Comparative Performance of GPT-4, RAG-Augmented GPT-4, and Students 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. Generative Pretrained Transformers (GPT) have significantly improved natural language processing, showcasing enormous versatility across diverse applications. Although GPT models have enormous potential, they frequently encounter issues such as mistakes and hallucinations, which may limit their practical use. Addressing these shortcomings, Retrieval-Augmented Generation (RAG) represents an innovative approach that potentially enhances the accuracy and reliability of these models by leveraging external databases to correct and enrich their outputs. In our study, a RAG-augmented GPT-4 model was tested within an AI-focused Massive Open Online Course (MOOC) and outperformed a standard GPT-4 model, achieving an 85% success rate compared to 81%. Notably, it also surpassed the average student performance, underscoring its ability to deliver precise and contextually relevant responses. These findings suggest the potential of RAG in enhancing AI models for educational use and indicate that instructors can leverage this technology to refine assessment methods and that students can achieve more personalized and engaging learning experiences.

Keywords: Large Language Models · GPT · Text Embeddings· MOOC · Online Education· Evaluation· Retrieval-Augmented Generation

1 Introduction

The introduction of Large Language Models (LLMs), such as the generative pre-trained transformer (GPT), has significantly transformed the field of artificial intelligence, specifically in the domain of natural language processing [1,2,3,4]. These models have shown impressive performance in diverse fields such as finance, technology, and healthcare [5,6,7]. Nevertheless, although large language models possess remarkable capabilities, they are not without limitations. They tend to 'hallucinate', whereby they produce content that may be factually inaccurate [8,9]. This phenomenon has the potential to give rise to information that may contradict established facts.

To overcome this issue, researchers developed the Retrieval Augmented Generation (RAG) technique, which was first proposed by Lewis et al. in 2020 [10]. RAG enhances LLMs by dynamically incorporating external knowledge during the generation process, which enhances the model's capacity to produce precise and pertinent responses. This marks a significant advancement in LLMs, especially for generative tasks [11,12]. Although RAG has demonstrated potential in different fields, its application in educational settings, particularly within Massive Open Online Courses (MOOCs), remains underexplored. This research gap motivated us to investigate how large language models enhanced with RAG could improve content accuracy in educational settings.

In our previous research, we explored how RAG enhances the performance of GPT models across four MOOC modules, finding significant improvements in response accuracy [13]. We continue the work initiated in [13] by conducting a more in-depth analysis, not only of accuracy but also of the relevance of the responses. Additionally, we are introducing a new personalized prompt designed to align the language model more precisely with the pedagogical requirements of learners. This innovation aims to foster richer and more engaging interactions between students and the GPT model, which is used as a learning companion in the online course. Such enhancements are expected to improve the effectiveness of autonomous learning.

Our study is guided by two main research questions. (RQ1): Does integrating RAG into the GPT-4 model improve the pedagogical quality (accuracy and relevance) of answers in MOOCs? Hypothesis 1 (H1): We hypothesize that GPT-4 enriched with RAG will exhibit improved performance over standard GPT-4 with respect to the pedagogical quality of generated responses. The second question (RQ2) is: How does the performance of RAG-augmented GPT-4 compare with that of students in MOOC exercises? We propose two hypotheses: (H2) the RAG-augmented GPT-4 will exceed the average scores of students, and (H3) the RAG-augmented GPT-4 will exceed the highest scores obtained by students.

2 Model

Our architecture enhances user interaction by augmenting GPT model with the Retrieval-Augmented Generation technique, as depicted in Figure 1. The workflow starts with the submission of a query by a user, which is subsequently converted into a vector representation using OpenAI's text-embedding-ada-002 model [14] for retrieval-augmented generation. This embedding model was selected based on its capacity to accurately capture the semantic details of text within contexts that are specific to a particular domain. After vectorizing the query, the system navigates through a database of domain-specific text embeddings to detect segments, referred to as "chunks", that have the highest cosine similarity scores. These selected segments were then combined with the initial user query to enhance the prompt by providing additional context. The prompts are based on White et al. [15]'s description of the Persona Pattern Prompt (see Figure 2). The goal of this strategy is to increase student engagement and facilitate learning in educational contexts. The enriched prompt is then fed into a Large Language Model such as GPT-4 to generate responses that are precise, contextually appropriate, and reflective of domain-specific knowledge. Our model aims to achieve superior response quality by utilizing both retrieval-based methods and the generative capabilities of the GPT-4.

