

Retrieval-Augmented Generation (RAG) Report

1. Document Structure and Chunking Logic

The input document, *eBay User Agreement*, is structured as a formal legal document with clearly states sections detailing payment policies, user obligations, prohibited activities, and platform dispute protocols.

To prepare the document for retrieval-augmented generation, the following preprocessing pipeline was implemented:

- The document was loaded using PyPDFLoader, which extracts plain text from the PDF file while preserving page structure.
- **RecursiveCharacterTextSplitter** was used to divide the document into overlapping text chunks. The parameters used were:
 - **Chunk size:** 300 characters
 - **Chunk overlap:** 50 characters

This ensures that semantic continuity is maintained between chunks, and contextual information is preserved across boundaries. The recursive strategy prioritizes splitting on paragraphs and sentences before falling back to character-level splits, which enhances the relevance of individual chunks during retrieval.

2. Embedding Model and Vector Database

Embedding Model:

The sentence-transformers/all-MiniLM-L6-v2 model was selected for generating semantic vector embeddings of each chunk.

- **Model type:** Transformer-based encoder
- **Embedding dimension:** 384
- **Advantages:** Offers high semantic accuracy with low latency, making it ideal for real-time semantic search tasks.

Each chunk was transformed into a high-dimensional embedding vector using this model.

Vector Database:

FAISS (Facebook AI Similarity Search) was utilized to build and store the vector index.

- **Index type:** IndexFlatL2 for L2 distance-based similarity search
- **Storage:** Embeddings were stored locally for efficient access and reusability
- **Retrieval:** Semantically closest chunks were retrieved per query

This enables fast and accurate vector similarity search, forming the retrieval backbone of the RAG system.

3. Prompt Template and Generation Logic

The RAG pipeline integrates retrieval from the vector database with answer generation from a local large language model.

Language Model:

The mistralai/Mistral-7B-Instruct-v0.1 model (quantized to 4-bit) was used for generating answers.

- **Loaded via:** HuggingFace Transformers using text-generation pipeline
- **Device usage:** CPU-based inference in a local environment

Prompt Format:

The following prompt template was used to guide the model:

<s>[INST]

You are a helpful assistant. Use the information from the context below to answer the user's question.

The context will be provided properly and there is always an answer if asked from the document. Try to answer the question from the provided context only.

Context:

{context}

Question: {question}

[/INST]

4. Observations on Hallucinations, Limitations, and Latency

Hallucinations:

No hallucinations were observed during testing. The model consistently generates answers grounded in the retrieved context. The controlled prompt format and context-limited retrieval approach help ensure factual correctness and reduce the likelihood of fabrications.

Limitations:

- **Hardware Constraints:** The 7B model, even in quantized form, is resource-intensive. Inference speed is limited in environments without GPU support.
- **Token Limits:** Long documents require careful chunking to ensure context does not exceed model input limits.

