Leveraging GPT-4 for Accuracy in Education: A Comparative Study on Retrieval-Augmented Generation in MOOCs

Fatma Miladi, Valéry Psyché, and Daniel Lemire

TELUQ University, 5800 rue Saint-Denis, Montreal, QC H2S 3L5, Canada
{fatma.miladi,valery.psyche,daniel.lemire}@teluq.ca

Abstract. Large Language Models (LLMs), such as Generative Pre-trained Transformers (GPTs), have demonstrated remarkable capabilities in natural language processing (NLP). However, these models often encounter challenges such as inaccuracies and hallucinations, which can undermine their utility. Retrieval-Augmented Generation (RAG) has emerged as a promising approach to enhance model accuracy and reliability by integrating external databases. This study investigates the use of RAG to improve the accuracy of GPT models in educational settings, particularly within the realm of Massive Open Online Courses (MOOCs). Through a comparative analysis of various GPT model iterations, we observed a significant improvement in accuracy, increasing from 60% with GPT-3.5 to 80% using the RAG-augmented GPT-4. This enhancement highlights the considerable potential of RAG-augmented GPT models in improving the accuracy of content generation. Such enhanced accuracy suggests revolutionizing assessment methodologies and learning experiences, fostering an educational environment that is more interactive and tailored to individual needs.

Keywords: Generative pre-trained transformers · GPT · Evaluation· MOOC · Online learning · Exercises assessments · Retrieval augmented generation

1 Introduction

The advent of Large Language Models (LLMs), such as the generative pre-trained transformer (GPT), has revolutionized the field of artificial intelligence, particularly in natural language processing (NLP) [1,2]. These models have demonstrated remarkable performance across various domains including finance, technology, and healthcare [3,4,5]. However, despite their impressive capabilities, large language models are not devoid of limitations. A significant challenge they face is their tendency to 'hallucinate', producing content that may not be factually accurate [6,7]. Such hallucinations can lead to the generation of information that is sometimes opposed to established facts, posing challenges for their reliable application in critical domains.

To address this issue, researchers developed a Retrieval-Augmented Generation (RAG) approach introduced by Lewis et al. in 2020 [8]. RAG aims to enhance LLMs by integrating external knowledge sources into the generation process. This integration not only improves the model's ability to generate accurate and relevant responses, but also represents a significant advancement within the realm of LLMs, particularly for generative tasks [9,10]. Although Retrieval-Augmented Generation has shown promise in various domains, its application in educational contexts remains largely unexplored. This research gap motivated our investigation into the potential of large language models augmented by RAG to enhance content accuracy in educational environments.

In this study, we investigated the potential of retrieval augmentation techniques to enhance the accuracy of traditional GPT models. Our primary focus was on educational settings, particularly in Massive Open Online Courses (MOOCs). This investigation is motivated by our central research question: How does integrating Retrieval-Augmented Generation with GPT models impact the accuracy of content in an educational context? To address this question, we formulated the following hypothesis (H1): The GPT-4 model, when enhanced with retrieval-augmented capabilities, will surpass both GPT-3.5 and its RAG-augmented version, as well as the standard GPT-4 model, in generating accurate responses. This hypothesis paves the way for a comparative analysis to understand the additional benefits of integrating RAG techniques with advanced GPT models in education.

2 Data

Our study utilized the MOOC focused on artificial intelligence (AI), developed by University TELUQ [11]. This course is divided into four main modules, each focusing on different aspects of AI. The first module introduces general AI concepts. The second module is dedicated to symbolic AI, whereas the third module covers connectionist AI. The final module discusses the application of artificial intelligence in education. These modules are supplemented by various learning resources including videos, texts, in-depth concepts, definitions, exercises, and illustrative images. The MOOC comprises 115 formative assessment exercises, encompassing a range of formats, including 24 true/false exercises, 24 multiple-choice questions (MCQs), 13 matching exercises, and 54 fill-in-the-blank exercises.

3 Models

In this study, we compare four AI models: RAG-augmented GPT-4, RAG-augmented GPT-3.5, GPT-3.5, and GPT-4. The augmented variants incorporated retrieval-augmented generation to enhance the accuracy.

As illustrated in Figure 1, the architecture of the augmented models utilizes a sophisticated workflow designed to enhance user interaction through a web interface. This process is initiated by the user query, which is converted

into a vector representation to encapsulate its semantic meaning. We employed OpenAI's text-embedding-ada-002 [12] for this purpose, enabling an effective retrieval-augmented generation [8]. Next, the system compares query embedding and a specialized database filled with text embeddings. Following this, it identifies and selects the text segments, or 'chunks', that demonstrate the greatest cosine similarity scores. The selected segments were integrated with the original query to provide additional context, thereby enriching the prompt. This enhanced prompt is then processed using a large language model such as GPT-4 to generate a comprehensive and relevant response, drawing on domain-specific knowledge to ensure accuracy.

