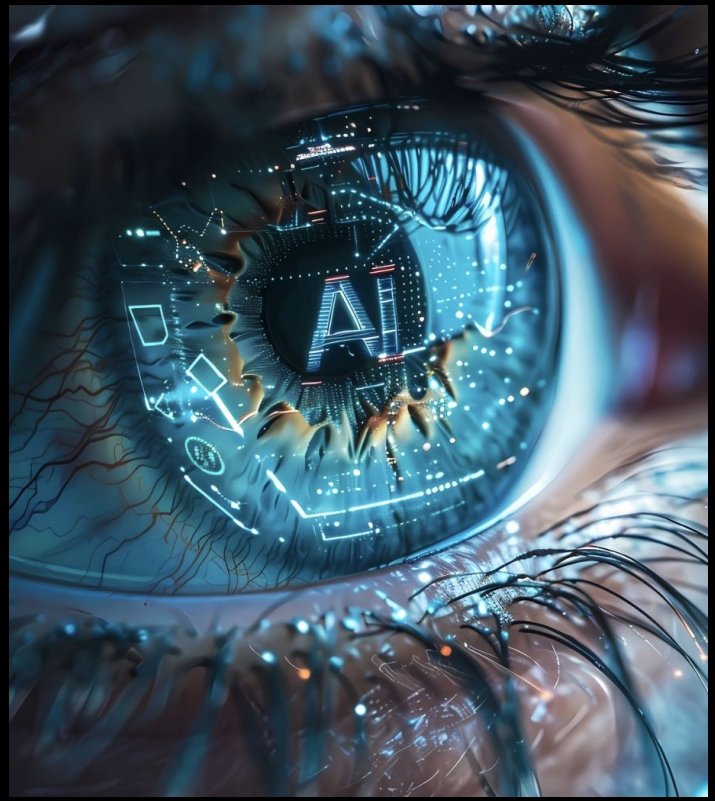


Unlocking the Power of Intel AI Laptops: Optimizing GenAI Workflows and LLM Inference

Topic Name :-

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

Team Members :-

1	Niraj Rajendra Patil
2	Yash Hemal Mehta
3	Dushyant Sanjay Pande
4	Ganesh Ishwar Patil
5	Sumeet Arvind Nikum

Description

We have developed a summary generation GenAI solution designed to generate concise and accurate summaries of cricket matches. This solution leverages advanced AI models to provide real-time summaries, making it easier for users to stay updated with match highlights. The use of Intel AI laptops and the Intel® OpenVINOTM toolkit ensures efficient model inference on CPUs, making the solution accessible and scalable.

Featured Offered

- Summary Generation: Generates detailed summaries for cricket matches, including key events, scores, and player performances.
- Our GenAI model is capable of generating detailed summaries of cricket matches, including key events, scores, and player performances. This feature allows users to quickly grasp the highlights of a match without needing to watch the entire game.

Process Flow

Steps:

1 . AI Model Development:

- Built using Hugging Face, TensorFlow, and PyTorch.
- We built an AI model using popular frameworks such as Hugging Face, TensorFlow, and PyTorch. These frameworks were chosen for their robustness and extensive support for natural language processing tasks.
- The model was trained on a dataset of cricket match descriptions and summaries, allowing it to learn the patterns and key elements necessary for generating accurate summaries.

2 . Frontend Development: Created a user-friendly interface.

