, yet they also reveal a clear opportunity to advance the field for European Portuguese by addressing persistent gaps. Building on these foundations, we present PHEB, a dataset that natively supports PT-PT and is explicitly aligned with the Portuguese national curriculum. It covers multiple school subjects, combines both multiple-choice and open-ended question formats, and emphasizes reasoning and language generation in authentic educational settings. In doing so, PHEB fills a gap left by existing resources: a curriculum-aligned benchmark that evaluates LLMs in native PT-PT across both factual and generative tasks.

3. PHEB Data Collection and Curation

Portuguese National Exams. We source the exams from the Instituto de Avaliação Educativa (IAVE; Institute for Educational Assessment), Portugal's official institute for educational assessment. IAVE designs and validates all external evaluations for Portuguese high school students, maintaining a comprehensive public archive of all national exams[2]. These exams are developed by experienced educators in alignment with the national curriculum and serve as university entrance requirements, ensuring both pedagogical validity and appropriate difficulty levels. Each exam includes detailed grading rubrics that specify evaluation criteria and point allocations for evaluating student responses.

3.1. Data collection and Processing

We systematically collected exams from 2006–2023 across 6 core subjects, spanning both the humanities and sciences: Portuguese Language and Literature, Mathematics, History, Geography, Philosophy, and Biology/Geology. The data collection process involved several stages:

Exams Data Extraction. We began by crawling the IAVE archive and downloading the exams and rubric PDFs for the target subjects. This results in a total of 279 exams. We then extracted all multiple-choice questions through a combination of automated PDF parsing and manual transcription, followed by complete manual verification of all question-answer pairs to ensure accuracy. We standardized all math notation to LaTeX format to ensure consistency across years and fairness when evaluating.

Grading Criteria Processing. Portuguese national exams include comprehensive grading rubrics that enumerate specific content points required for credit, along with their point values. We manually extracted these rubrics for open ended questions, allowing for LLM-as-a-judge evaluation.

3.2. Dataset Statistics

Table 1 presents the distribution of questions across subjects and question types. Our benchmark is composed of 3,575 questions, spanning 18 years of national exams. PHEB includes two types of question: **MCQs**, which always present four options, requiring models to select the correct answer, and **open-ended**, which vary from brief explanations and simple completions to multi-paragraph essays (see Figure 2 for examples). These questions are always accompanied by detailed grading criteria rubrics—used by professors to assess student performance—that enumerate the required elements and corresponding partial scores (Figure 3).

4. Experimental Methodology

PHEB allows researchers to benchmark NLP methods in terms of multiple choice questions, eliciting discriminative reasoning, and open-ended questions, eliciting structured reasoning and generation. The evaluation methodology is specific to each question type and requires different experimental protocols.

[2] https://iave.pt/

Apresentação da conclusão do argumento, de modo a constituir uma falácia *ad hominem*:
– A água de abastecimento público não deve ser enriquecida com flúor.

Níveis	Descritores de desempenho	Pontuação
2	Apresenta corretamente a conclusão do argumento, de modo a constituir uma falácia *ad hominem*. A resposta não contém elementos incorretos.	10
1	Mostra compreensão da falácia *ad hominem* sugerida nas premissas do argumento (por exemplo, referindo que o orador ataca pessoalmente o adversário), mas não apresenta a conclusão do argumento, ou apresenta-a incorretamente. Apresenta conteúdos irrelevantes ou incorretos, mas que não contradizem os conteúdos relevantes e corretos apresentados.	5

Figure 3: Example of an open-ended exam question grading rubrics, from a Philosophy exam.

