DAY – 67

6 November 2025

If you have **Ollama installed locally** (e.g., you can run ollama pull all-minilm or ollama pull mxbai-embed-large), then we can replace this part:

from langchain_huggingface import HuggingFaceEmbeddings
embeddings = HuggingFaceEmbeddings(model_name="all-MiniLM-L6-v2")

with Ollama like this:

from langchain_community.embeddings import OllamaEmbeddings
embeddings = OllamaEmbeddings(model="mxbai-embed-large")

Full Integration Notes

- It will use your **local Ollama server** (http://localhost:11434).
- Works **offline** — no API key needed.
- Faster for repeated queries if the model is already loaded.
- You can also switch to other models easily:

  embeddings = OllamaEmbeddings(model="nomic-embed-text")

Your new setup now does the following behind the scenes

What's Happening Now

1. **RAG loads only once**
   → The PDF (data/PRSC.pdf) is split into smaller 1000-character chunks and embedded **only on startup**, not every time you ask something.
   → All embeddings are saved in ./chroma_db, so even if you restart Flask, the knowledge base stays ready.

2. **Session memory works**

   → Each user (browser tab) gets a unique session_id, so their questions and context history are preserved in memory.

   → When you ask "Can you brief above?", the chatbot retrieves your previous topic (e.g., *Chakbandi Proforma*) and answers properly.

3. **Better understanding of follow-ups**

   → The prompt now explicitly handles words like "steps", "procedure", "above", and "brief", guiding the LLM to fetch *step-by-step* answers from your PDF.

4. **Punjabi–English bilingual mode**

   → The chatbot detects if you type in Punjabi (\u0A00–\u0A7F range).

   → It translates your question to English internally for accuracy, then translates the response back into Punjabi before displaying it.

5. **No accidental resets or wrong answers**

   → Because RAG isn't rebuilt each time, your answers remain consistent and context-aware across messages.

6. **Simple, stable Flask backend**

   → No Streamlit reloads, no session loss, no random resets — all requests stay linked to your ongoing chat.