PROBLEM STATEMENT 16:

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

Our goal is to use Intel AI Laptops and Intel® OpenVINO™ for efficient Generative AI and LLM inference on CPUs, especially for Technical Support Assistant Chabot. This configuration intends to improve performance, save costs, and ease the implementation of AI technologies. Using OpenVINO™ to fine-tune LLM models improves chatbot accuracy and responsiveness. It delivers fast and dependable technical help by optimizing resource use. Finally, it aims to increase customer happiness through quick, precise help.

UNIQUE IDEA BRIEF(Solution)

- Technical Support Assistant Chatbot

  Our Technical Support Assistant Chatbot utilizes the Intel® OpenVINO tools to optimize LLMs for CPU-based inference on Intel AI laptops. This chatbot provides real-time technical support and individualized advise to users. By using the efficiency of Intel®

  OpenVINO ensures that the model works efficiently on the CPU, making it both accessible and cost-effective.

- Objective

  Provide quick, accurate, and reliable technical support to users.

  - Leverage Intel AI Laptops and Intel® OpenVINO™ for efficient LLM inference on CPUs.
  - Fine-tune models for enhanced performance and responsiveness.
  - Ensure cost-effective and high-performance AI operations.
  - Improve user satisfaction by delivering timely and precise assistance.

FEATURES OFFERS

1. 24/7 Availability:

   Continually supports users by offering round-the-clock help to resolve problems at any moment.

2. Advanced Issue Diagnosis:

   This process, which frequently integrates with system logs and performance indicators, uses diagnostic tools and algorithms to determine the core cause of issues.

3. Comprehensive Solution Provision:

   Detailed, step-by-step troubleshooting instructions and answers, together with advice on setups and system changes, are provided by Comprehensive Solution Provision.

4. Knowledge Base Integration:

   Obtains and makes use of an extensive collection of articles, FAQs, and documentation to deliver precise and pertinent information in a timely manner.

5. Efficient Ticketing System:

   A well-organized ticketing system facilitates the tracking, prioritizing, and resolution of support requests. It also handles customer questions.

6. Real-Time Live Chat:

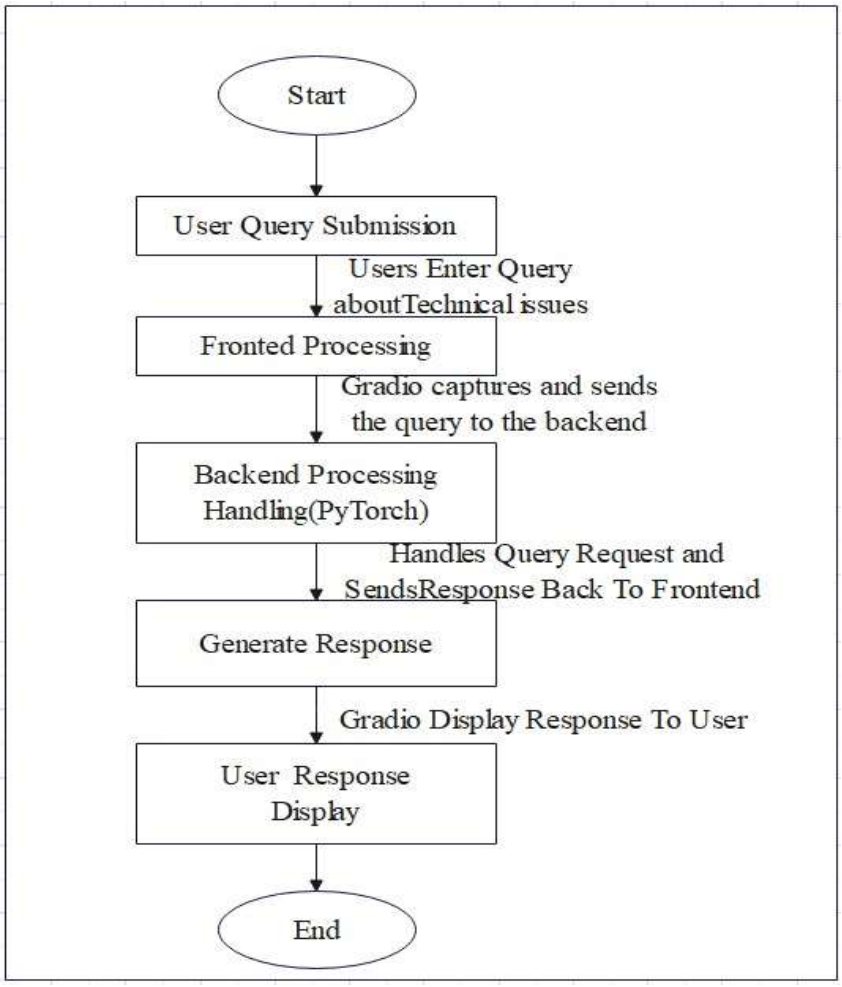
   Enables prompt replies and interactive problem-solving by providing instantaneous connection with people via live chat.