EN Translation: *Presentation of the conclusion of the argument, so as to constitute an ad hominem fallacy: – Public supply water should not be enriched with fluoride.* ***Level 2 - Score 10****: Correctly presents the conclusion of the argument, so as to constitute an ad hominem fallacy. The response contains no incorrect elements.* ***Level 1 - Score 5****: Shows understanding of the ad hominem fallacy suggested in the premises of the argument (for example, by mentioning that the speaker personally attacks the opponent), but does not present the conclusion of the argument, or presents it incorrectly. Presents irrelevant or incorrect content, but which does not contradict the relevant and correct content presented.*

4.1. Evaluation Protocols

Multiple Choice Questions (MCQ). For MCQ, the ground truth is clear and unambiguous. For each question there is a set of candidate answers and we evaluate the models using two approaches: likelihood-based evaluation, where we compute the probability of each answer option under the model, or chain-of-thought (CoT) prompting. Following established benchmarks, for **MCQ we compute accuracy**, and all models are evaluated using greedy decoding. While individual MCQs carry different point values in the original exams, we report the unweighted accuracy as our primary metric, as it provides a clearer measure of model performance standardized across questions and subjects.

Open-ended Questions. For open-ended questions, there is no single reference answer. Instead, there is a list of grading rubrics that need to be satisfied by the answer to receive the full score.

In practice, for **open-ended responses, we use manual human grading** on a representative subset of open-ended questions (minimum 10 per subject with each question evaluated across all models) following the official IAVE rubrics. Each rubric specifies the content, and associated point value, for each answer. We follow these rubrics, applying the same criteria used to evaluate student responses in actual exam settings. This approach yields weighted scores that reflect the varying difficulty and importance of questions within each subject. To assess the feasibility of scaling evaluation beyond manual grading, we conducted an experiment using Gemini-2.5-flash as an LLM-as-a-judge for a subset of open-ended questions (Section 5.1). This is aligned with recent work on LLMs-as-judge for different rubrics (Hashemi et al., 2024).

While initial results show promise for automated grading, developing robust evaluation approaches—including using multiple judges, pairwise ranking, and mathematics-specific evaluation strategies—is left as future work.

4.2. Baselines

We focus on open-weight multilingual foundational models in the 2–9B parameter range, prioritizing models with Portuguese or broad European language coverage to enable meaningful comparison. In MCQ, we evaluate **multilingual open models**, including Llama 3.1 (Grattafiori et al., 2024) (8B), OLMo 2 (OLMo et al., 2025) (7B), Mistral 7B (Jiang et al., 2023), Salamandra (Gonzalez-Agirre et al., 2025) (7B), and EuroLLM (Martins et al., 2024) (9B), which provide strong baselines trained on large-scale multilingual data including Portuguese. Lastly, we evaluate the **Portuguese-focused model** Tucano (Corrêa et al., 2025) (2B) which is specifically trained for Brazilian Portuguese. For open-ended questions, due to computational constraints in manual evaluation and LLM-as-a-judge grading, we focus our analysis on a representative subset: OLMo 2, Llama 3.1, EuroLLM, and Mistral 7B.

5. Results and Discussion

Multiple Choice Questions We evaluate MCQ performance using two methods that assess different model capabilities: generation-based with CoT prompting, and likelihood-based evaluation. In generation-based evaluation, we employ CoT prompting (Wei et al., 2022; Lewkowycz et al., 2022; Wang et al., 2024), instructing models to reason through each question before providing their final answer enclosed in \boxed{} notation. This approach assesses the model's ability to explicitly articulate its reasoning, mirroring how a student would approach the exam—nevertheless, we only evaluate whether the final choice is correct. For likelihood-based evaluation, we prepend the prefix

