Using Retrieval-Augmented Generation to improve Performance of Large Language Models on the Brazilian University Admission Exam

Leonardo de Campos Taschetto, Renato Fileto[1]

[1] Dept. of Computer Science, Universidade Federal de Santa Catarina, Florianópolis, Brasil

Abstract. *The Brazilian University Admission Exam (ENEM) presents a unique challenge for artificial intelligence. It requires deep mastering of knowledge from diverse fields. Recently, Language Models (LMs) with growing numbers of parameters have established the state-of-the-art performance on ENEM. However, techniques like Retrieval-Augmented Generation (RAG) can help further improvements, by exploiting trustfull knowledge bases to enhance contexts and reduce non-factual responses. This study investigates how RAG can improve LMs' performance on ENEM. The experiments reported in this article use up-to-date versions of four popular LMs, with and without RAG, on text-only and multi-modal data. The results reveal consistent gains using RAG with both kinds of data, across diverse fields, demonstrating the potential of RAG to improve LMs' performance on tasks requiring multidisciplinary knowledge.*

1. Introduction

University entrance examinations were first proposed as benchmarks for artificial intelligence (IA) over a decade ago [6]. Since 2017, the Brazilian University Admission Exam (ENEM) has been considered a proper test of success for IA advancements [11, 12]. Over the last three years, ENEM has become a standard test for increasingly large LMs [1, 7–9] in the Portuguese language. However, the rapid adoption of LMs in learning systems requires techniques to minimize hallucination-related mislearning [2].

The ENEM exam poses a significant challenge for LMs due to its coverage of four distinct knowledge fields: Languages, Human Sciences, Natural Sciences, and Mathematics. Recent studies have demonstrated that LMs can rely on both visual [9] and textual [7] components to solve the ENEM. However, these studies primarily focused on zero-shot prompting, and few-shot learning with Chain-of-Thought (CoT) [13].

Retrieval-Augmented Generation (RAG) [4] enables the integrated use of LMs with multidisciplinary knowledge from reliable external sources. This approach can enhance the contextual understanding and accuracy of models by grounding their responses in relevant retrieved information. However, to the best of our knowledge, this approach has not been applied to ENEM yet. This study evaluates performance gains of using RAG with specific LMs on the latest ENEM datasets.

2. Related Works

Several studies have established OpenAI's GPT-4 as the most effective LM in handling the interdisciplinary context of the ENEM [1, 7–9]. Open source models like Llama 70b, Claude 3 and Mistral reached accuracies between 65% and 80%.

Nunes et al. (2023) [7] evaluated the *gpt-4-0613* and *gpt-3.5-turbo-1106* text-only models on the ENEM 2022 dataset, excluding questions that contain images. They achieved the best results using the three-shot with CoT prompt strategy, which involves providing the model with a few examples that demonstrate the reasoning process required to solve similar problems [13]. This method resulted in the text-only version of GPT-4 attaining a mean accuracy of 87.29%, significantly higher than GPT-3.5 Turbo's 73.73%.

More recent work [9] by the same authors, conducted experiments with the multi-modal GPT-4-vision-preview model on the ENEM 2022 and 2023 datasets. Their findings indicated that this model performed consistently well across both years, with noticeable improvements in all fields of the exam, achieving a 100% accuracy in Human Sciences.

Our study builds on these works by investigating the use of RAG with four different LMs. Our experimental results, presented in section 4, show that RAG with 3-shot allows higher performance gains for answering the ENEM questions from most fields than using 3-shot with CoT.

3. Methodology

This study evaluates the LMs with and without RAG, using both zero-shot and 3-shot prompting strategies as baselines. It reproduces previous studies with current LM versions, and compares the baselines with strategies using RAG or CoT.

3.1. Model Selection

GPT-4o[1]**:** We opt for the flagship multimodal model *gpt-4o-2024-05-13*, which is distinguished by its context window of *128,000 tokens* and features for tasks requiring visual comprehension. Cutoff date: Oct 2023.

GPT-3.5[2]**:** The text-only model *gpt-3.5-turbo-0125* features a context window of *16,385 tokens*. This model is particularly significant as it currently offers *free of charge* use for most countries worldwide. Cutoff date: Sep 2021.

Llama 3 8b[3]**:** This open-source LM developed by Meta offers robust performance while maintaining computational efficiency. It supports a context length of up to *8,000 tokens*, with knowledge. Cutoff date: Mar 2023.

