DEVELOPING A PERSONALIZED COURSE RECOMMENDATION SYSTEM USING RAG ARCHITECTURE AND DOMAIN-SPECIFIC KNOWLEDGE BASES

(Technical Paper)

INTEGRATING LLM-BASED COURSE RECOMMENDATION SYSTEMS IN ACADEMIC ADVISING: IMPACTS ON STUDENT DECISION-MAKING AND TRUST

(STS Paper)

A **Thesis Prospectus** submitted to the

Faculty of the School of Engineering and Applied Science

University of Virginia | Charlottesville, Virginia

In partial fulfillment of the requirements of the degree

Bachelor of Science, School of Engineering

Alexander-Duc Nguyen

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Introduction

Selecting elective courses as a student in university involves two primary methods: flipping through online course catalogs with basic filter terms or relying on word-of-mouth recommendations. Both approaches can be time consuming given the amount of actions it takes and also limited in their ability to efficiently offer personalized guidance. For instance, browsing through course catalogs typically involves manually sifting through long lists of courses and reading their individual descriptions. On the other hand, relying on peers for advice on what elective courses to take might lead to incomplete information given that the available courses change for every semester/quarter. Therefore, both methods could leave students unaware of the valuable courses that align with their unique interests and academic goals.

This process is particularly challenging for students at universities like the University of Virginia (UVA), where courses are being added or unavailable through individual semesters. Therefore, given the diverse courses and frequently updated course offerings, it can make it difficult for students to actively discover the right courses for them. Several traditional online course catalog tools currently exist. A website, theCourseForum, was founded in 2005 by volunteer students from UVA. The purpose of the website is to filter classes' primary terms such as name, professor, and department. Additionally, the website allows students of the courses to give a rating to the professor/course and provides statistical information such as previous grade point averages (GPA). This application is great for gaining information about professors and course difficulty, but does not address the issue of efficiently finding courses that meet personal interests, due to the limited search capabilities (theCourseForum, 2005). Existing tools, such as course catalog websites with basic filtering systems, fail to provide an efficient way to curate personalized course information that aligns with the interests of students.

To address these challenges, I propose developing a course recommendation and scheduling system that is powered by Retrieval-Augmented Generation (RAG) architecture. This system will utilize the capabilities of local Large Language Models (LLMs) to retrieve and process information from UVA's course catalogs and other relevant academic data. The application will have the primary functionality of matching courses with a student's personal interests based on keywords. By integrating an external database or knowledge base with course catalogs, degree requirements, and scheduling rules, the system will be able to provide accurate and personalized guidance for students. This solution aims to save university students time by reducing decision-making stress and providing tailored, well-informed course information.

Technical Discussion

The goal of this technical project is to develop a RAG-based LLM application that assists students in selecting courses. This approach avoids the need to fine-tune or retrain existing LLMs, which tends to be resource heavy. The application avoids this issue by utilizing knowledge bases which are databases that represent and contain resources about a specific domain of knowledge. In this case, UVA courses and degree requirements. In this section, I will explore the parts of RAG, the creation of a domain-specific knowledge base using course catalogs, and the use of a local LLM such as the Llama model. In addition, I will briefly discuss the tools and frameworks such as Python and LangChain that will be used in this project.

There are several components of the RAG architecture that make it suitable for this project. The indexing component is able to embed documents while retaining their semantic meaning. Once this new information is embedded, it can then be stored into a vectorized database, which serves as an external knowledge base, outside of what an LLM was trained on. When users query within an application that uses RAG, a retrieval component is used to algorithmically search for relevant documents or datasets within the external knowledge base, which is then used as references for the generated responses given back to the user (Wang et al., 2024). This architecture enhances the adaptability of LLMs through swappable external knowledge bases, ensuring that users can have accurate and up-to-date information. In this context, this means that the application will retrieve and reference current UVA course offerings, descriptions, and requirements to provide personalized recommendations.

