**Project Report**

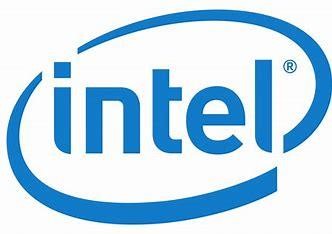
**On**

**INTRODUCTION TO GENAI AND SIMPLE LLM INTERFENCE ON CPU AND FINETUNING OF LLM MODEL TO CREATE A CUSTOM CHATBOT**

**Trainee College Mentor**

**DR.GIRISH DR. HARSHADA J PATIL**

**Vemana Institute of technology,kormangala**



# Table of Contents

|  |  |
| --- | --- |
| OBJECTIVE | 3 |
| ABSTRCT | 3 |
| FLOW DIAGRAM | 4 |
| DETAILED EXPLANATION | 4 - 5 |
| SOLUTION | 5-7 |
| FUTURE SCOPE | 8 |
| CONCLUSION | 9 |

**OBJECTIVE**

-To generate AI and simple LLM (large language Models)

-LLM models Inference on CPU

-Finetuning of LLM model to create a custom chatbot

**ABSTRACT**

The project is to generate AI and simple LLM(large language models)and LLM models interfence on cpu and finetuning of LLM model to create a

Custom chatbot. The large language models(LLMs) have is been attracting a lot of attention lately because of their extraordinary performance in various chatbot agents such as ChatGPT,GPT-4,Bard.

LLM are limited by the significant cost and time required to train or fine tune them.

## FLOW DIAGRAM

## Create Your Own Custom Chatbot. Train Large Language Models Quickly and… | by Intel(R) Neural Compressor | Intel Analytics Software | Medium

## DETAILED EXPLANATION

### genAI(Genetic Artifical intelligence)

GenAI involves the use of genetic alogorithums to create artificial intelligence systems.these alogrithum will mimic the process of natural selction to evolve solutions to comlex probems.

There are many application of genetic algorithum in AI can include optimization,learning and problem solving tasks.

Simple LLM inference on CPU

This refers to performing inference(making predictions)with a language model(LLM)on a central processing unit(cpu) it has opposed to graphics processing input(GPU) or other specialized hardware.

Finetuning of LLM model to create a custom chatbot

Finetuning involves taking a pre performed language model like GPT, BERT, etc and adapting it to perform a specific task, such as creating a chatbot.

-To create a custom chatbot you would finetune the language model on a dataset containing conversational data.This process will help the model to learn the nuances of natural language and context specific task at hand.

-Finetuning typically involves adjusting the models parameters and hyperparameters, as well as training it on the new dataset.

**SOLUTION**

To generative AI(GenAI) and Large Language Models (LLMs)for creating custom chatbot sand the CPU- based inference and finetuning

The following structured approach can be adopted:

-Choose an appropriate pre-trained LLM based on the complexity of your chatbot requirement and the available computational resources. Options include:

- GPT-3: Known for its versatility and robustness, but may require significant computational power.

-GPT-2: A smaller and more manageable model for CPU inference, suitable for less demanding applications.

-DistilGPT: A distilled version of GPT-2, offering faster inference and lower resource requirements.

2. Data Preparation

Collect and prepare a domain-specific dataset that reflects the kind of interactions your custom chatbot will handle:

- Data Collection: Gather conversation logs, customer queries, FAQs, and other relevant text data.

- Data Cleaning: Remove noise, correct errors, and standardize formats.

- Data Annotation: Label data if necessary to provide context and structure.

3. Environment Setup

Prepare your computational environment to support CPU-based inference and finetuning:

- \*Hardware Configuration\*: Ensure the CPU has adequate processing power and memory. Multi-core processors can improve performance.

- \*Software Installation\*: Install necessary libraries such as TensorFlow, PyTorch, and Hugging Face Transformers.

- \*Optimization Techniques\*: Use model quantization and pruning to reduce the model size and improve inference speed on CPUs.

