Enhancing classroom teaching with LLMs and RAG

Elizabeth Mullins
Americas High School
United States
emulli@sisd.net

Adrian Portillo
Northwest Early College High School
United States
adrianp77@gmail.com

Kristalys Ruiz Rohena
University of Texas at El Paso
United States
kruizrohena@miners.utep.edu

Aritran Piplai
University of Texas at El Paso
United States
apiplai@utep.edu

Abstract

Large Language Models have become a valuable source of information for our daily inquiries. However, after training, its data source quickly becomes out-of-date, making RAG a useful tool for providing even more recent or pertinent data. In this work, we investigate how RAG pipelines, with the course materials serving as a data source, might help students in K–12 education. The initial research utilizes Reddit as a data source for up-to-date cybersecurity information. Chunk size is evaluated to determine the optimal amount of context needed to generate accurate answers. After running the experiment for different chunk sizes, answer correctness was evaluated using RAGAs with average answer correctness not exceeding 50 percent for any chunk size. This suggests that Reddit is not a good source to mine for data for questions about cybersecurity threats. The methodology was successful in evaluating the data source, which has implications for its use to evaluate educational resources for effectiveness.

CCS Concepts

• **Applied computing** → **Education**; • **Information systems** → *Information retrieval.*

Keywords

Large Language Models, Retrieval Augmented Generation, Education

ACM Reference Format:
Elizabeth Mullins, Adrian Portillo, Kristalys Ruiz Rohena, and Aritran Piplai. 2024. Enhancing classroom teaching with LLMs and RAG. In *The 25th Annual Conference on Information Technology Education (SIGITE '24), October 10–12, 2024, El Paso, TX, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3686852.3687083

1 Background and Motivation

The recent significant advances in the efficacy of large language models (LLMs) have led to a boom in innovative approaches to using LLMs in educational settings and research [1]. Some uses are elaborate such as the creation of a simulated classroom powered by LLMs acting as teacher, assistant, and student agents managing to demonstrate interaction patterns like real classrooms between agents and other agents as well as between agents and students [8]. Others examined teacher-led interventions that enhance student interactions with LLMs, aiming to optimize the educational benefits of these tools [5].

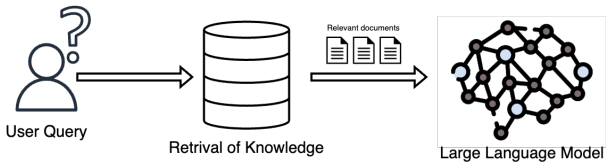

Figure 1: RAG Pipeline

With specialized training, LLMs have shown potential as personalized tutors able to break down problems and prompt students to come up with the answers as well as adjust to the needs of individual learners [2, 7]. LLMs remain difficult to train with one challenge being ensuring they have updated information to give accurate answers [2].

The merging of pre-trained parametric memory (state-of-the-art LLMs) and non-parametric memory, such as a vector database, in retrieval-augmented generation (RAG) models provides a method to give LLMs access to updated information without undergoing new training. These models deliver more specific and factual responses [6]. This methodology enhances the potential for LLMs to act as AI tutors [3] when connected to the appropriate sources.

2 Evaluating the utility of LLMs and RAG

2.1 Summary of current research

The research's purpose was to evaluate RAG as a method to use LLMs to provide up-to-date answers to cyber threat questions. In addition, we evaluated the effect of chunk size on the answer correctness of the LLM. We used Llama as our LLM because of its availability. The top 500 posts and the most upvoted answer were scraped from subreddits pertaining to cybersecurity and network security then stored in a Chroma vector database to serve as the data source for the RAG. 30 questions and ground truths were developed from information from CISA about the most recent threats.

SIGITE '24, October 10–12, 2024, El Paso, TX, USA

ACM ISBN 979-8-4007-1106-0/24/10
https://doi.org/10.1145/3686852.3687083

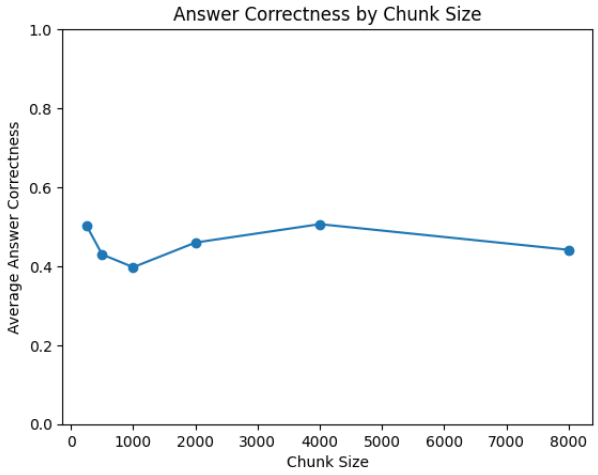

Figure 2: Average answers correctness by chunk sizes.

