RAG-based Interactive AI for MS in Applied Data Science Website

Group 7: Alex Tsourmas, Danylo Sovgut, Jane Lee, Kaylie Nguyen, Yeochan Youn
Due: 5/30/2025

Implementation

We began by collecting and web-scraping comprehensive content from the official MSADS program website. The collected data was preprocessed: we removed duplicates, cleaned out website artifacts, broke the text into concise, meaningful chunks with their titles and sections, and filtered out irrelevant or trivial passages. This process ensured our knowledge base was both high-quality and relevant.

For retrieval, we embedded each chunk of text using a transformer-based model, MiniLM, which captures semantic meaning. These embeddings were stored in a Chroma vector database. When a user submits a question, the system embeds the query in the same way and searches the database to find the most semantically similar pieces of information. Rather than relying on a single snippet, the system retrieves several top matches to ensure the answer covers all necessary details.

Once the relevant context is gathered, it is formatted into a prompt for the language model—GPT-4o-mini in our case. The prompt instructs the model to generate an answer strictly based on the retrieved passages, discouraging unsupported speculation. The language model then synthesizes a clear, fluent, and contextually accurate response, which is displayed to the user along with the supporting evidence.

To evaluate the system, we compiled a test set of frequently asked questions paired with human-written reference answers. We assessed the system on retrieval accuracy, answer relevance, fluency, and completeness. Our context recall is 0.75, answer correctness is 0.6535, and answer relevancy is 0.7641. Human evaluators confirmed that the system reliably retrieves the right context and produces accurate, readable answers, though some gaps can still occur if the knowledge base lacks certain details.

Flow Diagram