A Retrieval-Augmented Academic Assistant for East West University Students Using Semantic Search and LLMs

1. Problem Definition

Universities today generate an overwhelming amount of academic information, and much of it is delivered to students in the form of long PDF documents, departmental guidelines, program handbooks, and online pages spread across different platforms. Although this information is essential for students to understand their academic journey, it is often difficult to access quickly and efficiently. At East West University, for example, the undergraduate bulletin, departmental curriculum structures, course descriptions, and graduation regulations span hundreds of pages. New and returning students frequently struggle to find accurate answers to simple questions such as graduation requirements, prerequisite rules, or credit limits. These documents are dense, filled with academic terminology, and not optimized for natural language search. As a result, students often miss key details or misunderstand important policies, which leads to academic mistakes or misinformation.

Another major problem is that the information students need is scattered across many separate sources. A student may need to look at a PDF, browse the university website, check a departmental notice, and also consult seniors just to clarify one small question. Traditional keyword search inside a PDF usually fails because academic writing uses complex phrasing and context-dependent explanations. This is especially challenging for new students who are unfamiliar with academic structures. Meanwhile, faculty members and staff are repeatedly asked the same basic questions each semester because students cannot easily locate the answers on their own. This presents both an efficiency issue and a communication challenge within the academic environment. The core problem addressed in this paper is the absence of a centralized, intelligent, conversational system that can provide students with accurate academic information instantly by understanding their natural language queries and retrieving answers from official university documents.

2. Motivation

The motivation behind this project comes from observing how difficult and time-consuming it is for students to navigate through academic information, especially during the early stages of their degree. Students often rely on unofficial sources such as Facebook groups, seniors, or random online posts, which may not always be accurate. This dependency increases confusion and spreads misinformation, especially regarding prerequisites, course selection, retake rules, and graduation criteria. Many students avoid reading long PDFs simply because the documents feel

overwhelming. As a result, important details go unnoticed, leading to unnecessary stress and academic mistakes such as taking the wrong courses or misunderstanding degree requirements. A reliable academic assistant could solve these issues by providing instant clarification whenever students need it.

At the same time, modern students expect fast, user-friendly, and natural ways to access information. Traditional university documentation does not meet these expectations. With advancements in artificial intelligence, particularly in semantic search and large language models, it is now possible to build a system that reads and understands these long documents on behalf of the students. Retrieval-Augmented Generation (RAG) technology is especially appealing because it ensures that answers are grounded in official documents rather than relying solely on the language model's internal knowledge. This significantly reduces the risk of hallucination and improves trust in the system's responses. The project is also motivated by the desire to reduce repetitive workloads for faculty members. Since many student questions follow predictable patterns, an AI-based academic assistant can handle most initial inquiries, allowing teachers and staff to focus on more complex academic issues. Although this system was initially designed using datasets related to the CSE department, it has the potential to expand into a university-wide solution for all departments simply by adding their documents to the dataset. The long-term motivation is to develop a scalable, transparent, and efficient academic support system that benefits both students and faculty.

3. Proposed Methodology

The methodology used to build the academic assistant follows the principles of the Retrieval-Augmented Generation (RAG) architecture, which combines document retrieval and large language model reasoning. The first step involves collecting and organizing all relevant academic materials, such as undergraduate and graduate bulletins, departmental PDFs, course descriptions, and program structures. These documents are stored in a structured dataset and processed to extract clean, readable text. Since the original documents are lengthy, the text is divided into smaller overlapping pieces to preserve context even when split. This process allows the system to retrieve accurate portions of the documents when students ask questions.

Once the documents are segmented, each piece of text is converted into an embedding using a sentence-transformer model. These embeddings capture the semantic meaning of the text, enabling the system to match a student's question not by keywords but by deeper conceptual similarity. All embeddings are stored in a FAISS vector database that supports fast similarity search. When a student submits a question, the system generates an embedding for the question and uses FAISS to find the most relevant document segments. These selected segments are then combined into a contextual prompt for the large language model. The model used in this project is Groq's optimized LLaMA 3 version, which is capable of generating coherent, factual, and

context-aware responses. Because the model receives only the retrieved academic text as context, the answers remain grounded in real university documentation.

The final output is delivered through a web-based chatbot interface developed with Streamlit. The interface allows students to type questions naturally, view the answers, and see which document sections were used to generate the response. It also preserves conversation history, handles errors gracefully, and displays institutional branding if needed. This methodology ensures that the system is not only technically robust but also user-friendly and practical for everyday academic use. Even though it was initially built using CSE-related datasets, the architecture is fully adaptable to any department or university simply by adding new documents to the database.

Reference:

1. Vector Databases in Modern Applications: Real Time Search, Recommendations, and Retrieval Augmented Generation (RAG)

   https://ijaibdcms.org/index.php/ijaibdcms/article/view/257/260

2. **Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks**

   https://arxiv.org/abs/2005.11401

3. **PDFTriage: Question Answering over Long, Structured Documents**

   https://arxiv.org/pdf/2309.08872

4. **UniQA: an Italian and English Question-Answering Data Set Based on Educational Documents**

   https://ceur-ws.org/Vol-3877/paper16.pdf