After feeding the questions through the RAG pipeline, the questions, ground truths, and Llama answers were then fed through RAGAS [4] to get an answer correctness score. The process was repeated for chunk sizes of 250, 500, 1000, 2000, 4000, and 8000 characters. The average answer correctness score for each chunk size was close to 50 percent with chunk sizes 250 and 4000 yielding the highest accuracy.

These results suggest that Reddit is not a good source for a RAG database intended to assist with cybersecurity. Reddit's format introduces noise such as a variety of postings in a wide range of cybersecurity topics not specific to protecting against current vulnerabilities. In addition, the top posts on Reddit are constantly changing. The first response to a newly posted question can be generated by a bot directing the user to sources of relevant information. Depending on when the data is mined, this could appear multiple times in the context available to the LLM.

There was little variability in answer correctness by chunk size. Potential effects are thought to be better answers with smaller chunks due to smaller chunks containing more specific context or better answers with larger chunks because larger chunks contain more context. For these specific experiments, there were no generalizable effects. The question bears further examination. With a better aligned data source, chunk size could have a greater impact. Since Reddit did not provide appropriate data for the questions being asked, chunk size would not improve the context provided to the LLM.

2.2 Future Directions

This methodology has infinite variations, which is one way to extend it into the classroom. To continue to examine the question of cybersecurity, teachers of cybersecurity content can guide students through using this methodology with alternate sources to mine data for the RAG database. A different database could also lead to noticeable effects on answer correctness with different chunk sizes. This methodology can also be used by students to examine sources for their interests. What sites do they frequent? Can they use this method to mine their favorite sites to provide answers to their questions? To mine data from social media sites to take the trendiness on different topics? This is a methodology that supports student-led/centered research into any topic, making it adaptable to any content and therefore multidisciplinary.

In addition, this model can be used to develop a personalized tutor as demonstrated in the background. This model allows greater personalization based on student or teacher needs. Other models are designed to be generalized for universal purposes. This model allows for the variations that occur between teachers and students. A teacher who chooses to utilize AI as a teaching assistant, could build the database with their teaching materials, and make the model accessible to their students to query when they need help on a topic. This could be as broad as content for the year or as specific as information needed to complete a project. Changing the data file to create the vector database changes the purpose of the model. For students that do not have a technologically progressive teacher, they can use this model in the same way to create their own tutor.

The model can also be used by teachers to vet the appropriateness of teaching materials or the effectiveness of their assessment questions. Teachers can develop the vector database using their classroom resources such as textbooks and lecture notes or presentations. They can use their assessment questions and answers as the ground truth, then, feed their assessment questions through the RAG pipeline collecting the LLM's answers to the questions. Using RAGAs, teachers can analyze answer correctness of the LLM. If the answer correctness scores are high, this would suggest that the educational materials are appropriate for the learning goals. If the scores are low, this would suggest the materials should be realigned to the essential knowledge and skills.

This work is supported by National Science Foundation Award 2206982

References

[1] Ikpe Justice Akpan, Yawo M. Kobara, Josiah Owolabi, Asuama Akpam, and Onyebuchi Felix Offodile. 2024. An investigation into the scientific landscape of the conversational and generative artificial intelligence, and human-chatbot interaction in education and research. arXiv:2407.12004 [cs.CY] https://arxiv.org/abs/2407.12004

[2] Yulin Chen, Ning Ding, Hai-Tao Zheng, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Empowering Private Tutoring by Chaining Large Language Models. arXiv:2309.08112 [cs.HC] https://arxiv.org/abs/2309.08112

[3] Chenxi Dong. 2024. How to Build an AI Tutor that Can Adapt to Any Course and Provide Accurate Answers Using Large Language Model and Retrieval-Augmented Generation. arXiv:2311.17696 [cs.CL] https://arxiv.org/abs/2311.17696

[4] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated Evaluation of Retrieval Augmented Generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. ACL, 150–158.

[5] Harsh Kumar, Ilya Musabirov, Mohi Reza, Jiakai Shi, Xinyuan Wang, Joseph Jay Williams, Anastasia Kuzminykh, and Michael Liut. 2024. Impact of Guidance and Interaction Strategies for LLM Use on Learner Performance and Perception. arXiv:2310.13712 [cs.HC] https://arxiv.org/abs/2310.13712

[6] 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 *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 9459–9474. https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

[7] Shashank Sonkar, Naiming Liu, Debshila Mallick, and Richard Baraniuk. 2023. CLASS: A Design Framework for Building Intelligent Tutoring Systems Based on Learning Science principles. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 1941–1961. https://doi.org/10.18653/v1/2023.findings-emnlp.130

[8] Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhiyuan Liu, Lei Hou, and Juanzi Li. 2024. Simulating Classroom Education with LLM-Empowered Agents. arXiv:2406.19226 [cs.CL] https://arxiv.org/abs/2406.19226