RAG-Based Chatbot with Zephyr – Project Report

1. Project Overview

This project implements a Retrieval-Augmented Generation (RAG) pipeline to build an AI-powered chatbot capable of answering questions from a document. The solution integrates:

- Document ingestion and chunking
- Semantic search using embeddings and FAISS
- Response generation using the **`HuggingFaceH4/zephyr-7b-beta`** model
- Real-time response streaming via a **Streamlit app**

---

2. Document Structure & Chunking Logic

Input Document

- Source: PDF document
- Content: eBay User Agreement and related policies

Preprocessing Steps

- Extracted text using **PyPDF2**
- Cleaned by removing newlines, tabs, and unnecessary whitespaces
- Text split into logical **sentence-based chunks** (~200 words) using **NLTK's sentence tokenizer**

Output

- **Total Chunks**: ~65
- Stored in: `chunks/chunks.json`

---

3. Embedding Model & Vector Database

Embedding Model

- **Model**: `all-MiniLM-L6-v2` (Sentence-Transformers)
- **Vector Size**: 384
- Chosen for its speed, lightweight architecture, and strong performance on semantic similarity tasks

Vector Database

- **Library**: FAISS (Facebook AI Similarity Search)
- **Index Type**: `IndexFlatL2` (efficient L2 distance-based retrieval)
- All embeddings saved in: `vectordb/index.faiss` and `vectordb/index.pkl`

---

4. Prompt Format & Response Generation Logic

LLM Used

- **Model**: `HuggingFaceH4/zephyr-7b-beta`
- **Platform**: Hugging Face Transformers with local inference

Prompt Format

```
Answer the user's question based on the context below.

Context:
[Top Retrieved Chunks]

Question: [User Query]

Answer:
```

Generation Parameters

- **Max tokens**: 512
- **Temperature**: 0.2 (ensures factual, less creative responses)
- **Streaming**: Implemented via `TextIteratorStreamer` for real-time token generation

---

5. Example Queries, Responses & Evaluation

	User Query	AI Response
1	What is ebay?	Ebay is a marketplace...
2	What is the main purpose of eBay’s User Agreement?	The main purpose of eBay's User Agreement is to set out the terms and conditions governing access to...
3	What is the eBay International Shipping Program (EIS)?	The eBay International Shipping Program (EIS) is a service offered by...

---

6. Observations & Limitations

Success Cases

- Accurate answers when the question is covered in the document
- Real-time streaming improves user interaction

- Works well for policy documents and FAQs

Limitations & Failures

- **Hallucinations**: Occasional factual inaccuracies when no context is available
- **Speed**: Zephyr 7B is heavy; initial load is slow on low-end machines
- **Verbose Answers**: May over-generate content for short queries

---

7. Conclusion

This project demonstrates the effectiveness of RAG pipelines combined with instruction-tuned open-source LLMs for document-based Q&A. It showcases:

- Seamless ingestion, chunking, and embedding of large documents
- Fast and accurate similarity search with FAISS
- Grounded and coherent response generation using Zephyr
- A lightweight and interactive UI with Streamlit

Prepared By: *Kunal Kumar*
Date: *12th july 2025*