AI Legal Chatbot – RAG Pipeline Report

Candidate: Yash Jain
Role Applied: Junior AI Engineer
Company: Amlgo Labs
Submission Date: July 2025

1.Document Structure & Chunking Logic

The base document used was a **legal Terms & Conditions file** with approximately 10,000+ words. The structure included:

Header sections (e.g., Introduction, Definitions, User Responsibilities)

Subsections (e.g., Payment Terms, Privacy Policy, Contact Info)

Chunking Logic:

Used sentence-aware chunking via `nltk` and `textwrap`.

Each chunk was ~100–300 words.

Overlapping context was **not added** to keep chunking simple and independent.

Output was stored in `chunks.pkl` and later converted to embeddings.

2. Embedding Model & Vector DB Used

Component	Tool/Model Used	Why?
Embeddings	all-MiniLM-L6-v2 from sentence-transformers	Lightweight, fast, 384-dim vector
Vector DB	FAISS (Flat index)	Simple, fast, no external server

Process:

Each chunk was converted to a vector using the embedding model.

All vectors were indexed using FAISS.

At runtime, cosine similarity search was done to retrieve top-3 matching chunks.

3. Prompt Format & Answer Generation

Used google/flan-t5-base, an instruction-tuned, open-source LLM that runs on CPU.

Prompt Format:

php-template
CopyEdit
Context:
<retrieved chunk 1>
<retrieved chunk 2>
<retrieved chunk 3>

Question:<user query>

Answer:

Generation Logic:

Query matched against FAISS to fetch relevant chunks.

Chunks + user question were inserted into prompt.

Response was generated using generate() method from transformers.

Token limit set to max_new_tokens=256.

4. Query Examples (Success & Failure Cases)

Query 1:*"Where is my data stored?"*

Response: Your data is stored on secure servers in accordance with our Privacy Policy.
Success: Chunk about data storage was retrieved and grounded response was generated.

Query 2: *"What are the user responsibilities?"*

Response: Users must provide accurate information and avoid misuse of the platform.
Success: Captured the "User Obligations" section of the document correctly.

Query 3: *"Can I return a product after 15 days?"*

Response: Returns are only valid within 7 days of purchase.
Success with Limitation: Retrieved a related section even though the document was not about products.

Query 4: *"How can I reset my password?"*

Response: The document does not contain password-related policies.
Failure/Neutral: No matching chunks found — model guessed.

Query 5: *"Is there a live chat support option?"*

Response: Please refer to our support section for assistance.
Hallucination: No such section existed; model made up a general answer.

5.Limitations & Observations

Hallucinations:

When chunks are not retrieved correctly or context is missing, the model **hallucinates** generic answers.

This happened mostly when queries were **outside scope** of the provided document.

Latency:

flan-t5-base works on CPU, so **response time is ~1–2 seconds**.

Can be improved using GPU or smaller models (flan-t5-small or quantized versions).

Gated Models:

Attempted mistralai/Mistral-7B-Instruct initially, but was inaccessible due to Hugging Face permissions.

Summary

Feature	Status
RAG Pipeline Implemented	Yes
Semantic Search with FAISS	Yes
Streaming Response in UI	Yes
Open-source LLM used	Yes (flan-t5-base)

Feature	Status
Hallucinations Handled	Partially
Submission-Ready	Yes

Final Note:
This project is complete and meets the requirements of the Amlgo Labs assignment. Suggestions for next steps include multi-document ingestion, chat history, and support for Hindi queries.