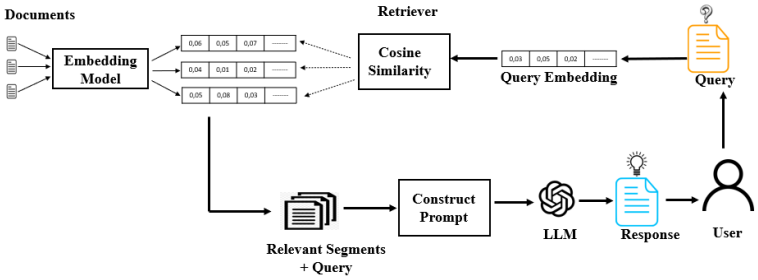

Fig. 1. Overview of the Model Architecture: From User Query Processing to Response Generation.

4 Experimental Design

To evaluate the effectiveness of the models in assessment exercises, we applied zero-shot and few-shot prompting, which are widely used in large language model studies for performance benchmarking [13,14]. Assessment questions were presented as they appear in the MOOC, utilizing prompt templates for true/false, multiple-choice, matching, and fill-in-the-blank questions, as shown in Table 1. We compared responses from the RAG-augmented GPT-4, RAG-augmented GPT-3.5, GPT-3, and GPT-4 models against correct answers, treating partially correct answers as incorrect, in line with MOOC standards. Our comparative analysis indicated that the outcomes of zero-shot and few-shot prompts were similar. Therefore, we chose zero-shot prompting as our primary evaluation method owing to its immediate practicality.

5 Results

In this study, we evaluated the performance of the RAG-augmented GPT-4 against GPT-3.5, RAG-augmented GPT-3.5, and the standard GPT-4 model, using a dataset of 115 French exercises from a MOOC on AI. The results summarized in Tables 2-5, indicate a progressive improvement in accuracy across

Table 1. Assessment exercises used in the performance evaluation of GPT models.

Type of Question	Sample Prompt
True/false	Indicate whether the following statement is true or false: An intelligent agent cannot adapt its actions to its environment nor act upon it. 1. True 2. False
Multiple Choice Question	Select the correct answer: According to Yann LeCun, making a machine intelligent allows it to: A. dream. B. memorize. C. learn. D. perceive.
Matching exercise	Match each definition with its corresponding term from the following: Definitions: 1. Various digital technology, mathematical, and other components that enable the design of an autonomous car. 2. The ability of a neural network to adjust itself, changing its behavior based on an environment, this ability can be used during the learning phase. 3. A robotic arm that has learned through trial-and-error manipulation to handle a Rubik's Cube. Terms: A. Artificial Intelligence B. Adaptability C. Intelligent Agent
Fill-in-the-blank	Fill in the blank: To pass the test of ..., the computer must be equipped with an artificial vision device to perceive objects and a robotic capability to manipulate objects and move.

generations of GPT models. The GPT-3.5 model achieved a baseline success rate of 60%. This was followed by the RAG-augmented GPT-3.5 that exhibited an improved success rate of 74%. The GPT-4 model further increased the accuracy to 77% and the RAG-augmented GPT-4 achieved the highest success rate of 80%.

True/False Exercises In the True/False exercises, as indicated in Table 2, the GPT-3.5 model demonstrated foundational capability with a 65% success rate. This was enhanced using the RAG-augmented GPT-3.5, which achieved an 85% success rate. Both the GPT-4 model and its augmented variant further improved the performance, reaching an 87% success rate, representing the highest level of accuracy among the models tested.

Multiple-Choice Questions (MCQs) As shown in Table 3, the performance on MCQs improved across the GPT model versions. The standard GPT-3.5 model began with a success rate of 60%, which was enhanced to 73% with the GPT-3.5 augmented model. Subsequently, both GPT-4 and its augmented version achieved a further increase in success rate, reaching 76%.

Matching Exercises The GPT-3.5 model achieves a 67% success rate, which is surpassed by the RAG-augmented GPT-3.5 at 75%, as shown in Table 4. The GPT-4 model continues this trend of improvement, reaching an 81% success rate, whereas the RAG-augmented GPT-4 achieves the highest accuracy at 87%.