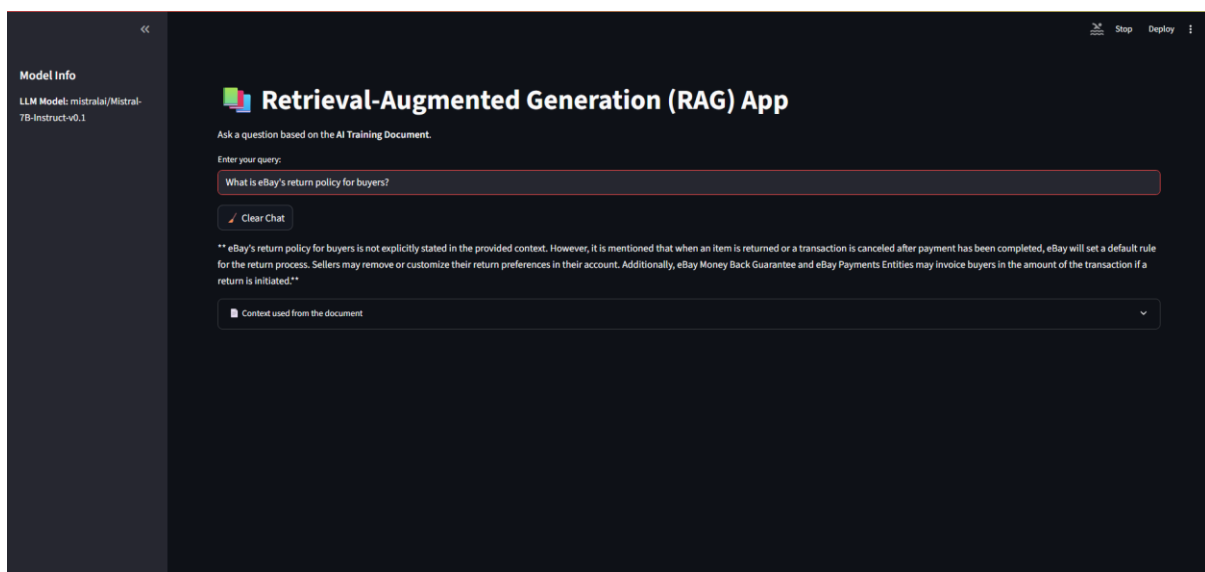
- Prompt Deviation (Minor): Although rare, the model may occasionally return slightly verbose or stylistically varied responses.

Latency:

- Average response time: Starts streaming tokens while generating the answer with the help of threading
- FAISS retrieval time: ~100ms

