Enhancing Large Language Model Performance on ENEM Math Questions Using Retrieval-Augmented Generation

João Superbi[1], Heitor Pereira[1], Emanoel Santos[1], Lucas Lattari[1], Bianca Castro[1]

[1]Instituto Federal do Sudeste de Minas Gerais - Campus Rio Pomba
{joao.superbi1608, hpff999, emanoelvisali}@gmail.com,
{lucas.lattari, bianca.portes}@ifsudestemg.edu.br

Abstract. *In this study, we explore the use of Retrieval-Augmented Generation (RAG) to improve the performance of large language models (LLMs), such as GPT-3.5 Turbo and GPT-4o, in solving ENEM mathematics questions. Our experiments demonstrate that RAG potentially provides significant improvements in accuracy by introducing relevant contextual information. With RAG, GPT-4o consistently outperforms GPT-3.5 Turbo, underscoring the potential of this technique to enhance educational AI tools. This research illustrates the potential of RAG-enhanced LLMs to advance educational applications and encourages further exploration in this field.*

1. Introduction

Advances in natural language processing (NLP) and deep learning (DL) have enabled the creation of large language models (LLMs) like ChatGPT, which are proficient in tasks such as translation, summarization, and question-answering [Devlin et al. 2019, Mercan et al. 2023]. Recent research has explored their application in education, especially for solving academic problems, including in Portuguese [Bordt and von Luxburg 2023, Choi et al. 2021, Pires et al. 2023, Silva et al. 2023, Nunes et al. 2023, Mendonça 2024].

Studies indicate that LLMs generally perform worse on math questions compared to other subjects [Nunes et al. 2023, Pires et al. 2023]. To address this challenge, our research evaluates state-of-the-art LLMs like ChatGPT on math problems from the *Exame Nacional do Ensino Médio* (ENEM), a key college entrance exam in Brazil known for its challenging math content. This study aims not only to test the capabilities of LLMs in solving ENEM math questions but also to explore their potential role in educational support, including exam preparation and identifying areas where students may need additional help.

Our study focuses on the effectiveness of ChatGPT, particularly the GPT-3.5 Turbo and GPT-4o models, using different versions and prompting strategies like Retrieval-Augmented Generation (RAG) [Lewis et al. 2020] and Chain-of-Thought (CoT) explanations. We also explore various configurations, including personas and temperature settings, to assess their impact on model accuracy.

Unlike prior studies [Pires et al. 2023, Nunes et al. 2023], we use RAG to integrate additional contextual information into prompts, aiming to enhance LLM accuracy. Our findings show that RAG generally improves performance over models not using it.

Our contributions include:

- Evaluating the effectiveness of RAG in enhancing LLM performance on ENEM mathematics questions.

- Comparing the performance of GPT-3.5 Turbo and GPT-4o with different prompting strategies.
- Exploring the impact of various configurations, such as temperature settings and the use of personas, on model accuracy.
- Providing an analysis of experiments to identify effective methods for solving ENEM math questions.

2. Related Works

Similar to our work, [Nunes et al. 2023] evaluated GPT-3.5 and GPT-4 on the ENEM exam. Using zero-shot, few-shot, and CoT prompts, GPT-4 achieved 87% accuracy on the 2022 exam, outperforming GPT-3.5 by 11 points. Building on previous work, [Pires et al. 2023] assessed GPT-4 on ENEM exams, incorporating both visual and textual comprehension. GPT-4 achieved 89.94% accuracy on the 2023 exam with image captions, indicating its vision potential and the need for improvement. Both works inspired our approach.

[Silva et al. 2023] tested GPT-3.5, GPT-4, and Llama 2 in answering agriculture-related questions with datasets from Brazil, India, and the USA. By incorporating RAG and ER techniques, GPT-4 achieved a 93% accuracy on agronomist certification exams, outperforming GPT-3.5. Their research inspired our use of RAG for answering questions in educational contexts.

3. Methodology

3.1. OpenAI Models

We used two ChatGPT models available via API to paying customers[1]: **GPT-3.5 Turbo** with a 16,385-token context window and training data up to September 2021, and **GPT-4o** with a 128,000-token context window and training data up to October 2023.

