Benchmarking Retrieval-Augmented Generation for Scientific Knowledge QA in European Portuguese

José Matos, Catarina Silva, Hugo Gonçalo Oliveira
CISUC, LASI, Portugal
DEI, University of Coimbra, Portugal
josematos@student.dei.uc.pt, {catarina,hroliv}@dei.uc.pt

Abstract

Retrieval-Augmented Generation (RAG) enables grounding of model outputs in external evidence, but its impact on European Portuguese (pt-PT) scientific question answering (QA) remains unclear. We present a controlled evaluation of RAG on pt-PT knowledge QA across different scientific domains using the Portuguese test split of the Global MMLU Lite dataset. As external evidence, we use a Portuguese scientific literature knowledge base containing over 32,000 documents converted to Markdown. We benchmark five instruction-tuned small language models (4-12B) and compare closed-book baselines against 16 RAG configurations that vary by: (i) dense retriever specialization (multilingual vs. Portuguese-specific), (ii) reranking (on/off), and (iii) number of retrieved chunks ($k \in \{1, 3, 5, 10\}$). Results suggest that RAG gains are model-dependent. Some models improve consistently, others are highly sensitive to retrieval choices, and some degrade under retrieval noise, especially at larger values of k. Findings highlight the importance of model-specific retrieval tuning and ensuring that the retriever and reranker languages and domains align when deploying RAG systems for Portuguese natural scientific language processing.

Keywords: Retrieval-Augmented Generation, Science, Evaluation, Portuguese, Small Language Models

1. Introduction

Large Language Models (LLMs) have demonstrated strong capabilities in question-answering (QA), even in highly technical scientific tasks (Zheng et al., 2025). However, as these models are employed in areas not covered by their training data, they are prone to issues such as *hallucinations* (Zhang et al., 2025). This problem compounds both when we move from LLMs with hundreds of billions of parameters to smaller models with limited parametric capacity, known as Small Language Models (SLMs), and when we move away from English to multilingual settings (Ul Islam et al., 2025). To overcome this, Retrieval-Augmented Generation (RAG) enables models to ground their responses in external knowledge (Lewis et al., 2020), helping to mitigate hallucinations by extending parametric knowledge and allowing models to reference human-verifiable sources.

For Portuguese, and European Portuguese (pt-PT) in particular, the recent appearance of open-source LLMs (Lopes et al., 2024; Martins et al., 2025; Ramos et al., 2026; Simplício et al., 2026), and native sentence encoders (Gomes et al., 2024) motivate the testing of whether these retrieval systems can effectively be transferred to pt-PT scientific knowledge QA. A limited number of works have studied the impact of RAG in Portuguese-specific settings. Finardi et al. (2024) reported a set of good practices for implementing and improving RAG systems in Brazilian Portuguese, by optimizing the retrieval pipeline and evaluating the performance of proprietary models on a QA dataset generated from a book. Instead of focusing solely on retrieval, Costa and Souza-Filho (2024) compared fine-tuning and RAG for adapting LLMs to specific domains for QA tasks in Brazilian Portuguese, finding that a combination of both is the best overall strategy. Both works focus on the Brazilian variety of Portuguese, highlighting the gap relative to pt-PT. We focus on a specific question:

How does RAG improve pt-PT scientific QA across open instruction-tuned SLMs, and how sensitive is it to retrieval setup choices?

To this end, we present a controlled empirical evaluation of retrieval-augmented multiple-choice scientific knowledge QA in Portuguese. Models are evaluated on the Portuguese split of Global MMLU Lite, whose categories correspond to disciplinary fields of science (e.g., STEM, Medical, and Social Sciences) and are paired with an external knowledge base of over 32,000 Portuguese scientific documents. We systematically benchmark five instruction-tuned SLMs, varying the *generator* and *retriever* models, retrieved context sizes, and reranking strategies. The goal is to establish a reproducible baseline for different RAG configurations in the pt-PT scientific knowledge QA setting. This work is part of a larger project aimed at developing a scientific reasoning LLM native to European Portuguese, including synthetic data generation, training, and evaluation techniques tailored to lower-resource languages.