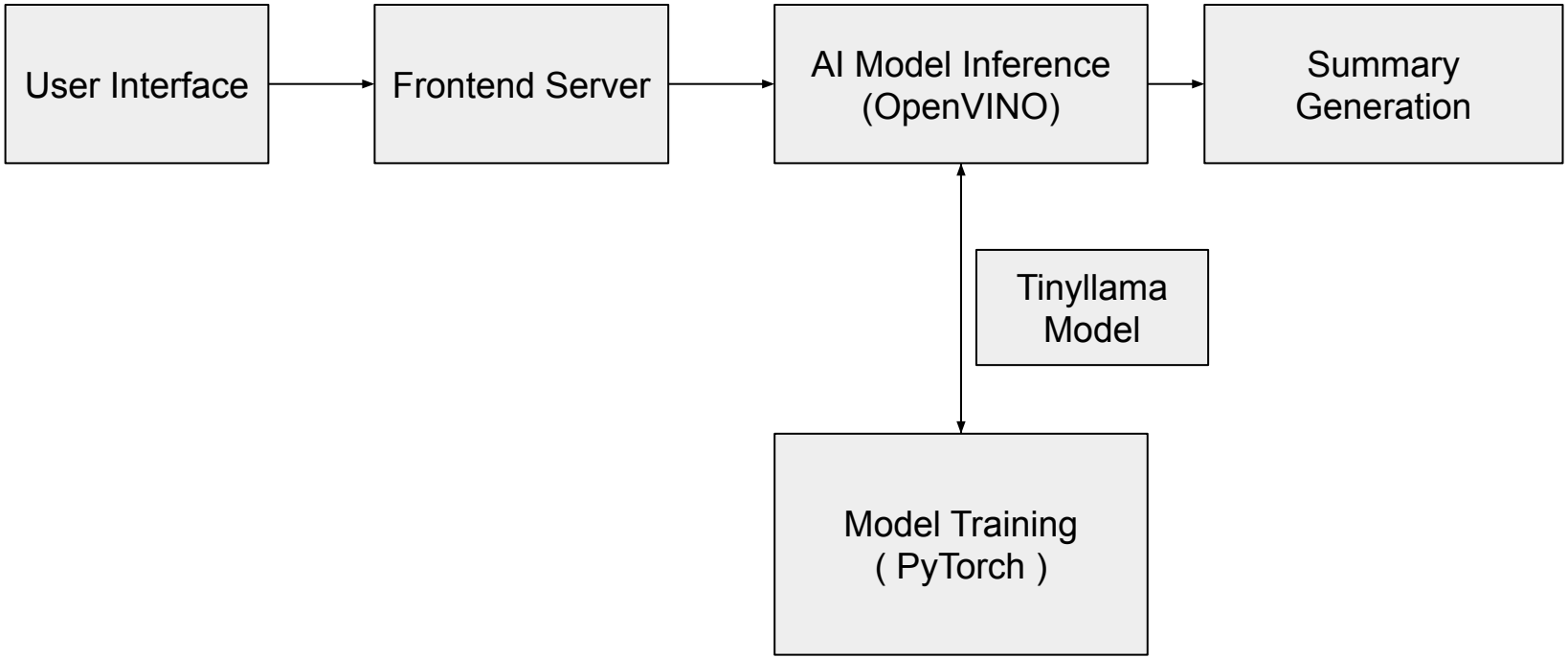
- A user-friendly frontend interface was developed using technologies such as React.js, HTML, CSS, and Flask. This interface allows users to interact with the GenAI model and view the generated summaries.

Steps:

3 . Integration to OpenVINO :

- Integrated with Intel OpenVINO to enhance performance.
- The AI model was integrated with Intel's OpenVINO toolkit to optimize performance and ensure efficient inference on CPU. OpenVINO provides a range of tools and libraries that help improve the speed and efficiency of AI model inference, making it a suitable choice for this project.

Architecture Diagram



Technologies Used

- **AI Frameworks** : PyTorch.
- **Frontend Development** : Flask, HTML, CSS.
- **Model Optimization** : Intel® OpenVINO™ Toolkit.
- **Model Formats** : ONNX, XML, BIN.
- **Model** : Tinyllama.
- **Hub** : Hugging Face.

Team Members and Contributions

1	Niraj Patil	Model building, Model training, Frontend Development , Project Management And Completion.
2	Yash Mehta	Model building, Backend, OpenVINO integration.
3	Ganesh Patil	Frontend development.
4	Dushyant Pande	helped OpenVINO integration.
5	Sumeet Nikum	Model conversion to BART/ONNX, Helped in Openvino Integration

Conclusion

Embracing the AI-Powered Future

Our project demonstrates the efficacy of GenAI for cricket match summaries using Intel AI laptops and OpenVINO. The collaboration ensured an accessible, scalable solution with efficient CPU inference. It serves as a foundation for future advancements in AI-powered summary generation.

Intel AI Laptops offer a transformative solution for optimizing your Generative AI and LLM workflows. By leveraging the powerful technologies within these laptops, you can unlock new levels of productivity, creativity, and innovation in your AI-driven projects.