RAG Chatbot Project Report

Author: Saksham
Date: July 13, 2025

1. Introduction to the Retrieval-Augmented Generation (RAG) Chatbot

This report details the implementation and functionality of a Retrieval-Augmented Generation (RAG) chatbot designed to answer user queries based on a specific set of documents. The RAG architecture combines the strengths of information retrieval systems with the generative capabilities of large language models (LLMs), ensuring that responses are grounded in factual information from the provided corpus, thereby reducing hallucinations common in pure generative models.

The chatbot leverages LangChain for orchestrating the RAG pipeline, Sentence Transformers for creating document embeddings, FAISS for efficient vector similarity search, and Streamlit for an interactive web-based user interface.

2. Document Structure and Chunking Logic

The chatbot operates on a corpus of text documents. For this project, the primary document used for demonstration is an "eBay User Agreement."

Document Structure:
The raw input documents are expected to reside in the data/ directory of the project. While specific file types (e.g., .txt, .pdf) can be supported through appropriate LangChain document loaders, the core processing focuses on their textual content.
Chunking Logic:
To effectively retrieve relevant information, large documents are broken down into smaller, manageable segments called "chunks." This process is handled within the 1_document_processing_and_embedding.ipynb notebook.
The chunking strategy typically involves:

- **Text Splitting:** Using a RecursiveCharacterTextSplitter (or similar from LangChain's text_splitter module) to divide the raw text into smaller pieces. This splitter attempts to split by paragraphs, then sentences, then words, ensuring that semantic units are kept together as much as possible.
- **Chunk Size and Overlap:** Chunks are defined with a specific chunk_size (e.g., 500 characters) and chunk_overlap (e.g., 50 characters). Overlap helps maintain context between chunks, preventing loss of information at chunk boundaries.

This chunking process ensures that when a user query is made, the retrieval system can pinpoint and extract highly relevant, smaller pieces of information rather than

entire large documents, which might contain a lot of irrelevant content.

3. Embedding Model and Vector Database

For the RAG system to find relevant document chunks, it needs a way to understand the semantic meaning of both the user's query and the document content. This is achieved through embedding models and a vector database.

Embedding Model:

- **Model Used:** SentenceTransformerEmbeddings from langchain_community.embeddings.
- **Specific Model ID:** "sentence-transformers/all-MiniLM-L6-v2".
- **Purpose:** This model converts text (both document chunks and user queries) into high-dimensional numerical vectors (embeddings). Texts with similar meanings will have vectors that are closer to each other in this high-dimensional space.

Vector Database (Vector DB):

- **Database Used:** FAISS (Facebook AI Similarity Search).
- **Purpose:** FAISS is an open-source library for efficient similarity search and clustering of dense vectors. After document chunks are converted into embeddings, these embeddings are stored in a FAISS index. When a user submits a query, its embedding is used to quickly find the most similar (i.e., most relevant) document chunk embeddings in the FAISS index.
- **Storage:** The FAISS index is saved locally to the ./vectordb/faiss_index directory, consisting of index.faiss (the vector data) and index.pkl (metadata and mappings).
- **Loading:** The DocumentRetriever class in src/retriever.py is responsible for loading this pre-built FAISS index and the embedding model.

4. Prompt Format and Generation Logic

The generation component of the RAG chatbot is handled by the ResponseGenerator class in src/generator.py. This component takes the user's question and the retrieved context, constructs a prompt, and sends it to a Large Language Model (LLM) to generate a coherent answer.

LLM Used:

- **Model:** "mistralai/Mistral-7B-Instruct-v0.2"
- **Integration:** Integrated via langchain_community.llms.Ollama, assuming a local Ollama server is running and the mistral model has been pulled. This allows for local, privacy-preserving LLM inference.

Prompt Format:
A ChatPromptTemplate is used to structure the input to the LLM. This template guides the LLM on its role and how to use the provided information.

```
prompt = ChatPromptTemplate.from_messages([
    ("system", "You are a helpful AI assistant. Use the following context to answer the question. If the answer is not in the context, state that you don't know."),
    ("human", "Context: {context}\nQuestion: {question}")
])
```

- **System Message:** Sets the persona and instructions for the AI, emphasizing the use of provided context and handling out-of-context queries.
- **Human Message:** Contains placeholders for the context (retrieved document chunks) and the question (user's query).