Our contributions include:

1. A pt-PT scientific QA baseline comparing closed-book with different retrieval configurations across five different open SLMs;

Proceedings of the 3rd Int. Workshop on Natural Scientific Language Processing (NSLP 2026) @ LREC 2026, pages 25–31

2. Controlled ablations of retriever language specializations, reranking, and context volume (k).

2. Methodology

In this section, we describe the data used for our evaluation and external knowledge base, followed by the specific SLMs selected for benchmarking.

2.1. Data

We evaluate models on the Portuguese test split of Global MMLU Lite[1], a compact version of Global MMLU, created by Singh et al. (2025), a benchmark combining machine translations for MMLU (Hendrycks et al., 2020) along with professional translations and crowd-sourced edits. The lite version comprises 400 test samples per language, spanning 18 languages (example test sample in Appendix A). For Portuguese, all samples were professionally translated. The dataset is split into the following categories: Humanities, STEM, Medical, Social Sciences, Business, and Other.

As the external knowledge base, we use over 32,000 open-access scientific documents in Portuguese, in Markdown format, ranging from book chapters to PhD theses, present in the CorEGe-PT[2] corpus (Kuhn et al., 2026). The texts in this dataset were sourced from a large academic repository in Portugal and span five scientific fields, with the majority overlapping with the subcategories in the evaluation benchmark: Exact and Natural Sciences, Engineering and Technology Sciences, Medical and Health Sciences, Social Sciences, and Humanities.

2.2. Models

We evaluate five non-quantized instruction-tuned SLMs at a comparable size range (4-12B): AMALIA (Simplício et al., 2026), EuroLLM-9B-Instruct (Martins et al., 2025), Qwen3-8B (Qwen Team, 2025), and finally Gemma3-4B-IT and Gemma3-12B-IT (Gemma Team, 2025).

AMALIA is a 9 billion parameter LLM trained with an emphasis on pt-PT. It is derived from EuroLLM-9B-Instruct by incorporating high-quality pt-PT data during the annealing phase of pretraining (Simplício et al., 2026). This makes it a clear choice to evaluate the impact of training on pt-PT specialized data. The pretraining phase of Qwen-3-8B includes 5 trillion high-quality STEM domain tokens (Qwen Team, 2025), making it important to test how strong science-oriented pretraining can impact the effects of retrieval on this task. The choice of Gemma3-4B-IT and Gemma3-12B-IT aims to assess the impact of model size on this task, while keeping the architecture family fixed.

3. Experimental Setup

In this section, we describe and discuss the experimental design choices for our evaluation, outlining the retrieval configurations tested and the evaluation protocol used.

3.1. Indexing and Retrieval Configurations

All gathered documents were indexed in a vector database. Given the maximum input length constraints of LLMs, and especially sentence encoders, documents are split into chunks in two phases. First, split by markdown section headers, and then recursively split into 500 character chunks with 50 character overlaps to preserve semantic continuity. After splitting, chunks are then embedded and stored in a *ChromaDB*[3] instance with cosine for computing similarity.

3.1.1. Encoder language specialization

To evaluate the impact of language specialization on indexing and retrieval, we compare two distinct dense embedding models for encoding corpus chunks and query passages. First, we use `paraphrase-multilingual-mpnet-base-v2`[4], a general-purpose multilingual encoder (278M parameters) trained on over 50 languages (Reimers and Gurevych, 2019). We compare this against `serafim-335m-portuguese-pt-sentence-encoder-ir`[5], a Portuguese-specific model from the Serafim family (Gomes et al., 2024). By selecting the 335M parameter version of Serafim, we maintain a comparable size to the multilingual baseline, effectively isolating the impact of language specialization over model scale.

3.1.2. Context selection and candidate pool

To assess the effects of context volume and potential noise, we evaluate different numbers of retrieved chunks ($k \in \{1, 3, 5, 10\}$). For efficiency and determinism in the retrieval step, we pre-compute the top 30 chunks per question for both encoders using the question stem as the query. During retrieval at inference, we slice the top-k chunks from

[1] https://huggingface.co/datasets/CohereLabs/Global-MMLU-Lite

[2] https://huggingface.co/datasets/NLP-CISUC/CorEGe-PT

[3] https://github.com/chroma-core/chroma

[4] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

[5] https://huggingface.co/PORTULAN/serafim-335m-portuguese-pt-sentence-encoder-ir

