Ema Intern Take-Home Challenge

Introduction

This report outlines the approach and implementation for the Ema Intern Take-Home Challenge, which involves building a Natural Language Query Agent over a small dataset of lecture notes and a table of LLM architectures. The system is designed to answer simple, conversational questions based on these documents, leveraging LLMs and vector indexing frameworks.

Approach

1. **Data Collection:**
    - **Lecture Notes:** Four lecture URLs were provided, each containing detailed content related to LLMs. The content was extracted using web scraping techniques.
    - **LLM Architecture Table:** A table of milestone LLM architectures was sourced from a GitHub page and similarly extracted.
2. **Data Processing:**
    - **Web Scraping:** The BeautifulSoup library was used to scrape and extract text from the provided lecture URLs.
    - **Text Extraction:** The extracted text was cleaned and organized into a list of lecture contents and a separate string for the architecture table.
3. **Embedding Generation:**
    - **Embedding Model:** The Google Generative AI model (Gemini) was used to generate embeddings for both the lecture contents and the architecture table.
    - **Embedding Creation:** The text data was converted into embeddings using the GoogleGenerativeAIEmbeddings class.
4. **Vector Store Creation:**
    - **Document Objects:** Each piece of text content was encapsulated into Document objects.
    - **FAISS Vector Store:** The FAISS library was used to create a vector store from the document embeddings, allowing for efficient similarity searches.
5. **Query Handling:**
    - **Query Embedding:** Incoming queries were converted into embeddings.
    - **Similarity Search:** The vector store was searched for the most relevant documents matching the query embedding.

- **Answer Generation:** A prompt was created by combining the content of the relevant documents, which was then passed to the LLM to generate a natural language answer.

Areas of Improvement

1. **Conversational Memory:**
   - **Current Limitation:** The system handles one query at a time without remembering previous interactions.
   - **Improvement Plan:** Implement a memory module to maintain context across multiple queries, enabling follow-up questions and a more interactive experience.
2. **Citation of References:**
   - **Current Limitation:** Answers are generated without explicit citations.
   - **Improvement Plan:** Enhance the answer generation process to include citations from the source documents, thereby providing more transparency and avoiding hallucinations.
3. **Scalability:**
   - **Current Limitation:** The current system is designed for a small dataset.
   - **Improvement Plan:** Develop a strategy for scaling the system to handle larger datasets, including more lecture notes and additional tables, ensuring efficient performance and storage.