AIML : IITI-Bot Proposal

Team Details:

Team Leader:
Name: Masabattula Teja Nikhil
Roll Number: 2404101014
GitHub: https://github.com/Tejanikhil
LinkedIn: https://www.linkedin.com/in/Tejanikhil/

Skills: C++, Python, PyTorch, CUDA, Git, Deep Learning, LLMs, RAGs, Computer Networks, Speech Processing, Web development.

Other Team Members:

Name: Ayush kumar Singh
Roll Number: 2404101011

Name: Jatin Sharma
Roll Number: 2404101003

Name: Ankit Kayastha
Roll Number: 2402102019

Name: Mohd Aamir
Roll Number: 2402101005

Name: Saransh Vashistha
Roll Number: 2404101005

Domain: AIML
PS Name: IITI-Bot
PS Number: AIML-11
Preference Number: 1

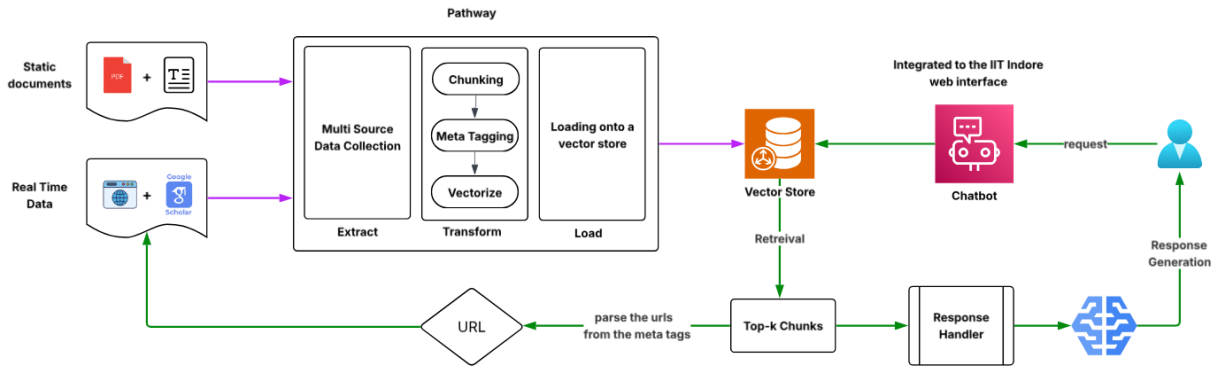

Figure 1: Tentative RAG pipeline

Project Solution:

The tentative pipeline for the project is as shown in figure 1 (**green** indicating the inference flow, **violet** indicating the development stage flow)

We have organized the development of the system into multiple clearly defined phases, each contributing to building a robust, real-time, and intelligent RAG (Retrieval-Augmented Generation) based chatbot for IIT Indore.

Phase 1: Dataset Collection

- **Objective :** Gather and clean foundational content from various static sources.
- Sources include PDFs, text documents, webscraped data, etc.
- It also involves data cleaning.

Phase 2: Pathway Integration & Real-time Streaming

- **Objective :** Enable real-time updates for timely updated data
- Real time streamed data includes timely varying information such as Fee structures, PhD/research openings, Faculty acheivements, Recent publications (Google scholar)
- Updated data gets streamed, and undergoes various transformations and will be stored in the vector DB for future retreival.
- The transformation phase includes :
  1. **Chunking** longer documents into manageable pieces
  2. **Tagging** the chunks with the labels such as Category (e.g., Admission, Faculty, Acheivements, etc.), Source URLs, Last updated time, etc.

3. **Categories** are created based on the diversity of data collected to improve the relevance of search and response.

- Vectorizing the chunks into embeddings using bert or sentence bert, or any other light weight models.

Phase 3: Response Generation & Reasoning

- **Objective:** Provide contextually accurate answers by combining multi sourced static knowledge and live updates.
- The flow for the response generation are as follows :
  1. Retrieve relevant chunks from the vector DB based on the query.
  2. Determine the URLs associated with those chunks.
  3. Use Pathway to stream the most recent version of those pages.
  4. Feed both vector-retrieved and streamed data to an LLM.
  5. Summarize, organize, and reason over this information to generate a coherent response
  6. Any new or updated content is added back to the vector DB for future optimization.
- Will test with hybrid retreival methods (dense retreival + sparse retreival) and employ strategies like score fusion or cascadings to improve the correctness of the response.

Phase 4: Chatbot Development

- **Objective**: Develop an interactive, responsive, and intelligent chatbot interface.
- It includes Response generation optimization, Efficient query classification and routing, Integration with backend RAG pipeline.

Phase 5: Frontend Integration

- **Objective:** Create a live demo for IIT Indore's website.
- Deliver a simple, functional frontend through which users can interact with the chatbot.

Highlights:

- Tagging the documents with category information for better retreival.
- Employ hybrid retreival strategies to ensure correctness of the responses.
- Deliver a frontend integrated with a chatbot.

Project Timeline:

- **Week 1:** Completion of phase 1 and phase 2 parallelly creating the front-end for demo.
- **Week 2:** Work on generating coherent responses with reasoning.
- **Week 3:** Completion of chatbot creation and frontend integration.