Problem Statement

Running GenAI on Intel AI Laptops and Simple LLM Inference on CPU and fine-tuning of LLM Models using Intel® OpenVINO™ for enhanced performance and efficiency on CPU-based systems.

Objective:

- Improve GenAI and LLM inference performance on Intel AI laptops.
- Utilize Intel® OpenVINO™ for effective fine-tuning of LLM models.
- Enhance efficiency and scalability on CPU-based systems.

Unique Idea Brief (Solution)

1.<u>Use Large Pre-trained LLM Model:</u>

Utilize pre-trained models like **tiny-llama-1b-chat / Phi-3-mini-4k-instruct-gguf** for enhanced NLP tasks through extensive training on diverse datasets.

2.<u>Efficient LLM Inference on CPU:</u>

Optimize LLM inference on CPUs to reduce costs and increase deployment flexibility, making AI accessible on more hardware platforms.

3.<u>Provide Effective User Interface:</u>

Design an intuitive, responsive UI for seamless interaction with LLMs, ensuring a positive user experience.

4.<u>Effective Fine-tuning of LLM Models:</u>

Fine-tune models with domain-specific data to enhance accuracy and relevance for specific applications.

Unique Idea Brief (Solution)

Solutions:-

1.Compress Models with Optimum-CLI:

Use Optimum-CLI to reduce model size, making them manageable and deployable on systems with limited resources.

3.CPU-Optimized Inference with OpenVINO:

Use Intel OpenVINO to optimize LLM inference on **Intel® CPUs for smooth and fast performance.**

4.Fine-tuning with OpenVINO Tools:

Leverage **OpenVINO tools** for **efficient fine-tuning, reducing training time and resource consumption.**

5.Model Extraction with Hugging Face:

Utilize **Hugging Face** for easy access to pre-trained models, facilitating quick integration and deployment

Features Offered

1. **Effective UI:**
   - Intuitive and user-friendly design
   - Clear navigation and real-time feedback
   - Responsive for seamless interaction

2. **Natural Language Query:**
   - Supports conversational user interaction
   - Improves accessibility for non-technical users
   - Provides accurate, contextually appropriate responses

3. **Multiple Language Support:**
   - Broadens usability and reach
   - Handles linguistic nuances and cultural contexts
   - Ensures high performance and accuracy across languages

