RAG Chatbot Project - Evaluation Report

1. Document Structure and Chunking Logic

The source document used in this project was a legal agreement provided by eBay. It consisted of structured sections outlining responsibilities, limitations, and procedures for dispute resolution. To prepare it for retrieval, the text was cleaned and split into sentence-aware chunks between 100 to 300 words. This allowed for meaningful semantic indexing while maintaining context for accurate responses.

2. Embedding Model and Vector Database

The sentence-transformers model 'all-MiniLM-L6-v2' was used to generate embeddings for each chunk. It offers a balance of speed and semantic accuracy, ideal for real-time applications. These vectors were stored using FAISS, a high-performance similarity search library. This enables fast retrieval of relevant document segments based on user queries.

3. Prompt Format and Generation Logic

A concise prompt template was created to combine user input and the top-k retrieved chunks. This structured format encourages the language model to provide grounded, context-aware answers and to avoid hallucinations when information is missing from the source.

4. Evaluation and Examples

The chatbot was tested with a variety of factual and ambiguous queries. Examples include:
- 'Who is the contracting entity in the UK?' -> Correctly retrieved from the document.
- 'Does eBay offer 24/7 phone support?' -> Returned a cautious answer noting lack of evidence in context.
- 'Does eBay support AI-generated listings?' -> Identified that the document does not contain such information.
- 'What are the penalties for listing counterfeit items?' -> Avoided hallucination by stating that no explicit penalties are defined in the source.

The chatbot showed strong performance in maintaining factual consistency and handling edge cases.

5. Limitations and Recommendations

The system depends heavily on embedding quality and chunk granularity. Broad or vague queries may not retrieve meaningful context, resulting in fallback or generic answers. Additionally, response latency could be optimized by caching frequent queries and preloading model components. Future versions can benefit from more advanced rerankers or hybrid retrieval techniques.

Two hallucination-prone questions tested:
- 'What is eBay's refund policy for cryptocurrency transactions?'
- 'What are the penalties for listing counterfeit luxury items?'

In both cases, the chatbot correctly responded that the provided document did not contain relevant information, demonstrating that it avoided hallucinating unsupported answers.