Course Assistant Bot - Data | KPIs | Stakeholders

Author: Reggie Bain

Data

The data for this project will be drawn, primarily, from real course material developed for my own courses and materials from other courses with explicit permission from the course developers/instructors. Data sources will primarily focus on course syllabi (that are designed to be extremely detailed) but will, hopefully, also incorporate course slides and data extracted from course Canvas modules. The goal will be to create a tool for users to search and query course material in a fast, accurate, and easy-to-use way. As needed, we will also utilize publicly available datasets on question answering such as:

- Stanford Q&A Dataset: https://www.kaggle.com/datasets/stanfordu/stanford-question-answering-dataset
- Science Test LLM Competition Dataset: https://www.kaggle.com/competitions/kaggle-llm-science-exam/data
- Other datasets relevant to the benchmarks found on Hugging Face leaderboards: https://huggingface.co/collections/open-llm-leaderboard/the-big-benchmarks-collection-64faca6335a7fc7d4ffe974a.

Stakeholders

- Students enrolled in college/advanced high school courses, especially those in large enrollment college courses that have a lot of course infrastructure
- Instructors of courses who often field hundreds of emails regarding procedural course questions.
- Centers for Teaching Excellence and other administrative organizations who assess teaching quality and provide resources to improve course resources

Key Performance Indicators (KPIs)

- Model outperforms a baseline of answering queries using an LLM without added context.
- Successfully create a database of vectorized embeddings of real course content.
- Identify key metrics for benchmarking fine-tune chat-bots, where accuracy of responses and limiting the scope of answerable questions are critical.