Due to certain restrictions, including financial ones, we focused our experiments on these models, which serve as baselines for our study. GPT-3.5 Turbo provides a more limited capability, while GPT-4o represents state-of-the-art performance, allowing us to effectively assess the impact of the RAG approach.

3.2. Datasets

Our reference set consists of 2,439 ENEM math questions from 2019 to 2022 and their solutions, sourced from math-focused YouTube channel transcriptions and the Brasil Escola website[2]. To enhance RAG, we also included image captions of ENEM questions from [Pires et al. 2023].

For evaluation, we used 44 math questions from the orange booklet of ENEM 2023, which were not part of the reference set. This setup allows us to assess the performance of our approach on unseen data, as the GPT models did not have access to these questions since the exam occurred in November 2023.

To avoid confusion with standard machine learning terms, we use “reference set” and “evaluation set” instead of “train set” and “test set”, as our methodology does not follow the typical ML training and testing framework.

[1]https://platform.openai.com/docs/overview

[2]https://exercicios.brasilescola.uol.com.br/exercicios-matematica

3.3. RAG implementation

Our RAG method develops dynamic prompts using Few-Shot learning and CoT explanations, where selected questions and their solutions from the reference set are included in prompts. This helps the LLM break down complex problems into smaller parts, improving performance compared to zero-shot prompts.

Unlike other few-shot approaches that use random questions [Nunes et al. 2023], we retrieve questions semantically similar to the target question, introducing a non-parametric memory to enhance LLM accuracy and performance.

Our RAG approach involves the following steps:

1. **Data compilation and structuring:** Compiling all reference questions into a CSV file with separate columns for statements and alternatives, followed by a manual inspection to check for errors and improve writing quality.
2. **Pre-processing text:** Standard tasks from NLP problems were applied, including converting text to lowercase, removing punctuation, numbers, stopwords and repetitive words like "ENEM", "2023", "2022", "question" among others.
3. **Generating embeddings:** Each question was converted into an embedding using a Sentence-Transformers model based on BERT[Reimers and Gurevych 2019]. This BERT model, trained on Portuguese, maps sentences and paragraphs to a 768-dimensional dense vector. We performed the same task for the evaluation set (math questions from ENEM 2023) to enable semantic search and retrieval of questions similar to the target question.
4. **Comparing embeddings:** We compared the embeddings from the evaluation set with those from the reference set using cosine similarity. Cosine similarity is suitable for comparing texts that can vary in length and magnitude.
5. **Retrieving similar questions:** For each question from the evaluation dataset, we selected the 3 most similar questions based on the similarity score. We applied a similarity threshold of 75%, ensuring that only relevant questions meeting this threshold are included, minimizing potential noise. This threshold was chosen empirically based on testing.
6. **Constructing prompts:** The final prompt in RAG approach is described as follows: *Your task is to help solve ENEM math multiple-choice questions. Below are some solved questions from previous editions, with a step-by-step explanation of how to achieve the correct answer. Use these examples as a reference to solve the last question, which is the target question.* This is followed by the most similar questions.

3.4. Comparison of Prompting Strategies

We tested two primary prompting strategies to evaluate the performance of GPT-3.5 Turbo and GPT-4o on ENEM math questions. The non-RAG strategy used a Chain-of-Thought (CoT) approach, where the model reasoned step-by-step without retrieving external context. In contrast, the RAG strategy applied few-shot learning, dynamically retrieving relevant questions and solutions from the reference set to provide similar examples for solving the target question, thereby enhancing answer accuracy.

We did not consider zero-shot or other variations due to differences highlighted in previous studies and the financial constraints of API usage.

3.5. LLM Call

After generating the prompts, we created a Google Colab environment to read and process the prompts for each question. For each generated output, we used regular expressions to extract the alternative selected by the LLM. We then compared it with the correct alternative from each question of ENEM 2023 to compute the accuracy metrics for this study.

3.6. Experiments

We conducted eight experiments using GPT-3.5 Turbo and GPT-4o to evaluate the impact of RAG, varying prompt styles (with or without persona) and temperature settings (0.0, 0.3, 0.6, and 1.0) (Table 1). All experiments were conducted with a maximum token output limit of 1024 tokens.

