pranavkumaragrawal02@gmail.com Pranav Kumar Agrawal

AI-ChatBot

1. Description of Document Structure and Chunking Logic

- The input PDF document (AI Training Document.pdf) contains approximately 10,000+ words.
- The document is loaded using PyPDFLoader from LangChain.
- Each page's content is cleaned to remove newlines and extra spaces.
- Text is split using RecursiveCharacterTextSplitter:
    - **Chunk size**: 300 characters
    - **Overlap**: 50 characters
- This ensures sentence-aware splitting with some redundancy across chunks to preserve semantic continuity.

2. Explanation of Embedding Model and Vector DB

- **Embedding Model**: all-MiniLM-L6-v2 (from sentence-transformers)
    - A lightweight transformer model ideal for semantic similarity tasks.
    - Converts each text chunk into a 384-dimensional dense vector.
- **Vector Database**: FAISS (Facebook AI Similarity Search)
    - Efficient in-memory indexing of vector embeddings.
    - Enables rapid semantic search to retrieve the most relevant chunks for any user query.

3. Prompt Format and Generation Logic

- Retrieved chunks and user query are injected into this fixed prompt template:

makefile

CopyEdit

Use the following context to answer the question.

If unknown, say "I don't know".

Context:

{context}

Question:

{question}

Answer:

- This prompt is passed to a local LLM (distilgpt2) using Hugging Face's pipeline("text-generation").
- The chatbot simulates streaming by displaying the generated response word-by-word using time.sleep() delays.

4. Example Queries and Responses

Query	Result	Response Summary
What is the objective of this training?	**Failure**	Repetitive, incoherent hallucination. Source chunks were unrelated (about arbitration).
What is Federated Learning?	**Success**	Correctly retrieved a chunk defining federated learning. Factual and concise.
What is the role of the retriever?	**Success**	Answered accurately that retriever performs semantic search and fetches top matching chunks from vector DB.
Who is the President of India?	**Success**	Correct fallback: *"I don't know."* Model did not hallucinate.
Explain Gradient Descent	**Partial**	Retrieved related content, but response lacked technical depth due to model limitations.

5. Notes on Hallucinations, Model Limitations, and Speed

- **Hallucination Risk**: Occurs when:
  - Irrelevant chunks are retrieved (legal text, disclaimers)
  - Model tries to "guess" answers when no good context is available
- **Model Weakness**:
  - distilgpt2 is a small decoder-only model not trained for instruction-following
  - Lacks coherence and may repeat phrases under low-context scenarios
- **Speed**:
  - Fast response generation locally
  - Simulated streaming using time.sleep(0.03) works smoothly
- **Suggestions**:
  - Use Mistral or OpenAI GPT models for more accurate, reliable outputs
  - Add chunk filters or score thresholds to avoid including off-topic context