Problem Statement

Introduction to GenAI and Simple LLM Inference on CPU and finetuning of LLM Model to create a Custom Chatbot

- AI systems having general artificial intelligence (GenAI) abilities can do any cognitive task.

- CPU-Based Large Language Models (LLMs): exhibiting a CPU's fundamental capabilities through simple inference.

- Making LLM more tailored for a chatbot: fitting a pre-trained LLM model to a particular task or dataset, allowing for the effectiveness of specific duties like customer service, translation, and entertainment

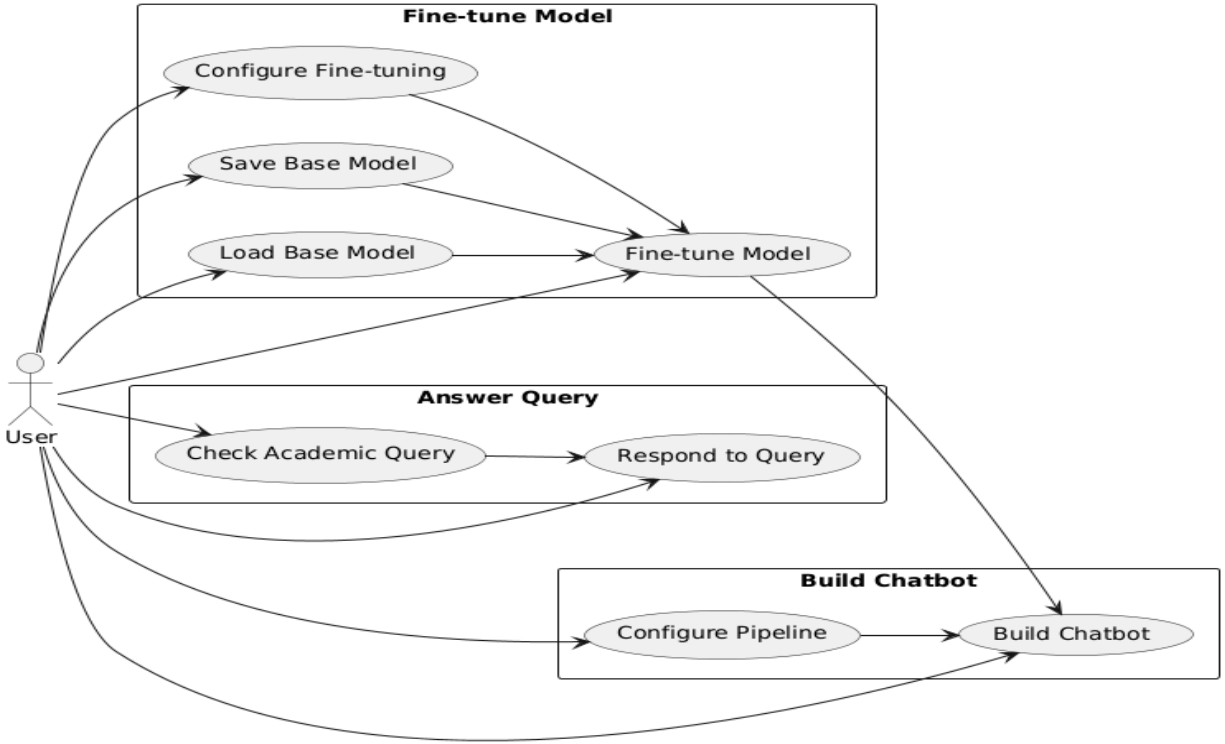
Unique Idea Brief (Solution)

- A pre-trained Language Model (LLM) ought to be adopted.
- Build a dataset with the input queries and responses.
- Modify the LLM model.
- Improve the UI user-friendly.
- Create an API endpoint and deploy it in the cloud.
- Assess and enhance the chatbot

