

REPORT

1. Document Structure & Chunking Logic

The uploaded PDF is first parsed using PyMuPDF and cleaned to remove line breaks and irregular spacing.

Text is tokenized into sentences using nltk.sent_tokenize() and grouped into chunks of 100–300 words to maintain semantic consistency.

Each chunk becomes a retrievable unit during user queries.

2. Embedding Model & Vector Store

We use all-MiniLM-L6-v2 from sentence-transformers to generate fixed-size dense embeddings for each chunk.

These embeddings are stored in ChromaDB, a lightweight and efficient local vector database.

Collection name: "rag_chunks"

Database folder: vectordb/

3. Prompt Format & Response Generation

For every user query, the top k=5 chunks (most relevant) are retrieved via vector similarity.

These chunks are inserted into a prompt template and streamed to a **Groq-hosted LLaMA 3 (8B)** model using OpenAI-compatible APIs.

Example:

The screenshot shows a chatbot interface titled "RAG Chatbot with Streaming". A user query is displayed: "explain me legal document in 100 words". Below the query, a retrieved context chunk is shown: "The provided legal document outlines the terms and conditions for using eBay's services. It outlines the rules and policies for buyers, including the transfer of ownership and payment terms. The document also establishes an arbitration process for resolving disputes between eBay and its users, including a mandatory informal dispute resolution process and waiver of jury trial and class actions. Additionally, the document provides for severability and opt-out procedures. Overall, the document aims to establish a clear and binding agreement between eBay and its users." At the bottom, there is a section labeled "Retrieved Context Chunks".

4. Observations & Limitations

LLM responds only with grounded context — no hallucinations observed.

Responses may stream slowly if Groq is rate-limited.

Model can't answer casual or out-of-document queries (e.g., "How are you?")

Works well with structured documents but may miss scattered info across pages.