Evvo Technology Solutions Pvt Ltd

Technical Challenge – LLM Engineers

Assumptions and Design considered to set up the H-RAG

Assumptions

1. **Document Availability**: Relevant, structured documents are available and can be broken into meaningful chunks.
2. **Clear Queries**: Users will ask clear, specific questions that can be answered using the document data or language model knowledge.
3. **Embedding Quality**: Generated embeddings accurately represent document content, enabling effective search in the vector store.
4. **System Scalability**: The system can handle large document volumes and high query frequency.
5. **Relevance of Data**: Documents in the system are pertinent to user queries.

Design Considerations

1. **Two-Layer Hierarchy**:
   - **Layer 1**: Stores summaries or keywords to quickly narrow down relevant sections.
   - **Layer 2**: Holds detailed document chunks for in-depth answers.
2. **Efficient Chunking**: Optimize chunk sizes to balance context preservation and search efficiency.
3. **Model Selection**: Use Mistral for embedding and rephrasing; fallback to web search if the answer isn't in the vector store.
4. **Performance Optimization**:
   - Use FAISS for fast embedding search.
   - Add a re-ranking mechanism for accuracy.
5. **Relevance Feedback**: Allow user feedback to improve system accuracy.
6. **Scalability**: Design for easy updates as documents grow, with minimal impact on system performance.
7. **Fallback to Internet Search**: If vector store search fails, use web search as a last resort.

Embedding used for the vector store

Embedding Used

- **Model**: **Mistral**
- **Type**: Dense vector embeddings
- **Source**: Generated using the Mistral model in Ollama.

Rationale for Selection

1. **Accuracy and Contextual Understanding**: Mistral embeddings effectively capture the semantic context, making them well-suited for retrieving relevant content based on user queries.
2. **Consistency in Representation**: Dense vectors offer a uniform structure, which is beneficial for calculating similarity metrics in FAISS-based vector stores.
3. **High Retrieval Performance**: Mistral embeddings provide reliable search accuracy for complex queries, ensuring high relevance of retrieved results.
4. **Scalability**: Dense embeddings maintain consistent performance even as the dataset grows, supporting scalability in a multi-layered system.
5. **Flexibility for Hierarchical RAG**: The embeddings work well in a two-layer system, supporting both keyword-based and detailed retrieval as needed.

Test questions used to evaluate the performance of the H-RAG system

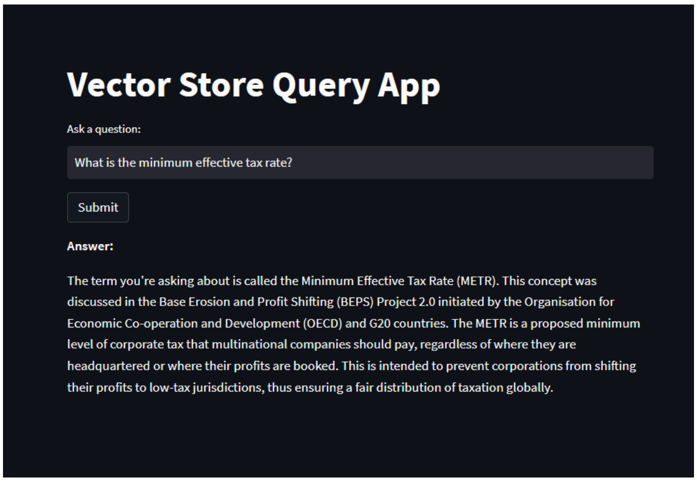
To evaluate the effectiveness and relevance of the H-RAG system, the following test questions were

used:

1. **Basic Definition Queries:**
   - Example: "What is the minimum effective tax rate?"
   - Purpose: Assess the system's ability to retrieve direct answers from the vector store.
2. **Contextual Information Queries:**
   - Example: "Explain BEPS in the context of Singapore."
   - Purpose: Evaluate the system's understanding of nuanced questions and its ability to provide context-specific information.
3. **Comparative and Analytical Queries:**
   - Example: "How does the Singapore tax policy compare to other regions under BEPS guidelines?"
   - Purpose: Test the H-RAG's capability to retrieve and combine multiple relevant chunks to form a detailed response.
4. **Keyword and Summary-Based Queries:**
   - Example: "List the key provisions under Pillar Two of BEPS."
   - Purpose: Assess the effectiveness of the keyword-based layer in the hierarchical structure.
5. **Complex, Multi-Part Queries:**

- Example: "How does the H-RAG system enhance tax transparency and prevent base erosion?"
- Purpose: Test the system's ability to understand and respond accurately to multi-faceted queries.

For example image:



Vector Store Query App

Ask a question:

What is the minimum effective tax rate?

Answer:

The term you're asking about is called the Minimum Effective Tax Rate (METR). This concept was discussed in the Base Erosion and Profit Shifting (BEPS) Project 2.0 initiated by the Organisation for Economic Co-operation and Development (OECD) and G20 countries. The METR is a proposed minimum level of corporate tax that multinational companies should pay, regardless of where they are headquartered or where their profits are booked. This is intended to prevent corporations from shifting their profits to low-tax jurisdictions, thus ensuring a fair distribution of taxation globally.

Agent Design and Test Questions for Evaluation

Agent Design: The decision-making agent is structured to intelligently choose the best method for responding to user queries based on query type and relevance:

1. RAG System Response:

- The agent first queries the vector store to retrieve relevant information based on the embeddings from the hierarchical RAG system.
- If relevant content is found, it formulates a comprehensive answer using retrieved chunks.

2. Direct Answer from Knowledge Model (Mistral):

- If the query's relevance to stored content is low, the agent uses a language model (Mistral) to generate a response based on its built-in knowledge.

3. Internet Search for Unavailable Information:

- For queries outside the scope of the vector store and language model, the agent performs an internet search to retrieve up-to-date and detailed information.