Generation Logic:
The generate_response method in ResponseGenerator creates a LangChain "chain" by piping the prompt, the llm instance, and a StrOutputParser. This chain takes the question and context, formats them according to the prompt, sends them to the LLM, and then parses the LLM's raw output into a simple string response.

5. Example Queries and Responses

Here are example interactions with the RAG chatbot, demonstrating its ability to answer questions based on the provided eBay User Agreement document excerpts.

Document Excerpts Used (Simplified for brevity):

- **Chunk 1:** "User Agreement ... The entity you are contracting with is: eBay Inc., 2025 Hamilton Ave., San Jose, CA 95125, if you reside in the United States; eBay (UK) Limited, 1 More London Place, London, SE1 2AF, United Kingdom, if you reside in the United Kingdom; eBay GmbH, Albert-Einstein-Ring 2-6, 14532 Kleinmachnow, Germany, if you reside in the European Union; eBay Canada Limited, 240 Richmond Street West, 2nd Floor Suite 02-100, Toronto, ON, M5V 1V6, Canada, if you reside in Canada; eBay Singapore Services Private Limited, 1 Raffles Quay, #18- 00, Singapore 048583, if you reside in India;"
- **Chunk 2:** "Singapore Services Private Limited, 1 Raffles Quay, #18- 00, Singapore 048583, if you reside in India; and eBay Marketplaces GmbH, Helvetiastrasse 15/17, CH-3005, Bern, Switzerland, if you reside in any other country. In this User Agreement, these entities are individually and collectively referred to as "eBay," "we," or "us." Read this User Agreement carefully as it contains provisions that govern how claims you and we have against each other are resolved (see "Disclaimer of Warranties; Limitation of Liability" and "Legal Disputes" provisions

below). It also contains an Agreement to Arbitrate which will, with limited exception, require you to submit claims you have against us or related third parties to binding and final arbitration, unless you opt out of the Agreement to Arbitrate in accordance with section 19.B.9 (see Legal Disputes, Section B ("Agreement to Arbitrate")). If you do not opt out: (1) you will only be permitted to pursue claims against us or related third parties on an individual basis, not as a plaintiff"

- **Chunk 3:** "permitted to pursue claims against us or related third parties on an individual basis, not as a plaintiff or class member in any class or representative action or proceeding; (2) you will only be permitted to seek relief (including monetary, injunctive, and declaratory relief) on an individual basis; and (3) you are waiving your right to pursue disputes or claims and seek relief in a court of law and to have a jury trial. 2. About eBay eBay is a marketplace that allows users to offer, sell, and buy goods and services in various geographic locations using a variety of pricing formats. eBay is not a party to contracts for sale between third-party sellers and buyers, nor is eBay a traditional auctioneer. Any guidance eBay provides as part of our Services, such as pricing, shipping, listing, and sourcing is solely informational and you may decide to follow it or not. We may use artificial intelligence or AI-powered tools and products to provide and improve our Services, to offer you a customized and personalized experience, to provide you with enhanced customer service, and to support fraud"
- **Chunk 4:** "powered tools and products to provide and improve our Services, to offer you a customized and personalized experience, to provide you with enhanced customer service, and to support fraud detection; availability and accuracy of these tools are not guaranteed. We may help facilitate the resolution of disputes between buyers and sellers through various programs. Unless otherwise expressly provided, eBay has no control over and does not guarantee: the existence, quality, safety, or legality of items advertised; the truth or accuracy of users' content or listings; the ability of sellers to sell items; the ability of buyers to pay for items; or that a buyer or seller will actually complete a transaction or return an item. 3. Using eBay In connection with using or accessing our Services you agree to comply with this User Agreement, our policies, our terms, and all applicable laws, rules, and regulations, and you will not: • breach or circumvent any laws, regulations, third-party rights or our systems, Services, policies, or determinations of your account status; • use our Services if you are not able to form legally binding contracts (for example, if you are"

Example Query 1 (Success Case)

Query: "What is this document about?"

