NCERT Book Chatbot Implementation

1. Choice of Model and Architecture:

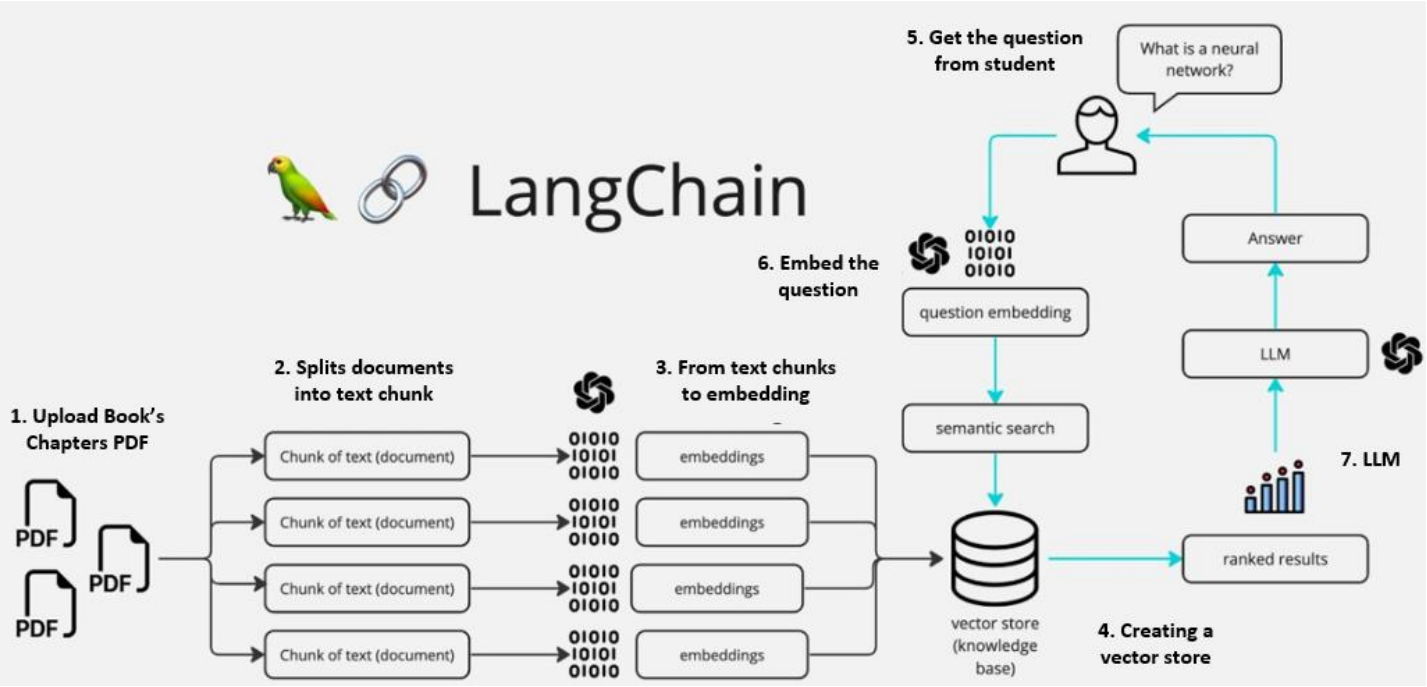

Figure 1: Project Flow

Model Choice

The RAG architecture leverages OpenAI's **GPT-4o-mini** model, which is a powerful transformer model known for its capabilities in conversational AI and question answering tasks. GPT-4o-mini is optimized for use cases where a balance between performance and cost is essential, making it an ideal choice for applications like this one, where accuracy and responsiveness are key but resource efficiency is also important.

The model is integrated with **OpenAIEmbeddings**, which is used to generate dense vector embeddings for text from the uploaded PDFs. These embeddings are essential for the vector search in the FAISS store, which retrieves relevant text chunks based on the user's query.

Architecture

The architecture follows the retrieval-augmented generation (RAG) approach, which uses external knowledge (in this case, the text from uploaded PDFs) to generate more accurate and contextually relevant answers.

Langchain is used as the underlying framework to combine the GPT-4o-mini model with tools like FAISS (for vector search) and OpenAIEmbeddings (for generating embeddings).

FAISS (Facebook AI Similarity Search) creates a vector store to index the extracted text chunks. The FAISS vector store allows for efficient and scalable similarity searches to retrieve the most relevant sections of the text in response to a user's question.

2. Evaluation Methodology & Result Analysis:

The evaluation strategy revolves around checking if the chatbot generates responses based on the curriculum content (in-context questions) and ensures that it properly responds with a default message for out-of-context questions.

NCERT Book's Chapter: https://ncert.nic.in/textbook.php?

In-Context Questions: These questions are directly related to the curriculum material and should receive responses derived from the content of the uploaded PDFs. There are 10 in-context question that has been made for test as below:

"What is the focal length of a convex lens with a radius of curvature of 20 cm?"
"Define the principal focus of a concave mirror."
"Why do convex mirrors provide a wider field of view?"
"How does a concave mirror form a real image?"
"What is the power of a lens with a focal length of 50 cm?"
"Explain the relationship between the radius of curvature and the focal length of a mirror."
"What are the two laws of reflection?"
"How is the refractive index related to the speed of light?"
"What happens to light when it moves from air to water?"
"List some uses of convex lenses."

