

LLM Chatbot for Cornell's MPS in Info Science

Proposal

Problem or question

Prospective and current students researching Cornell's Master of Professional Studies (MPS) in Information Science might struggle to find clear, consolidated, and personalized information regarding the program and its policies. For example, the official Cornell websites contain extensive content about curriculum options, concentrations, admissions requirements, career outcomes, degree requirements, and differences between programs. However, this information is fragmented across many pages. The existence of a single chatbot that could consolidate all information and provide users with a single master resource would be extremely helpful.

Given this need, our goal is to create an LLM-based conversational assistant/chatbot that guides a prospective or current student through the process of exploring the Cornell MPS in Information Science specifically, helping them understand information like whether the program fits their background, whether they meet application requirements (like prerequisites), how the curriculum works (core, elective, concentrations), career outcomes, and differences between this program and other graduate programs in Bowers CIS.

This tool is important as graduate program exploration might be overwhelming, especially for underserved users who might not be as familiar with graduate school; e.g., first-generation, international, and career-switching students. Providing an interactive and personalized interface can reduce confusion, increase transparency, and support more informed decisions. For example, users or applicants would get clear answers to common questions like, e.g., "What is the cost of tuition?" or "How many credits do I need to graduate?" This tool could thus also reduce any administrative friction in the Information Science Department or Bowers CIS departments since they would need to answer less student queries, which can often become repetitive.

Data sources

Our chatbot’s knowledge base will be constructed from official Cornell University Information Science Department websites, specifically focused on the MPS in Information Science program. These sources provide authoritative and comprehensive information that prospective and current students need. The primary data sources include:

1. **MPS Program Overview** (<https://infosci.cornell.edu/master-professional-studies-information-science>): This page contains essential program details including the program structure (2-3 semester options), optional focus areas (User Experience, Data Science, Interactive Technologies, Networks, Crowds, and Markets), degree requirements (minimum 30 credits, course distribution requirements), career outcomes (employment statistics, salary data), and the signature MPS Project component.
2. **Application Information** (<https://infosci.cornell.edu/master-professional-studies-information-science/apply>): This source provides critical information for prospective applicants including application deadlines (Fall and Spring), required materials (statement of purpose, diversity statement, resume, transcripts, letters of recommendation, video interview), ideal candidate profile and prerequisites (programming experience, mathematical background, writing skills), international student requirements (TOEFL/IELTS scores), and financial support options (scholarships and GTRS positions).
3. **Academic Planning Resources** (<https://infosci.cornell.edu/programs/master-professional-studies-information-science/academic-planning>): This page details the curriculum structure including Information Systems (IS) and Human and Social Systems (HSS) course requirements (9 credits each), the MPS Project requirement (INFO 5900), Professional Career Development course (INFO 5905), elective options, and planning tools like the course selection spreadsheet.

These official sources will be scraped and processed to create a comprehensive knowledge base for our RAG system. The chatbot will use this verified content to answer questions about admissions, curriculum, requirements, career outcomes, and program policies, ensuring all responses are grounded in official Cornell information with proper citations and links back to source pages.

Final product

We will build a Shiny-based interactive chatbot, deployed on Shinyapps.io or Posit Connect Cloud. This chatbot application will provide a conversational experience tailored specifically to those exploring or wanting to know more about the MPS in Information Science. The chatbot will use RAG to ground all responses in verified Cornell content and links. This tool will be prompt engineered to answer common questions most effectively.

Ethical concerns

1. Accuracy & Misinformation

Our chatbot answers questions about admissions, curriculum, and program requirements for the Cornell MPS in Information Science. While we rely on publicly available information from official Cornell websites, LLMs may still generate inaccurate or outdated responses (“hallucinations”). This creates ethical risks because prospective students may make important academic decisions based on incorrect guidance.

To mitigate this, we: 1. Restrict the chatbot’s knowledge to scraped or manually curated official content. 2. Design retrieval-augmented prompts that force the model to ground answers in sourced text. 3. Include disclaimers encouraging users to double-check key information with official Cornell pages.

2. Outdated Information Due to Yearly Program Changes

Graduate program requirements change every year. Course offerings, prerequisites, tuition fees, deadlines, and even visa rules may all be updated by the university. Since our dataset is collected at one specific time, the chatbot may give information that becomes outdated later.

To reduce this problem, we add clear timestamps to the data, plan regular updates, and remind users that the system may not always reflect the newest official information.

3. Bias & Fairness

LLMs may generate responses that unintentionally privilege certain applicant backgrounds (e.g., U.S. vs. international students, CS vs. non-CS backgrounds). Biased wording or differential encouragement could influence user confidence or perceived eligibility.

To reduce unfair outcomes, we: - Avoid using model-generated interpretations of applicant competitiveness. - Rely strictly on factual program descriptions. - Manually review responses for demographic favoritism, exclusionary language, or cultural bias.

4. Privacy & Data Handling

Although our system does not require personally identifiable information, users may voluntarily enter details about GPA, nationality, work experience, or visa status when asking admissions-related questions. These inputs may carry sensitive academic or immigration-related information.

To address this, we avoid storing user queries, do not collect personal data, and ensure that any logs used for debugging are anonymized and deleted regularly. We clearly state that users should not submit personal data.

5. Intellectual Property & Web Scraping Ethics

Although we only use publicly available content from Cornell's official website, scraping must still respect terms of use, avoid abusive request frequency, and acknowledge content ownership.

We ensure that: - Scraping is limited, rate-controlled, and compliant with robots.txt. - Cornell is credited as the original source of information. - No proprietary or restricted documents are included in the dataset.

Project timeline

- Nov. 20th: Proposal due
- Nov. 25th: Data generation
- Nov. 28th: RAG construction
- Dec. 3rd: Chatbot construction
- Dec. 4th: Draft product due
- Dec. 9th: Product improvement and report labor assignment
- Dec. 15th: Report refinement
- Dec. 18th: Final product and report due