Problem Statement

PS-16 : 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)

1. **Use of OpenVINO:** Introduce Intel’s OpenVINO as a solution for optimizing and accelerating neural network models on Intel architecture.

2. **Integration with Hugging Face:** Detail how integrating OpenVINO with Hugging Face models leverages both cutting-edge AI models and advanced optimization techniques.

Features Offered

- **Model Quantization:** Explain the implementation of various model quantizations (FP16, INT8, INT4) to reduce model size and improve performance.

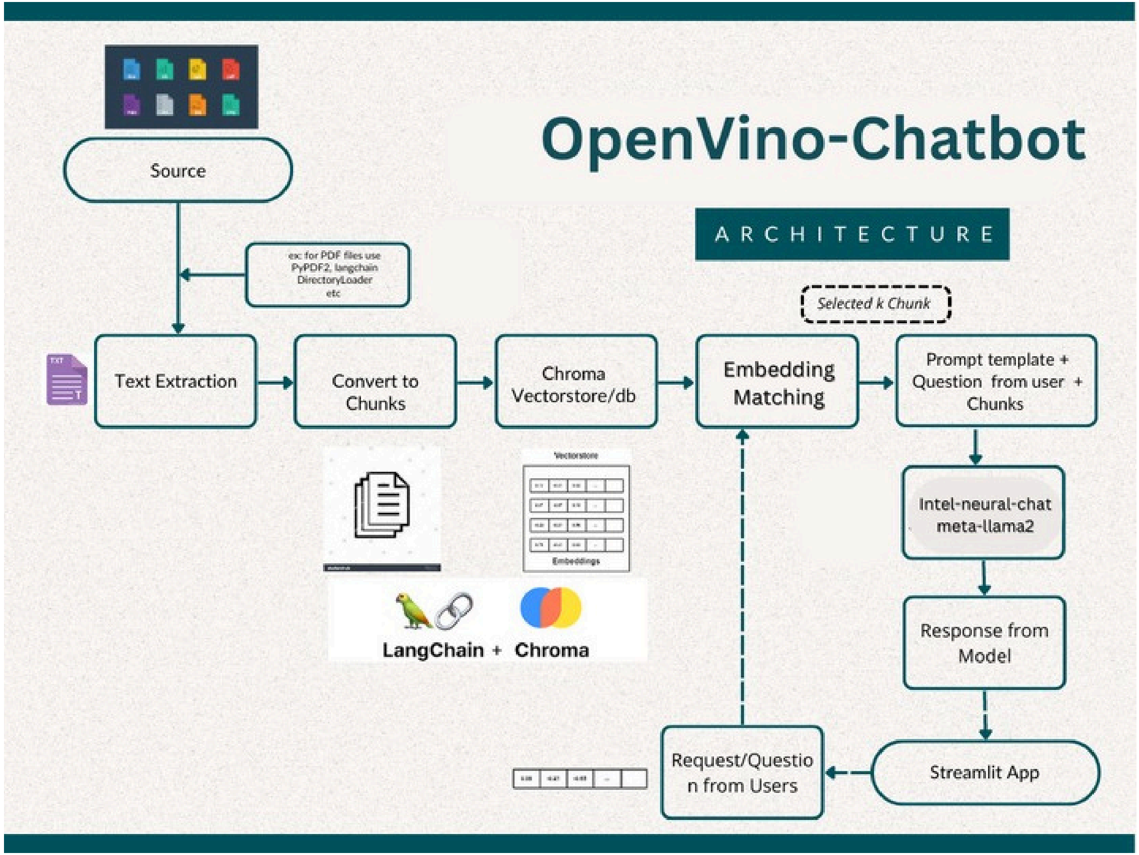
- **Client-Server Architecture:** Describe the setup of a retrieval-augmented generation (RAG) system using a client-server architecture to handle queries and compute responses efficiently

Process flow

- **From Model Download to Deployment:** Outline the process flow starting from model downloading, optimization, setup in the server environment, and interaction through the client application.

- **Detailed Steps:** Include steps from the scripts llm-model-downloader.py for model preparation and openvino-rag-server.py for server setup.

Architecture Diagram



Technologies used

- **List of Technologies:** Enumerate technologies like OpenVINO, Hugging Face, LangChain, Chroma, FastAPI, and Streamlit.

- **Purpose and Integration:** Briefly explain the role each technology plays in the project, enhancing understanding of the system’s complexity and integration depth.

Team members and contribution:

Name : Aditya Raj

Individual project

Mentor : Dr. Shilpa Suresh

Conclusion

- **Project Achievements:** Summarize the key achievements of the project, focusing on performance improvements and efficiency gains.

- **Benefits:** Emphasize the benefits such as reduced computational requirements, faster response times, and the ability to deploy advanced AI on standard laptop configurations.