Llama 3 70b[4]**:** It is, according to Meta, *"the most capable openly available LM to date"*[5]. Cutoff date: Dec 2023.

3.2. Datasets

We used two datasets: one corresponding to ENEM 2022 and another one from 2023. Both comprise 180 questions distributed across the four knowledge fields mentioned in the introduction, with 45 questions per field. One question from the Mathematics field was annulled in each edition. Hence, these datasets have a total of 358 multiple-choice questions. They encompass all information about the questions (text, images, tables, and correct answers), for experiments with both text-only LMs and multi-modal LMs.

The 2022 dataset was created by [7], and later updated [9] to allow the evaluation of vision models. The authors of [9] also created the ENEM 2023 dataset[6], and incorporated the official textual descriptions of the visual elements for both exams, allowing a

[5] `https://ai.meta.com/blog/meta-llama-3/`

[6] Datasets are available at `https://huggingface.co/datasets/maritaca-ai/enem`

fair evaluation of textual-only models in questions requiring image comprehension. Our study uses the latest dataset versions for both years, feeding all models with text-only data (including image descriptions), and only the vision-enabled GPT-4o with images too.

3.3. External Corpora

For our experiments, we compiled two corpora relevant to the ENEM exams, employing each of the evaluated LMs as generators. To ensure comprehensive coverage of the topics addressed in ENEM, we selected two sources for these corpora:

1. **Wikipedia articles**: Starting from the ENEM syllabus[7], we derived a list of topics, reaching a total of 443 topics relevant to the syllabus. After determining the corresponding article in the Portuguese version of Wikipedia [8], we collected each article and prepared it for chunking.
2. **Classroom workbooks**: We started from a collection of 12 copyrighted workbooks in PDF format, one workbook for each one of the following subjects: Portuguese Language, Literature, Foreign Language (English), Mathematics, Biology, Physics, Chemistry, Geography, History, Philosophy and Sociology.

Wikipedia offers a comprehensive and regularly updated collection of articles covering a wide range of subjects, making it ideal for interdisciplinary research. In contrast, textbooks provide curriculum-specific contents tailored to the Brazilian educational context. By combining the collaboratively maintained information from Wikipedia with the educational material from textbooks, we enhance the relevance of our corpora.

3.4. Experiments

To evaluate the models' performance, we extend the experiment proposed by [9], available in a GitHub repository[9], by integrating RAG. Their experiment addresses text-only and multi-modal questions using APIs and the datasets described in subsection 3.2. Only minor changes were made to the original repository to include the RAG retriever results within the prompt. The prompts used in this experiment were identical to those employed in [9]. It ensures consistent evaluation, allowing direct comparison of the results.

Hyperparameters (i.e. parameters whose values are not learned in the training process) were set in accordance with previous studies [7, 9], as follows:

- `frequency_penalty`: **0** *(Disables the penalty applied to tokens based on their frequency in the training data.)*
- `max_tokens`: **512**
- `presence_penalty`: **0** *(Disables the penalty applied to tokens based on their presence in the input.)*
- `temperature`: **0** *(Reduces the randomness in token selection, making the output more predictable, while not strictly deterministic.)*

RAG Setup: The RAG technique involves two main components: a retriever and a generator. The retriever searches a document corpus to find relevant information based on the input query, and the generator combines the retrieved information

[7] `https://download.inep.gov.br/download/enem/matriz_referencia.pdf`
[8] `https://pt.wikipedia.org/`
[9] `https://github.com/piresramon/gpt-4-enem`

with the original query to generate responses [4]. We developed a custom retriever and compiled the document corpora described in subsection 3.3, using each one of the evaluated LMs as generators.

Chunking: In RAG, *chunking* refers to the process of dividing a document into smaller, manageable pieces, to enhance retrieval and processing efficiency [4]. We employed two distinct chunking strategies, tailored to specific types of document in our corpora. These strategies were chosen to balance retrieval efficiency and content relevance. Other strategies could be considered, but the ones we chose yielded the best results.

- **Wikipedia chunking**: Articles were divided into smaller sections by flattening nested subsections into tuples of parent subtitles and subsection texts. Irrelevant sections and blank entries were filtered out.
- **Workbook chunking**: Text from PDF workbooks was split by page separators. Trailing spaces, additional line breaks, page numbers, and external content references were removed to create manageable chunks.