Model	Base (CB)	Best RAG (R)	Best RAG+R (R+R)	Δ_{BEST}(pp)
AMALIA	63.0 ± 2.4	64.7 ± 2.4 (*pt*, k=10)	65.7 ± 2.4 (*pt*, k=1)	+2.7
EuroLLM-9B-Instruct	62.0 ± 2.4	63.5 ± 2.4 (*pt*, k=1)	64.2 ± 2.4 (*pt*, k=10)	+2.2
Qwen3-8B	73.0 ± 2.2	72.0 ± 2.2 (*ml*, k=3)	71.8 ± 2.3 (*pt*, k=1)	-1.0
Gemma-3-4B-IT	50.7 ± 2.5	60.3 ± 2.4 (*ml*, k=3)	60.5 ± 2.4 (*ml*, k=5)	**+9.8***
Gemma-3-12B-IT	55.5 ± 2.5	70.5 ± 2.3 (*pt*, k=1)	71.0 ± 2.3 (*pt*, k=1)	**+15.5***

Table 1: **Overall accuracy (%) ± standard error (pp) on Global MMLU Lite (PT split)**. k is the number of chunks added in context. Δ_{BEST} calculated with the best-performing RAG configuration (with or without reranking) minus baseline (CB), in pp. (*pt* - Portuguese sentence encoder; *ml* - Multilingual sentence encoder; * $p < 0.05$ against CB, two-proportion Z test and McNemar).

this pool. For evaluations involving a *reranking* step, we keep a separate pool of chunks for each question and embedding model, previously reranked using a fixed multilingual cross-encoder reranker[6].

3.1.3. Retrieval variants

To further isolate the impact of different components of the retrieval pipeline on performance, we evaluate three configurations: (i) Closed-Book (CB) uses only parametric and question knowledge, serving as a reference for the model's base performance on the task. (ii) Retrieval (R) appends the top-k retrieved chunks as context and (iii) Retrieval+Reranking (R+R) uses the reranked top-k chunks.

3.2. Evaluation Protocol

We use *lm_evaluation_harness*[7] to evaluate the selected models and configurations using a 3-shot prompting technique. Few-shot exemplar selection is deterministic (fixed seed, default for framework), so that comparisons across different retrieval configurations use the same demonstrations, attributing any differences to the retrieval setup rather than exemplar variation. Scoring is based on option log-likelihood (standard for multiple-choice evaluation) rather than decoding, thus sampling parameters such as temperature do not affect the selected answer. For consistency, we keep the prompt format fixed across configurations, varying only the presence and amount of retrieved context (k) and the application of reranking. The prompt template used for every test item is provided in Appendix A. All experiments were run using a single NVIDIA RTX A6000 GPU.

[6] https://huggingface.co/Alibaba-NLP/gte-multilingual-reranker-base

[7] https://github.com/EleutherAI/lm-evaluation-harness

4. Results and Discussion

We evaluate a grid of 16 retrieval configurations per model (2 sentence encoders × 4 values of k × reranking). For compactness, Table 1 reports the maximum (upper-bound) accuracy ± standard error observed within these configurations. To evaluate robustness across all configurations, Table 2 summarizes the distribution of Δ accuracy from the baseline (CB) across all tested configurations.

Model	Δ_{WORST}	Δ_{BEST}	Mean ± SD Δ
AMALIA	-1.50	2.75	0.53 ± 1.30
EuroLLM-9B-Instruct	-2.00	2.25	0.08 ± 1.04
Qwen3-8B	-4.75	-1.00	-2.89 ± 1.18
Gemma-3-4B-IT	6.25	9.75	8.41 ± 1.08
Gemma-3-12B-IT	**11.75**	**15.50**	**13.61 ± 1.08**

Table 2: **Robustness across configurations.** Distribution of Δ accuracy (pp) over the 16 RAG configurations per model.

For Gemma models, the positive effect of RAG is clear. Accuracy increases range from 9.8 percentage points (pp) in the 4B to 15.5 pp in the 12B model, with gains consistent across all tested components and retrieval settings. Results from Table 2 further support this impact, indicating gains are strongly positive across all tested configurations for these models (8.41 ± 1.08 pp for the 4B, and 13.61 ± 1.08 pp for the 12B version). The best retrieval configurations show +2.7 pp increase for AMALIA and +2.2 pp for EuroLLM-9B-Instruct (Table 1), while Table 2 shows that the average gains for these models near zero with around 1 pp of standard deviation, suggesting they are heavily sensitive to retrieval configurations, and can even, in some cases, degrade their performance under RAG. Qwen3-8B achieves the highest baseline accuracy but shows slight degradation across all retrieval configurations, with larger context volumes (k) yielding worse results (Figure 1), suggesting that retrieval noise may be detrimental to this model's performance.