Out-of-Context Questions: These questions are unrelated to the curriculum. The expected behavior is that the chatbot will indicate that the information is not available in the curriculum. There are 10 out-of-context question that has been made for test as below:

"Who discovered the laws of reflection?"
"What is quantum theory?"
"How does a telescope work?"
"What is the speed of light in a vacuum in miles per second?"
"Can a plane mirror form a magnified image?"
"What is the history of concave mirrors?"
"How does diffraction affect light?"
"What are the applications of fiber optics?"
"What is the cost of a convex mirror?"
"How do prisms split light into colors?"

Evaluation Metrics: The confusion matrix is calculated based on the number of correctly and incorrectly classified responses.

True Positives (TP): The number of in-context questions that the chatbot correctly classifies and provides relevant answers.

True Negatives (TN): The number of out-of-context questions that the chatbot correctly classifies and returns a message indicating the information is not in the curriculum. The chatbot correctly identified **9 out-of-context questions** as irrelevant, demonstrating that it is effective at rejecting irrelevant queries.

False Positives (FP): The number of out-of-context questions that the chatbot wrongly answers as if they were related to the curriculum. The chatbot made **2 false positive** errors, where it classified out-of-context questions as in-context.

False Negatives (FN): The number of in-context questions that the chatbot fails to answer with relevant information from the curriculum. The chatbot made **1 false negative** error, where it failed to identify an in-context question and missed providing the relevant information.

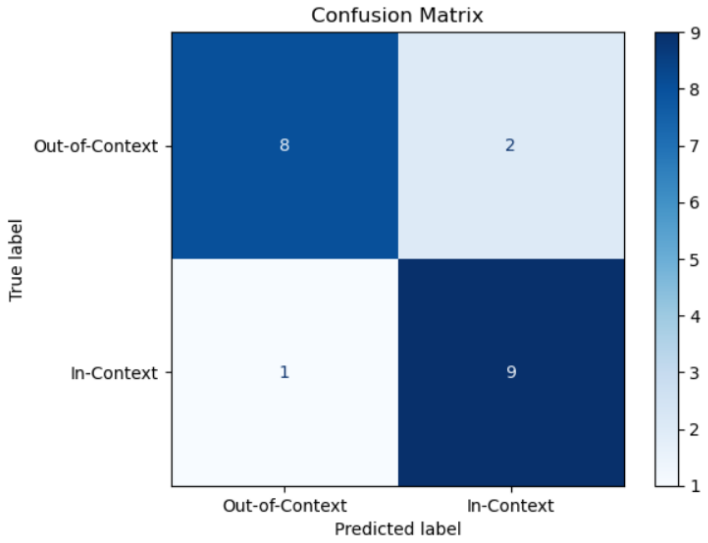

Figure 2: Confusion Matrix

F1-Score: The chatbot has an **F1-Score of approximately 84.21%**, which is a good balance between precision and recall. The chatbot is performing **reasonably well** with an overall accuracy of **85%**, but there can be more improvement, particularly in reducing false positives and increasing recall to ensure more accurate responses to in-context questions.

3. Suggestions for Improvement

Text Extraction Quality: The text extraction from PDFs is highly dependent on the quality of the PDF file itself. Using alternative libraries such as **pdfplumber** or **PyMuPDF** may improve the extraction of content from complex PDFs. For paid options, **Azure Document AI** stands out as an excellent choice, particularly for handling unstructured data, including complex and dynamic table. Chunkr.ai provides advanced tools/api for parsing and extracting data from documents, including unstructured and semi-structured formats. This solution can help further streamline text extraction, especially when dealing with documents that contain various types of data presentation.

Enhanced Contextualization: The conversation history could be further optimized to handle more complex multi-turn dialogues. This would ensure that the assistant can better refer to earlier parts of the conversation when answering new questions.

Performance with Large Datasets/ PDF: As more PDFs are uploaded, the size of the vector store increases, which may impact retrieval times. Implementing dimensionality reduction or introducing more advanced search techniques could help maintain performance with large datasets.

- Instead of the traditional chunking technique, which divides text into smaller pieces for processing, it's possible to optimize this step by using more intelligent **metadata-based filtering**. This can reduce the overhead of irrelevant chunks while maintaining high-quality results.
- **FAISS** is a popular choice for open source local vector stores, there are paid alternatives like **Pinecone** and **Weaviate** that provide optimized, scalable solutions for vector databases.

Prompt Optimization: Optimizing prompts is critical for improving the accuracy, relevance, and consistency of the generated responses in a RAG-based system.

Agentic RAG: It could autonomously fetch multiple PDFs (if they cover different chapters) and combine information from different sources, improving the depth of answers. Conversation history could be used to adjust the responses dynamically, making the assistant feel more context-aware and relevant as users ask follow-up questions.

Multimodal RAG: Incorporate more data types such as **graphs, images, and charts** along with text to make the assistant's answers more robust. For example, if the PDFs contain diagrams or educational charts, these could be included in the

answers to provide **visual explanations** alongside text-based answers, which could significantly enhance the educational value.