Chatbot with RAG(Retrieval Augmented Generation) + Voice Features

1 About RAG

- RAG stands for Retrieval-Augmented Learning.
- It improves the capabilities of LLMs by gathering necessary information from a large collection of data and feeding it to the model.
- This allows it to answer questions with more context and accuracy.

2 About Agentic RAG

- Enhances standard RAG systems by introducing autonomy and proactivity.
- The agent can make decisions on how to retrieve information, manage retrieved data, and apply intelligent strategies.
- Results in a more self-directed and optimized RAG pipeline.

3 Knowledge Base used

- Harry Potter and the Sorcerer's Stone.

4 Libraries Used

- streamlit

- os
- pyttsx3
- faiss
- langchain
- duckduckgo_search

5 About Embeddings

- Machine Learning algorithms cannot directly process plain text.
- Text is split into numbers and represented as numeric vectors.
- Captures semantic and contextual relationships.

6 RAG Pipeline

1. Extract text from PDF into several chunks.
2. Convert chunks into vector embeddings.
3. Store vector embeddings in a vector store (knowledge base).
4. Convert user prompt into embeddings.
5. Conduct semantic search on vector database.
6. Retrieve necessary information and pass it to an LLM for response.

7 LLM Used

- The LLM used in the program is LLaMA 3.2.

