RUNNING GENAI ON INTEL AI LAPTOP

Leveraging openvino for efficient inference and fine-tuning

Presented By,

INTELlect Innovators
Farhan khan K A
Sowmya G M
PESITM , Shivamogga

Problem statement:

GenAI on intel AI laptops and simple LLM inference on CPU and fine tuning of LLM models using intel Openvino.

- Challenges in running GenAI models on local systems due to compatibility and hardware constraints.
- Need for efficient inference and fine-tuning of large language models on accessible hardware.
- Requirement for optimization techniques to enhance performance and efficiency.

UNIQUE IDEA BRIEF (SOLUTION)

- Leveraging Intel Edge Developer Cloud for compatibility and performance.
- Utilizing Intel OpenVINO for model optimization.
- Focus on CPU-based inference and fine-tuning to avoid the need for expensive GPUs.
- Focus on converting models to FP16, INT8, and INT4 formats for improved performance.

FEATURES OFFERED

Performance Enhancement:

FP16, INT8, and INT4 optimizations for faster and more efficient execution.

Resource Efficiency:

CPU optimization reducing reliance on GPUs

Cost-Effectiveness:

- Utilization of existing Intel CPU resources, broad deployment options.
- Efficient inference using neural-chat-7b-v3-1 model.
- Fine-tuning capabilities with Meta-LLama-2-7b-hf model.
- Model optimization using FP16, INT8, and INT4 formats for better performance.

PROCESS FLOW

Environment Setup:

- Intel Edge Developer Cloud platform
- Python 3.10 (OpenVINO Notebooks 2024.1.0)

Model Selection and Configuration:

- Inference Model: neural-chat-7b-v3-1
- Fine-Tuning Model: Meta-LLama-2-7b-hf

Optimization:

- FP16, INT8, and INT4 conversions
- Inference and Fine-Tuning Processes:
- Tokenization, training setup, model export, and benchmarking