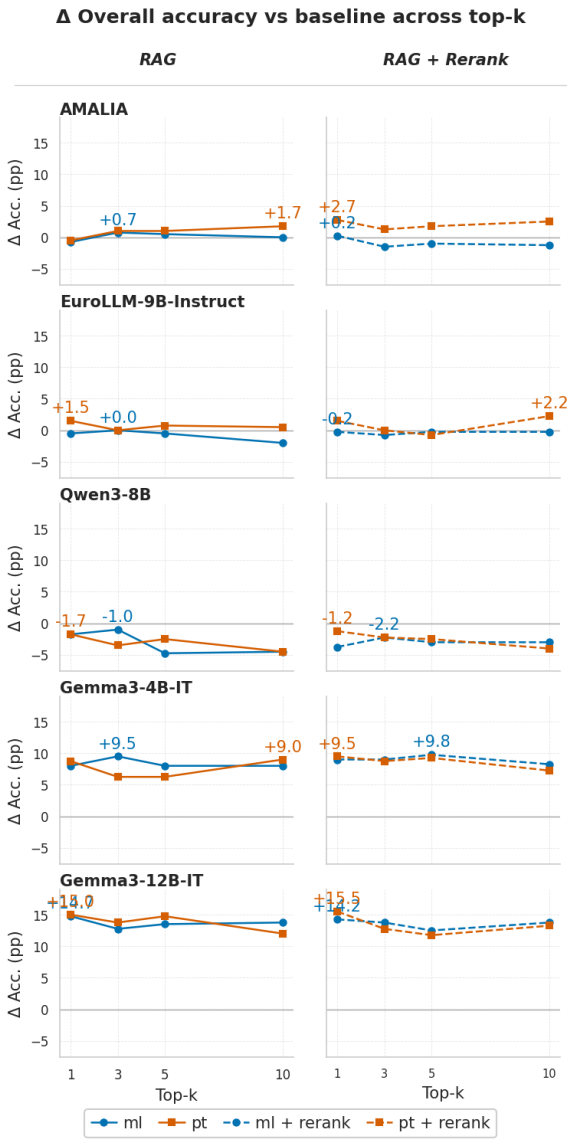

Figure 1: **Effect of context volume k on accuracy.** Δ Overall accuracy (pp) relative to the closed-book (CB) baseline as a function of $k \in \{1, 3, 5, 10\}$. Rows correspond to different generator models.

Paired tests across subcategories and proportion-based tests on overall accuracy show that only the Gemma models achieve statistically significant gains over closed-book, while AMALIA and EuroLLM-9B-Instruct show non-significant gains, and Qwen shows a consistent yet non-significant decrease.

Figure 1 suggests that performance does not scale linearly with added context. For some models, performance peaks at lower values of k, and then plateaus/decreases as additional chunks are added to the context. This aligns with the trade-off between useful evidence and noisy context as recall increases. Reranking does not guarantee a consistent accuracy improvement, but can shift the best-performing values of k toward smaller contexts, improving efficiency.

Without reranking, when using the Portuguese-specific sentence encoder, models often outperform the multilingual baseline, particularly in the Business, STEM, and Medical categories. Figure 2 indicates that language specialization interacts with domains and context size rather than acting as a universal improvement to performance. These three categories are the most technical in the group and include specialized terminology, suggesting the advantage provided by the Portuguese-specific encoder could be tied to the training data being more representative of these domains, compared to the vast multilingual encoder's training data.

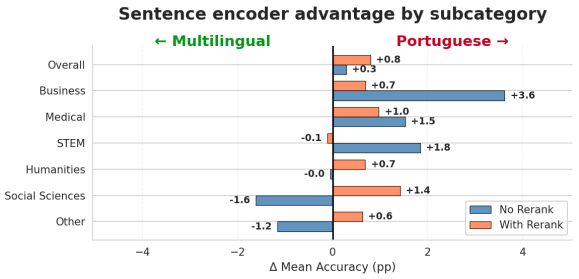

Figure 2: **Δ Accuracy between RAG configurations using different sentence encoders.** Results are reported with and without reranking.