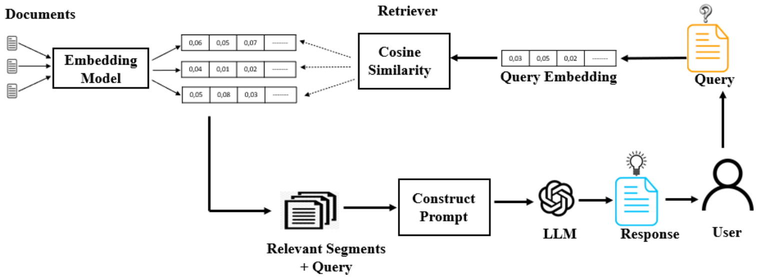
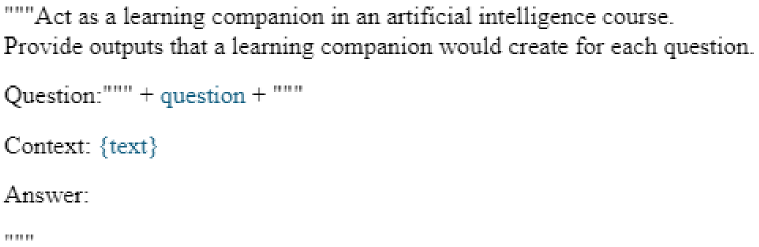


Fig. 1. Overview of the model architecture: From user query processing to response generation, taken from [13].


"""Act as a learning companion in an artificial intelligence course.
Provide outputs that a learning companion would create for each question.

Question:""" + question + """

Context: {text}

Answer:

"""

Fig. 2. Prompt configuration for RAG-augmented GPT-4.

3 Methods

In this section, we first describe the dataset used in our study. Following this, we outline two experiments designed to evaluate the performance enhancements that Retrieval-Augmented Generation technology provides to GPT-4 model, particularly within MOOC environments.

3.1 Dataset

Our research employs the AI MOOC focused on artificial intelligence developed by University TELUQ [16]. This course is segmented into four modules, each focusing on different AI aspects: general AI concepts, symbolic AI, connectionist

AI, and AI applications in education. Our analysis focuses on the first module, which includes 33 assessment exercises for the purposes of this study. There are eight true/false questions, seven multiple-choice questions, four matching exercises, and fourteen fill-in-the-blank exercises in this set of exercises. Table 1 presents examples of these questions to provide a better understanding of the assessment types used in evaluating GPT models.

Table 1. Examples of exercises used for GPT models evaluation.

Question Type	Example
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.

3.2 Experiment Design

Our experimental approach consists of two distinct components, each designed to provide comprehensive insights into the effectiveness of the RAG-augmented GPT-4 model. The first experiment evaluates the efficacy of the RAG-augmented GPT-4 comparing it to the standard GPT-4 model. The second experiment aims to conduct a comparative analysis between the responses of students and those of the RAG-augmented GPT-4, focusing specifically on the same MOOC exercises.

Experiment 1: GPT-4 Models Performance Evaluation (RQ1)

Objective In Experiment 1, the objective is to evaluate the impact of incorporating the retrieval technique on GPT-4 model outputs, which involves integrating external information to enhance response quality. Specifically, we aim to compare the accuracy and the relevance of responses produced by the RAG-enhanced GPT-4 model with those generated by the conventional GPT-4 model.