Table 1. Overview of experiments with their configurations, such as GPT models, Temperature, and Prompt Types.

Experiment	Models	Persona	Temperature
1	GPT-3.5, GPT-4o	Yes	0.0
2	GPT-3.5, GPT-4o	No	0.0
3	GPT-3.5, GPT-4o	Yes	0.3
4	GPT-3.5, GPT-4o	No	0.3
5	GPT-3.5, GPT-4o	Yes	0.6
6	GPT-3.5, GPT-4o	No	0.6
7	GPT-3.5, GPT-4o	Yes	1.0
8	GPT-3.5, GPT-4o	No	1.0

In each experiment, tests were performed both with and without the RAG technique. For RAG, a few-shot strategy included retrieved questions and their solutions from the reference dataset, while without RAG, a zero-shot approach was used. All experiments employed the CoT strategy to enhance the reasoning in the outputs.

For experiments 1, 3, 5, and 7, the prompts included a persona, stating: *"You are a highly competent math expert"*, to guide the LLM in generating more precise and relevant answers. In contrast, experiments 2, 4, 6, and 8 used prompts without a persona, directly instructing the model to solve a specific ENEM math question step-by-step and conclude with the correct answer choice.

Initially, we aimed to test all ENEM subjects and available GPT models, but the financial constraints of the API limited our scope and influenced the parameter selection.

4. Results

This section presents the results of experiments on ENEM 2023 math questions, evaluating GPT-3.5 Turbo and GPT-4o with and without RAG, and exploring prompt variations and different temperature settings.

4.1. Overall

Table 2 shows the results for GPT 3.5-Turbo and GPT-4o on 44 ENEM 2023 math questions. GPT-4o consistently outperformed GPT-3.5 Turbo, with accuracies ranging from 0.66 to 0.89 for GPT-4o and 0.36 to 0.55 for GPT-3.5 Turbo. Mean accuracies were 0.445 (GPT-3.5) and 0.761 (GPT-4o) without RAG, and 0.486 (GPT-3.5) and 0.833 (GPT-4o) with RAG.

Table 2. Accuracy (Assertiveness) of GPT Models on ENEM Math Questions.

Experiment	GPT-3.5	GPT-3.5 with RAG	GPT-4o	GPT-4o with RAG
1	0.36	**0.50**	0.77	**0.82**
2	0.45	0.45	0.77	**0.89**
3	0.50	0.50	0.77	**0.86**
4	0.43	**0.45**	0.77	**0.84**
5	0.50	**0.55**	0.80	**0.84**
6	0.48	**0.55**	0.82	**0.84**
7	**0.41**	0.39	0.66	**0.82**
8	0.43	**0.50**	0.73	**0.75**
Average	0.445	**0.486**	0.761	**0.833**

For GPT-3.5, RAG improved accuracy in six out of eight experiments. It resulted in draws in two experiments (2 and 3) and showed a minor decline in one (Exp. 7). For GPT-4, RAG consistently increased accuracy in all experiments.

Figure 1 depicts an error bar graph with 95% confidence intervals for both models, GPT-3.5 and GPT-4o, with and without RAG. The graph illustrates the mean accuracy for each model configuration and highlights the impact of using RAG on performance.

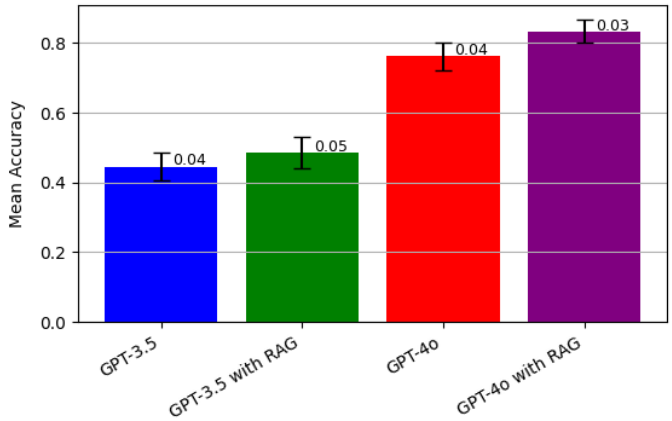

Figure 1. Error Bar Graph of Mean Accuracy for GPT-3.5 and GPT-4o with and without RAG. The error bars indicate 95% confidence intervals for the mean accuracies.