Model	Portuguese	Math	History	Geography	Phil.	Bio/Geology	Average
Generative - CoT							
Llama 3.1	0.66	0.56	0.78	0.73	0.63	0.54	0.65
OLMo-2	0.50	0.37	0.56	0.49	0.46	0.40	0.46
Mistral 7B	0.59	0.32	0.68	0.61	0.52	0.45	0.53
EuroLLM	0.52	0.39	0.66	0.63	0.57	0.46	0.54
Salamandra	0.49	0.19	0.42	0.40	0.45	0.26	0.37
Tucano	0.23	0.12	0.23	0.27	0.28	0.27	0.23
Discriminative - Likelihood:	*"The answer is" ⟨letter⟩*						
Llama 3.1	0.66	0.36	0.79	0.74	0.61	0.55	0.62
OLMo-2	0.56	0.21	0.59	0.42	0.40	0.34	0.42
Mistral 7B	0.61	0.27	0.74	0.68	0.57	0.47	0.56
EuroLLM	0.61	0.25	0.78	0.66	0.57	0.47	0.56
Salamandra	0.55	0.15	0.62	0.61	0.52	0.37	0.47
Tucano	0.24	0.24	0.29	0.24	0.23	0.24	0.25

Table 2: Main MCQ results using generation with CoT (Wei et al., 2022) and likelihood-based evaluation.

Model	Port.	Phil.	History	Geogra.	Bio/Geology	Average
EuroLLM	**67.8**	60.3	72.2	80.0	33.9	**62.8**
Llama 3.1	56.1	**63.5**	**74.4**	76.0	38.7	61.7
Mistral 7B	51.7	42.9	71.4	**80.7**	**46.0**	58.5
OLMo 2	46.1	51.7	71.9	64.7	41.9	55.3

Table 3: Human-graded performance (as a % of maximum score) on open-ended questions. Each model is evaluated on 10 questions per subject. Biology/Geology performance is notably lower across all models due to stricter grading criteria requiring precise scientific terminology.

"A resposta é" (*"The answer is"*) to each possible answer, and select the option with the highest decoding likelihood (Clark et al., 2018), which directly probes the model's implicit knowledge, while being efficient to evaluate. The pre-pending strategy addresses the specific bias of Portuguese, where models exhibit severe bias towards "A", Section 5.2.

Table 2 presents accuracy across all subjects for both evaluation approaches. The results reveal substantial performance differences between evaluation methods. Generation-based evaluation generally outperforms likelihood, with the most pronounced gaps in mathematics. This aligns with the intuition that explicit reasoning steps improve model performance on questions which require multi-step problem solving (Wei et al., 2022). As expected, Llama achieves the most consistent performance across subjects. Notably, Portuguese and European-focused models underperform despite their language specialization, suggesting that PHEB benefits from general reasoning capability alongside language-specific adaptation.

Open Questions Table 3 presents human-graded performance on open-ended questions across subjects. Human evaluation was conducted by three native European Portuguese speakers with domain expertise (two linguists graded the humanities questions, and one biologist for Biology/Geology content). We evaluated each model on complete exams, resulting in at least 10 questions per subject, with all models evaluated on the same questions. Results show that model performance varies considerably across subjects, with lower scores in Biology/Geology compared to other disciplines. Questions in this domain demand precise factual knowledge, with the grading rubrics enumerating the exact content required to achieve specific point values, leaving little room for approximation. Models that generate well-written but imprecise responses receive low scores, regardless of fluency. Conversely, subjects like Portuguese and History will frequently ask students to analyze longer passages, identify implicit meanings, and construct arguments grounded in the documents. Therefore, stronger models are expected to at least receive partial credit.

Overall, all models perform relatively well in the humanities, but struggle in the sciences. This discrepancy appears to stem from rubrics which contain unambiguous correct content, such as Mathematics, Geography, and Biology, grading answers specifically on whether they adhere to it. In contrast, in humanities, the grading is based on fluency, coherence, and language understanding, allowing for multiple valid interpretations, as long as they

Model Prompt Formatting	Llama-3 PT	Llama-3 EN	OLMo-2 PT	OLMo-2 EN	EuroLLM PT	EuroLLM EN
CoT + \boxed{X}	0.66	0.69	0.48	0.62	0.51	0.56
With prefix						
"The answer is", letter	0.63	0.68	0.41	0.64	0.56	0.56
"The answer is", number	0.51	0.54	0.37	0.57	0.46	0.46
"Answer:", letter	0.33	0.62	0.40	0.65	0.37	0.53
"Answer:", number	0.52	0.48	0.44	0.52	0.45	0.49
Without prefix						
Number	0.33	0.33	0.34	0.35	0.26	0.37
Letter	0.22	0.41	0.25	0.45	0.22	0.34

Table 4: Accuracy across answer format variations on 250 MCQ samples. CoT with explicit answer boxing substantially outperforms all other formats. Among prefix-based approaches, "The Answer is" with letter options achieves the best performance. Without prefix constraints, models exhibit severe positional bias in Portuguese (see Figure 4), with letter-based extraction performing particularly poorly.

are well argued.

