Data Progress Report: Course Syllabus RAG System

Cade Boiney, Ken Lam, Ognian Trajanov, Benjamin Zhao

Data Collection Progress

We are gathering syllabi from a variety of departments across the college. Currently, all collected syllabi are downloaded into a shared Google Drive folder organized by department (Syllabi). Sources include PDFs from departmental websites and syllabi sent directly by faculty in response to emails. Some departments, such as Economics, have been more responsive, while others have shown lower participation, leaving gaps in the dataset. Several professors who are not currently teaching courses to our team have not responded.

Challenges Encountered

Collecting syllabi has been more difficult than anticipated. Limited accessibility and inconsistent faculty responsiveness slow progress, making data acquisition the primary challenge.

Next Steps and Proposed Solutions

To improve data collection efficiency, we plan to work with department chairs to request department-level syllabus archives, allowing bulk collection rather than emailing individual faculty. Additionally, we are exploring the use of Student Planning data, either via an official API or direct database access from the Registrar's Office. This would provide structured course information—including course descriptions, schedules, instructors, and potentially syllabus attachments—which could serve as a foundation for the RAG system.

If API or database access is not available, a secondary option is web scraping the Student Planning interface to extract publicly available course information. Any scraping would be conducted carefully in coordination with the Registrar to comply with usage policies.

Once more syllabi and course data are collected, we will begin cleaning and structuring PDFs into a machine-readable format suitable for embedding and retrieval. Completing this phase will require additional time, but it is essential to ensure a sufficiently comprehensive dataset to make the RAG system accurate and meaningful.