Process flow

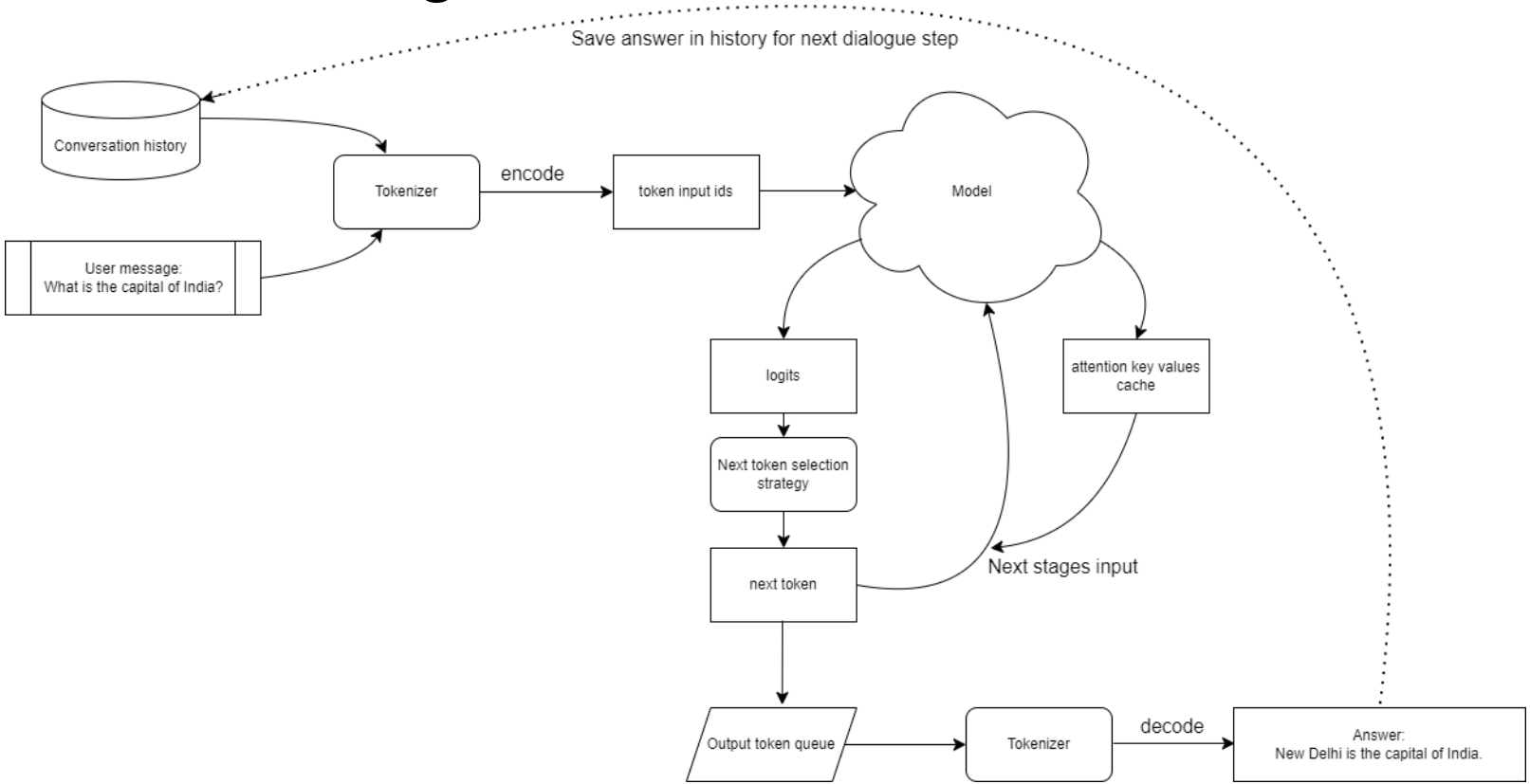
1.<u>Query Upload:</u>

- User enters a query through the effective UI.
- The system processes the natural language input, accommodating English, Chinese, or Japanese.

2.<u>Response Generation:</u>

- The system uses the optimized LLM to generate a response.
- Response is displayed to the user through the UI in the corresponding language.

Architecture Diagram



Technologies used

- **Python**: Primary programming language ,version 3.10.8
- **Hugging Face Transformers**: For accessing pre-trained LLM
- **Intel OpenVINO**: For optimization of model using inference on CPU
- **Gradio**: For creating the user interface for interaction
- **Google Colab**: Development and deployment platform, powered by Google

Team members and contribution:

Subhadip Ghosh (Project lead**):** LLM integration , model evaluation , OpenVINO optimization, instantiate model.

Jitu Pradhan:- LLM integration , model evaluation , OpenVINO optimization, instantiate model.

Akash Ghosh: Architecture design and Documentation of complete project.

Sumita Das: User interface development and integration

Conclusion

The project leverages large pre-trained language models to enhance NLP tasks, offering multilingual support (English, Chinese, and Japanese) and efficient CPU-based inference using Intel's OpenVINO toolkit. By utilizing tools like Optimum-CLI for model compression and Hugging Face for model extraction, the system ensures high performance and accessibility. An effective user interface facilitates natural language queries, enabling seamless interaction for users across different languages. The process flow from user query upload to response generation highlights the system's streamlined approach to delivering accurate and contextually relevant outputs. Overall, this project aims to provide a powerful, efficient, and user-friendly solution for natural language processing on Intel AI laptops.