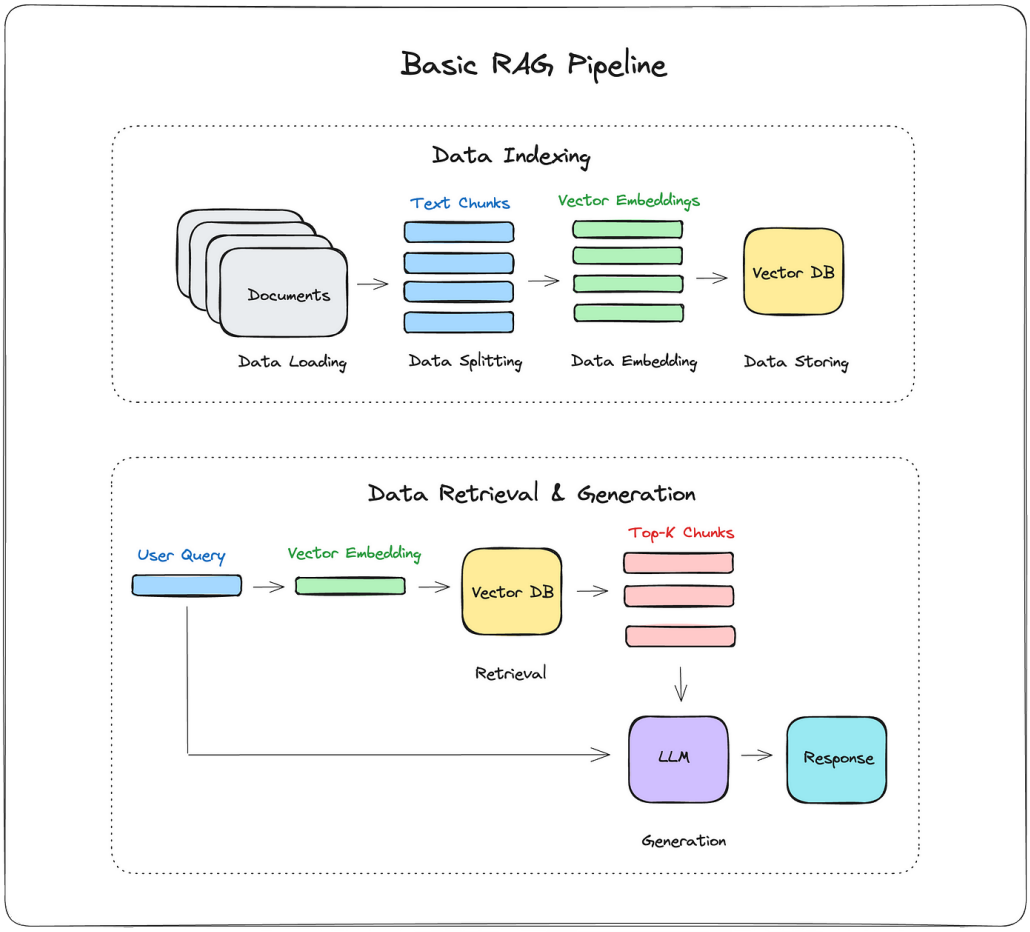

Figure 1: RAG Pipeline

8 Program Flow

1. Define directories for the PDF file and vector indices.
2. Convert text into numerical vectors for semantic search.
3. Check if FAISS index exists:
    - If it exists, load it.
    - Else, load the PDF using PDFPlumberLoader and split text into smaller chunks (1000 characters each with a 200 character overlap).
    - Store as embeddings.
4. Define function `retrieve_docs` to search VectorDB based on user prompt.
5. Define function `answer_question` to:
    - Merge retrieved text and use LLaMA 3.2 to generate response.
    - If no data in DB, perform Web Search.
6. Define tools:
    - FAISS Search: Searches the vector database.
    - Web Search: Fetches information from DuckDuckGo.
7. Initialize AI agent to decide whether to use DB or Web Search.

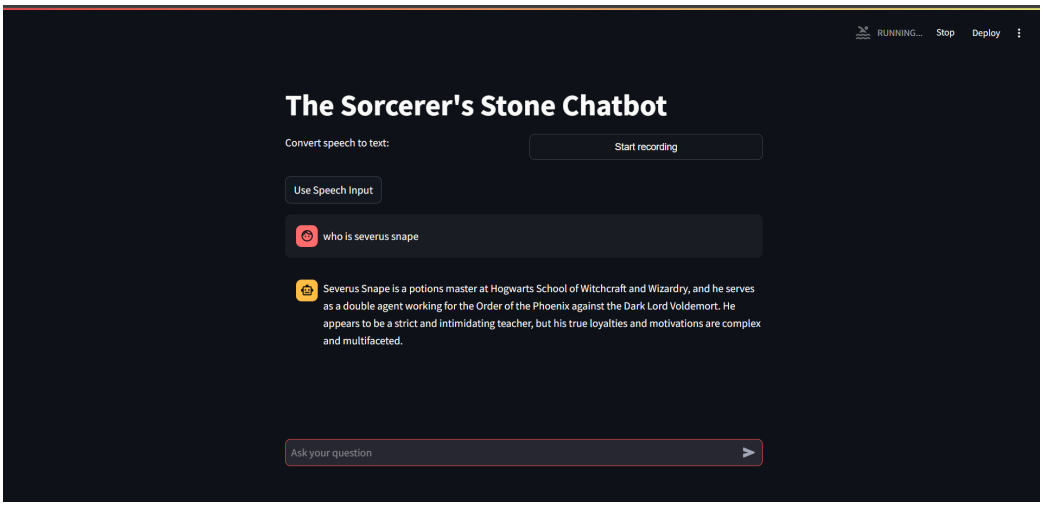

Figure 2: Using Basic RAG

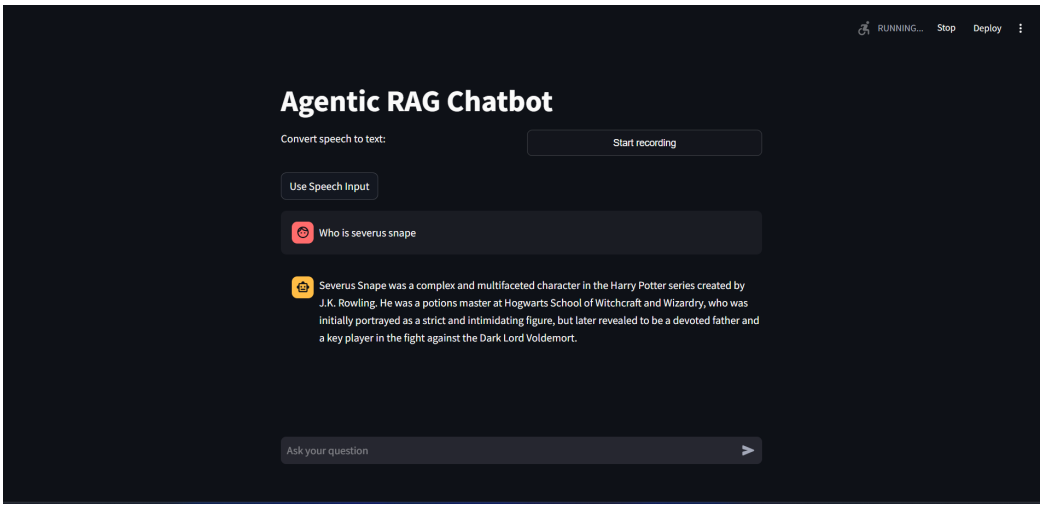

Figure 3: Using Agentic RAG