Fill-in-the-Blank Exercises In the fill-in-the-blank exercises, as shown in Table 5, there was a noticeable progression in the model performance. The GPT-3.5 model begins with a success rate of 48%, which is significantly enhanced to 63% with the RAG-augmented GPT-3.5. Following this, the GPT-4 model achieved a success rate of 65%, with the RAG-augmented GPT-4 further improving performance to 72%.

Table 2. True/False exercises assessments results.

Module Topic of MOOC	True/False Exercises			
	GPT-3.5	RAG-augmented GPT-3.5	GPT-4	RAG-augmented GPT-4
General AI concepts	7/8 (87%)	8/8 (100%)	8/8 (100%)	8/8 (100%)
Symbolic AI	3/4 (75%)	3/4 (75%)	4/4 (100%)	4/4 (100%)
Connectionist AI	3/6 (50%)	5/6 (83%)	4/ 6 (67%)	5/6 (83%)
AI applications in education	3/6 (50%)	5/6 (83%)	5/6 (83%)	4/ 6 (67%)
Total	65%	85%	87%	87%

Table 3. MCQ exercises assessments results.

Module Topic of MOOC	MCQ Exercises			
	GPT-3.5	RAG-augmented GPT-3.5	GPT-4	RAG-augmented GPT-4
General AI concepts	4/7 (57%)	5/7 (71%)	5/7 (71%)	5/7 (71%)
Symbolic AI	5/7 (71%)	5/7 (71%)	5/7 (71%)	5/7 (71%)
Connectionist AI	5/8 (62%)	4/8 (50%)	5/8 (62%)	5/8 (62%)
AI applications in education	1/2 (50%)	2/2 (100%)	2/2 (100%)	2/2 (100%)
Total	60%	73%	76%	76%

6 Discussion

Our findings indicate that the RAG-augmented GPT-4 not only exhibited marked proficiency in navigating various exercises from the MOOC but also consistently outperformed the standard GPT-3.5, the RAG-augmented GPT-3.5, and the standard GPT-4 model. This superior performance aligns with our initial hypothesis (H1), substantiating the claim that the RAG-augmented GPT-4 is highly effective in producing accurate responses.

In our analysis, including fill-in-the-blank exercises, we demonstrated the effectiveness of the GPT augmented models. By leveraging retrieval-augmented

Table 4. Matching exercises assessments results.

Module Topic of MOOC	Matching Exercises			
	GPT-3.5	RAG-augmented GPT-3.5	GPT-4	RAG-augmented GPT-4
General AI concepts	2/4 (50%)	2/4 (50%)	3/4 (75%)	4/4 (100%)
Symbolic AI	2/2 (100%)	2/2 (100%)	2/2 (100%)	2/2 (100%)
Connectionist AI	2/4 (50%)	2/4 (50%)	2/4 (50%)	2/4 (50%)
AI applications in education	2/3 67%	3/3 (100%)	3/3 (100%)	3/3 (100%)
Total	67%	75%	81%	87%

Table 5. Fill in the blank exercises assessments results.

Module Topic of MOOC	Fill in the Blank Exercises			
	GPT-3.5	RAG-augmented GPT-3.5	GPT-4	RAG-augmented GPT-4
General AI concepts	9/14 (64%)	8/14 (71%)	11/14 (79%)	10/14 (71%)
Symbolic AI	8/13 (61%)	11/13 (85%)	9/13 (69%)	13/13 (100%)
Connectionist AI	5/13 (38%)	7/13 (54%)	7/13 (54%)	8/13 (62%)
AI applications in education	4/14 (29%)	6/14 (43%)	8/14 (57%)	8/14 (57%)
Total	48%	63%	65%	72%

capabilities, our results are in alignment with the findings of Mao et al. [15], specifically highlighting that RAG significantly enhances the accuracy of open-domain question answering. This underscores the utility of RAG for navigating complex question formats.

Despite the promising outcomes of our study, acknowledging its limitations is crucial. Our research was conducted exclusively on a single MOOC platform and focused on assessments in French, including multiple choice, true/false, matching, and fill-in-the-blank questions. This specialization may limit the generalizability of our findings to other MOOCs, especially to those that utilize a diverse array of assessment types and languages. To mitigate these limitations, future research should include more diverse exercises that involve different MOOCs. Addressing these limitations could provide a more comprehensive and fair comparison.

Our study primarily focused on the capabilities of GPT models, particularly those enhanced by Retriever-Augmented Generation, instead of conducting a broad examination of every generative AI technology, such as Gemini or Copilot. To explore GPT-4's augmented capabilities to produce accurate and contextually appropriate responses, we sought to uncover their transformative impact on education for both educators and students.

