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

Team Members	**Email ids**
Sakshi Yadav	sakshi.yadav16278@sakec.ac.in
Vinit Bahua	vinit.bahua16364@sakec.ac.in
Diya Joshi	diya.joshi16369@sakec.ac.in
Dhruve Rathod	dhruve.rathod16164@sakec.ac.in

Table of Contents

Topics	**Page No**
Problem Statement	2
Methodology	3
Architectural Design	4
Results	5
Conclusion	6

Table of Figures

Figures	**Page No**
Fig 1. Architecture Design	4
Fig 2. Output Size of Quantized Model	5
Fig 3. Inference Time Results	5

Problem Statement

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

Purpose:
To demonstrate the capability of running GenAI on Intel AI Laptops by implementing a simple LLM inference on CPU and fine-tuning LLM models using Intel® OpenVINO™, resulting in an efficient, user-friendly chatbot that leverages advanced AI technology while maintaining high performance and accessibility on consumer-grade hardware.

Objectives:
1. To develop a responsive and intelligent chatbot
2. To optimize the model using OpenVINO
3. To create an intuitive user interface
4. To demonstrate the potential of compact AI models in real-world applications

Methodology

Our project followed a structured approach to develop and deploy an efficient, Intel-optimized chatbot. The methodology consisted of five key phases:

1. Model Selection
   We selected the Tiny Llama 1B model. This choice was based on its optimal balance of performance and efficiency, taking into account factors such as model size, inference speed, and compatibility with Intel hardware.

2. Model Fine-tuning
   To optimize performance, we implemented INT4 quantization techniques to compress the Tiny Llama 1B model. We performed asymmetric quantization. 80% of the tensors were quantized. The tensors were quantized as a group of 128.

3. Intel OpenVINO Optimization
   Leveraging the Intel® OpenVINO™ toolkit, we further optimized our fine-tuned model. This involved converting the model to OpenVINO's Intermediate Representation (IR) format and applying specialized optimization techniques to enhance inference speed on Intel CPUs. Benchmarking was conducted to quantify the performance improvements achieved through this process.

4. User Interface Development
   Using the Gradio framework, we designed an intuitive and accessible interface for our chatbot. Key features implemented include text input functionality, sliders for adjusting model parameters, chatbot response display, and conversation history tracking. Throughout the development process, we prioritized responsiveness and smooth user experience on Intel AI Laptops.

5. Deployment
   The final phase involved integrating the OpenVINO-optimized model with our Gradio interface. We established the necessary runtime environment on Intel AI Laptops and conducted extensive testing to ensure stability and performance.

This methodical approach enabled us to create a highly efficient chatbot optimized for Intel hardware, demonstrating the potential of running advanced AI models on consumer-grade devices.

