Fine-Tuned RAG Chatbot with Streaming Responses

Project Name: Fine-Tuned RAG Chatbot with Streaming Responses

(Amlgo Labs - Junior AI Engineer Assignment)

Objectives: Build a smarter chatbot that can accurately respond to user queries using legal documents (such as T&C, Privacy Policy) through a fine-tuned Retrieval-Augmented Generation (RAG) pipeline and streaming responses using Streamlit.

Technologies used:

- Language Model: zephyr-7b-beta (Hugging Face)

- Embedding Model: all-MiniLM-L6-v2

-Vector Store: FAISS

- Frontend UI: Streamlit (real-time streaming as well)

- Libraries: Transformers, Sentence Transformers, LangChain, FAISS, PyPDF2

Pipeline Design & Implementation Document Chunking:

- Used PyPDF2 to extract text from a PDF.

Applied sentence-aware chunking (300 tokens with 50 overlap) by LangChain's RecursiveCharacterTextSplitter.

Embedding & Indexing:

- Took the chunks and represented them with semantic vectors with all-MiniLM-L6-v2.

- Cached the vectorized chunks with FAISS for quick retrieval.

Format:

You are an Associate. Refer to the passage below to answer the question.

Context:

Question:

Answer:

RAG Workflow:

1. Extract + chunk + embed doc

2. User inputs a query -> collect top-k relevant chunks, re-ranked ones by collecting top-k chunks using relevance search (3) from the output of FiG 2 from FAISS

3. Prompt with chunks and version 1/2/3 + query

4. Message passed to the Zephyr-7B model

5. Final answer streamed to UI

Sample Queries & Results:

Query	Answer Summary	Success
What happens in the refund process?	"But no refund for normal use..."	Yes
What about private information?	"User data is encrypted and is not shared..."	Yes
About uploaded content ownership?	"User still owns User Objects"	Yes

| Can I erase my information? | | “Yes, you can delete your account at any time.” | Yes |

| What is the AI model? | | Hallucination occurred? not in paper | No |

Known Limitations:

- Sometimes the LLM may hallucinate if the context is not sufficient.

- Large model requires >16 GB RAM to work properly.

- OCR is not supported yet on scanned images or PDFs.

Warning/error usually appears when:

1. PyTorch tries to inspect or instantiate a C++ torch.classes extension class.
2. You're not using any custom C++ operator, so this message appears unnecessarily.
3. You’re loading models like **Zephyr, Mistral, and LLaMA**, which might internally look for optional TorchScript classes.

Conclusion:

This chatbot also serves as a successful demonstration of employing the RAG pipeline with open-source tools and LLMs. The tool is composed of modular components and is accessible via a user-friendly Streamlit UI.