Embeddings are dense vector representations that capture semantics and context [4]. In this study, embeddings were obtained using the **text-embedding-3-large**. This model was selected because its 3072 dimension embeddings are currently the best performing for OpenAI models.

Vector stores are databases optimized for storing and querying vectors efficiently in a high-dimensional space, using specialized data structures and indexing [3, 4]. Each of our chunks was embedded using the described model and the resulting embeddings were persisted in vector stores for future retrieval.

Retrieval: We measured the similarity between each query vector and document vectors using the cosine similarity.

4. Results and Discussion

The performance of the models was evaluated with and without RAG, in both zero-shot and 3-shot configurations, to determine the impact of RAG on each model's performance. Figure 1 compares the models' accuracy using distinct strategies on ENEM 2022 (above) and ENEM 2023 (below), respectively. Notice that RAG allows performance gains over the baselines, namely 0-shot and 3-shot, for almost all areas in both datasets. RAG also provides superior gains than Chain-of-Thought (CoT) prompting. In addition, the 3-shot+RAG strategy outperforms the other approaches in all areas. Notably, the combination 3-shot+RAG shows the highest performance gains in Math and Natural Sciences.

Ablation Study [5, 10]: We use the 0-shot strategy as the baseline to asses the RAG impact on the models, measuring the improvements obtained in the 3-shot and 3-shot+RAG for each model and performance metric. Figures 2 and 3 present the ablation study results for ENEM 2022 and ENEM 2023, respectively. The 3-shot+RAG strategy consistently yielded the highest ablation values across all models, particularly in Mathematics and Natural Sciences. Similar patterns of moderate improvements are observed in Language and Human Sciences.

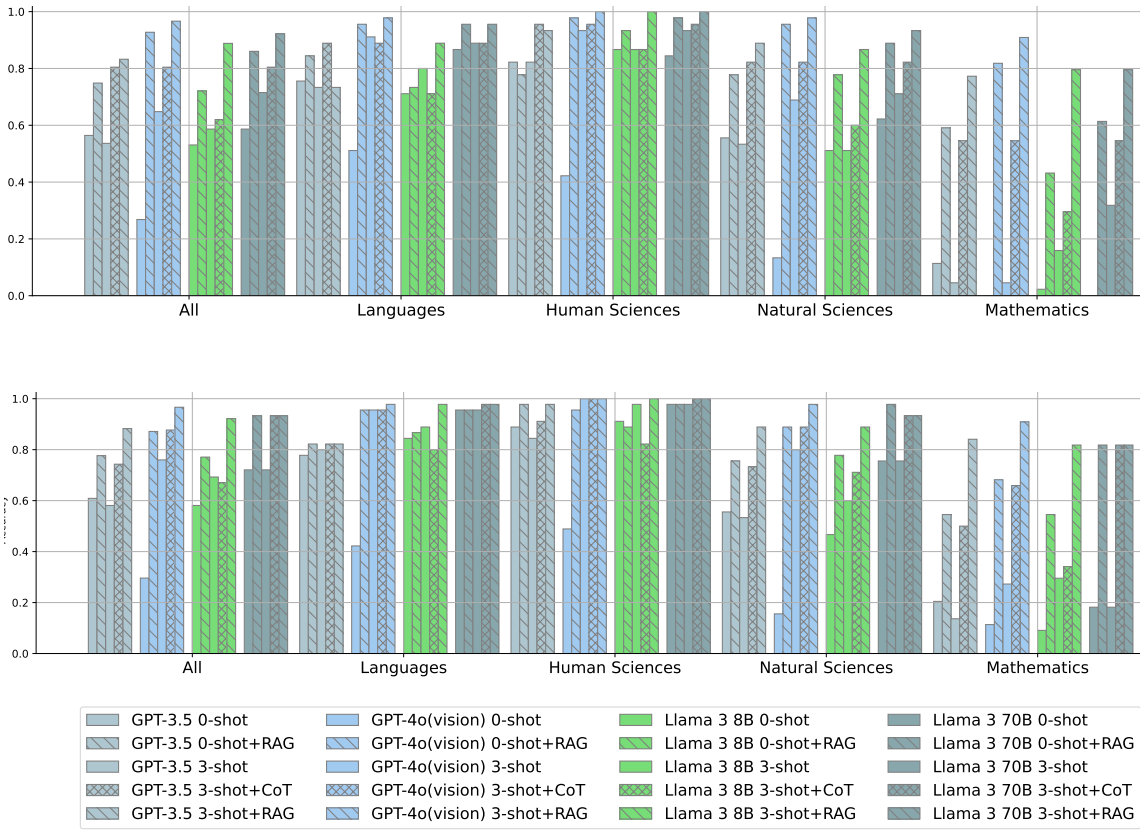

Figure 1. Accuracies on ENEM 2022 (above) and ENEM 2023 (below).

