Emory Infobot: A LLM-powered Virtual Assistant for University Students

Darren Ni Aaron Zaiman Yutong Hu Jinho D. Choi
Department of Computer Science
Emory University
Atlanta, GA, USA
{darren.ni, aaron.zaiman, yutong.hu, jinho.choi}@emory.edu

Abstract

We introduce Emory Infobot, a question answering system tailored for college students at Emory University. The Emory Infobot aims to be a centralized information system capable of answering all questions related to the college experience. The system employs a Retrieval Augmented Generation (RAG) system over a synthetically generated set of question answer pairs from Emory-related web domains to provide accurate information. To evaluate performance, we conduct a series of tests with real Emory University students, using a diverse set of queries ranging from academic schedules, campus facilities, to student life and extracurricular activities. The results indicate a high level of accuracy and exceptional user satisfaction showcasing the system's ability to effectively answer a wide range of college-related queries.

1 Introduction

The rise of LLMs are revolutionizing the way we interact with and process information. Chat applications such as ChatGPT are pioneering more natural interactions with digital information, moving beyond the traditional search engine methods. These new technologies allow users to engage in conversational dialogues with AI, enabling a more intuitive and accessible way to find answers. We wanted to extend the motivation behind this technology to develop a tool for students designed to facilitate more natural interactions with important university-related information, so we created Emory Infobot – a centralized AI information system designed to streamline the university student experience.

2 Data Collection

We compile a dataset of university related information using a synthetic data generation pipeline that extracts publicly available information from Emory-related web domains and access the quality of the dataset by utilizing GPT as an evaluator.

2.1 Web Crawling

We implement a custom web crawler to scrape textual data from university-related websites. We base our search algorithm from ECS, an Emory-specific directory containing a diverse set of web links ranging from Academic Support to Campus Facilities, compiling information from 3.5K+ websites spanning over 50+ departments.

2.2 Synthetic Data Generation

We utilize synthetically generated data to create high quality data points for our RAG system (Lu et al., 2023). We construct a pipeline to generate a hypothetical question answer pair dialogues based off previously scraped web content to simulate likely conversations. To do this, we employ a two-pass data generation technique (Figure 1).

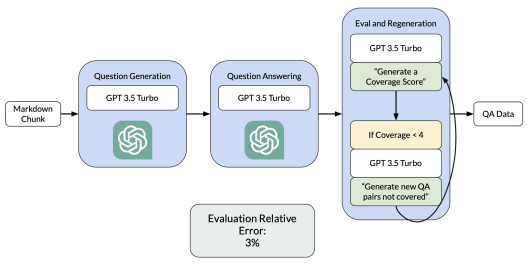

Figure 1: Overview of QA Generation Pipeline

We first split parsed markdown text from each page into chunks with a sliding window approach to encourage cross-chunk coverage. Each markdown chunk is then fed to GPT-3.5-Turbo with a question generation task, whose output is passed back into GPT with the original content chunk with a question answering task. We found the quality of the results of this two-pass method to be better than a single QA generation pass method. Using this pipeline over our full dataset, we generated 180K+ QA pairs.

2.3 Data Quality Evaluation

We construct a evaluation pipeline, utilizing GPT as an evaluator to rate how well the generated QA covers the content from which it was derived from on a scale of 1-5. For results below a score of 4, we task GPT to regenerate QA pairs. The results from GPT were cross evaluated against human evaluated scores across 100 chunks, which yielded a 3% post-evaluation error. Further, our results yielded a score of 4.65 across all chunks.

3 Model

We choose to use a RAG system as a lightweight alternative to fine-tuning a LLM (Gao et al., 2023). Recent works also show that the RAG architecture performs better than parametric-only-S2S models and task-specific retrieve-and-extract architectures(Lewis et al., 2020). We implement our RAG system to perform semantic search over our synthetic dataset and utilize careful prompt templates to generate tailored responses to user queries, giving us a robust, interpreatble model.

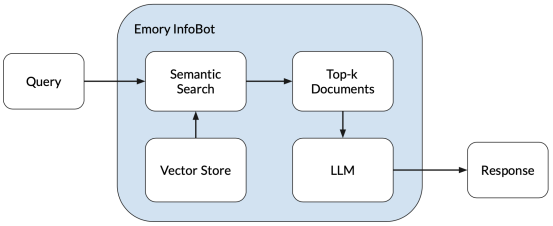

Figure 2: Overview of Infobot RAG System

3.1 Embeddings

We embed each QA pair in our dataset using the BAAI/bge-large-en model with a vector normalization step inbetween. Since user query embeddings could remseble either parts of our dataset questions more than answers or vice versa, we attach equal weights to both the question and answer component of each QA pair to create our vector store.

3.2 Search Engine

We construct a search engine to perform a cosine similarity search between user query and our QA dataset, which is accelerated with FAISS, a GPU-optimized similarity search library.

3.3 Prompt Templates

We prompt engineer a template designed to summarize retrieved documents prioritizing accuracy of information and minimizing hallucinations. At inference time, retrieved documents are put into the template and fed to GPT and the output is displayed as the system response.

3.4 Safeguards

To ensure the reliability and relevance of the responses generated by our system, we implement two safeguards. First, we compare the cosine similarity score between our system-generated response and each retrieved document to provide a general heuristic for us to determine whether the generated response is sufficiently relevant. Second, we provide the source links of every document used to generate each system response to encourage model interpretability.

4 Evaluation

To evaluate the Infobot's performance and accuracy, we generate and test on a hybrid dataset of human-generated queries and GPT-generated queries.

4.1 Validation Set

We create our own diverse dataset of student questions to have a validation set for Emory-specific queries. The hybrid dataset consists of questions from real students and hypothetical questions generated from GPT, whose topic scope ranges from financial aid to academic policies to athletics.

4.2 Results

We tasked human annotators to rate the generated responses of the validation set from 1-5. The results yielded an average score of 4.63, indicating high user satisfaction and accuracy of information.

5 Limitations and Future Work

The current iteration of the Infobot is robust and capable of answering most student queries, but we would like to put more efforts into making the Infobot more conversational. To do this, we plan to utilize a Dialogue State Generation (DSG) approach, where extracted utterance-level context can be used to generate significantly more tailored responses. Further, we plan to add more safeguard mechanisms to address ethical challenges associated with deploying AI-based assistants (Piñeiro-Martín et al., 2023).

References

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv*.

Yingzhou Lu, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, and Wenqi Wei. 2023. Machine learning for synthetic data generation: A review. *arXiv*.

Andrés Piñeiro-Martín, Carmen García-Mateo, Laura Docío-Fernández, and María del Carmen López-Pérez. 2023. Ethical challenges in the development of virtual assistants powered by large language models. *Electronics*, 12(14):3170.