The introduction of RAG for both models showed numerical improvements but no statistical significance, likely due to sample heterogeneity (multiple variations of parameters) and the small sample size (n=8). In future experiments, we expect to use larger and more homogeneous samples to better evaluate the impact of RAG.

4.2. Prompt Variation (Persona)

Introducing a Persona prompt (e.g., "You are a highly competent math expert.") had minimal impact on performance. GPT-3.5 had mean accuracies of 0.4425 (with Persona) and 0.4475 (without). For GPT-4o, the mean accuracies were 0.75 (with) and 0.7725 (without). These results suggest that other factors, such as RAG use, temperature, and question complexity, have a more substantial impact.

4.3. Temperature

Varying the temperature in GPT-3.5 Turbo and GPT-4o showed that moderate temperatures yield better accuracy than extreme low or high settings. Table 3 shows that the average accuracy of GPT-3.5 Turbo improves from 0.405 to 0.465 when the temperature is adjusted from 0 to 0.3, achieving 0.49 at 0.6. GPT-4o follows a similar pattern, improving from 0.77 to 0.81 at a temperature of 0.6. However, at a temperature of 1.0, accuracy decreases significantly for both models.

Table 3. Effect of Temperature Variation on LLM Accuracy.

Temperature	GPT-3.5 (Mean Accuracy)	GPT-4o (Mean Accuracy)
0.0	0.405	0.77
0.3	0.465	0.77
0.6	**0.49**	**0.81**
1.0	0.42	0.695

At a temperature of 0.0, outputs are deterministic (the output is always the same), but higher temperatures introduce variability, impacting consistency. Therefore, while intermediate temperatures show better accuracy in our tests, further experiments with larger samples are needed to validate these findings. These insights highlight the need for more comprehensive research, particularly involving other LLMs, which we plan to investigate further.

4.4. Discussion

Although RAG generally improved accuracy, there were questions where the basic GPT model outperformed GPT with RAG and vice versa. This section addresses these cases, aiming to analyze the patterns that make RAG perform better or worse.

4.4.1. Frequent Errors in RAG vs. Basic Model

Analysis of the experimental results revealed that basic GPT models correctly answered more questions than GPT with RAG in the areas of Geometry and Trigonometry (136, 166, 168, and 176), Financial Calculations (144 and 160), Statistics and Probability (159, 165, and 172), Graph Interpretation (177), Arithmetic, Ratio and Proportion, Basic Mathematics (163, 175, and 178).

GPT-4o outperformed GPT4-o with RAG on questions 136, 144, and 150, while GPT-3.5 Turbo made errors on most other questions, with both models failing on questions 136 and 144 (Table 4).

Table 4. Questions Incorrectly Answered by GPTs with RAG but Correctly by Basic GPTs.

Model	**Question Numbers**
GPT-3.5 Turbo	136, 139, 141, 144, 146, 150, 159, 160, 163, 165, 168, 172, 175, 178
GPT-4o	136, 144, 150

Many errors were due to the visual interpretation required for questions such as 136, 139, 141, 150, 159, and 172. Despite descriptive image captions, if RAG does not retrieve similar examples involving geometric patterns and graphs, noise can be introduced. For example, question 139 required calculating surface areas, but the retrieved questions

involved more complex geometric shapes and conversions. This added unnecessary complexity, increasing the potential for errors in interpretation and calculation.

Similarly, the relevance of retrieved questions is also crucial. For instance, question 179 involved converting Mayan calendar dates, which is not adequately covered in the training dataset. Thus, RAG returned questions about solar cycle calculations and arithmetic progression, causing errors due to the irrelevant context.

Effective curation of the training dataset is essential. Retrieved questions with distinct or unnecessary contexts and advanced complexity in their solutions introduce noise, increasing the chance of errors. However, GPT-4o handled this better than GPT-3.5 Turbo, which struggled more with irrelevant or complex examples.

