BOARD # 92: WIP: Generative AI-based Learning Tutor for Biomedical Data Science (GAIL Tutor BDS)

Katie Vu, University of Michigan

Katie Vu, an University of Michigan computer science undergraduate Class of 2026, is interested in natural language processing research with an interdisciplinary focus. She has a particular interest in LLMs.

Mr. Avery Mitchell Maddox, University of Michigan

Avery Maddox is an MD/PhD student at the University of Michigan with a great interest in educating future bioinformaticians and medical technology innovators. As a member of Dr. Arvind Rao's Systems Imaging and Bioinformatics Lab, he has been the co-instructor of an innovative project-based course, Diagnostic Intelligent Assistance for Global Health, that exists as part of the University of Michigan's Multidisciplinary Design Program.

Caleb William Tonon, University of Michigan
Guli Zhu, University of Michigan

Guli Zhu is a graduate student in the Health Data Science program at the University of Michigan. His research interests include machine learning, large language models, and multimodal learning, particularly in the context of healthcare applications.

Tyler Wang, Stony Brook University
Rafael Mendes Opperman, University of Michigan

Rafael Opperman is a second-year undergraduate pursuing a B.S.E. in Industrial & Operations Engineering at the University of Michigan. Through the Multidisciplinary Design Program he has developed a research interest in operations, optimization, generative AI, and logistics.

Qiuyi Ding, University of Michigan

Undergraduate student

Zifei Bai, University of Michigan
zhanhao liu, University of Michigan
Ziyi Wang, University of Michigan
Arvind Rao, University of Michigan
Daniel Yoon, University of Michigan

WIP: Generative AI-based Learning Tutor for Biomedical Data Science (GAIL Tutor BDS)

Introduction

The recent and rapid development of generative large language models (LLMs) is driving the production of tools that promise to change the way people write, think, and learn. Such tools, trained on dozens to hundreds of terabytes of data and built around multi-billion parameter transformer-based architectures, are capable of interpreting and generating language at an unprecedented level of similarity to humans. Available in a wide array of readily accessible interfaces, from mobile applications to web sites, and now fully integrated into the operating systems of the latest smart phones [1], LLMs are being increasingly adopted into more-and-more aspects of human life.

One of the most active areas of LLM adoption has been in education. Like a personal tutor, LLMs can provide direct answers to complex questions, a distinct jump in ease-of-use from course material and search engines where the user must search through resources related to their query. Also like a personal tutor, their utility is dependent upon their ability to draw from their knowledge base to give accurate responses. OpenAI promotes the breadth of knowledge in their latest model, GPT-4 [2], underpinning their general-purpose chat application ChatGPT [3], which is capable of scoring a 5, the best score possible, on advanced placement exams for art history, biology, macroeconomics and more [2]. This high level of performance has led to a massive increase in use by students across academic disciplines, with mixed acceptance at the university and department level. How to ensure that students gain experience with these tools, which are likely to be essential within their future professions, while protecting their development in other core concepts and ensuring academic integrity remains an open question that is actively being investigated [4], [5], [6], [7].

Despite their promising potential as educational tools, there exist several important limitations in mainstream general purpose LLM-based chat applications like ChatGPT, Claude [8], and Google Gemini [9]. In highly advanced, domain specific applications, the body of data upon which the models are trained is often insufficient to provide a useful or even accurate response. This is partly due to the massive volume of training data and the computational demands of model training and tuning, leading to prolonged delays between updates, which prevents the utilization of the most up-to-date information. Responsible users also want to ensure that the responses they receive are based on verifiable sources of information, however, most applications lack the ability to directly cite the sources of their response [10], [11]. Concerningly, LLMs also have a propensity to give nonsensical responses based upon faults in their internal reasoning, so-called model hallucinations, which can be difficult for users to detect, especially learners in the absence of cited sources in the application's response [10], [11]. Finally, the diverse corpus of information upon which popular general purpose LLM-based applications are trained inhibits their ability to provide the most contextually appropriate and informative answers when the correct response is highly setting specific [10], [11].

Retrieval-augmented generation (RAG) is a technique that can be used with LLMs to address these issues by connecting the LLM to a database of embedded resources. This design,

the RAG-LLM, enables the LLM to draw upon a specially designed database for context in its response generation. These databases can be customized to strengthen LLM responses in niche knowledge domains, and give more appropriate responses based upon the user's specific setting. The additional context reduces the LLM propensity for hallucination and provides the LLM with resources it can cite directly in its response [10], [11]

For the last four years, our lab has been running a project-based experiential learning course, named Diagnostic Intelligent Assistance for Global Health (DIAG), as part of the University of Michigan's Multidisciplinary Design Program [12]. The mission of the course is to educate undergraduate and first-year master's students in biomedical engineering, computational medicine and bioinformatics through multi-year longitudinal real-world projects focused on addressing global healthcare issues in resource limited settings. DIAG's projects incorporate knowledge across fields such as the biomedical sciences, computer science, bioinformatics, and AI. As part of DIAG's inclusionary and interdisciplinary approach, students from widely varying backgrounds in these fields are encouraged to join. While this diversity comes with countless benefits, it makes it challenging to provide each student with the specific learning resources and support they need to efficiently and confidently get started and progress on new projects.