ARCHITECTURE

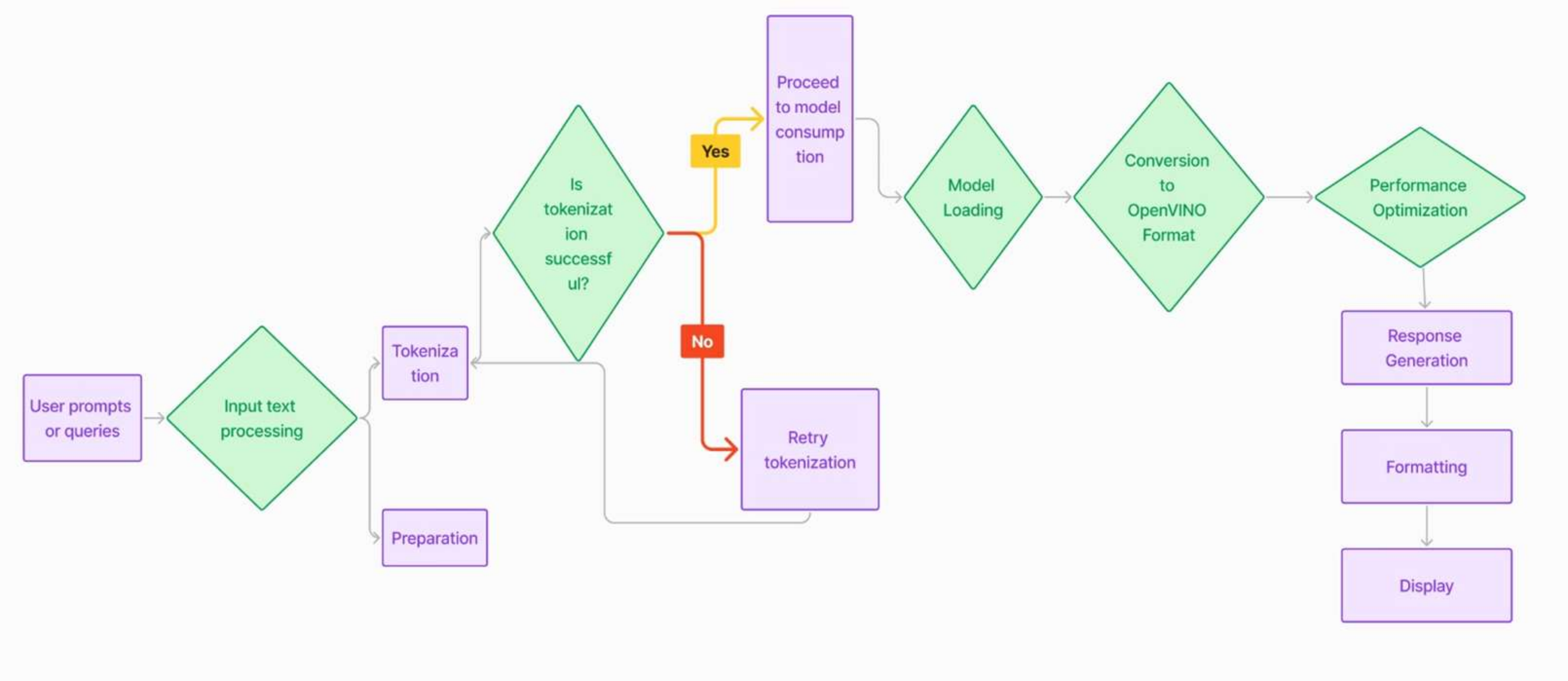


figure 1: Inference Process

ARCHITECTURE

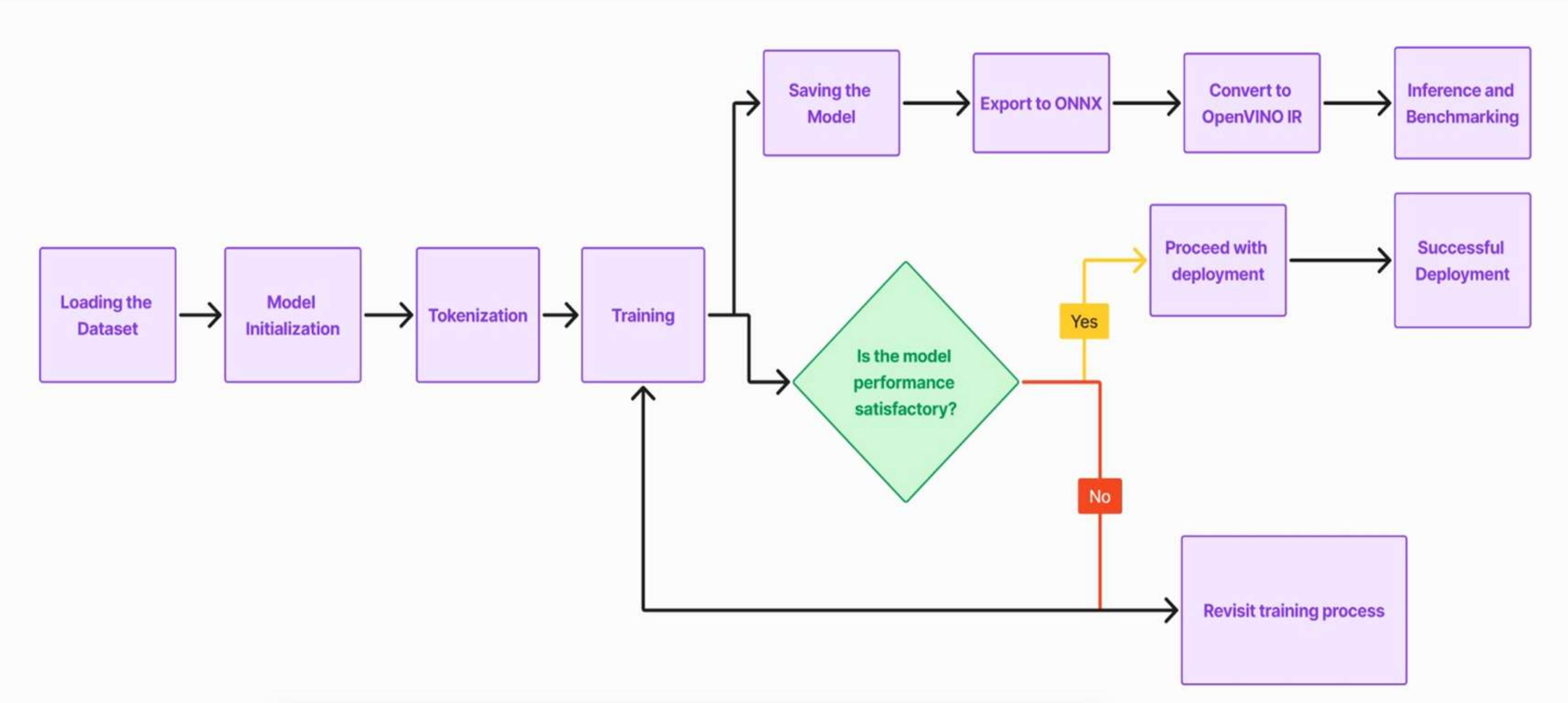


Figure 2: Fine-TuningProcess