7. Remote Assistance Capabilities:

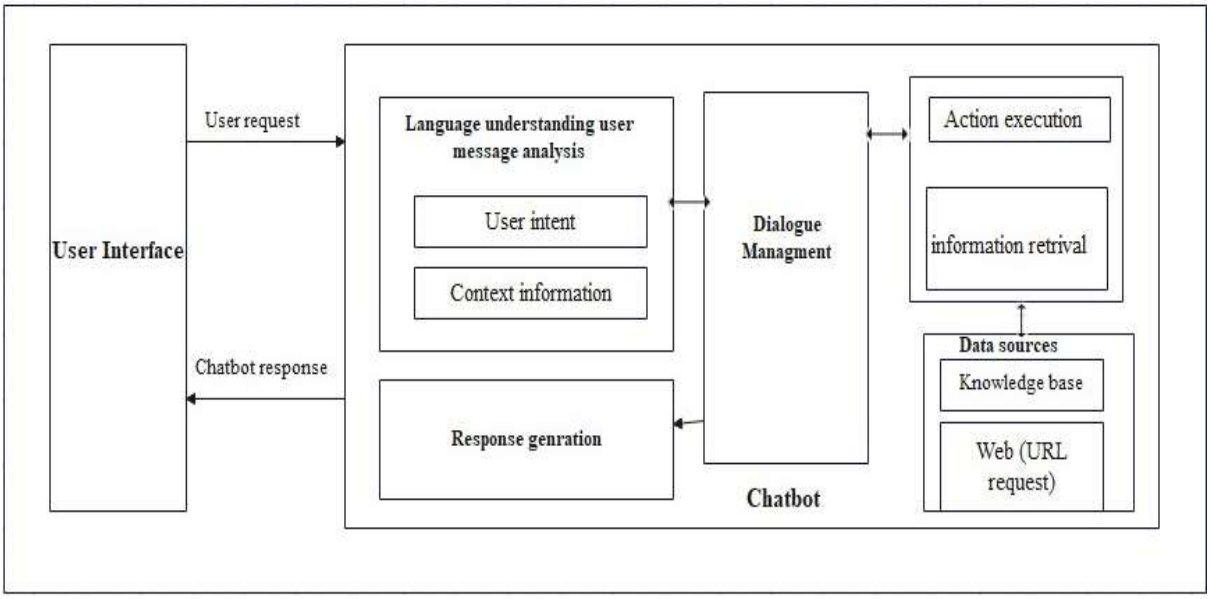
   Provides remote access to users' systems for hands-on troubleshooting and support, which can include screen sharing or remote control.

   These features ensure a robust and effective technical support experience, addressing user issues efficiently and comprehensively.

PROCESS FLOW



ARCHITECTURE DIAGRAM



TECHNOLOGIES UESD

1. Frontend Technology: Gradio

Frontend development utilize Gradio, especially for developing interactive interfaces for machine learning models. It offers a straightforward method for creating web-based user interfaces (UIs) that let users interact with models, enter data, and view real-time outcomes. Gradio is particularly helpful for sharing and rapidly developing machine learning applications.

2. Backend Technology: PyTorch

Pytorch is a lightweight Python web application framework. The architecture prioritizes simplicity and flexibility, enabling developers to build intricate backend services with little setup. Pytorch handles server-side functionality and facilitates communication between the frontend and machine learning.

3. Model Used: Hugging Face, Fine-Tuned

Hugging Face is a company and platform specializing in Natural Language Processing (NLP) and machine learning. It provides the Transformers library, which includes a wide array of pre-trained models for tasks such as text classification, translation, and questionanswering. Users can fine-tune these models on specific datasets to adapt them for particular tasks or domains. Fine-tuning involves training a pre-trained model on a new dataset with a smaller learning rate to adjust its weights for the new task. This process leverages the model's existing knowledge while making it more relevant to specific use cases.

4. Optimization: Intel® OpenVINO ™

Intel® OpenVINO™ is a toolkit for optimizing deep learning models, enhancing their performance for Intel hardware. It supports various frameworks and allows for efficient inference across CPUs, GPUs, and other accelerators. The toolkit helps accelerate model deployment and improves execution speed while maintaining accuracy.

5. Hardware: Intel AI Laptops

Intel AI laptops are high-performance computer systems that can do artificial intelligence and machine learning activities. They use Intel processors with inbuilt AI

acceleration, such as Intel Core and Xeon CPUs, and frequently add Intel GPU technologies for increased computing capability. These laptops are designed for activities like as model training, inference, and data processing, with strong support for AI frameworks and libraries. They give developers and data scientists strong tools for building and implementing AI systems on the fly.

CONCLUSION

The Technical support Assistant Chatbot utilizes Intel® OpenVINO and fine-tuned language models (LLMs) to provide real-time technical help. This technology, which runs on Intel AI laptops and uses CPU-based inference, is affordable and easily available, making it a feasible option. A tool for people seeking technical support. This study showcases how Intel's AI technology may transform technical applications and benefit both users and providers. Our chatbot revolutionizes technical issue resolution with rapid processing, individualized advice, and an easy-to-use interface. The chatbot's powerful AI and efficient hardware provide speedy and precise replies, improving user experience and confidence. Using Gradio for the front end and Pytorch for the back end creates a seamless experience. The combination of Gradio and Pytorch creates a seamless interaction platform. The hugging face model, optimized with OpenVINO, allows efficient inference on standard CPUs, making powerful AI accessible to a wider audience. This initiative highlights the future of technical assistance, in which technology and AI play critical roles in providing individualized service.