Procedure

- Data: We utilized the 33 assessment exercises from the first module of the MOOC as queries for both models, as detailed in Table 1.
- Prompt Configuration for RAG-augmented GPT-4: The prompt depicted in Figure 2 is structured as follows:
  - Persona Prompt: ”Act as a learning companion in an artificial intelligence course. Provide outputs that a learning companion would create for each question.”
  - Query and Contextual Integration: Each exercise’s text replaces the ’question’ placeholder, and the ’context’ placeholder is filled with segments that are contextually appropriate. These segments are selected based on high cosine similarity scores from a domain-specific text-embedding database, as detailed in Section 2.
- Standard GPT-4 Model Configuration: The standard model uses a similar persona prompt but without additional contextual data integration.

Evaluation Process The answers from both models were processed sequentially as they appeared in the MOOC, without feedback on incorrect responses. An independent reviewer, a Ph.D. candidate in cognitive computing, scored the answers using a binary scale (0-1) based on two criteria: accuracy and relevance.

For accuracy, a score of ”1” was assigned to answers that were factually correct, while a score of ”0” was given to those that were factually incorrect or contained fabrications, referred to as hallucinations.

Relevance was assessed based on how well an answer addressed the specific question posed; accurate answers received a score of ”0” if they did not directly relate to the question.

This binary scoring system simplified the assessment process, enabling a direct comparison of the two models’ performance on the same set of exercises, consistent with the MOOC’s established grading standards.

Experiment 2: Comparison of RAG-augmented GPT-4 and Student Performance (RQ2)

Objective This experiment aims to assess the effectiveness of the GPT-4 model augmented with Retrieval-Augmented Generation by comparing it to the performance of students on the same set of 33 MOOC exercises.

Participants The sample for this study consisted of twenty students, with a gender distribution of seven males and thirteen females. The participants’ ages ranged from 22 to 35 years. The study included participants with diverse educational backgrounds, with 60% pursuing computer science, 25% pursuing education, and 15% pursuing engineering. All participants voluntarily participated in this study.

Procedure The procedure involved conducting 33 exercises under controlled, exam-like conditions without any external assistance for both the students and the GPT-4 model augmented with RAG. This approach guarantees that the evaluation closely matches real testing situations, allowing for an equitable comparison of performance between human learners and the AI model.

Evaluation Process The exercises were part of the course evaluations for the students, who completed them in a conventional academic environment. In parallel, the RAG-augmented GPT-4 produced responses under identical conditions to ensure consistency. We then conducted a direct comparison between the students' responses and those generated by the augmented model, assessing both for accuracy. This comparative approach facilitated a fair evaluation and provided valuable insights into the practical efficacy of the RAG-enhanced GPT-4 model in an educational setting.

4 Results

In this section, we present the results of using Retrieval-Augmented Generation with the GPT-4 model in MOOC environments. We organize the results around the main research questions.

4.1 Research Question 1 (RQ1): Does integrating RAG into the GPT-4 model improve the pedagogical quality (accuracy and relevance) of answers in MOOCs?

In addressing RQ1, we evaluated the performance of the standard GPT-4 model and its enhancement with Retrieval-Augmented Generation technology using a dataset of 33 exercises from an AI MOOC. The results indicate a slight improvement in both accuracy and relevance when the GPT-4 model is augmented with RAG. The standard model achieved a success rate of 81% in both accuracy and relevance, while the augmented model showed an improvement to 85% in accuracy and 84% in relevance. Table 2 summarizes the comparative results of the exercise assessments in module 1 of the MOOC.

Table 2. Exercise assessment results on module 1.

Exercise Type	GPT-4		RAG-augmented GPT-4	
	Accuracy	Relevance	Accuracy	Relevance
True False	8/8 (100%)	8/8 (100%)	8/8 (100%)	8/8 (100%)
Multiple-Choice	5/7 (71%)	5/7 (71%)	5/7 (71%)	5/7 (71%)
Matching	3/4 (75%)	3/4 (75%)	4/4 (100%)	4/4 (100%)
Fill in the blank	11/14 (79%)	11/14 (79%)	10/14 (71%)	9/14 (64%)
Total	81%	81%	85%	84%

4.2 Research Question 2 (RQ2): How Does the Performance of RAG-augmented GPT-4 Compare to that of Students in MOOC Exercises?