Here we propose the DIAG student-lead development and assessment of custom LLM-based applications to assist students from diverse educational backgrounds in confidently and efficiently getting started and progressing on team projects. This provides students with the opportunity to learn about LLMs and their limitations through the process of building them rather than simply using them, while ultimately producing a tool they can use to advance their team's work on other projects.

As the team produces working applications, all code and data will be shared freely and in such a way that enables others to re-implement our applications and validate our assessments. If resources are available, we intend to host the application online to enhance its usability by interested groups.

Methods

As a structured project-based learning experience for the DIAG course, 6 students were assigned to a sub-team for this project and worked on it over the course of a semester. Students started with the goal of designing and building a useful chat application to help with onboarding to team projects. Students met once per week for meetings with the full DIAG team and at least one additional time per week with their sub-team. At each meeting, students (1) presented their work and results (2) analyzed and discussed results with input from the course instructor and other team members, (3) were assigned study objectives, and (4) planned the next steps for the project. Students were evaluated using a hybrid contract approach [13] where they receive an automatic grade of "A-" if they actively participated in their team and in meetings and met their course-credit based time commitments. The motivation for this approach was to incentivize students to actively engage in the project and creative problem solving without feeling limited in what they had to produce in order to receive a good grade.

Chat applications were built around the LLAMA3-8B model [14], [15], an 8 billion parameter LLM utilizing a transformer architecture released on April 18th, 2024. The LLAMA3-8B model was loaded pre-trained and optimized for helpfulness and safety through supervised fine-tuning and reinforcement learning with human feedback. LLAMA3-8B was chosen for its high benchmark performance, support with other programming tools, open-source design, and intended use for assistant-like chat applications [14], [15].

The database for use with the RAG-LLM was creating using the latest release of articles from the Biorxiv pathology collection [16]. Pathology was chosen as the initial area of focus because it is the area where most DIAG students have the least experience, while also being central to most DIAG projects. The licensing of all articles was programmatically inspected and only articles with licenses permitting reuse and remixing of material for non-commercial purposes without author permission were used [17]. Due to resource constraints, only the first 100 articles were used for data base construction. The selected articles were then embedded and stored using the vector stores suite of tools from LangChain [18] and Chroma [19]. For embedding, article text was split and embedded every 500 characters. As a result of resource constraints, the RAG technique used only the top 5 matching embeddings to provide context for response generation. The resultant RAG-LLM chat application design is shown in Figure 1a.

Chat applications with and without the addition of RAG were assessed using a set of 50 pathology related questions. Questions were initially generated through prompts to ChatGPT requesting pathology related questions. Questions were then refined by faculty and students based on what they found most interesting and useful for their work in the DIAG course. Both applications were graded on a scale of 0 to 1 for the faithfulness and relevance of their answers. The RAG-LLM application was also graded on a scale of 0 to 1 on the relevance of retrieved context. Grading of responses was performed by a single medical student DIAG faculty member in a blinded fashion regarding which model generated the responses.