5.1. LLM-Human Grading Agreement

To assess the potential of automated grading, we evaluated inter-rater agreement between the expert human graders and Gemini-2.5-flash (Comanici et al., 2024) as a judge, on a subset of over 50 open-ended questions spanning five subjects, each answered by four different LLMs (over 250 total scored responses, similarly to Section 5). To account for varying question weights (maximum values per question range from 5 to 30 points), we normalized all scores to percentages of the maximum possible score. We then computed Pearson's correlation between both scores, obtaining $r = 0.70$ ($p < 0.001$), indicating moderately strong agreement. The LLM judge was slightly more lenient, assigning higher mean scores than human graders (mean scores of 66% vs. 62%, RMSE = 0.27), possibly reflecting greater human sensitivity to linguistic quality beyond factual correctness. While individual score predictions show higher variance, LLM-as-a-judge, using the detailed rubrics as scaffolding, may capture meaningful performance differences, enabling scalable evaluation across the full dataset.

5.2. Ablation Studies

Unbiased MCQ. In the multiple-choice question likelihood setting, we observed a systematic bias among Portuguese LLMs towards selecting "*A*" more frequently than expected. This effect, while partially addressable with heuristics and normalizations, aligns with previous reported work (Zheng et al., 2024), but appears more prevalent. We partially attribute this phenomenon to a structural property of Portuguese grammar. In likelihood-based MCQ evaluation, we compute the probability of each answer continuation given the prompt. In English, the construction "*The answer is B*" clearly distinguishes the answer prefix from the option label. The equivalent Portuguese phrase, "*A resposta é B*" begins with the definite article "A", which creates token-level ambiguity with option "A". Then, when computing likelihoods, this may systematically bias them towards option "A". This structural bias is distinct from positional bias, as is specific to Portuguese, Figure 4. The usage of different prefixes when computing likelihood mitigates this effect, with different prefixes showcasing different effects. In addition, Table 4 reports different prompting techniques, where we can confirm that all models suffer from this language-induced structural bias and how it distorts the performance of likelihood-based MCQ. In general, CoT prompting is the most unbiased technique, allowing the model to focus on the question reasoning inference and select the correct choice. Hence, this reinforces the need for language-specific approaches to guarantee an unbiased evaluation of models.

MT-derived Benchmarks. An important question for non-English evaluation is whether translation-based benchmarks adequately capture language-specific capabilities. To investigate this, we translate questions to English and back to Portuguese using Tower+ 72B (Rei et al., 2025), comparing model performance across three conditions: original Portuguese, English translation, and round-trip translation. The results (Table 5) reveal two key findings. First, all models show substantial and consistent improvement when evaluated on English translations. This improvement is present even in models with substantial Portuguese training data (EuroLLM), suggesting that modern LLMs may have internalized English reasoning patterns even when the required knowledge remains consis-