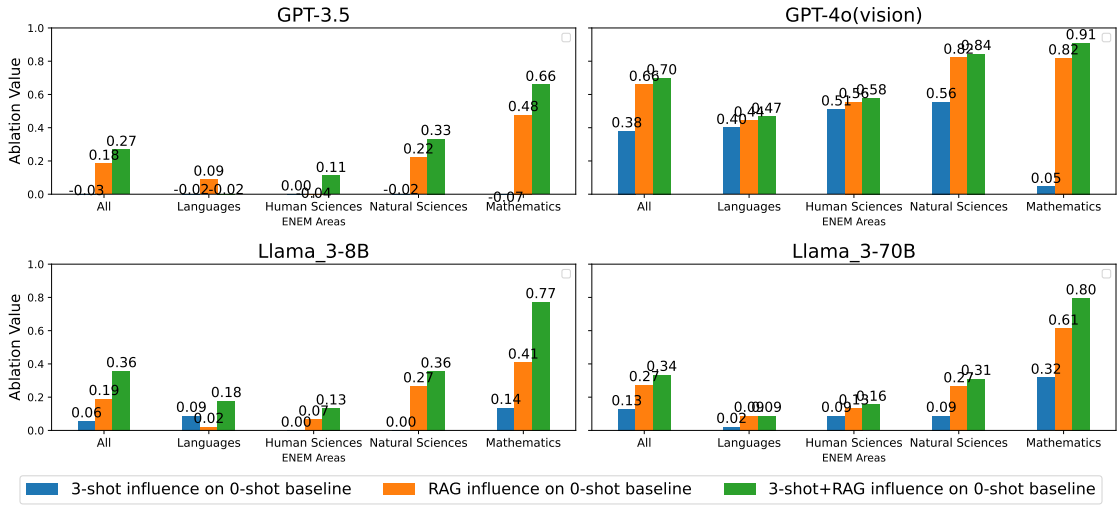

Figure 2. Ablation study results for ENEM 2022.

Discussion:

Some negative ablation values were observed in comparison with the 3-shot baseline, particularly in the Human Sciences category for GPT-3.5 and Llama 3-8B. This suggests that the 3-shot approach may occasionally mislead the model or provide inadequate context for certain questions, highlighting the complexity of designing effective few-shot examples.

