BITS BRIGADE
Group No. :

Intel® Unnati Industrial Training Program 2024.

Problem Statement :

- Introduction to GenAI and Simple LLM Inference on CPU and fine tuning of LLM Model to create a Custom Chatbot

Unique Idea Brief (Solution) :

- Our project focuses on the innovative utilization of a pre-trained Language Learning Model (LLM) to develop a bespoke chatbot.
- By implementing straightforward LLM inference on a CPU and fine-tuning the model, we aim to create a highly adaptable chatbot capable of addressing a variety of needs.
- This approach leverages advanced AI techniques to enhance interaction capabilities, including text generation, summarization, and voice communication, thereby providing a comprehensive solution for modern communication challenges.

Features Offered :

- **Text Chat Training:** Our chatbot is meticulously trained to handle text-based interactions, ensuring fluid and coherent conversations.

- **Voice Chat Training:** Equipped with voice processing capabilities, the chatbot can seamlessly manage and respond to voice inputs, offering a more interactive user experience.

- **Mixed Precision (BF16) Optimization:** The incorporation of mixed precision optimization accelerates model training and inference, significantly boosting performance without compromising accuracy.

- **Plugins for Text-to-Speech (TTS) and Automatic Speech Recognition (ASR):** These plugins enhance the chatbot's functionality, enabling it to convert text to speech and recognize spoken language, thereby broadening its application scope.

- **Custom Fine-Tuning:** The model is fine-tuned for specialized tasks such as code generation, summarization, and text generation, ensuring it meets specific user requirements with high efficiency.

- **Integrated Tools and Libraries:** Utilizing a robust suite of tools and libraries, including Intel Extension for Transformers, pytorch, peft, tensorflow, and accelerate, we optimize performance and streamline the development process.

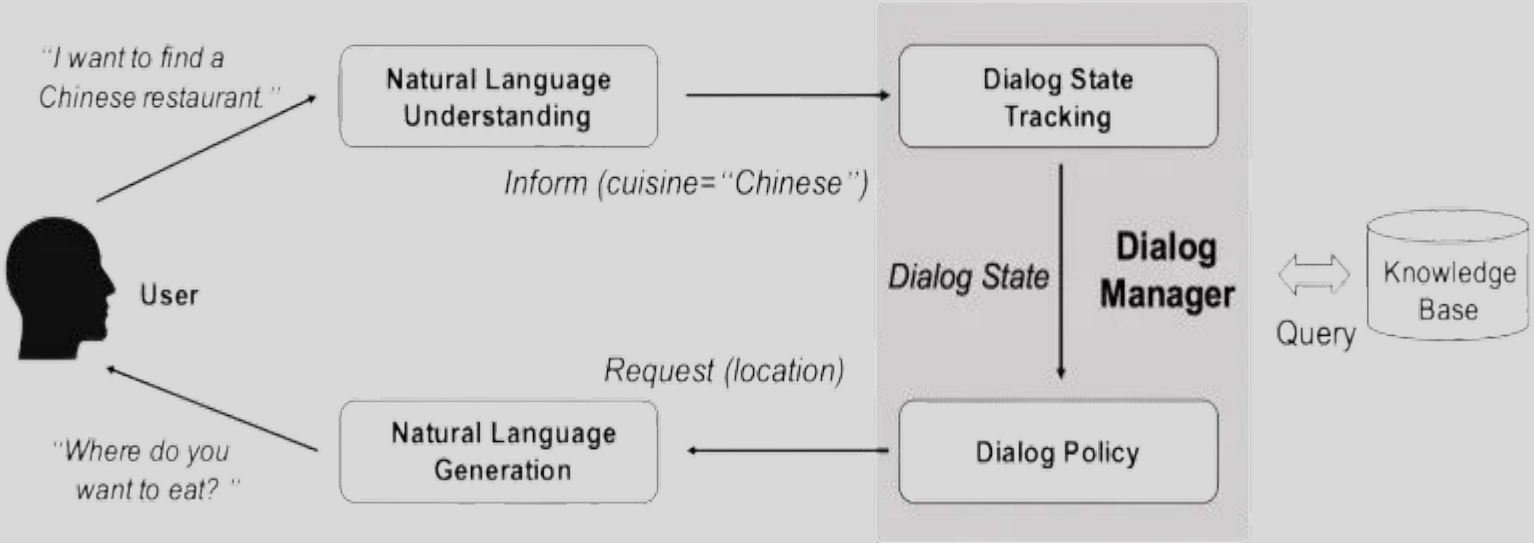
Process Flow :

1. Importing Required Libraries :
2. Configuring BF16 Optimization
3. Enabling Plugins for Text-to-Speech (TTS) and Automatic Speech Recognition (ASR)
4. Reapplying BF16 Optimization for a Different Query
5. Finetuning of Custom dataset