4.4.2. Frequent Errors in Basic Model vs. RAG

When considering the questions where RAG appears to have contributed, certain patterns are observed. These questions cover topics such as Geometry and Trigonometry (questions 150, 166, 176, and 179), Functions and Graphs (questions 147 and 158), Probability and Combinatorics (questions 152, 155, and 162), and Arithmetic, Ratio, and Proportion (question 175). Table 5 provides a summary.

Table 5. Questions Correctly Answered by GPTs with RAG but Incorrectly by Basic GPTs.

Model	Question Numbers
GPT-3.5 Turbo	147, 158, 166, 175, 179
GPT-4o	150, 152, 155, 156, 162, 178, 179

It appears that RAG retrieved examples closely matching the target questions, which probably contributed to the correct responses. For example, question 152 presented a similar problem structure, focusing on the inclusion of white balls in an urn. Both questions dealt with Probability and Combinatorics and had similar solution steps, which suggests a relevance that positively influenced the results.

A similar observation can be made about question 166, which involves calculations related to water consumption and storage volume. RAG retrieved a question from the 2020 ENEM (question 48 from the digital exam) with a similar structure and topic. The close match in terms and detailed calculations, including unit conversions, emphasizes the importance of matching terms for the success of RAG.

Handling questions with images is also difficult. Among the 11 questions analyzed in this section, only 3 contained images. Incorporating well-described image-based support questions in RAG to match the target questions is challenging. The retrieved questions depend on having a very similar structure, including images, and when this similarity is missing, errors can occur.

The RAG approach may be particularly effective for straightforward questions. For instance, question 175 and those retrieved by RAG involve basic arithmetic, multiplication, and unit conversion. Thus, questions without additional complexity are more likely to be part of the RAG training dataset, increasing the probability of correct answers when the LLM responds to a target question.

Finally, the more previous ENEM questions are incorporated into the dataset, the

higher the potential chance for a LLM using RAG to correctly guide the response generation accurately.

5. Conclusions

This study demonstrates the potential of integrating RAG with LLMs like GPT-3.5 Turbo and GPT-4o for solving complex mathematical problems on the ENEM exam. Our findings reveal that GPT-4o consistently outperforms GPT-3.5 Turbo, with RAG slightly improving accuracy by providing relevant context and examples. The integration of relevant contextual examples proves beneficial, although it emphasizes the necessity for meticulous data curation. These findings indicate that LLM-driven educational tools can significantly impact problem-solving efficiency, personalized learning, and overall educational outcomes.

Future research should extend this approach to other ENEM subjects, improve retrieval algorithms by considering factors beyond semantic similarity, and use larger datasets.

References

Bordt, S. and von Luxburg, U. (2023). Chatgpt participates in a computer science exam. arXiv preprint arXiv:2303.09461.

Choi, J. H., Hickman, K. E., Monahan, A. B., and Schwarcz, D. (2021). Chatgpt goes to law school. *J. Legal Educ.*, 71:387.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Mendonça, N. C. (2024). Evaluating chatgpt-4 vision on brazil's national undergraduate computer science exam. *ACM Trans. Comput. Educ.* Just Accepted.

Mercan, Ö. B., Cavsak, S. N., Deliahmetoglu, A., and Tanberk, S. (2023). Abstractive text summarization for resumes with cutting edge nlp transformers and lstm. *2023 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–6.

Nunes, D., Primi, R., Pires, R., Lotufo, R., and Nogueira, R. (2023). Evaluating gpt-3.5 and gpt-4 models on brazilian university admission exams. arXiv preprint arXiv:2303.17003.

Pires, R., Almeida, T. S., Abonizio, H., and Nogueira, R. (2023). Evaluating gpt-4's vision capabilities on brazilian university admission exams. arXiv preprint arXiv:2311.14169.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Silva, B., Nunes, L., Estevão, R., Aski, V., and Chandra, R. (2023). Gpt-4 as an agronomist assistant? answering agriculture exams using large language models. arXiv preprint arXiv:2310.06225.