**1. Document Structure & Chunking Logic**

**PDFs are converted to plain text using robust extraction (pypdf or PyMuPDF).**

* Noise like headers/footers is removed with regex-based filters.
* Cleaned text is **split into overlapping, sentence-aware chunks** (default: 250 characters, 50 character overlap) using LangChain’s RecursiveCharacterTextSplitter.
* Chunks are stored along with metadata indicating the source document and chunk index.

**2. Embedding Model & Vector DB**

* **Embeddings:** all-MiniLM-L6-v2 from Sentence Transformers. Chosen for its speed, modern pretraining, and accurate semantic similarity on short passages.
* **Vector Database:** ChromaDB (local persistent mode) stores [chunk, embedding, metadata, id] objects for fast cosine-similarity search. This enables rapid retrieval of relevant passages in response to user queries.

**3. Prompt Format & Generation Logic**

* Given a user query, the query is embedded via the same embedding model.
* ChromaDB retrieves the top K most relevant chunks (by default, 5).
* The prompt to the LLM follows this template:

*You are an AI assistant.  
Answer the user's question ONLY using the provided context from documents below.  
If the answer is not present, reply:  
“I could not find an answer in the documents.”*

*Context 1: [chunk 1]  
Context 2: [chunk 2]  
...  
User question: [user input]  
Answer:*

* The LLM is then invoked in streaming mode, emitting the response to the user as it is generated.

**4. Example Queries (Success & Failure Cases)**

| **User Query** | **Chatbot Response Example** | **Outcome** |
| --- | --- | --- |
| What is the conclusion of the report? | "The report concludes that..." *[from relevant passage]* | Success |
| Who are the co-authors? | "The document lists Jane Doe, John Smith as co-authors." | Success |
| What page is the bibliography on? | "I could not find an answer in the documents." | Correct Failure |
| List all survey years covered. | "The document covers surveys conducted in 2018, 2019, and 2020." | Success |
| What is the capital of France? | "I could not find an answer in the documents." | Correct Failure |

**5. Notes: Hallucinations, Limitations, and Performance**

* **Hallucinations (Fabricated Answers):**
  + Mitigated by prompt constraints; LLM instructed to only answer from seen context.
  + Still possible if chunks are ambiguous or insufficient.
* **Model Limitations:**
  + If the answer is not found in the top K chunks, it will not be retrieved (recall limited by chunking overlap/size and vector search accuracy).
  + Streaming generation slows with large LLMs on low-powered machines.
* **Performance:**
  + Mistral-7B is performant on modern GPUs (10GB+ VRAM); use TinyLlama on smaller CPUs/GPUs.
  + Initial indexing of large PDFs can take several minutes but is required only once per corpus update.