Problem Statement

Running GenAI on Intel AI Laptops and Simple LLM Inference on CPU and fine-tuning of LLM Models using Intel® OpenVINO™

Unique Idea Brief (Solution)

Our goal is to create a chatbot (we named it as SoulCare) that provides mental health support to patients. By fine-tuning a large language model (LLM), we have optimized it to understand and respond effectively to mental health-related queries. Our solution ensures efficient performance and a user-friendly experience, making mental health support more accessible.

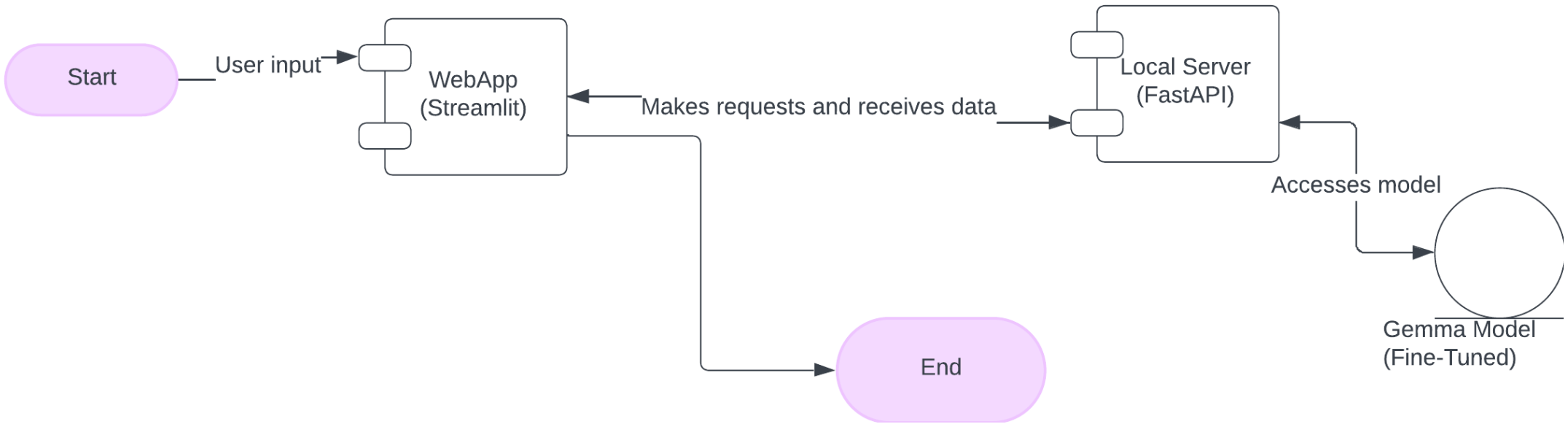
Features Offered

- **Fine-Tuned Mental Health Model:** Customized using specialized datasets to address mental health queries with empathy and accuracy.
- **Efficient Inference:** Leveraged OpenVINO for optimized performance on Intel CPUs to ensure quick and efficient responses.
- **Local Deployment with FastAPI:** Utilized FastAPI to deploy the chatbot server locally.
- **Interactive Web Interface:** Streamlit-based web app provides a user-friendly and intuitive platform for interacting with the chatbot.

Process Flow

1. Data Collection and Preprocessing:
   - Collected mental health-related from various HuggingFace repositories.
   - Preprocessed the data to ensure it is suitable for fine-tuning the model.
2. Model Fine-Tuning:
   - Fine-tuned the LLM model using the preprocessed data to specialize it for mental health support.
3. Model Selection:
   - Trained Gemma-2b and Llama 2 and based on performance of the model after fine tuning, Gemma-2b model is selected.
4. Model Optimization:
   - Converted the fine-tuned model to OpenVINO format for optimized inference performance on Intel CPUs.
5. Local Deployment:
   - Deployed the optimized model locally using FastAPI to serve the chatbot on localhost at port 8000.
6. Web Interface Development:
   - Developed a user-friendly web application using Streamlit to provide an intuitive interface for users to interact with the chatbot.

Architecture Diagram



Technologies Used

- **HuggingFace:** Used for accessing various repositories and datasets.
- **Google Gemma 2B:** The base model that is fine-tuned to provide specialized mental health support.
- **OpenVINO:** Optimizes the model for efficient inference on Intel CPUs, enhancing performance.
- **FastAPI:** Deploys the model locally, providing a robust and efficient server for the chatbot.
- **Streamlit:** Develops a user-friendly web application for easy interaction with the chatbot.
- **Python:** The primary programming language used for data preprocessing, model fine-tuning, and server and web app development.

Team members and contribution

Jefi Ryan (Team Leader):

- Collected data from HuggingFace and preprocessed and created a separate HuggingFace repository containing preprocessed data.
- Fine-tuned Gemma-2b model with the preprocessed data for 4000 steps (chosen based on hardware limitations) using PEFT and LoRA methods for parameter-efficient fine-tuning and did 4-bit quantization with bits and bytes.
- Converted the fine-tuned model to OpenVINO model.
- Created a local server for LLM inference with FastAPI.
- Developed a webapp with Streamlit for user-friendly inference.

Jason Jacob (Team Member):

- Fine-tuned Llama 2 model with preprocessed data using PEFT and LoRA methods for parameter-efficient fine-tuning and did 4-bit quantization with bits and bytes.

Meenakshi Sundaram (Team Member):

- Attempted to integrate LangChain for managing chat history, with the aim of making the AI responses more contextually aware and coherent. Although the feature did not fully work out as intended, it provided significant learning and paved the way for future improvements. This experience underscored the importance of context in AI-driven conversations and highlighted areas for further development and refinement in our project.

Conclusion

SoulCare, our mental health chatbot, demonstrates a practical application of AI in mental health support. By fine-tuning a powerful language model and optimizing it for efficient performance, we have developed a chatbot for users seeking mental health assistance. Currently, the model is deployed locally using FastAPI, and an intuitive web interface created with Streamlit ensures that SoulCare delivers an effective and user-friendly experience.

We extend our sincere thanks to Intel for their support and resources, which have been instrumental in bringing this project to fruition. This initiative not only demonstrates our technical expertise but also highlights the potential of AI in providing meaningful mental health support. We look forward to future improvements and further contributions to enhance this vital resource.