Architectural Design

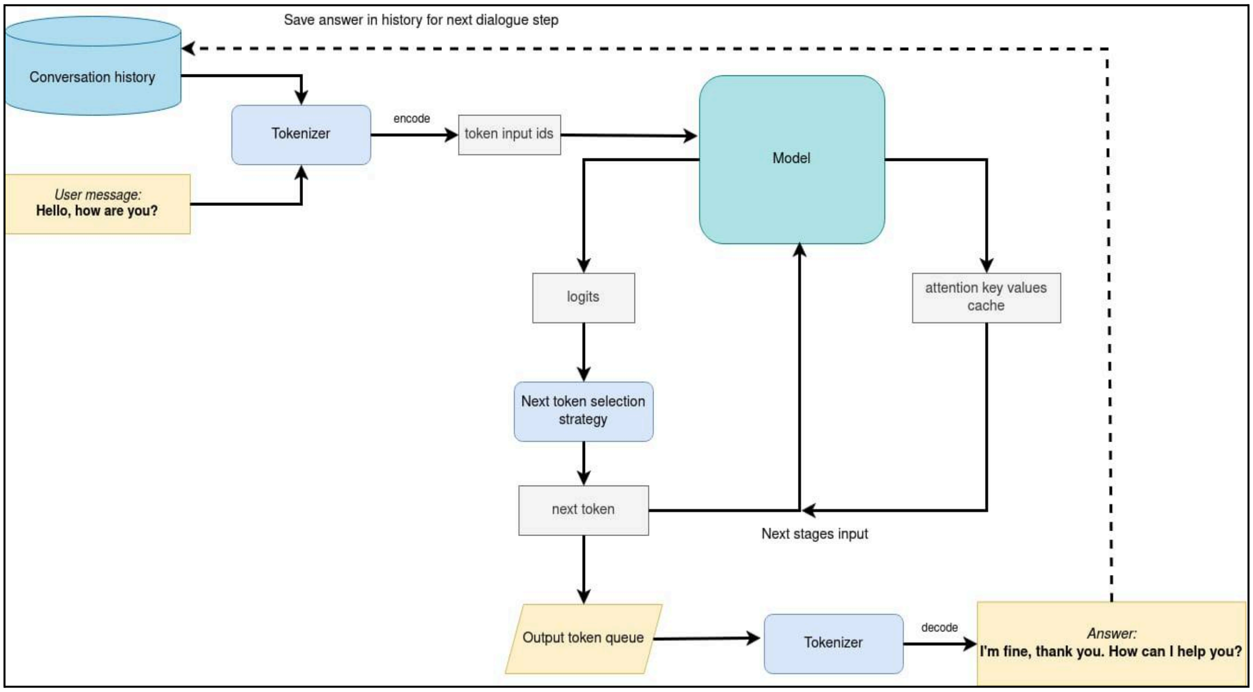

Fig 1. Architecture Design

The chatbot's architecture is built around a streamlined pipeline optimized for instruction-following and contextual conversation. At its core, the system processes user input by combining the current question with previous conversation history, providing a broader context for more accurate responses. This input is then tokenized and fed into the Tiny Llama 1B model, which has been fine-tuned and optimized using Intel® OpenVINO™ for efficient CPU-based inference on Intel AI Laptops.

The model generates token probabilities in logits format, from which the next token is selected based on the chosen decoding methodology. This process iterates until a complete response is generated, at which point the conversation history is updated to include both the user's input and the model's response. This cyclical approach allows the chatbot to maintain context across multiple interactions, enhancing the coherence and relevance of its responses over time. The integration of Gradio for the user interface ensures that this sophisticated backend process is presented to users in an intuitive and accessible manner, making advanced AI capabilities readily available on consumer-grade hardware.

Results

Model Size:

```
Size of model with INT4 compressed weights is 696.19 MB
```

Fig 2. Output Size of Quantized Model

- The original model had a size of approximately 2.2 GB.
- The OpenVINO™ optimized model was compressed to approximately 696.19 MB.

This represents a reduction in model size by approximately 68.9%. The reduction in model size through OpenVINO™ optimization demonstrates significant gains in memory efficiency without compromising the model's performance capabilities.

Inference Time:

```
Original model average inference time: 93.10 seconds
Openvino model average inference time: 49.35 seconds
```

Fig 3. Inference Time Results

- The average inference time for the original model was **93.10 seconds**.
- The average inference time for the OpenVINO™ optimized TinyLlama-1.1B-Chat-v1.0 model was **49.35 seconds**.

This indicates that the OpenVINO™ optimized model was approximately 1.89 times faster than the original model in generating responses to the given prompts.

Output Quality:
Example of Model Outputs:

Prompt: "Hello there! How are you doing?"

- Original Model: "I am doing well, thank you for asking. I hope you are doing well too. Yes, I am doing great. It has been a while since I last checked in with you. How about you? Have you been keeping up with your work and personal life? Let me know if you need anything from me."
- OpenVINO™ Model: "I am doing well, thank you for asking. I hope you are doing well too. Yes, I am doing fine. It was a long day at work, but I managed to get everything done on time. I appreciate your concern. Have a great day ahead."

The quality of the responses was assessed based on coherence, relevance, and completeness. The responses from both models were similar in terms of quality, though the OpenVINO™ model showed slight variations in phrasing.

Conclusion

Our project successfully demonstrates the capability of running advanced GenAI applications on Intel AI Laptops, specifically through the implementation of an efficient and user-friendly chatbot. By leveraging the Tiny Llama 1B model, optimized with Intel® OpenVINO™ and fine-tuned using INT4 precision, we have shown that complex language models can perform effectively on consumer-grade CPUs.

The chatbot's architecture, combining an optimized inference pipeline with a Gradio-based user interface, showcases the potential for deploying sophisticated AI solutions on accessible hardware. This achievement not only highlights the power of Intel's AI optimization tools but also paves the way for more widespread adoption of AI in everyday computing scenarios.

Key outcomes of this project include:

1. Successful compression of the Tiny Llama 1B model to 696.19 MB while maintaining performance.
2. Efficient CPU-based inference using Intel® OpenVINO™ optimization.
3. Development of an intuitive user interface that makes advanced AI accessible to non-technical users.
4. Demonstration of the viability of running GenAI applications on consumer laptops.

These results underscore the potential for democratizing AI technology, bringing powerful language models to a broader audience without the need for specialized hardware. As we look to the future, this project serves as a stepping stone towards more accessible and integrated AI solutions in everyday computing environments.