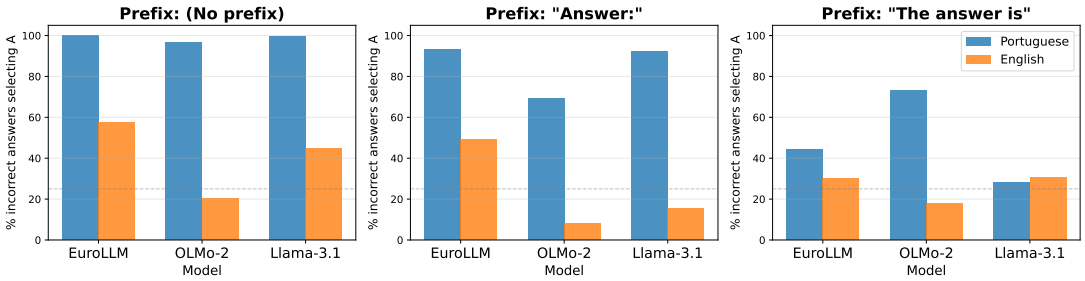


Figure 4: Selection rate for option "A" in incorrect responses across Portuguese and English evaluation. All models exhibit severe positional bias in Portuguese, with over 95% of incorrect answers defaulting to "A", compared to 20–60% in English translations of identical questions.

Model	Port.	Math	Hist.	Geogra.	Phil.	Bio/Geology	Average
				Original			
Llama	**0.66**	**0.56**	**0.78**	**0.73**	**0.63**	**0.54**	**0.65**
EuroLLM	0.52	0.39	0.66	0.63	0.57	0.46	0.54
OLMo	0.50	0.37	0.56	0.49	0.46	0.40	0.46
				Original → English			
Llama	**0.72**	**0.57**	**0.78**	**0.75**	**0.70**	**0.60**	**0.69**
EuroLLM	0.61	0.29	0.75	0.62	0.57	0.53	0.56
OLMo	0.66	0.49	0.74	0.66	0.63	0.54	0.62
				Original → English → Portuguese			
Llama	**0.66**	**0.52**	**0.76**	**0.69**	**0.63**	**0.55**	**0.64**
EuroLLM	0.51	0.34	0.63	0.60	0.57	0.45	0.52
OLMo	0.46	0.46	0.62	0.51	0.46	0.42	0.49

Table 5: Performance of generative MCQ when translating to English and then back to Portuguese

tent. Second, round-trip translation (PT→EN→PT) produces surprisingly modest effects overall, with most subjects showing performance comparable to the original. The exception is the Portuguese subject, which consistently degrades across all models. This pattern is expected, as these exams test linguistic nuance, idiomatic expressions, literary analysis, and language-specific conventions that are vulnerable to translation errors accumulating through both translation steps. Subjects like Mathematics, Biology, Geology and Geography show a much lower sensitivity. This reflects the necessity of native language benchmarks for realistic evaluation. While back-translated benchmarks may be enough for measuring factual knowledge in some domains, they cannot capture whether models can effectively process authentic native language content as it naturally occurs—the actual deployment scenario for which these systems are intended. We note that the analysis is limited to MCQ, and we expect this effect would be more pronounced in open-ended questions.

6. Conclusions

We introduced PHEB, a benchmark of 3,575 questions from 18 years of Portuguese national high school exams across six core subjects. Unlike translation-based resources, our benchmark provides European Portuguese evaluation with MCQ and open-ended questions, aligned with the Portuguese national curriculum. Our evaluations reveal several key findings: first, models exhibit systematic bias in Portuguese likelihood-based evaluation which needs to be addressed explicitly. Second, translating to English consistently improves benchmark performance, and translating these back to Portuguese causes unstable results. Third, automated grading using LLM-as-a-judge shows moderate agreement with human graders, though substantial work is still needed to achieve reliable automated evaluation. In summary, PHEB establishes a foundation for rigorous European Portuguese LLM evaluation in educational contexts. Future work should explore robust automated grading approaches and expand coverage to additional subjects.

7. Acknowledgements

This work was supported by the AMALIA project under Measure RE-C05-i08 of the Portuguese national Programa de Recuperação e Resiliência. We also acknowledge the support of Fundação para a Ciência e Tecnologia (FCT). Finally, we thank the Barcelona Supercomputing Center (BSC) for providing the computational resources that made this work possible.

