Technical Report: Legal RAG Chatbot

1. Document Structure & Chunking

The chatbot is built on a single user agreement PDF document. The preprocessing script (located at notebooks/preprocess.py) uses LangChain's RecursiveCharacterTextSplitter to chunk the document into 300-word segments with 50-word overlap.
This allows semantic retrieval while preserving context. The chunked segments are stored in chunks/chunked_docs.txt for reference.

2. Embeddings & Vector Database

The project uses the HuggingFace embedding model 'all-MiniLM-L6-v2' for converting text chunks into dense vectors. These vectors are indexed using the FAISS vector store, which enables fast semantic similarity searches. The FAISS index is stored locally under vectordb/ as index.faiss and index.pkl.

3. Prompt Format & Generation Logic

The chatbot uses a Retrieval-Augmented Generation (RAG) approach. When a user submits a question, the top-k relevant chunks are retrieved and injected into a prompt using the following template:

Prompt Template:
Answer the following question using the provided context.

Context: {retrieved_chunks}

Question: {user_query}

The LLM used is google/flan-t5-small, running via the HuggingFace transformers pipeline for text2text generation.

4. Sample Queries & Responses

- Q: Who am I entering into contract with when using eBay?
  A: eBay Inc. (or the respective regional entity listed in the User Agreement).

- Q: What happens if a seller doesn't ship the item?
  A: The user agreement allows eBay to take actions including limiting the seller's account and refunding the buyer.

- Q: What is eBay's responsibility regarding listed vehicles?
  A: eBay is not responsible for vehicle inspections, ownership transfer, or condition.

- Q: What are eBay's policies on international shipping?
  A: eBay offers global shipping via EIS and requires compliance with local laws.

- Q: Can I cancel an order as a buyer?
  A: Buyers can request cancellations, but sellers have 3 days to accept or decline the request.

5. Limitations & Observations

- The FLAN-T5-small model is lightweight and fast, but not ideal for complex legal understanding.
- The chatbot can hallucinate if the retrieved context is weak or irrelevant.
- FAISS does not perform re-ranking; chunk quality affects relevance significantly.
- Streaming responses are simulated through Streamlit updates, not true token-level streaming.