Again, in the technical domains, it is worth noting that reranking can degrade the performance gains achieved with the Portuguese-specific encoder. This degradation is likely due to a mismatch in language in the retrieval pipeline, as the multilingual cross-encoder used for reranking may lack the granular understanding of specialized pt-PT terminology needed to accurately score the highly technical chunks. Moreover, in less technical domains like Social Sciences, the multilingual reranker successfully corrects the initial retrieval, from a -1.6 pp deficit to a +1.4 pp advantage. This highlights a central design consideration in selecting reranking models, suggesting that coupling a language-specific retriever with a general multilingual reranker can erase domain-specific gains. Ultimately, further work should explore this by coupling a language-specific retriever and reranker.

5. Conclusion

We established a baseline for scientific knowledge QA in pt-PT using RAG. Evaluating 5 SLMs of similar size reveals that the effects of RAG are highly model-dependent. By testing two different sentence encoders for chunk indexing and retrieval

purposes, one multilingual and one specific to Portuguese, we find that the impact of language specialization can vary depending on the knowledge domain being targeted. We also test the isolated effect of reranking after retrieval, and the influence of different top-k sizes on performance.

Future work should evaluate these systems beyond performance on broad scientific knowledge multiple-choice QA tasks and address open-response questions, particularly in highly technical domains, including scientific reasoning tasks. In addition, these findings motivate further efforts toward grounding-oriented, task-specific training and evaluation methods, as well as effective and efficient selection of relevant context from scientific documents, for natural scientific language processing in Portuguese.

Limitations

Due to the lack of Portuguese evaluation resources (and pt-PT specifically), we resorted to the Global MMLU Lite dataset. The performance of the systems we currently aim to develop, the ones capable of effective reasoning over scientific knowledge, may not be fully captured by this type of benchmark. Future work should focus on creating language-specific datasets in pt-PT to assess this impact. Additionally, because it is not directly targeted for pt-PT, linguistic biases toward the Brazilian variety are likely present in the questions, and consequently affect the retrieval step.

Another limitation is that retrieval quality is not directly evaluated at the chunk level. Instead, we use task accuracy as a proxy for this measure. Furthermore, only a single chunking strategy was used. Alternative approaches, such as semantic chunking or larger windows, may result in different outcomes.

Additionally, for this specific task, multiple-choice log-likelihood evaluation is appropriate for assessing knowledge capabilities. Still, further work should explore the capacity of these models for pt-PT open-ended answer generation, possibly employing prompting techniques such as Chain-of-Thought (Wei et al., 2022), which would be more aligned with the real-world use of the targeted systems.

Finally, we note that other models could serve as additional size comparisons, such as Qwen3-4B, which may be explored in future work.

Code & Data Availability

The code and data used for the experiments, as well as the results, are available at https://github.com/NLP-CISUC/Benchmark-ScientificQA-RAG-pt-PT.

Acknowledgments

This work was partially supported by the AMALIA project, funded by FCT/IP in the context of measure RE-C05-i08 of the Portuguese Recovery and Resilience Program; by the Portuguese Recovery and Resilience Plan through project C645008882-00000055, Center for Responsible AI; and by national funds through FCT – Foundation for Science and Technology I.P., in the framework of the Project CISUC (UIDB/00326/2025 and UIDP/00326/2025).

6. Bibliographical References

Leandro Costa and João Baptista de Oliveira e Souza-Filho. 2024. Adapting LLMs to New Domains: A Comparative Study of Fine-Tuning and RAG strategies for Portuguese QA Tasks. In *Proceedings of the 15th Brazilian Symposium in Information and Human Language Technology*, pages 217–227, Belém do Pará, Brazil. Association for Computational Linguistics.

Paulo Finardi, Leonardo Avila, Rodrigo Castaldoni, Pedro Gengo, Celio Larcher, Marcos Piau, Pablo Costa, and Vinicius Caridá. 2024. The Chronicles of RAG: The Retriever, the Chunk and the Generator. *arXiv Preprint arXiv:2401.07883*.

Gemma Team. 2025. Gemma 3 Technical Report. *arXiv Preprint arXiv:2503.19786*.

