Project Executive Summary - Course Assistant Bot

Reggie Bain

This project is a proof of concept of what we envision as a broad class of AI-based assistants for course development. It focuses on a tool for building effective course syllabi, which serve a variety of key functions in course development. The syllabus can not only serve to outline the key components of the course, but also serves as a "contract" between the students and the instructional staff. In recent years, the rise of AI use in educational contexts (and resulting issues with academic integrity violations) has made the contract aspect of syllabi particularly important. Course policies must be enforced equitably and worded clearly, particularly in cases where litigation could be involved.

This project is an example of a simple suite of tools that can help instructors, departments, or colleges to quickly assess how effective syllabi are at answering student questions by actually simulating asking those questions using Large Language Models (LLMs) and so-called Retrieval Augmented Generation (RAG). In short, RAG is a process of providing queries to LLMs that are supplemented with semantically relevant context from source material. Source material could range from scraped internet resources from Wikipedia to aid in more general questions to specific documents/databases that user's may want to query for precise details. We will be exploring the latter in an effort to allow LLMs to provide accurate and precise answers to important questions students and others have about college courses.

Using tools such as HuggingFace transformers (to access pre-trained LLMs) and LangChain (to quickly handle inputs outputs to the LLMs) we constructed a pipeline to query documents in various formats. We also created an evaluation pipeline where LLMs create synthetic question sets, perform quality checks on these questions, and reconstruct the true answers to these questions based on the context. We then created another set of answers using our RAG pipeline and evaluated the generated answers using GPT-4 and a detailed evaluation prompt to generate a "correctness" score of 1-5 for each Q&A pair. Comparing our results with a baseline of asking general purpose LLMs questions that many students would normally want to "ask" a syllabus, our RAG model (somewhat obviously) performs far better.

Finally, we created a Streamlit app to showcase our system that allows users to upload and query their own documents. We tested this on a variety of questions where we knew the syllabi contained correct answers and were quite pleased with the results. Obviously, financial, compute, and storage limitations played a significant role as many open source LLMs (via the HuggingFace/OpenAI/Meta APIs) imposed strict rate limits for free-tier users. With additional resources, we could explore a much larger hyperparameter space, use more powerful models with many billions of parameters, and turn our prototype application into a full fledged tool we plan to deploy in the near future.