TECH STACK USED:

Platform:

- Intel Edge Developer Cloud

Kernel:

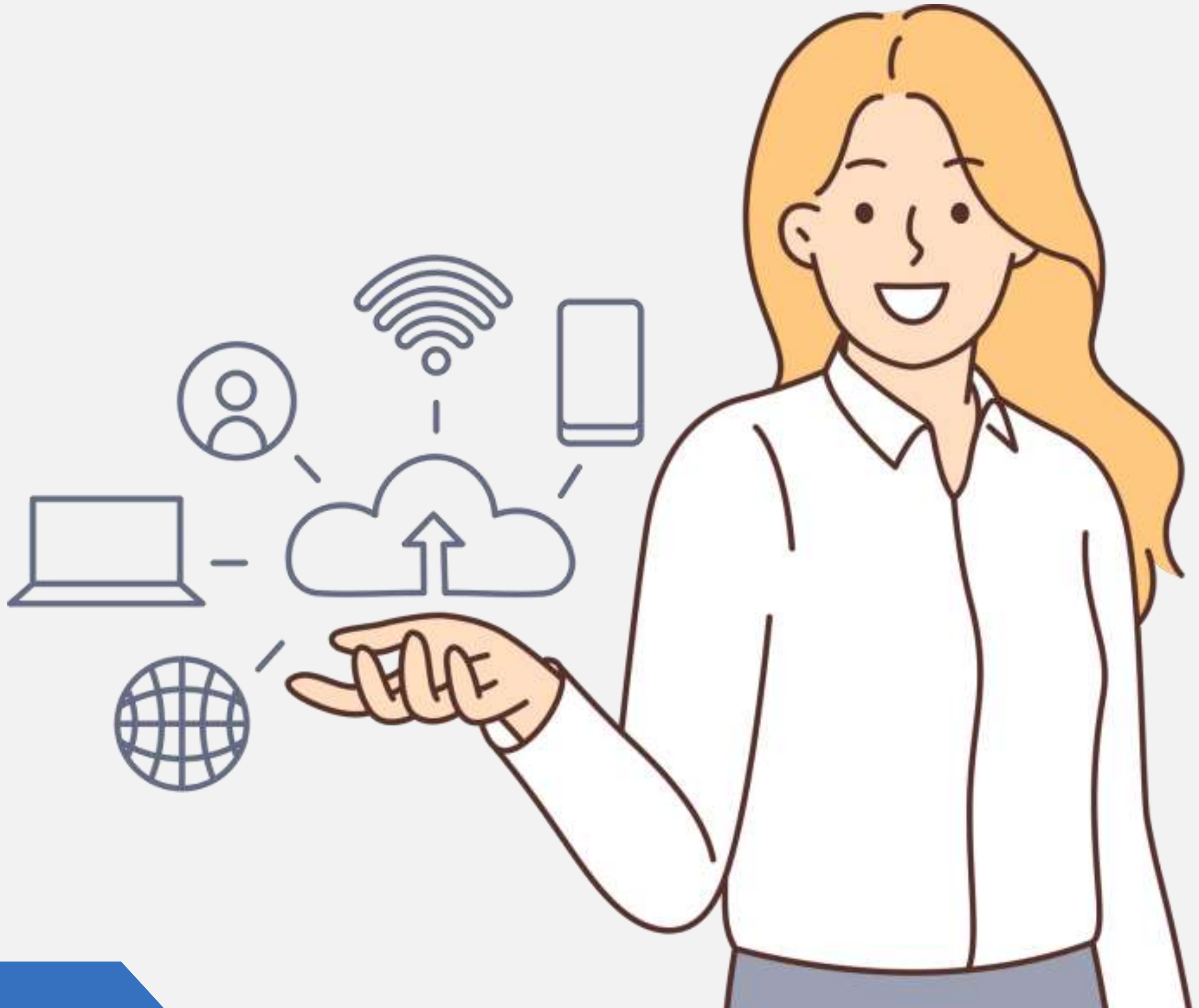
- Python 3.10 (OpenVINO Notebooks 2024.1.0)