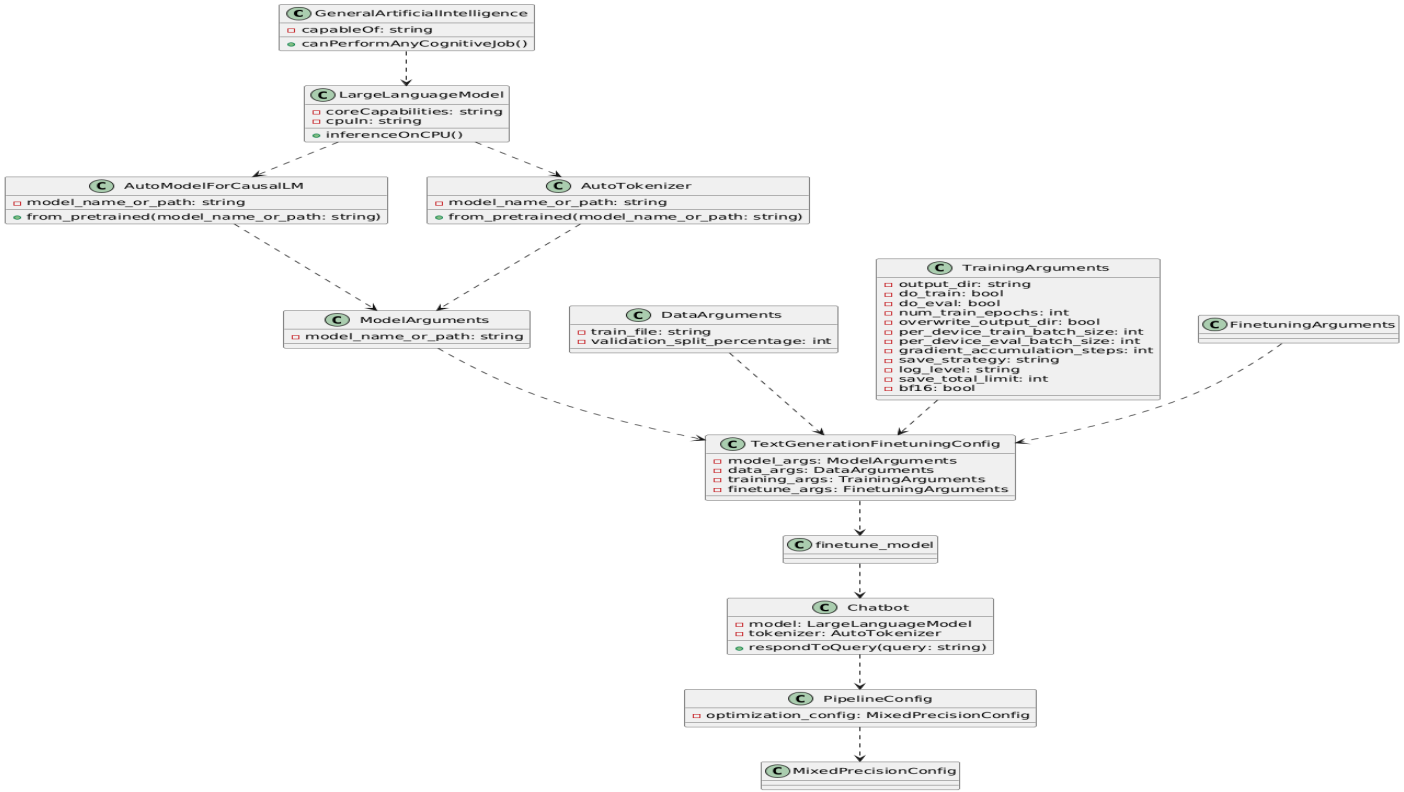
Features Offered

- For simple interaction, this system provides a conversational interface.
- It allows modification to meet particular needs.
- The solution seamlessly integrates with other systems.
- It improves efficiency by eliminating repetitive activities.
- Analytics is offered by the solution for tracking performance.
- It provides data protection and security.
- The system can be expanded in order to meet rising demand.
- It is economical and lowers operating expenses.

Process flow



Architecture Diagram



Technologies used

Transformers Library:

- Hugging Face Transformers: A library that provides general-purpose architectures for Natural Language Understanding (NLU) and Natural Language Generation (NLG) with pre-trained models.
    - TrainingArguments
    - AutoModelForCausalLM
    - AutoTokenizer

Intel Extension for Transformers:

- Intel Extension for Transformers (IEXT): Enhances the performance of Transformer-based models with Intel hardware optimizations.
    - ModelArguments
    - DataArguments
    - FinetuningArguments
    - TextGenerationFinetuningConfig
    - finetune_model

- build_chatbot
- PipelineConfig

Model and Tokenizer:

- Meta LLaMA: A series of large language models developed by Meta (formerly Facebook), specifically Llama-2-7b-chat-hf.

Python Standard Library:

- Functions: Basic Python functions and structures to define and process data, such as def, return, if.

Data Handling:

- Handling JSON data files (train_file="alpaca_data.json") for training.

Model Fine-tuning:

- Techniques for fine-tuning pre-trained models for specific tasks using various configurations (TrainingArguments, TextGenerationFinetuningConfig).

Team members

S Varshita	(URK22CS2005)
V Jaswanthini	(URK22CS5093)
Arpudhaa CA	(URK22CS5075)
V Siva Sankari	(URK22CS2039)
Anointina A	(URK22CS1107)

Conclusion

- Highlighted about generative AI and LLMs, emphasizing the advantages of CPUs and NLP applications.
- Evaluated fine-tuning LLMs for custom chatbots beforehand.
- Fine-tuning models to suit individual requirements and boosts performance.
- Chatbots and NLP can be modified by generative AI and LLMs.
- Customized chatbots for distinct use cases are made feasible by fine-tuning.