The knowledge base for this application will be constructed using UVA's online course catalogs, degree requirements, and other relevant course information found throughout the websites. This involves several steps of data collection, preparation, and embedding and storage. To collect data, content related to course descriptions or other academic materials will be scraped from UVA's websites. Then, the collected documents will be chunked into manageable text segments for embeddings. This step ensures that each document retains its semantic meaning while also retaining computational efficiency by limiting the size of each input. Lastly, the prepared data will then be embedded, using a Python library, and stored in a database that serves as the external knowledge base for the RAG system (Khder, 2021).

In this project I will utilize Llama, a LLM model that I can run locally on my machine. Llama has all the necessary capabilities of natural language processing (NLP) and generative outputs. In addition, since Llama can be deployed locally, it offers more control over privacy, resource utilization, and adaptability (Malisetty & Perez, 2024). Therefore, a local model such as

Llama is an ideal choice for a university-specific application that could handle sensitive user data such as transcripts, while also maintaining high performance and cost efficiency.

Python will serve as the primary programming language for implementing the application due to its useful collection of libraries that will help with web crawling and RAG pipeline designing. LangChain is a Python framework specifically designed for aiding in developing applications with LLMs. This framework will be used to streamline the integration of the RAG architecture. LangChain is able to simplify tasks by having functionalities such as prompt templates, designing the retrieval process, and also connecting relevant knowledge bases. Therefore, Python and LangChain will be the base of development for the technical project.

In this next section, I will discuss the desired user workflow when using this application. First, the user will input a query, such as asking for recommended electives based on their major and personal interests for their next semester. The retrieval component of the application will then embed the user's query and semantically search for all the relevant course descriptions that can be used in the response back to the user. After the documents are retrieved, then the sources are passed to the Llama model, which generates a response directly sourced from the information retrieved. In hopes, the response presented to the user will accurately reference all relevant courses applicable to the query.

This approach offers several advantages to students and other users. First, it allows for up-to-date knowledge bases that ensures that the recommendations reflect the most current course offerings. Second, the integration and utilization of a local Llama LLM aids in any privacy concerns and cost control while developing and maintaining the application. Finally, by providing a holistic search of courses, it will reduce the time it takes students to plan out their next semester by providing tailored course recommendations efficiently. In addition, by leveraging tools such as RAG and LangChain, this project not only addresses the challenges that students face, but could serve as an example of how LLM applications can fit into higher education or education in general.

STS Discussion

The use of an LLM driven recommendation system in academic settings introduces significant social and ethical considerations. In regards to academic advising, and any decision-making influence, users must be able to fully trust that the information that they are provided is accurate. Although the introduction of this technology in an academic setting might offer promising benefits, it also raises important questions about its role and limitations for students and advisors.

One of the primary socio-technical questions that I will be exploring in this project is how LLM driven tools, such as a course recommendation system, will influence the decision making of students. Academic advisors hold an important position in universities as students seek guidance on either factual information such as course requirements, or more abstract advice on mentorships and career advice. In the context of this project, LLMs have the potential to automate routine advising tasks such as helping students identify elective courses that align with their interests. However, LLM applications are not fully capable of replacing human advisors. This is because Advisors bring expertise, emotional empathy, and the ability to address complex issues such as navigating personal challenges, especially in higher degree education. Current trends in academic technology show how artificial intelligence (AI) or LLMs are more of a complementary tool rather than a replacement for human teachers and advisors (Shao et al., 2024). Therefore, by having supporting LLM technology deal with handling repetitive tasks, human advisors can focus more on complex and individualized interactions, creating a system that enhances efficiency without diminishing the advisor's essential roles.