Concerning RQ2, we evaluated the performance of the RAG-augmented GPT-4 across 33 different types of exercises within an MOOC and compared its results with those of an average student. The model demonstrated superior performance, surpassing the average for all exercise types. It achieved an overall success rate of 85% compared with a student average of 60%, as indicated in Figure 3. Notably, the model scored 100% in the True/False and matching exercises, whereas its performance in the Fill in the blank tasks was 71%, which is still well above the student average of 35%.

As detailed in Table 3, the performance analysis of students reveals diverse outcomes across various exercise types. Both the median and mode in True False exercises achieve a perfect score of 100%. In multiple-choice question (MCQ) exercises, the median success rate decreases to 57%, while the most frequent score (mode) is even lower at 43%. The matching exercises demonstrate a median success rate of 75% and a mode of 100%. Fill in the blank exercises are particularly challenging, as indicated by a median and mode score of 43%.

Table 3. Median and mode for each exercise type.

Exercise Type	Median Success Rate (%)	Mode Success Rate (%)
True False	100	100
MCQ	57	43
Matching	75	100
Fill in the blank	43	43

Further analysis of the RAG-augmented GPT-4's performance indicates its superior capabilities compared to students. For instance, the model's 52th percentile ranking in the True/False and Multiple Choice exercises suggests competitive performance, with the model either exceeding or matching the scores of more than half of the students. Regrading the matching exercises, the RAG-augmented GPT-4's percentile ranking of approximately 67th demonstrates its superior performance, surpassing the scores of the majority of students. Moreover, the model's 95th percentile ranking in the Fill in the Blank exercises highlights its exceptional performance, placing it within the top 5% of student scores. This indicates that the RAG-augmented GPT-4 outperformed 95% of students in Fill in the Blank exercises.

5 Discussion

The findings obtained from experiment 1 demonstrate that the GPT-4 model when enhanced with Retrieval-Augmented Generation, attains an average accuracy

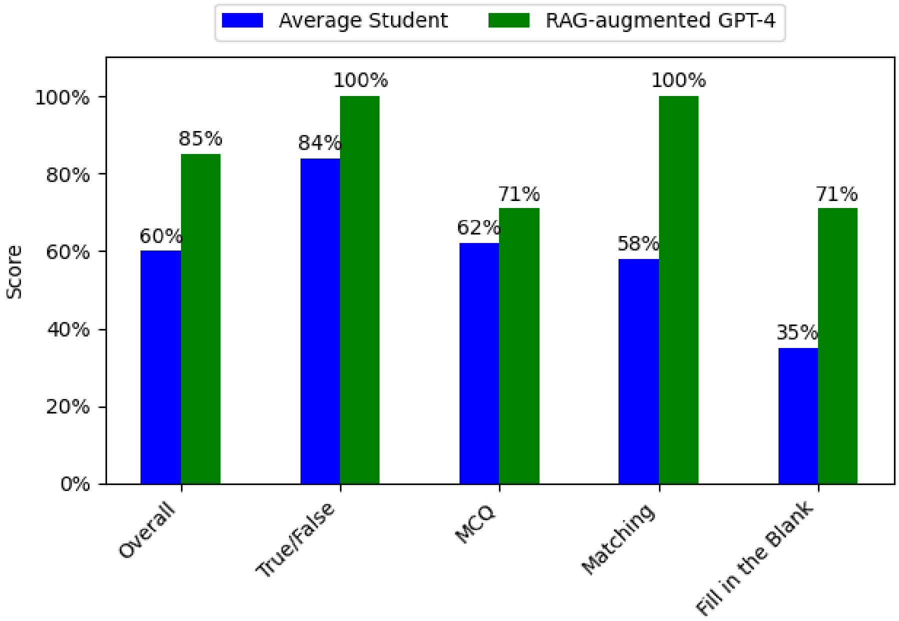


Fig. 3. Average student (blue) and RAG-augmented GPT-4 (green) percentage scores across different types of exercises.

rate of 85% and an average relevance rate of 84%. In contrast, the standard GPT-4 model recorded a success rate of 81% for both accuracy and relevance. The findings of this study provide evidence in support of Hypothesis 1 (H1), which claimed that incorporating retrieval capabilities would enhance the performance of the GPT-4 model in generating responses within MOOC settings.