For educators, the use of this advanced technology may become a key element in developing effective course content. By adjusting the complexity of the course materials, educators can strike a perfect balance of difficulty, ensuring that each lesson aligns with the diverse learning abilities of their students. This allows for a more engaging and interactive learning experience, in which students are neither overwhelmed by excessive challenges nor bored by tasks that fail to stimulate their intellect.

For students, the GPT-4-augmented model may transform educational experience into immersive and interactive dialogue. Serving as a sophisticated 'learning companion,' as envisaged by Chan and Baskin [16], it can provide instant clarifications and offer detailed explanations tailored to the student's current level of understanding. This personalized interaction not only encourages active learning and critical thinking, but also allows students to explore subjects at their own pace and according to their interests. Moreover, embedding a technology's ability to facilitate contextually rich interactions can further enhance retention and motivation, ultimately improving learning outcomes.

7 Conclusion and Future Work

Our study evaluated the GPT-4-augmented model by leveraging the retrieval-augmented capabilities through 115 assessment exercises. The achievement of a notable success rate of 80% highlights its potential in an educational context.

Looking forward, we aim to deepen our understanding of the GPT-4-augmented model's impact on online learning experiences. To this end, we plan to conduct case studies involving students from different countries. These studies will focus on evaluating various aspects of the learning experience, including student motivation, feelings of isolation, knowledge acquisition, and retention.

The potential effects of integrating the RAG-augmented GPT-4 into education could provide important insight regarding the future possibilities of AI-supported learning. By providing educators with advanced tools to customize the curriculum and offering students an immersive and tailored learning experience, this model has the potential to establish a new standard for educational technology. Therefore, the integration of Retrieval-Augmented Generation into GPT-4 has the potential to not only enhance the teaching and learning methods but also mark the beginning of a new era characterized by interactive and personalized education.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems, 30 (2017)
2. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9 (2019)

3. Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Mann, G.: Bloomberggpt: A large language model for finance. arXiv preprint arXiv:2303.17564 (2023)
4. Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J.,Zaremba, W.: Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374 (2021)
5. Yang, X., Chen, A., PourNejatian, N., Shin, H. C., Smith, K. E., Parisien, C., Wu, Y.: A large language model for electronic health records. NPJ digital medicine, 5(1), 194 (2022)
6. Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Shi, S.: Siren's song in the AI ocean: a survey on hallucination in large language models. arXiv preprint arXiv:2309.01219 (2023)
7. Zhou, C., Neubig, G., Gu, J., Diab, M., Guzman, P., Zettlemoyer, L.,Ghazvininejad, M.: Detecting hallucinated content in conditional neural sequence generation. arXiv preprint arXiv:2011.02593 (2020)
8. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N.,Kiela, D.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33, 9459-9474 (2020).
9. Shi, W., Min, S., Yasunaga, M., Seo, M., James, R., Lewis, M., Zettlemoyer, L., Yih, W.: REPLUG: Retrieval-Augmented Black-Box Language Models. ArXiv, abs/2301.12652. https://doi.org/10.48550/arXiv.2301.12652 (2023)
10. Liu, J., Jin, J., Wang, Z., Cheng, J., Dou, Z., Wen, J.: RETA-LLM: A Retrieval-Augmented Large Language Model Toolkit. ArXiv, abs/2306.05212. https://doi.org/10.48550/arXiv.2306.05212 (2023)
11. Clom-motsia, `https://clom-motsia.teluq.ca/`, last accessed 2024/01/17
12. Neelakantan, A., Xu, T., Puri, R., Radford, A., Han, J. M., Tworek, J., Weng, L.: Text and code embeddings by contrastive pre-training. arXiv preprint arXiv:2201.10005 (2022).
13. Bommarito, J., Bommarito, M., Katz, D. M., Katz, J.: GPT as knowledge worker: a zero-shot evaluation of (AI) CPA capabilities. arXiv preprint arXiv:2301.04408 (2023)
14. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Amodei, D.: Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901 (2020)
15. Mao, Y., He, P., Liu, X., Shen, Y., Gao, J., Han, J., Chen, W.: Generation-Augmented Retrieval for Open-Domain Question Answering. , 4089-4100. https://doi.org/10.18653/v1/2021.acl-long.316 (2020).
16. CHAN, Tak-Wai and BASKIN, Arthur B.: Studying with the prince: The computer as a learning companion. In Proceedings of the International Conference on Intelligent Tutoring Systems(1988).