8. Bibliographical References

Hugo Queiroz Abonizio, Thales Sales Almeida, Thiago Soares Laitz, Roseval Malaquias Junior, Giovana K. Bonás, Rodrigo Nogueira, and Ramon Pires. 2024. Sabiá-3 Technical Report. *CoRR*, abs/2410.12049.

Thales Sales Almeida, Hugo Queiroz Abonizio, Rodrigo Nogueira, and Ramon Pires. 2024. Sabiá-2: A New Generation of Portuguese Large Language Models. *CoRR*, abs/2403.09887.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *CoRR*, abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, et al. 2024. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities.

Nicholas Kluge Corrêa, Aniket Sen, Sophia Falk, and Shiza Fatimah. 2025. Tucano: Advancing Neural Text Generation for Portuguese. *Patterns*.

José Ángel González, Ian Borrego-Obrador, Álvaro Romo Herrero, Areg Mikael Sarvazyan, Mara Chinea-Rios, Angelo Basile, and Marc Franco-Salvador. 2026. IberBench: LLM evaluation on Iberian languages. *Comput. Speech Lang.*, 96:101899.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, Mario Mina, Adrián Rubio, Alexander Shvets, Anna Sallés, Iñaki Lacunza, Iñigo Pikabea, Jorge Palomar, Júlia Falcão, Lucía Tormo, Luis Vasquez-Reina, Montserrat Marimon, Valle Ruíz-Fernández, and Marta Villegas. 2025. Salamandra Technical Report.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, et al. 2024. The Llama 3 Herd of Models.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5427–5444, Online. Association for Computational Linguistics.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. LLM-Rubric: A Multidimensional, Calibrated Approach to Automated Evaluation of Natural Language Texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13806–13834, Bangkok, Thailand. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay V. Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving Quantitative Reasoning Problems with

Language Models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Ricardo Lopes, Joao Magalhaes, and David Semedo. 2024. GlórIA: A generative and open large language model for Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 441–453, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. EuroLLM: Multilingual Language Models for Europe.

Hugo Gonçalo Oliveira, Ricardo Rodrigues, Bruno Ferreira, Purificação Silvano, and Sara Carvalho. 2024. BATS-PT: Assessing Portuguese masked language models in lexico-semantic analogy solving and relation completion. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 207–217, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, et al. 2025. 2 OLMo 2 Furious.

Tomás Freitas Osório, Bernardo Leite, Henrique Lopes Cardoso, Luís Gomes, João Rodrigues, Rodrigo Santos, and António Branco. 2024. PORTULAN ExtraGLUE datasets and models: Kick-starting a benchmark for the neural processing of Portuguese. In *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024*, pages 24–34, Torino, Italia. ELRA and ICCL.

Miguel Angel Peñaloza Perez, Bruno Lopez Orozco, Jesus Tadeo Cruz Soto, Michelle Bruno Hernandez, Miguel Angel Alvarado Gonzalez, and Sandra Malagon. 2025. AI4Math: A Native Spanish Benchmark for University-Level Mathematical Reasoning in Large Language Models. *arXiv preprint arXiv:2505.18978*.

Ricardo Rei, Nuno Miguel Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, M. Amin Farajian, and André F. T. Martins. 2025. Tower+: Bridging Generality and Translation Specialization in Multilingual LLMs. *CoRR*, abs/2506.17080.

Sagi Shaier, George Arthur Baker, Chiranthan Sridhar, Lawrence Hunter, and Katharina Von Der Wense. 2025. MALAMUTE: A Multilingual, Highly-granular, Template-free, Education-based Probing Dataset. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4051–4069, Vienna, Austria. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, André F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2025. Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 18761–18799. Association for Computational Linguistics.

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, and Mehdi Ali. 2024. Towards Multilingual LLM Evaluation for European Languages. *CoRR*, abs/2410.08928.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. In *Advances in Neural Information Processing Systems 38:*

Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: a multilingual, multimodal, multilevel benchmark for examining large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large Language Models Are Not Robust Multiple Choice Selectors. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.