Retrieved Context (Simulated):

User Agreement
1. Introduction
This User Agreement, the Mobile Application Terms of Use, and all policies and additional terms
posted on and in our sites, applications, tools, and services (collectively "Services") set out the terms
on which eBay offers you access to and use of our Services. You can find an overview of our policies
here. The Mobile Application Terms of Use, all policies, and additional terms posted on and in our
Services are incorporated into this User Agreement. You agree to comply with all terms of this User
Agreement when accessing or using our Services.
The entity you are contracting with is: eBay Inc., 2025 Hamilton Ave., San Jose, CA 95125, if you
reside in the United States; eBay (UK) Limited, 1 More London Place, London, SE1 2AF, United
Kingdom, if you reside in the United Kingdom; eBay GmbH, Albert-Einstein-Ring 2-6, 14532
Kleinmachnow, Germany, if you reside in the European Union; eBay Canada Limited, 240 Richmond
Street West, 2nd Floor Suite 02-100, Toronto, ON, M5V 1V6, Canada, if you reside in Canada; eBay
Singapore Services Private Limited, 1 Raffles Quay, #18- 00, Singapore 048583, if you reside in India;...
Singapore Services Private Limited, 1 Raffles Quay, #18- 00, Singapore 048583, if you reside in India;
and eBay Marketplaces GmbH, Helvetiastrasse 15/17, CH-3005, Bern, Switzerland, if you reside in
any other country. In this User Agreement, these entities are individually and collectively referred to
as "eBay," "we," or "us."
If you reside in India and you register for our Services, you further agree to the eBay.in User
Agreement.

Read this User Agreement carefully as it contains provisions that govern how claims you and we have
against each other are resolved (see "Disclaimer of Warranties; Limitation of Liability" and "Legal
Disputes" provisions below). It also contains an Agreement to Arbitrate which will, with limited
exception, require you to submit claims you have against us or related third parties to binding and
final arbitration, unless you opt out of the Agreement to Arbitrate in accordance with section 19.B.9
(see Legal Disputes, Section B ("Agreement to Arbitrate")). If you do not opt out: (1) you will only be
permitted to pursue claims against us or related third parties on an individual basis, not as a plaintiff...

Response:
This document is primarily about the eBay User Agreement, outlining the terms and conditions for accessing and using eBay's Services. It covers various policies, defines the contracting eBay entities based on geographic location, and details legal provisions regarding claims and arbitration.

Example Query 2 (Success Case)

Query: "Where is eBay Inc. located for users in the United States?"

Retrieved Context (Simulated):

User Agreement
1. Introduction
This User Agreement, the Mobile Application Terms of Use, and all policies and additional terms
posted on and in our sites, applications, tools, and services (collectively "Services") set out the terms
on which eBay offers you access to and use of our Services. You can find an overview of our policies
here. The Mobile Application Terms of Use, all policies, and additional terms posted on and in our
Services are incorporated into this User Agreement. You agree to comply with all terms of this User
Agreement when accessing or using our Services.
The entity you are contracting with is: eBay Inc., 2025 Hamilton Ave., San Jose, CA

95125, if you
reside in the United States; eBay (UK) Limited, 1 More London Place, London, SE1 2AF, United
Kingdom, if you reside in the United Kingdom; eBay GmbH, Albert-Einstein-Ring 2-6, 14532
Kleinmachnow, Germany, if you reside in the European Union; eBay Canada Limited, 240 Richmond
Street West, 2nd Floor Suite 02-100, Toronto, ON, M5V 1V6, Canada, if you reside in Canada; eBay
Singapore Services Private Limited, 1 Raffles Quay, #18- 00, Singapore 048583, if you reside in India;...

Response:
For users residing in the United States, eBay Inc. is located at 2025 Hamilton Ave., San Jose, CA 95125.

Example Query 3 (Success Case)

Query: "Does eBay guarantee the quality of items advertised?"

Retrieved Context (Simulated):

powered tools and products to provide and improve our Services, to offer you a customized and
personalized experience, to provide you with enhanced customer service, and to support fraud
detection; availability and accuracy of these tools are not guaranteed. We may help facilitate the
resolution of disputes between buyers and sellers through various programs. Unless otherwise
expressly provided, eBay has no control over and does not guarantee: the existence, quality, safety,
or legality of items advertised; the truth or accuracy of users' content or listings; the ability of sellers
to sell items; the ability of buyers to pay for items; or that a buyer or seller will actually complete a
transaction or return an item. 3. Using eBay In connection with using or accessing our Services you
agree to comply with this User Agreement, our policies, our terms, and all applicable laws, rules, and
regulations, and you will not: • breach or circumvent any laws, regulations, third-party

rights or our
systems, Services, policies, or determinations of your account status; • use our Services if you are not
able to form legally binding contracts (for example, if you are...