CPU:

- Intel(R) Xeon(R) Gold 5218R CPU @ 2.10GHz

Libraries and Tools:

- OpenVINO
- Hugging Face transformers library
- torch.onnx.export



Models:

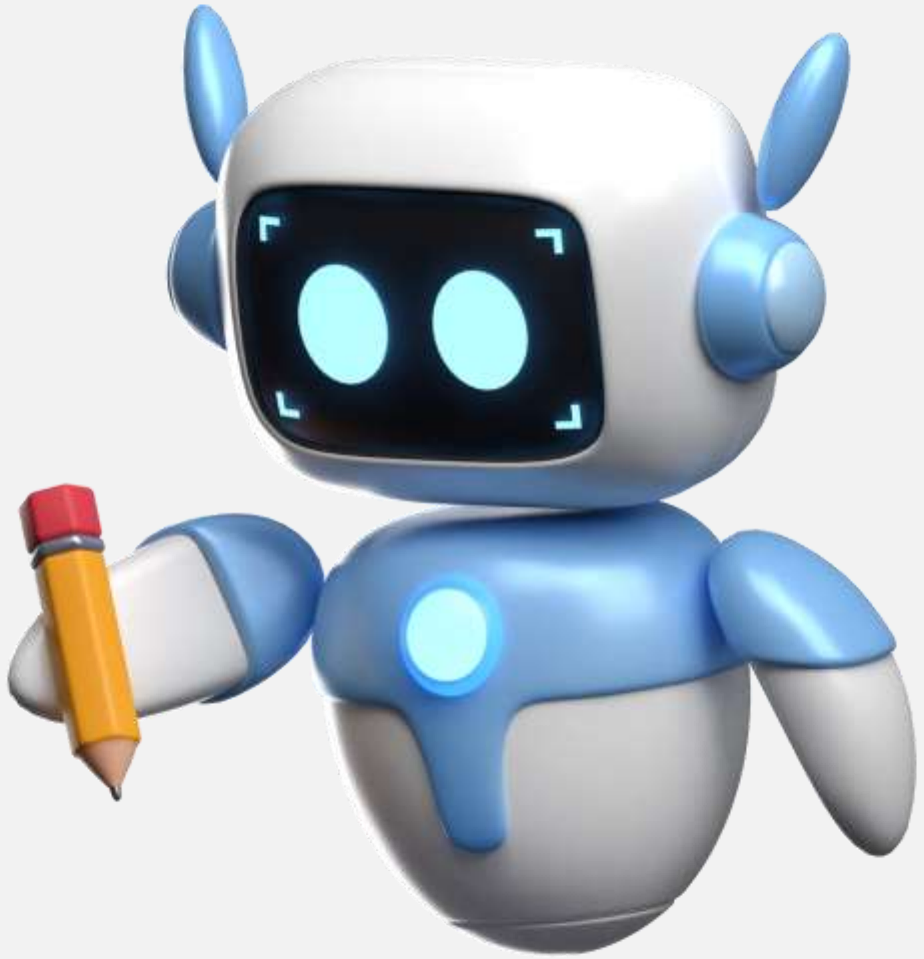
- Inference Model: neural-chat-7b-v3-1
- Fine-Tuning Model: Meta-LLama-2-7b-hf

Dataset:

- Wikitext-2-raw-v1

Transformers Library :

- (AutoTokenizer, OVModelForCausalLM)
- Hugging Face Trainer API
- ONNX (Model Export)
- PyTorch (Model Training)



TEAM MEMBERS AND CONTRIBUTION:

Farhan Khan:

- He set up the fine-tuning environment using Intel OpenVINO and fine-tuned the Meta-LLama-2-7B-HF model for specific tasks.
- He evaluated the performance of the fine-tuned model and documented the entire fine-tuning process and results