Luís Gomes, António Branco, João Silva, João Rodrigues, and Rodrigo Santos. 2024. Open Sentence Embeddings for Portuguese with the Serafim PT* Encoders Family. In *Progress in Artificial Intelligence: 23rd EPIA Conference on Artificial Intelligence, EPIA 2024, Viana Do Castelo, Portugal, September 3–6, 2024, Proceedings, Part III*, pages 267–279, Berlin, Heidelberg. Springer-Verlag.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

Tanara Zingano Kuhn, José Matos, Bruno Neves, Daniela Pereira, Elisabete Cação, Ivo Simões, Jacinto Estima, Delfim Leão, and Hugo Gonçalo Oliveira. 2026. CorEGe-PT: Compiling a Large Corpus of Academic Texts in Portuguese. In *Proceedings of 15th Language Resources and Evaluation Conference*, LREC 2026, page Accepted. ELRA.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal,

Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 9459–9474, Red Hook, NY, USA. Curran Associates Inc.

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 Linguistics.

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. 2025. EuroLLM: Multilingual Language Models for Europe. *Procedia Computer Science*, 255:53–62.

Qwen Team. 2025. Qwen3 technical report. *arXiv Preprint arXiv:2505.09388*.

Miguel Moura Ramos, Duarte M. Alves, Hippolyte Gisserot-Boukhlef, João Alves, Pedro Henrique Martins, Patrick Fernandes, José Pombal, Nuno M. Guerreiro, Ricardo Rei, Nicolas Boizard, Amin Farajian, Mateusz Klimaszewski, José G. C. de Souza, Barry Haddow, François Yvon, Pierre Colombo, Alexandra Birch, and André F. T. Martins. 2026. EuroLLM-22B: Technical Report. *arXiv Preprint arXiv:2602.05879*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Afonso Simplício, Gonçalo Vinagre, Miguel Moura Ramos, Diogo Tavares, Rafael Ferreira, Giuseppe Attanasio, Duarte M. Alves, Inês Calvo, Inês Vieira, Rui Guerra, James Furtado, Beatriz Canaverde, Iago Paulo, Vasco Ramos, Diogo Glória-Silva, Miguel Faria, Marcos Treviso, Daniel Gomes, Pedro Gomes, David Semedo, André Martins, and João Magalhães. 2026. AMALIA Technical Report: A Fully Open Source Large Language Model for European Portuguese. In *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)*.

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, Andre 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)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Saad Obaid Ul Islam, Anne Lauscher, and Goran Glavaš. 2025. How Much Do LLMs Hallucinate across Languages? On Realistic Multilingual Estimation of LLM Hallucination. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29077–29098, Suzhou, China. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed 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*, volume 35, pages 24824–24837. Curran Associates, Inc.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2025. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *Computational Linguistics*, 51(4):1373–1418.

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. 2025. From Automation to Autonomy: A Survey on Large Language Models in Scientific Discovery. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17733–17750, Suzhou, China. Association for Computational Linguistics.

A. Additional Information

Example from the pt split of the Global MMLU Lite Dataset

subject_category: "STEM"

question: "Os materiais usados no dissipador térmico devem ter:"

option_a: "alta condutividade térmica."
option_b: "grande superfície."
option_c: "alto ponto de fusão."
option_d: "Todas as anteriores."

answer: "D"

Example from the pt split of the Global MMLU Lite Dataset (translated)

subject_category: "STEM"

question: "The materials used in the heat sink must have:"
option_a: "high thermal conductivity."
option_b: "large surface area."
option_c: "high melting point."
option_d: "All of the above."

answer: "D"

Figure 3: Example from the Portuguese test split of the Global MMLU Lite benchmark (*electrical_engineering/test/78*)
.

Prompt template for evaluation

<few_shot_example_1>
<few_shot_example_2>
<few_shot_example_3>

Contexto Adicional:
<retrieved_chunk_1>;
<retrieved_chunk_2>;
(...)
<retrieved_chunk_K>;

Agora responde à pergunta seguinte. Se necessário utiliza o contexto adicional:
<question>
<options>
Resposta: (A/B/C/D)

Prompt template for evaluation (translated)

<few_shot_example_1>
<few_shot_example_2>
<few_shot_example_3>

Additional Context:
<retrieved_chunk_1>;
<retrieved_chunk_2>;
(...)
<retrieved_chunk_K>;

Now answer the following question. If necessary, use the additional context:
<question>
<options>
Answer: (A/B/C/D)

Figure 4: Prompt used for evaluation on the Global MMLU Lite benchmark.