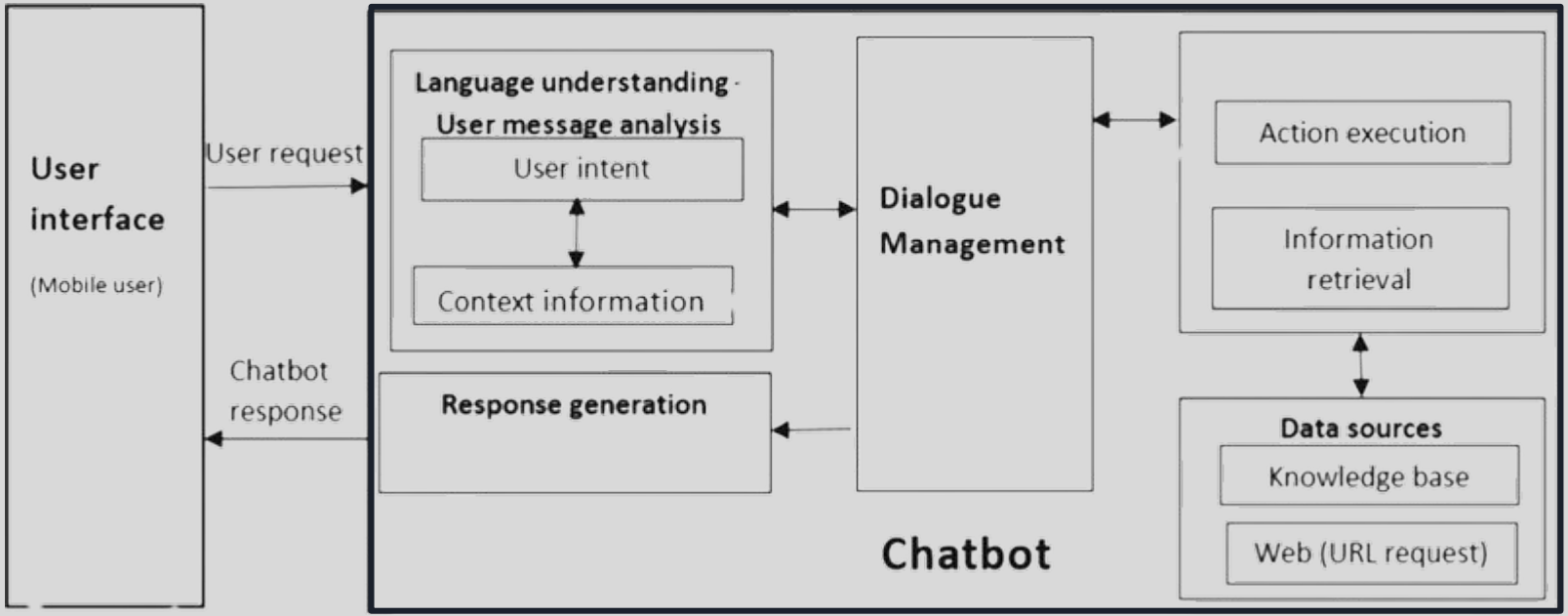
Process Flow :

1. **Importing Required Libraries:** Prepare the environment by importing the necessary modules.
2. **Configuring BF16 Optimization:** Set up mixed precision optimization for improved performance.
3. **Enabling Plugins for Text-to-Speech (TTS) and Automatic Speech Recognition (ASR) :** Set up plugins for additional functionalities like TTS and ASR.
4. **Reapplying BF16 Optimization for a Different Query :** Optionally configure and make predictions with different queries or configurations.
5. **Finetuning of custom dataset:** Finetuning of alpaca_cleaned_dataset for meta-llama/Llama-2-7b-chat-hf model.

Architecture Diagram :



Architecture Diagram :



Technologies Used :

- **Intel Extension for Transformers (IET):** (NeuralChat)
  - Purpose: Optimizes transformer models for Intel hardware.
- **Mixed Precision (BF16):**
  - Speed up model training and inference without significantly affecting accuracy.
- **Chatbot Framework:** (build_chatbot, PipelineConfig)
  - Purpose: Builds and runs the chatbot model.
- **Plugins:** (Text-to-Speech(TTS), Automatic Speech Recognition (ASR))
  - Purpose: Extends the chatbot functionality to include speech features.
- **Tools:**
  - Intel Developer Cloud, Huggin Face, GitHub
- **Libraries :**
  - Pytorch , accelerate , transformers 4.35.2 , huggingface_hub , tensorflow 2.13.0 , etc.

Neural Chat

HUGGING FACE

PyTorch TensorFlow

Team Members and Contribution:

- ★ Atharva Date (Lead)
  - Voice Chat Training for building chatbot,
  - Fine tuning model for Code Generation
- ★ Anshu Vairagade
  - Text Chat Training for building chatbot,
  - Fine tuning model on Summarization task
- ★ Yash Ukirde
  - Low Precision Optimization for building chatbot,
  - Fine tuning model for Text Generation task

Conclusion :

In summary, our project exemplifies the strategic application of a pre-trained LLM, optimized for CPU-based inference and meticulously fine-tuned for distinct tasks. The integration of advanced features like mixed precision optimization and speech processing plugins positions our chatbot as a versatile and high-performance solution. This initiative not only showcases the potential of custom fine-tuning in enhancing language models but also underscores its applicability across diverse domains, thus providing a significant advancement in AI-driven communication technologies.