Sowmya:

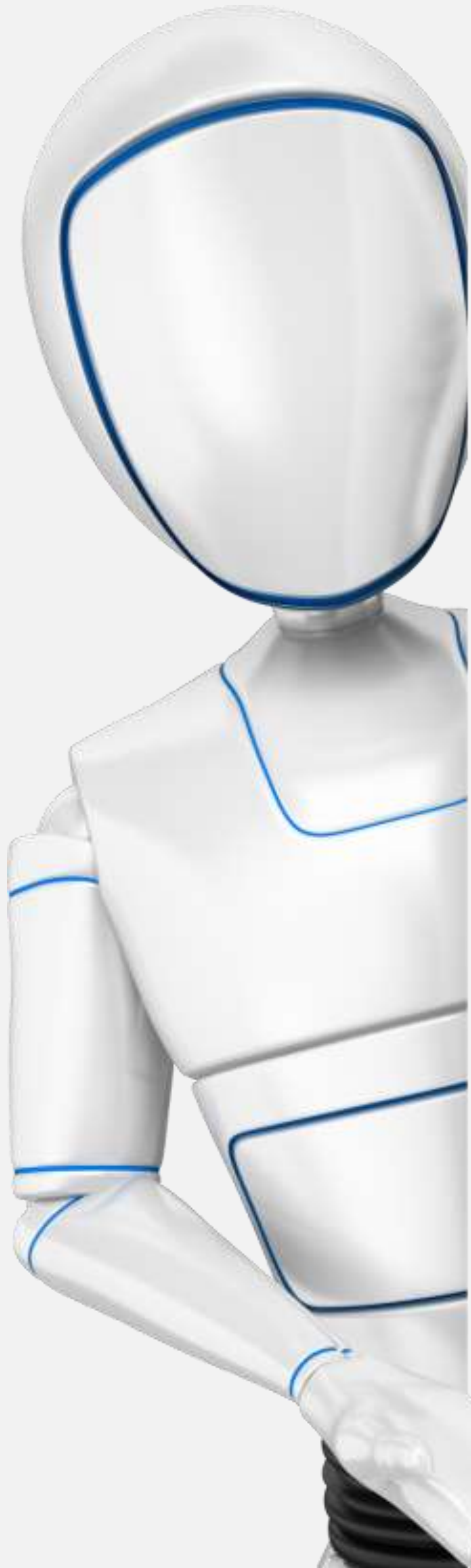
- She set up the environment on Intel Developer Cloud Edge and implemented LLM inference using the Neural-Chat-7B-v3-1 model.
- She tested and validated the model’s performance on CPU and successfully demonstrated inference with test data.

CONCLUSION:

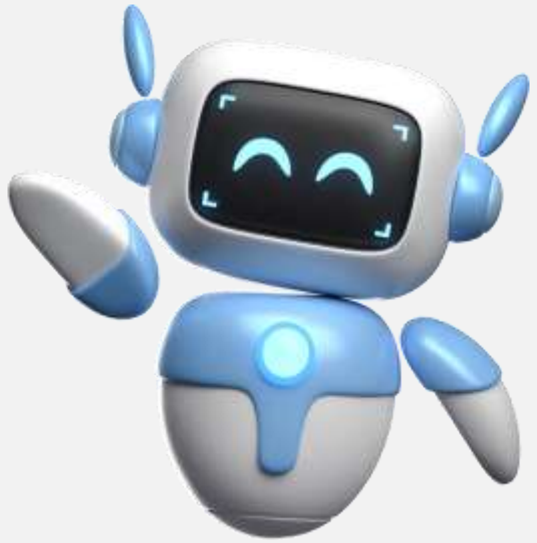
- Significant improvements in inference performance and efficiency using OpenVINO.
- Viable solution for deploying AI models on existing hardware infrastructure.
- Future work includes further optimizations, real-world application integration, and comprehensive benchmarking
- Successful implementation of GenAI models on Intel laptops using openvino.

FUTURE DIRECTIONS:

- Explore additional OpenVINO optimizations.
- Integrate optimized models into real-world applications.
- Conduct comparative studies for performance evaluation.
- Integrate models into real-world applications



VIDEO DEMO LINK



Click Here!!