Another critical concern in the adoption of LLM driven systems in education is trust. Students must be able to trust all of the responses and recommendations that the system provides, and that the information has no inaccuracies or biases. Transparency plays a key role in building this trust. Since RAG based applications have LLMs that build their responses from a limited pool of directly, referenced sources, this increases the transparency of where the information is coming from. For example, when suggesting courses, the application has the ability to indicate where they retrieved the UVA course descriptions. However, transparency in an application will not resolve all the concerns of trust in decision-making with LLM systems. The potential of students over-relying on any response given by the system as absolute must be mitigated. Therefore, it is important to be able to restrict the LLM application in only responding with answers that can be retrieved from its knowledge base, else it should tell the user that it cannot find an answer. Overall, this highlights the importance of being able to educate users on the limitations and reinforcing the idea of using the tool as a support, not to replace human expertise (Schmidt et al., 2020)

Beyond addressing the challenges university students face with selecting elective courses, the RAG LLM application that will be developed has the potential to benefit various areas of society. This technology can serve a broader audience, other than just specialized professional fields. For instance, educators could use such an application to new academic trends. Another example would be for students to build up their own knowledge bases to aid in studying and keeping track of valuable sources. This concept aligns with the idea of universal design in regards to how technology built for a specific set of users can actually benefit a larger audience

(McMahon & Walker, 2019). Therefore, this project aims not only to build a tool for university students in selecting courses, but to also create a versatile tool for a wider population.

Research Question and Methods

I will be addressing the following research question: "How do LLM driven course recommendation systems influence student decision-making and trust in academic advising, and how can these systems complement human advisors to enhance the academic advising process?" To help answer the question, the approach I will use is split into addressing whether or not the developed LLM application is accurate and the responses will not hallucinate. This will be done by comparing the recommended course descriptions and the keywords within the response and query. In addition, I will be conducting surveys and interviews on test users, UVA students, on whether or not they find the recommendations useful. On the other hand, I will address in which areas will the LLM application support human advisors. This will be done by creating an evaluation that compares a range of simple to complex commonly asked advising questions with the expected responses from advisors and the generated response of the LLM application. This will give us greater insight on the overall limitations of the system. Overall, these methods will help discover insights on the implications of integrating LLM applications into academic decision-making processes.

Conclusion

The technical deliverable for this project is a RAG-based LLM application that is able to support students in their course selection process by being able to match course descriptions to the student's interests and academic needs. The tool will integrate LangChain with a local Llama model to create a personalized system. The STS deliverable will explore the societal implications in academic advising, particularly with student decision-making, trust, and complementary roles with human advisors. By combining both quantitative and qualitative evaluations, the analysis will be able to provide an understanding of how such LLM tools can be implemented ethically and effectively.

References

Khder, M. (2021). Web scraping or web crawling: State of Art, Techniques, approaches and application. *International Journal of Advances in Soft Computing and Its Applications*, *13*(3), 145–168. https://doi.org/10.15849/ijasca.211128.11

Malisetty, B., & Perez, A. J. (2024). Evaluating Quantized Llama 2 models for IOT Privacy Policy Language generation. *Future Internet*, *16*(7), 224. https://doi.org/10.3390/fi16070224

McMahon, D. D., & Walker, Z. (2019). Leveraging Emerging Technology to design an inclusive future with Universal Design for Learning. Center for Educational Policy Studies Journal, 9(3), 75–93. https://doi.org/10.26529/cepsj.639

Schmidt, P., Biessmann, F., & Teubner, T. (2020a). Transparency and trust in Artificial Intelligence Systems. Journal of Decision Systems, 29(4), 260–278. https://doi.org/10.1080/12460125.2020.1819094

Shao, A., Lu, Z., Zhong, B., Liu, S. Q., & Lu, W. (2024). Evaluating the adoption and effectiveness of human educators vs. AI tools in Education. *Computers in Human Behavior*, 108492. https://doi.org/10.1016/j.chb.2024.108492

Wang, Z., Zhao, S., Wang, Y., Huang, H., Xie, S., Zhang, Y., Shi, J., Wang, Z., Li, H., & Yan, J. (2024, October 3). Re-task: Revisiting LLM tasks from capability, Skill, and knowledge perspectives. arXiv.org. https://arxiv.org/abs/2408.06904

Your guide to courses at UVA. theCourseForum. (2005). https://thecourseforum.com/