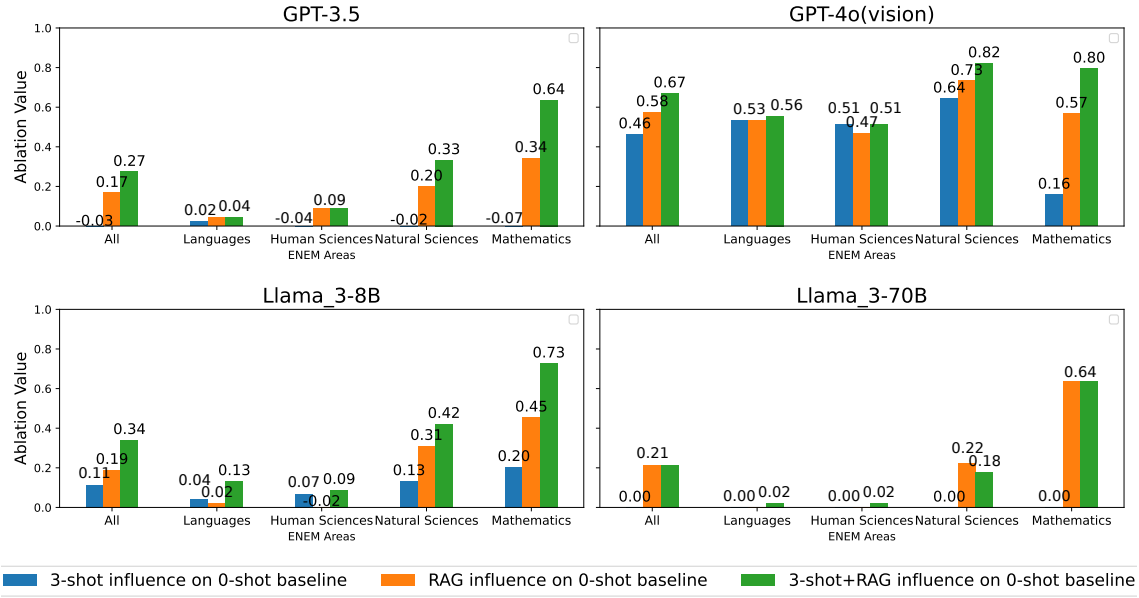

Figure 3. Ablation study results for ENEM 2023.

However, the 3-shot strategy was crucial for the GPT-4o (vision) model performance.

5. Conclusion and Future Work

The major contribution of this work is the assessment of the RAG potential to improve LMs' performance in a multi-disciplinary exam. Our findings indicate that integrating RAG with a 3-shot prompting strategy significantly enhances the performance of these models across all ENEM fields, particularly in Mathematics and Natural Sciences. It has broader implications, far beyond ENEM. The results in this paper underscore the importance of leveraging external knowledge to augment the contextual understanding of language models. AI applications in many other environments where multidisciplinary knowledge and accurate information retrieval are critical, such as future intelligent tutoring and recommendation systems, may also benefit from approaches like RAG.

We are currently working on the following extensions of this research: (i) exploiting RAG with other models for the Portuguese language, and open source and relatively small LMs, as they present more opportunities for performance gains and can be used on-premises with limited resources and confidential data; (ii) enhancing the robustness of semantic enriching training data from public and private sources; (iii) evaluating the relevance of data enrichment sources separately to determine their individual contributions to performance, and (iv) applying intelligent agents to assess the results and provide feedback to improve performance. These future directions aim to optimize the application of RAG and few-shot learning strategies in educational contexts and beyond, ultimately enhancing the accuracy and reliability of language models in complex, real-world scenarios.

Acknowledgements

This work has been supported by a 2022 CNPq Universal grant, FAPESC grant 2021TR1510, the Print CAPES-UFSC Automation 4.0 Project, and indirectly by the Céos project, financed by the Public Ministry of Santa Catarina State (MPSC).

References

[1] Thales Sales Almeida, Hugo Abonizio, Rodrigo Nogueira, and Ramon Pires. Sabiá-2: A new generation of portuguese large language models, 2024.

[2] Zhang et al. Siren's song in the ai ocean: A survey on hallucination in large language models. *ArXiv*, abs/2309.01219, 2023.

[3] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *16th Conf. of the European Chapter of the ACL*, pages 874–880, Online, April 2021. Association for Computational Linguistics (ACL).

[4] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[5] Richard Meyes, Melanie Lu, Constantin Waubert de Puiseau, and Tobias Meisen. Ablation studies in artificial neural networks. *arXiv preprint arXiv:1901.08644*, 2019.

[6] Yusuke Miyao and Ai Kawazoe. University entrance examinations as a benchmark resource for NLP-based problem solving. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1357–1365, Nagoya, Japan, October 2013. Asian Federation of Natural Language Processing.

[7] Desnes Nunes, Ricardo Primi, Ramon Pires, Roberto Lotufo, and Rodrigo Nogueira. Evaluating gpt-3.5 and gpt-4 models on brazilian university admission exams, 2023.

[8] Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. Sabiá: Portuguese large language models. In *Brazilian Conference on Intelligent Systems*, pages 226–240. Springer, 2023.

[9] Ramon Pires, Thales Sales Almeida, Hugo Abonizio, and Rodrigo Nogueira. Evaluating gpt-4's vision capabilities on brazilian university admission exams. *Semantic Scholar*, abs/2311.14169, 2023.

[10] School of Electrical Engineering Sheikholeslami, Sina. KTH and Computer Science (EECS). Ablation programming for machine learning, 2019.

[11] Igor Cataneo Silveira and Denis Deratani Mauá. University entrance exam as a guiding test for artificial intelligence. In *2017 Brazilian Conference on Intelligent Systems (BRACIS)*, pages 426–431. IEEE, 2017.

[12] Igor Cataneo Silveira and Denis Deratani Mauá. Advances in automatically solving the enem. In *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 43–48. IEEE, 2018.

[13] Jason Wei et al. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.