The results of experiment 2 not only confirm that the RAG-augmented GPT-4 performed better than the average student, thus validating Hypothesis 2 (H2), but they provide partial support for Hypothesis 3 (H3). This evidence indicates that the model performs exceptionally well in different types of exercises, especially in fill in the blank exercises, where it consistently achieved scores within the top 5% of student performance. These tasks, which are characterized by the need to both remember factual information and apply it in a specific context, often pose a challenge for students, as indicated by their relatively low median and mode scores. Nevertheless, the RAG-augmented GPT-4 shows exceptional proficiency in these exercises. Its advanced understanding and data processing abilities are specifically designed to perform effectively in tasks that frequently present significant challenges for human learners.

In our study, we focused specifically on the capabilities of GPT models enhanced with Retrieval-Augmented Generation. This technique enables the GPT models to access external databases or documents, retrieving information that is then incorporated into their responses. This process enhances the accuracy and relevance of the content generated. The results of our study demonstrate the potential of GPT-4 augmented with RAG to improve educational experiences, providing advantages for students as well as instructors.

For students, the Persona Prompt, designed to act as a 'learning companion', could significantly enhance student engagement and improve learning outcomes. This role enables the technology to guide learners effectively by providing constructive feedback, posing thought-provoking questions, and encouraging the exploration of new perspectives. Consequently, the integration of RAG-augmented GPT-4, which utilizes the Persona Prompt, has the potential to further enrich educational experiences through the creation of immersive and dynamic dialogues. It serves as a sophisticated 'learning companion', as envisioned by Chan and Baskin [17], offering students instant clarifications and in-depth explanations that are customized to their levels of understanding.

However, it is crucial to recognize that the RAG-augmented GPT-4 is not a substitute for human instructors. It lacks the deep understanding that human instructors have of their learners' backgrounds, cultures, and individual learning needs. While this technology can deliver accurate answers and facilitate learning, it cannot replace the nuanced understanding and adaptability of an instructor.

Instructors should adopt a more sustainable assessment approach that prioritizes the cultivation of higher-order cognitive abilities, including critical thinking [18], problem-solving, and creativity [19]. These skills are less susceptible to replication by GPT models. It is recommended that instructors prioritize the improvement of learning and skill acquisition over solely preparing students for traditional evaluations[19]. This can be achieved by integrating more complex assessments such as problem-solving activities and collaborative learning that necessitate active demonstration of comprehension. These methods enhance students' understanding and utilization of knowledge, cultivating a more comprehensive educational approach.

In Retrieval-Augmented Generation systems, the effectiveness of the retrieval component is crucial, as it sources relevant information for content generation. Enhancing this component, particularly through the Hybrid Retrieve approach, has significantly improved RAG systems [20]. Hybrid Retrieve combines dense and sparse retrieval methods, capitalizing on dense retrieval's ability to capture semantic similarities and sparse retrieval's efficiency in keyword matching. This combination results in a more effective retrieval process, enhancing the quality and relevance of the information provided to the generator [21,22]. Consequently, RAG systems based on Hybrid Retrieve could become more capable of producing accurate and contextually relevant outputs, paving the way for advancements in educational technologies and beyond.

Although our study yielded encouraging results, it is crucial that we recognize and address its limitations in order to attain a thorough comprehension. The relatively small sample size (N=20) may affect the statistical robustness and generalizability of our findings. Additionally, our research was conducted on a single MOOC platform and exclusively assessed French language exercises, including multiple choice, true/false, matching, and fill-in-the-blank questions. This specialization may limit the generalizability of our findings to other educational contexts that utilize a diverse array of assessment types and languages.

