RAG Chatbot: Technical Report

1. Description of Document Structure and Chunking Logic

Uploaded documents (PDF/TXT) are first extracted and cleaned. The text is then split into sentence-aware chunks, each containing approximately 200 words. This chunking approach ensures that each segment is semantically meaningful and suitable for embedding, while also maintaining context for downstream retrieval and generation.

2. Explanation of Embedding Model and Vector DB Used

The system uses the 'all-MiniLM-L6-v2' model from the sentence-transformers library to generate dense vector embeddings for each chunk. These embeddings are stored in a FAISS (Facebook AI Similarity Search) vector database, which enables efficient similarity search and retrieval of relevant chunks at query time.

3. Prompt Format and Generation Logic

When a user submits a query, it is embedded using the same model. The top-K most similar chunks are retrieved from the FAISS index. These chunks are concatenated and used as context in a prompt template, which is then passed to a language model (e.g., distilgpt2) for answer generation. The response is streamed to the user in real time.

4. Example Query and Response

https://drive.google.com/file/d/1Iazbi8IxASH4Fmrk_ENY-YQAMCKJ3v_4/view?usp=sharing

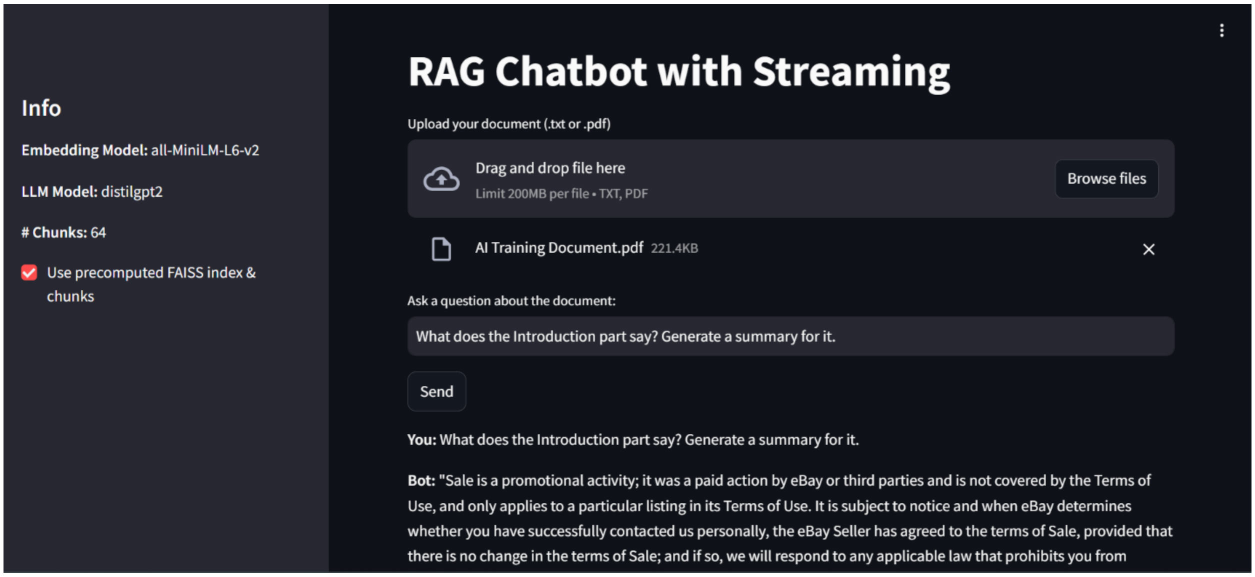
5. Notes on Hallucinations, Model Limitations, or Slow Responses

- The language model may occasionally hallucinate or generate plausible-sounding but incorrect answers, especially if the retrieved context is insufficient or ambiguous.

- The system is limited by the quality and coverage of the uploaded documents; out-of-domain queries may not be answered accurately.

- Response time may be slow for large documents or complex queries, as embedding, retrieval, and generation steps are performed sequentially.

- For best results, ensure documents are well-structured and queries are specific.



Info

Embedding Model: all-MiniLM-L6-v2

LLM Model: distilgpt2

# Chunks: 64

☑ Use precomputed FAISS index & chunks

RAG Chatbot with Streaming

Upload your document (.txt or .pdf)

Drag and drop file here
Limit 200MB per file • TXT, PDF

AI Training Document.pdf 221.4KB

Ask a question about the document:

What does the Introduction part say? Generate a summary for it.

You: What does the Introduction part say? Generate a summary for it.

Bot: "Sale is a promotional activity; it was a paid action by eBay or third parties and is not covered by the Terms of Use, and only applies to a particular listing in its Terms of Use. It is subject to notice and when eBay determines whether you have successfully contacted us personally, the eBay Seller has agreed to the terms of Sale, provided that there is no change in the terms of Sale; and if so, we will respond to any applicable law that prohibits you from