Response:
No, eBay explicitly states that, unless otherwise expressly provided, it does not guarantee the existence, quality, safety, or legality of items advertised on its marketplace.

Example Query 4 (Failure Case - Out of Context)

Query: "What is the weather like in London today?"

Retrieved Context (Simulated):

User Agreement
1. Introduction
This User Agreement, the Mobile Application Terms of Use, and all policies and additional terms
posted on and in our sites, applications, tools, and services (collectively "Services") set out the terms
on which eBay offers you access to and use of our Services. You can find an overview of our policies
here. The Mobile Application Terms of Use, all policies, and additional terms posted on and in our
Services are incorporated into this User Agreement. You agree to comply with all terms of this User
Agreement when accessing or using our Services.
The entity you are contracting with is: eBay Inc., 2025 Hamilton Ave., San Jose, CA 95125, if you
reside in the United States; eBay (UK) Limited, 1 More London Place, London, SE1 2AF, United
Kingdom, if you reside in the United Kingdom; eBay GmbH, Albert-Einstein-Ring 2-6, 14532
Kleinmachnow, Germany, if you reside in the European Union; eBay Canada Limited, 240 Richmond
Street West, 2nd Floor Suite 02-100, Toronto, ON, M5V 1V6, Canada, if you reside in Canada; eBay
Singapore Services Private Limited, 1 Raffles Quay, #18- 00, Singapore 048583, if you reside in India;...

Response:
I'm sorry, I can only provide information based on the provided document excerpts. The document does not contain information about the weather in London.

6. Notes on Hallucinations, Model Limitations, and Performance

Hallucinations

The RAG architecture significantly reduces the likelihood of hallucinations (LLMs generating factually incorrect or unsupported information). By forcing the LLM to answer *only* from the retrieved context, the chatbot is less likely to "make things up." As demonstrated in Example Query 4, when the answer is not in the context, the model is prompted to state that it doesn't know, rather than fabricating a response.

Model Limitations

- **Context Window:** The effectiveness of the RAG system is limited by the context window of the underlying LLM. If the answer requires synthesizing information from many chunks that exceed the LLM's context window, the quality of the response may degrade.
- **Retrieval Quality:** The quality of the generated response is directly dependent on the quality of the retrieved documents. If the embedding model fails to retrieve truly relevant chunks, or if the relevant information is not present in the document corpus, the LLM will not be able to provide an accurate answer.
- **Chunking Strategy:** The chosen chunking strategy can impact retrieval. Too small chunks might lose context, while too large chunks might dilute relevance.
- **LLM Capabilities:** Even with perfect context, the LLM's inherent reasoning and summarization abilities play a role. A less capable LLM might struggle to synthesize complex information even if it's provided in the context.

Slow Responses

- **Embedding Model Loading:** The sentence-transformers/all-MiniLM-L6-v2 model needs to be downloaded and loaded into memory on its first use. This can cause a noticeable delay during the initial startup of the Streamlit application or the first query. Subsequent queries benefit from caching.
- **FAISS Index Loading:** Similarly, loading the FAISS index (especially for very large datasets) can contribute to initial startup time.
- **LLM Inference Time:** The primary factor for query response time is the LLM inference speed.
    - **Local LLMs (like Ollama/Mistral):** Performance depends heavily on the local hardware (CPU, RAM, GPU if available). Larger models require more resources

and will be slower.
    - **API-based LLMs (like OpenAI GPT):** Response times depend on API latency, network speed, and the LLM provider's server load.
- **Network Latency:** Downloading models (Hugging Face) or making API calls (OpenAI) introduces network latency.

Performance Improvements & Future Work

- **Optimized Chunking:** Experiment with different RecursiveCharacterTextSplitter parameters or other chunking strategies (e.g., semantic chunking).
- **Advanced Retrieval:** Explore more sophisticated retrieval methods beyond basic similarity search, such as HyDE (Hypothetical Document Embeddings) or RAG-Fusion.
- **LLM Optimization:** Quantization of local LLMs or using smaller, more efficient models can improve inference speed.
- **Asynchronous Operations:** For Streamlit, implementing asynchronous loading and processing can improve perceived responsiveness.
- **Error Handling and User Feedback:** Enhance the Streamlit UI to provide more detailed feedback on loading status, errors, and potential solutions.

This RAG chatbot provides a robust foundation for building grounded question-answering systems, with clear avenues for future optimization and expansion.