To overcome these limitations, future research should incorporate a larger sample size and a broader selection of MOOCs featuring varied exercise formats and languages. Addressing these limitations will enhance our understanding of the effectiveness of GPT-4 augmented with RAG across different educational settings.

6 Conclusion and Future Work

Our study has demonstrated the enhanced capabilities of the GPT-4-augmented model in handling diverse assessment exercises within a MOOC setting. By integrating Retrieval-Augmented Generation with GPT-4, we have not only enhanced the accuracy and relevance of responses but also demonstrated clear superiority over the average performance of students across various types of exercises.

Looking ahead, we plan to extend our research by conducting two detailed case studies involving students from diverse geographical backgrounds. These studies aim to analyze the broader impact of our RAG-augmented GPT-4 model on the online learning experience. We will evaluate not only academic performance but also factors such as students' motivation, engagement, and feelings of isolation. The ability of the RAG-augmented GPT-4 to serve as a learning companion through the Persona Prompt and provide contextually relevant information can enhance the learning experience, making it more engaging and motivating. Furthermore, by delivering timely and relevant responses, the augmented model can help reduce feelings of isolation, particularly in online learning environments. This comprehensive analysis will offer deeper insights into how such advanced AI tools can be effectively integrated into various educational settings, ultimately improving learning experiences.

The integration of RAG-augmented GPT-4 into educational environments promises transformative changes. This innovative technology can revolutionize traditional teaching methods by equipping instructors with advanced tools to enhance student engagement and enrich the learning experience. The RAG-augmented GPT-4 enables customized learning experiences by adjusting educational content to meet the specific needs of individual students. This ensures that learning is not standardized but personalized, addressing each individual's needs. Additionally, the model promotes a more engaging and interactive learning environment through its natural language processing capabilities. As a sophisticated learning companion, the RAG-augmented GPT-4 can significantly enhance student engagement, motivation, and learning outcomes.

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. Devlin, J., Chang, M. W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
4. 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)
5. 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)
6. 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)
7. 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)
8. 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)
9. 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)
10. 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).
11. 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)
12. 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)
13. Miladi, F., Psyché, V., Lemire, D.: Evaluating Generative Pre-trained Transformers in MOOC Assessments: A Comparative Study of GPT Models. In International Conference on Artificial Intelligence in Education (2024)
14. 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)
15. White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Schmidt, D. C.: A prompt pattern catalog to enhance prompt engineering with chatgpt. arXiv preprint arXiv:2302.11382 (2023)
16. Clom-motsia, `https://clom-motsia.teluq.ca/`, last accessed 2024/03/17
17. CHAN, Tak-Wai et BASKIN, Arthur B.: Studying with the prince: The computer as a learning companion. In Proceedings of the International Conference on Intelligent Tutoring Systems (1988).
18. Smolansky, A., Cram, A., Raduescu, C., Zeivots, S., Huber, E., Kizilcec, R. F.: Educator and student perspectives on the impact of generative AI on assessments in higher education. In Proceedings of the tenth ACM conference on Learning@ Scale (pp. 378-382) (2023)

19. Savelka, J., Agarwal, A., An, M., Bogart, C., Sakr, M.: Thrilled by your progress! Large language models (GPT-4) no longer struggle to pass assessments in higher education programming courses. In Proceedings of the 2023 ACM Conference on International Computing Education Research-Volume 1 (pp. 78-92) (2023)
20. Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., Cui, B.: Retrieval-Augmented Generation for AI-Generated Content: A Survey. arXiv preprint arXiv:2402.19473 (2024)
21. Wang, W., Wang, Y., Joty, S., Hoi, S. C.: Rap-gen: Retrieval-augmented patch generation with codet5 for automatic program repair. In Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (pp. 146-158)(2023)
22. Lu, S., Duan, N., Han, H., Guo, D., Hwang, S. W., Svyatkovskiy, A.: Reacc: A retrieval-augmented code completion framework. arXiv preprint arXiv:2203.07722 (2022)