# Fine-Tuned RAG Chatbot with Streaming Responses – Amlgo Labs

## Candidate: Rajesh Kumar Dogra

**Objective:**

To build an AI chatbot that can answer questions based on a long policy document using a Retrieval-AugmentedGeneration (RAG) pipeline with real-time streaming responses.

**Technologies:-**

- LangChain, FAISS, Streamlit, HuggingFace Transformers

- Embeddings: sentence-transformers/all-MiniLM-L6-v2

- Vector DB: FAISS

- UI: Streamlit with token-by-token streaming

**Document Processing:-**

-PDF loaded using PyPDFLoader

- Chunked using RecursiveCharacterTextSplitter (300 tokens, 50 overlap)

- Embeddings generated and stored in FAISS

**Folder Structure:-**

- /data – raw PDF file

- /chunks – processed segments

- /vectordb – saved vector store

- /src – model and pipeline scripts

- app.py – main Streamlit application

- requirements.txt, README.md

# RAG Architecture, Prompting, and Streamlit UI

## RAG Pipeline:-

- Vector store is queried with top-k semantic search (k=3)

- Results passed to prompt for LLM response generation

## Model:-

- LLM: google/flan-t5-base via HuggingFacePipeline

## - Prompt Template:-

"You are an assistant answering questions based only on the context below. If the answer is not in the context, say 'I don't know'."

## Streamlit Features:-

- Chat interface with message memory
- Real-time token-by-token response streaming
- Expandable section for source references
- Sidebar displays model name and document stats

**Response Example (Prompt + Context Injected):**

**User Question:** "What is eBay’s return policy?"

**Context:** Retrieved chunks...

**Answer:** "Returns depend on seller's policy, typically within 30 days."

# Evaluation: Success & Failure Cases, Observations

**Successful Queries:**

1. Q: What is eBay's return policy?

A: eBay allows returns based on seller-defined timelines.

2. Q: Who owns user data?

A: eBay retains some rights, but user owns the content.

3. Q: How long are eBay records stored?

A: Usually retained up to 7 years.

**Failure/Hallucination Cases:**

4. Q: What is refund percentage for damaged products?

A: Incorrect guess; not present in context.

5. Q: Does eBay allow crypto payments?

A: Incorrect 'yes'; document has no such mention.

**Model Limitations:-**

- Initial slow response due to model warm-up
- Hallucinations when context insufficient
- Limited reasoning for ambiguous or multi-hop questions

**Suggestions:-**

- Add user feedback loop
- Optimize token generation
- Add support for multiple documents in pipeline