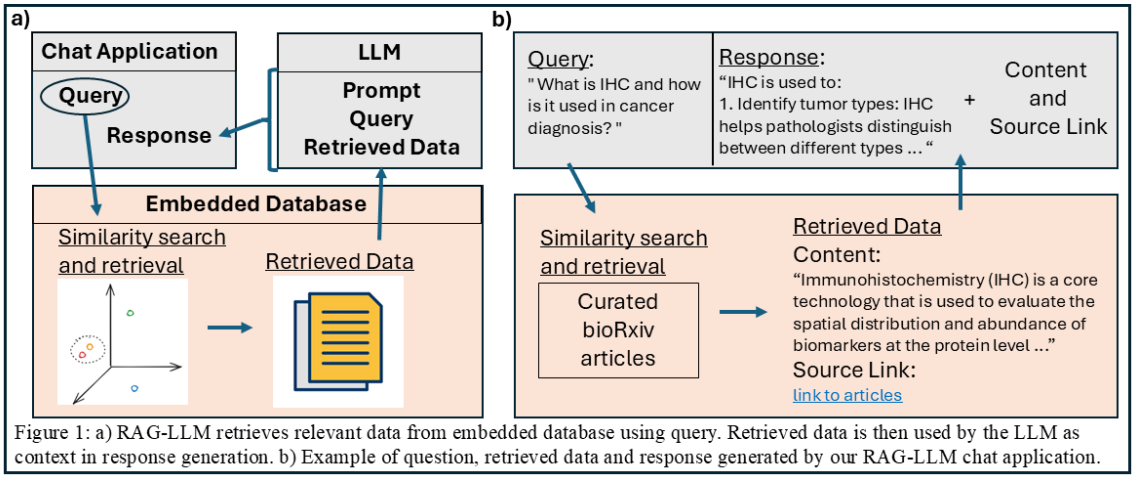

Figure 1: a) RAG-LLM retrieves relevant data from embedded database using query. Retrieved data is then used by the LLM as context in response generation. b) Example of question, retrieved data and response generated by our RAG-LLM chat application.

Results

Both applications achieved a perfect score in answer relevance, where every response was completely relevant to the question asked. This highlights how LLAMA3-8B, and likely other modern LLMs, are sufficiently advanced to understand the context of the questions being asked and respond accordingly with or without RAG assistance. Likewise, answer faithfulness was similar between the two approaches. The LLM-based application and RAG-LLM-based application achieved a mean answer faithfulness score of 0.94 (sd. 0.086) and 0.954 (sd. 0.073) respectively with a t-test p-value of 0.383 when comparing the two sets of responses. Disappointingly, the RAG-LLM-application performed poorly on context relevance, with a mean score of 0.352 (sd. 0.247), representing partially relevant but uninformative context retrieval. Of note, the context for responses of the RAG-LLM application was only directly related to the source article in 3 responses, with an abbreviated example shown in Figure 1b. Where the context was relevant, it was typically drawn from the article abstracts and introductions.

Discussion

Our initial assessment of answer faithfulness and relevance for the LLM and RAG-LLM-based chat applications highlights the promising capabilities of these tools to provide useful and informative responses for learners. At the same time, the poor performance of the RAG-LLM on its context relevance scores highlights the challenges in creating a database that adds to existing LLM capabilities. While keeping resource usage minimal during this initial phase of application development, the limitations to the size of the database and context retrieval likely contributed to the issue of uninformative additions by the RAG system to the application. In addition, the poor relation of the context returned by the RAG technique to its source material highlights that the information contained in the databases' research articles is likely too limited in scope for the questions asked. Finally, using only appropriately licensed sources for the database was also limiting and prevented the use of the major textbooks containing well-established information of significant depth in the subject matter.

These issues are being addressed in ongoing work on this project. Now that the initial applications have been developed, computational resources have been allocated to run the RAG-LLM application with the full database and optimize the context retrieval. Also, additional programmatic and manual searches are being performed to add a wider range of open-source materials to the database including reviews, surveys, and online textbooks and courses. In addition, a database using university materials is being developed with appropriate limitations on use with regards to the materials' respective copyrights and licensing.

In addition, the value of using an LLM fine-tuned for medical knowledge tasks will be investigated. Our team has recently completed LLM and RAG-LLM applications built around the OpenBioLLM-Llama3 model, a specialized variant of the Llama3 model fine-tuned for medical knowledge domain coverage with state-of-the-art scores across a wide range of medical benchmark challenges [20], [21].

After the chat applications reach acceptable levels of performance on their standardized question-based assessment, they will be evaluated as teaching tools for the DIAG course. New students and students starting on new projects will be assigned to use one of the applications to help with their onboarding and project work. We will assess the amount of time needed before starting a project; their attitude towards the onboarding process; and other self-reported measures including their confidence and competency before and after extended use of the application.

Finally, ongoing work with improving the application provides several exciting projects for DIAG students to work on. We intend to follow the general approach developed here to develop applications for assisting with other subjects including data science, AI, and programming. As these specialized systems are developed, we plan to integrate them into a single application of interacting expert agents using LangGraph. We are also investigating application designs that will allow for the model to continuously learn and improve through user feedback. Finally, creating an accessible, effective and enjoyable user interface is another important area of ongoing development. It should be easy for students to make their queries, adjust application settings, add new materials to the database, and provide feedback to the model.

Conclusion

LLM-based applications are an important and rapidly developing technology that are likely to play an important role in every field of work, including education. Here we presented our progress on a project for DIAG students to learn about LLMs and help future students through building LLM-based chat applications. In the initial phase of this project, students were able to successfully build and implement an LLM and RAG-LLM chat application. At each step along the way, students learned the key concepts and essential tools related to LLM design and use while exercising creative reasoning and problem solving. Over the course of the project, students became adept in a wide range of new skills from web-scraping, to text embedding and storage, and finally to building and running RAG-LLM chat applications. Initial assessment of the applications highlights their potential as a teaching tool, while also shedding light on their limitations and potential pitfalls in their design and and implementation. Further work will address these shortcomings to build better applications for the team and assess them as teaching tools for new DIAG students.

