Problem Statement

Introduction to GenAI and Simple LLM Inference on CPU and finetuning of LLM Model to create a Custom Chatbot.

GenAI refers to General Artificial Intelligence, a field focused on creating AI systems that can perform a wide range of tasks similar to human intelligence. Simple LLM (Large Language Model) Inference on CPU refers to leveraging language models like T5 (Text-To-Text Transfer Transformer) on Central Processing Units (CPUs) for tasks such as text summarization.

Unique Idea Brief (Solution)

Your unique idea involves fine-tuning the T5 model for text summarization. This involves adapting a pre-trained T5 model to specifically generate concise summaries of given documents, which is crucial for applications like news aggregation, document understanding, and more.

1. Understanding T5 Model and Text Summarization

- **T5 Model Overview:** Begin with an overview of the T5 model architecture, emphasizing its ability to perform various NLP tasks via text-to-text transformation.
- **Text Summarization:** Explain the task of text summarization and its importance in extracting essential information from lengthy documents.

Unique Idea Brief (Solution)

2. Dataset Preparation

- **Data Collection:** Describe the process of collecting relevant datasets suitable for training the T5 model for text summarization.
- **Preprocessing:** Detail the steps involved in preprocessing the raw text data to ensure compatibility with the T5 model input format.

3. Model Fine-Tuning

- **Fine-Tuning Strategy:** Outline the methodology used to fine-tune the T5 model for the text summarization task. This includes parameter tuning, learning rate scheduling, and batch size optimization.
- **Training Process:** Discuss the training process, highlighting key metrics monitored during training (e.g., loss function, convergence criteria).

4. Evaluation Metrics

- **Evaluation Criteria:** Define the metrics used to evaluate the performance of the fine-tuned model, such as ROUGE scores (Recall-Oriented Understudy for Gisting Evaluation) for summarization quality assessment.

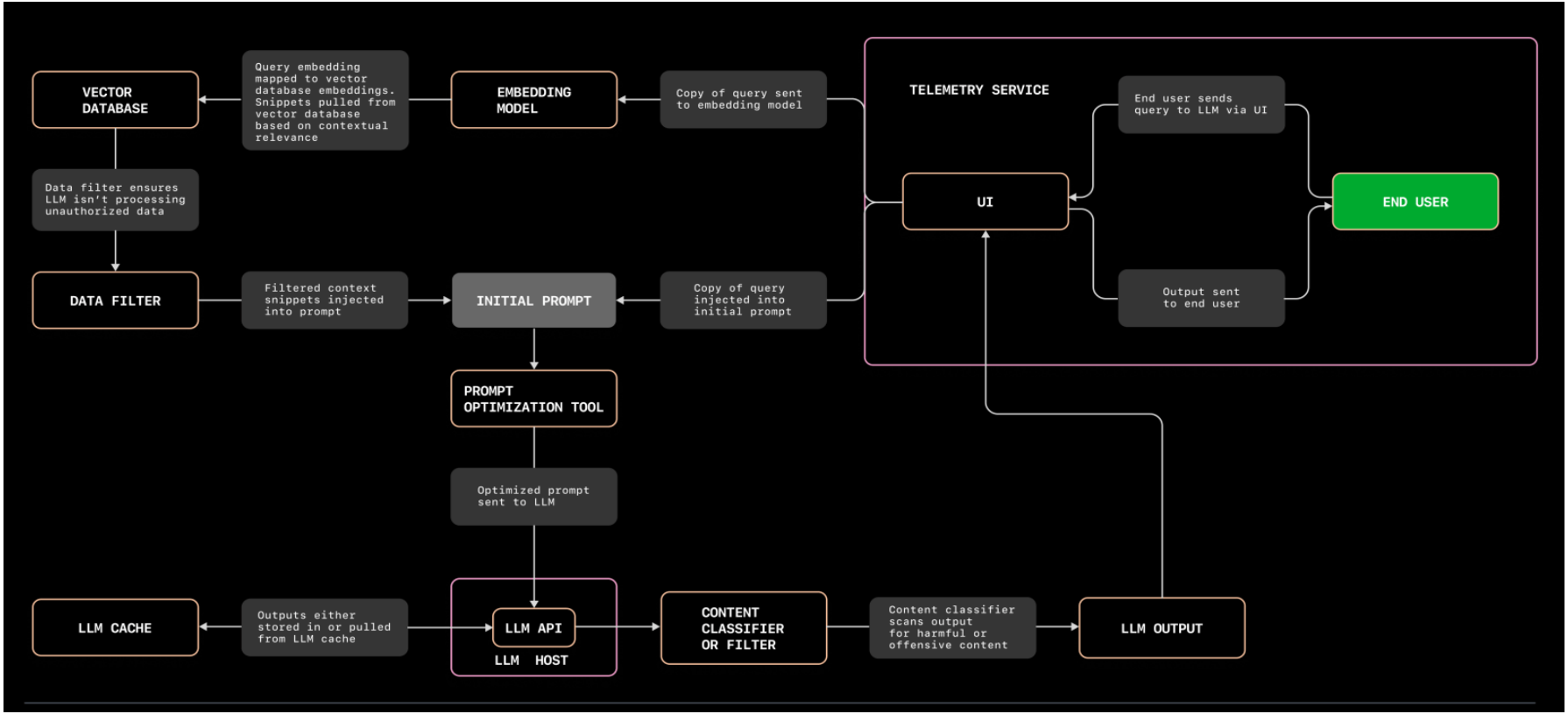
Features Offered

- **Text Summarization**: The ability to generate summaries of input articles from the CNN/Daily Mail dataset.
- **GPU/CPU Compatibility**: Flexibility to run on both GPU and CPU depending on availability.
- **Training and Evaluation**: Structured setup for training and evaluating the summarization model.
- **Model and Tokenizer Saving**: Capability to save both the trained model and the tokenizer for future use.

Process flow

- **Data Loading**: Load the CNN/Daily Mail dataset.
- **Tokenization**: Tokenize the dataset for input to the T5 model.
- **Model Preparation**: Load the pre-trained T5 model and define necessary training configurations.
- **Training**: Fine-tune the T5 model on the summarization task using Seq2SeqTrainer.
- **Saving**: Save the trained model and tokenizer for deployment and future use.

Architecture Diagram



Technologies used

- **Python Libraries**: `transformers`, `datasets` (for managing datasets), `accelerate` (for efficient training).
- **Machine Learning Framework**: PyTorch (for training the T5 model).

Team members and contribution:

SAARANSH GUPTA 22BAC10027 Project Developer

Conclusion

In conclusion, by leveraging the T5 model through fine-tuning, you've created a robust solution for text summarization, capable of operating on both GPU and CPU platforms. This approach not only demonstrates effective use of modern AI techniques but also provides a scalable solution for various text summarization applications.