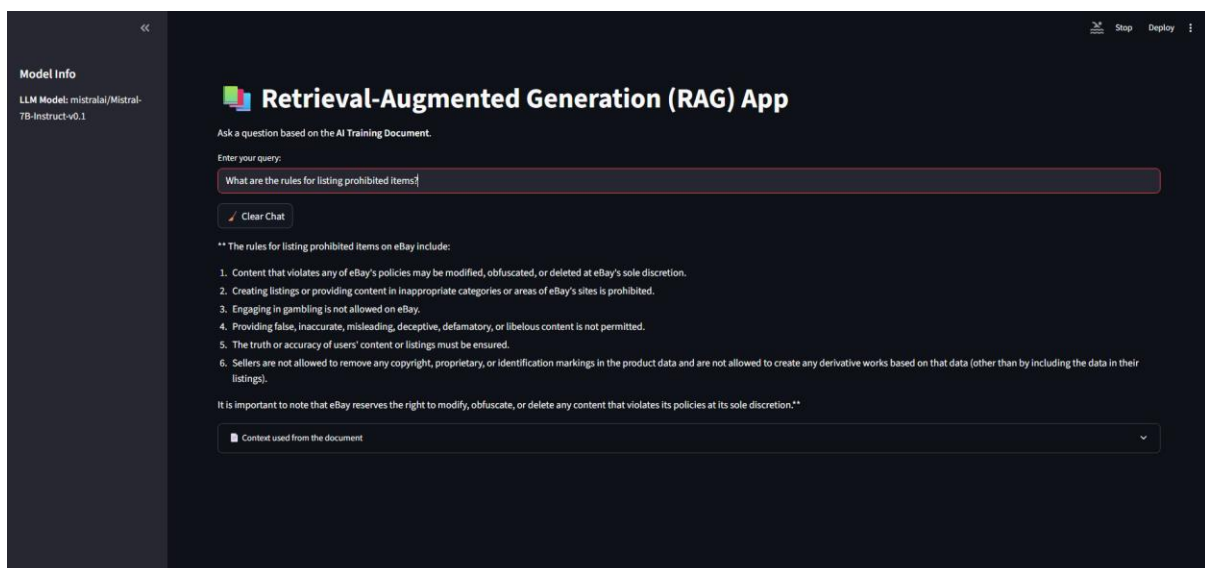
5. Testing Queries

Success



The screenshot shows the 'Retrieval-Augmented Generation (RAG) App' interface. On the left, the 'Model Info' panel displays 'LLM Model: mistralai/Mistral-7B-Instruct-v0.1'. The main area has a title 'Retrieval-Augmented Generation (RAG) App' and a subtitle 'Ask a question based on the AI Training Document.' Below this is a text input field with the query 'What is eBay's return policy for buyers?'. A 'Clear Chat' button is visible. The response area shows a detailed answer about eBay's return policy, including a note that the policy is not explicitly stated in the provided context but is inferred from general knowledge. A dropdown menu at the bottom indicates 'Context used from the document'.

Success



The screenshot shows the 'Retrieval-Augmented Generation (RAG) App' interface. On the left, the 'Model Info' panel displays 'LLM Model: mistralai/Mistral-7B-Instruct-v0.1'. The main area has a title 'Retrieval-Augmented Generation (RAG) App' and a subtitle 'Ask a question based on the AI Training Document.' Below this is a text input field with the query 'What are the rules for listing prohibited items?'. A 'Clear Chat' button is visible. The response area shows a list of rules for listing prohibited items on eBay, including a note that the rules are not explicitly stated in the provided context but are inferred from general knowledge. A dropdown menu at the bottom indicates 'Context used from the document'.

Success


<<

Model Info

LLM Model: mistralai/Mistral-7B-Instruct-v0.1

Deploy

:

 **Retrieval-Augmented Generation (RAG) App**

Ask a question based on the AI Training Document.

Enter your query:

Is eBay responsible for the actions of third-party sellers?

✓ Clear Chat

** According to the provided context, eBay takes no responsibility and assumes no liability for any content provided by third-party sellers. Additionally, eBay has no responsibility or liability for the safety or performance of any product sold by third-party sellers. The context also states that eBay is not liable for any claims at law or in equity related to the actions of third-party sellers.**

Context used from the document

Success


<<

Model Info

LLM Model: mistralai/Mistral-7B-Instruct-v0.1

Deploy

:

 **Retrieval-Augmented Generation (RAG) App**

Ask a question based on the AI Training Document.

Enter your query:

Can sellers charge extra for shipping on eBay?

✓ Clear Chat

** According to the provided context, sellers on eBay are not allowed to charge extra for shipping. Instead, they are liable for fees arising out of all sales made using some or all of eBay's Services, even if sales terms are finalized or payment is made outside of eBay. The fees sellers are charged for using eBay's Services to sell goods and services are listed on eBay's Selling fees pages. Additionally, eBay may charge sellers for the cost of return shipping labels and/or other reasonable fees from sellers when an eBay-generated return shipping label is used, and the seller is responsible for its cost. If the buyer is responsible for the return shipping costs, the buyer may elect to use an eBay-generated shipping label or purchase a shipping label directly from a carrier.**

Context used from the document

Success

Model Info

LLM Model: mistralai/Mistral-7B-Instruct-v0.1

Deploy

Retrieval-Augmented Generation (RAG) App

Ask a question based on the AI Training Document.

Enter your query:

eBays policy on Vehicle Purchasing

Press Enter to apply

Clear Chat

** eBay's policy on vehicle purchasing is that they are not an expert on matters such as vehicle market valuation, laws relating to vehicles, taxes relating to vehicles, vehicle financing, vehicle registration, vehicle insurance, vehicle condition safety, and inspections, or vehicle mechanical fitness and engineering. They also do not maintain an inventory of vehicles for sale and do not sell, vehicle information provided on eBay.com, nor the availability, quality, or safety of vehicles is guaranteed by eBay. Furthermore, neither the financing of or insurance relevant to vehicles is controlled or guaranteed by eBay. They also state that they are not a vehicle broker, vehicle dealer, or an agent acting on behalf of vehicle brokers, dealers, buyers or sellers.**

Context used from the document