References

[1] "Apple Intelligence," Apple. Accessed: Jan. 12, 2025. [Online]. Available: https://www.apple.com/apple-intelligence/

[2] OpenAI *et al.*, "GPT-4 Technical Report," Mar. 04, 2024, *arXiv*: arXiv:2303.08774. doi: 10.48550/arXiv.2303.08774.

[3] "Introducing ChatGPT." Accessed: Jan. 15, 2025. [Online]. Available: https://openai.com/index/chatgpt/

[4] A. Abd-alrazaq *et al.*, "Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions," *JMIR Med. Educ.*, vol. 9, p. e48291, Jun. 2023, doi: 10.2196/48291.

[5] M. Bernabei, S. Colabianchi, A. Falegnami, and F. Costantino, "Students' use of large language models in engineering education: A case study on technology acceptance, perceptions, efficacy, and detection chances," *Comput. Educ. Artif. Intell.*, vol. 5, p. 100172, Jan. 2023, doi: 10.1016/j.caeai.2023.100172.

[6] E. Kasneci *et al.*, "ChatGPT for good? On opportunities and challenges of large language models for education," *Learn. Individ. Differ.*, vol. 103, p. 102274, Apr. 2023, doi: 10.1016/j.lindif.2023.102274.

[7] J. G. Meyer *et al.*, "ChatGPT and large language models in academia: opportunities and challenges," *BioData Min.*, vol. 16, no. 1, p. 20, Jul. 2023, doi: 10.1186/s13040-023-00339-9.

[8] "Claude." Accessed: Jan. 15, 2025. [Online]. Available: https://claude.ai

[9] "Gemini - chat to supercharge your ideas," Gemini. Accessed: Jan. 15, 2025. [Online]. Available: https://gemini.google.com

[10] "Retrieval Augmented Generation: What Is It and How Do Enterprises Benefit?," AI Search Blog. Accessed: Jan. 12, 2025. [Online]. Available: https://www.coveo.com/blog/retrieval-augmented-generation-benefits/

[11] "What is Retrieval-Augmented Generation (RAG)?," Google Cloud. Accessed: Jan. 12, 2025. [Online]. Available: https://cloud.google.com/use-cases/retrieval-augmented-generation

[12] "About MDP | Multidisciplinary Design Program." Accessed: Jan. 15, 2025. [Online]. Available: https://mdp.engin.umich.edu/about-mdp/

[13] J. Danielewicz and P. Elbow, "A Unilateral Grading Contract to Improve Learning and Teaching," *Coll. Compos. Commun.*, vol. 61, no. 2, pp. 244–268, Dec. 2009, doi: 10.58680/ccc20099471.

[14] "Introducing Meta Llama 3: The most capable openly available LLM to date," Meta AI. Accessed: Jan. 14, 2025. [Online]. Available: https://ai.meta.com/blog/meta-llama-3/

[15] “meta-llama/Meta-Llama-3-8B · Hugging Face.” Accessed: Jan. 14, 2025. [Online]. Available: https://huggingface.co/meta-llama/Meta-Llama-3-8B

[16] “bioRxiv.org - the preprint server for Biology.” Accessed: Jan. 14, 2025. [Online]. Available: https://www.biorxiv.org/

[17] “bioRxiv.org - FAQs.” Accessed: Jan. 14, 2025. [Online]. Available: https://www.biorxiv.org/about/FAQ

[18] “Vector stores | LangChain.” Accessed: Jan. 14, 2025. [Online]. Available: https://python.langchain.com/docs/concepts/vectorstores/

[19] “Chroma.” Accessed: Jan. 14, 2025. [Online]. Available: https://www.trychroma.com/

[20] C. Markham, “Introducing OpenBioLLM-Llama3-70B & 8B: Saama’s AI Research Lab Released the Most Openly Available Medical-Domain LLMs to Date!,” Saama. Accessed: Jan. 15, 2025. [Online]. Available: https://www.saama.com/introducing-openbiollm-llama3-70b-8b-saamas-ai-research-lab-released-the-most-openly-available-medical-domain-llms-to-date/

[21] “aaditya/Llama3-OpenBioLLM-70B · Hugging Face.” Accessed: Jan. 15, 2025. [Online]. Available: https://huggingface.co/aaditya/Llama3-OpenBioLLM-70B