Evvo Technology Solutions Pvt Ltd

Technical Challenge – LLM Engineers

General Instructions

- Use only Mistral and phi3 for this assessment. **Do not use Langchain or LlamaIndex.**
- Place all codes in a private Gitlab repository.

Question 1: Implement a Hierarchical Retrieval-Augmented Generation (RAG)

Build a Hierarchical RAG system using the provided dataset. The first vector store layer should contain the keywords/summaries linked to the actual documents stored in the second layer.

Note:

1. State clearly the assumptions and design considerations in setting up the H-RAG
2. State clearly the embedding used for the vector store, including the rationale for their selection.
3. List the test questions used to evaluate the performance of the H-RAG system.

Question 2: Design a Decision-Making Agent

Create an agent that determines the best approach to handle a query, using the RAG system above, answering directly, or fetching information from the internet.

Note:

1. State clearly the test questions used to evaluate the system.

Choose one of the following options (3a or 3b)

Question 3a: Deploy the Model and Demonstrate Functionality

Write the scripts to deploy the above system on CPU using llama.cpp and with vllm on GPU.

Question 3b: Implement Guardrails

Develop the guardrails to address hallucination and off-topic questions (tax related) and responses.

Additional Question: Fine-Tuning

Fine-tune phi3 model on a dataset of your choice using the unsloth library. Track and monitor the training process using Wandb.

Note:

1. Clearly specific the dataset used for fine-tuning.
2. Explicitly detail the fine-tuning method and all parameters used.