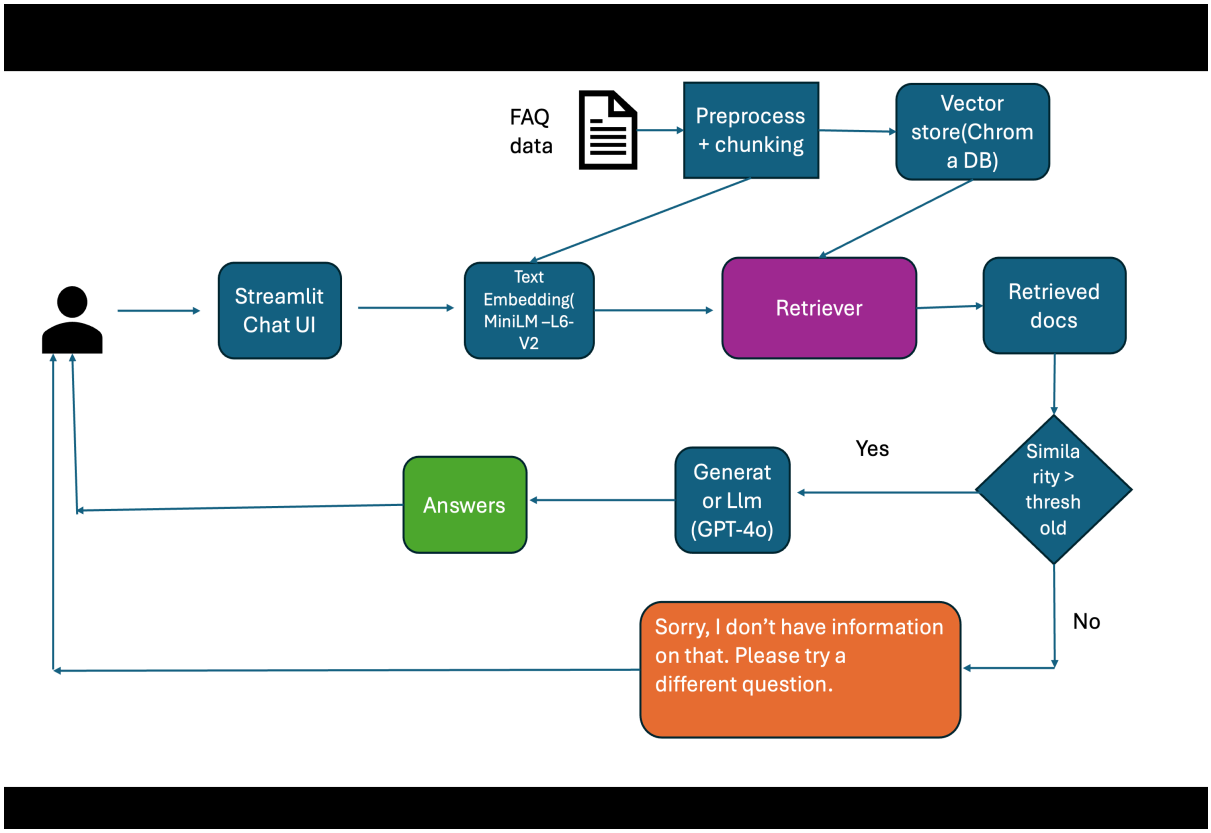
Overall Architecture

Below diagram is the overall architecture of the RAG-Llm-Chatbot.



Similarity Threshold Selection

I have chosen 0.7 as the baseline threshold for the retrieved documents. The main reasons are, it is a balanced point and has a good trade-off between relevant recall and avoiding hallucination. Also, it ensures the below points while fetching the most matching documents.

1. Responds helpfully to natural paraphrasing
2. Avoids hallucinating answers when the query is off topic
3. Maintains a good balance between recall and precision for FAQ-based RAG

Chunking Strategy

Since this dataset mainly contains question and answers with a limited length, and also in a structured way (Json format), traditional chunking strategy will not be suitable for embedding, instead full questions are used for embedding without splitting its original context. If the data is in fully unstructured format like in pdf or word docs, then we can apply the various forms of chunking before embedding.