4. Finetuning the Model

Finetune the selected pre-trained LLM on your custom dataset:

- \*Training Setup\*: Configure the training script to utilize CPU. Set appropriate batch sizes and learning rates to accommodate the CPU’s limitations.

- \*Training Process\*: Monitor the training process closely to avoid overfitting. Use early stopping and validation techniques to ensure the model generalizes well.

- \*Evaluation\*: Assess the finetuned model using metrics like perplexity, BLEU score, and user satisfaction surveys.

5. Deployment

Deploy the finetuned model as a chatbot:

- \*API Integration\*: Use frameworks like FastAPI or Flask to create an API endpoint for your chatbot.

- \*User Interface\*: Develop a user-friendly interface for interaction, such as a web chat interface or integration with messaging platforms like Slack or WhatsApp.

- \*Scaling and Monitoring\*: Implement scaling strategies to handle varying loads and monitor the chatbot’s performance to ensure reliability.

6. Continuous Improvement

Continuously improve the chatbot based on user feedback and performance metrics:

- \*User Feedback\*: Collect and analyze user feedback to identify areas for improvement.

- \*Model Updates\*: Periodically retrain and update the model with new data to keep it relevant and accurate.

- \*Feature Enhancement\*: Add new features and capabilities based on emerging user needs and technological advancements.

ools and Technologies

- \*Hugging Face Transformers\*: For accessing pre-trained LLMs and finetuning.

- \*TensorFlow/PyTorch\*: For model training and inference.

- \*FastAPI/Flask\*: For deploying the chatbot as an API.

- \*NLTK/Spacy\*: For data preprocessing and natural language understanding.

1. \*Select Model\*: Choose GPT-2 for its balance between performance and resource requirements.

2. \*Prepare Data\*: Collect customer service logs, clean and annotate them.

3. \*Set Up Environment\*: Install Hugging Face Transformers, TensorFlow, and necessary libraries.

4. \*Finetune Model\*: Use Hugging Face’s Trainer API to finetune GPT-2 on the prepared dataset.

5. \*Deploy Chatbot\*: Develop a Flask API to serve the chatbot and integrate it with a web interface.

6. \*Improve Iteratively\*: Collect user interactions, retrain the model with new data periodically, and enhance features based on feedback.

By following this structured approach, you can effectively create a custom chatbot using GenAI and LLMs, leveraging CPU-based inference and finetuning to ensure accessibility, efficiency, and relevance to your specific application needs.

**FUTURE SCOPE**

Future Scope of GenAI and Custom Chatbots

Advanced Personalization:

Development of chatbots that can adapt to individual user preferences and contexts, providing highly personalized interactions.

Improved Natural Language Understanding:

Enhancements in models that allow for more accurate comprehension and generation of human language, including understanding nuances, slang, and idioms.

Integration with Multimodal Systems:

Combining text, voice, image, and video inputs to create more comprehensive and interactive AI systems.

Domain-Specific Expertise:

Creation of highly specialized chatbots for fields such as medicine, law, finance, and education, offering expert advice and information.

Ethical AI and Bias Mitigation:

Ongoing research and implementation of methods to ensure AI systems are fair, unbiased, and ethically sound.

Scalability and Efficiency:

Continued improvements in the efficiency of LLMs, enabling more widespread and scalable deployment, including on resource-constrained devices.

Human-AI Collaboration:

Developing AI systems that can effectively collaborate with humans, augmenting their capabilities rather than replacing them.

By addressing these areas, the future of GenAI and custom chatbots holds significant promise for enhancing human-computer interactions and providing valuable tools across various domains.

**CONCLUSION**

The evolution of Generative AI (GenAI) and Large Language Models (LLMs) has significantly transformed how we interact with technology, especially in natural language processing (NLP) applications. The ability to perform simple LLM inference on a CPU and finetune LLMs for specific tasks has opened up vast possibilities for creating custom chatbots tailored to unique requirements.

### 