Date of Submission: 08/07/2024

Name: Hari Venakataraman

Institute: SRM Institute of Science and Technology Kattankulathur

Branch/Specialization: B. Tech - Artificial Intelligence [Dept: Computational Intelligence]

Register Number: RA2211047010116

Internal Mentor: Dr. Sumathy G.

External Mentor: Dr. Vasudha Kumari (AI Software Solutions Engineer, Intel)

Simple LLM Inference on CPU: Fine Tuning a Chatbot

In Intel's Industrial Training Program, the problem statement assigned to me was the Simple LLM (Large Language Models) statement focusing on fine tuning chatbots and increasing the accuracy of the responses and training the language model with more variety of prompts. My task was to run the designated Jupyter Notebook cells for building a chatbot on spr and fine tuning a chatbot on a single node.

Large Language Models, like GPT-3, are powerful tools capable of understanding and generating human-like text. Inference refers to the process of using a trained model to make predictions or generate responses. Typically, LLMs require significant computational resources, often utilizing GPUs for efficient processing. However, running inference on CPUs can be beneficial in certain scenarios, such as when GPUs are not available or when deploying models on edge devices.

The primary task was to run a pre-trained language model on the CPU. The results of running the build_chatbot_on_spr notebook upon running each cell are as follows:

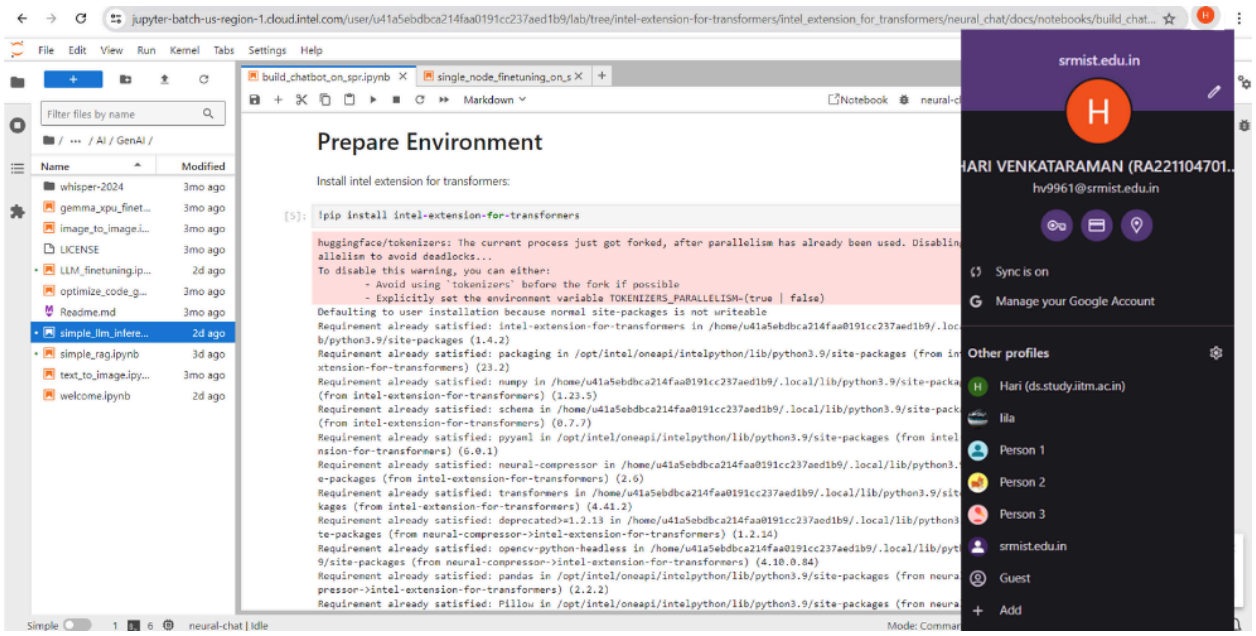

Fig 1: Installing the necessary libraries

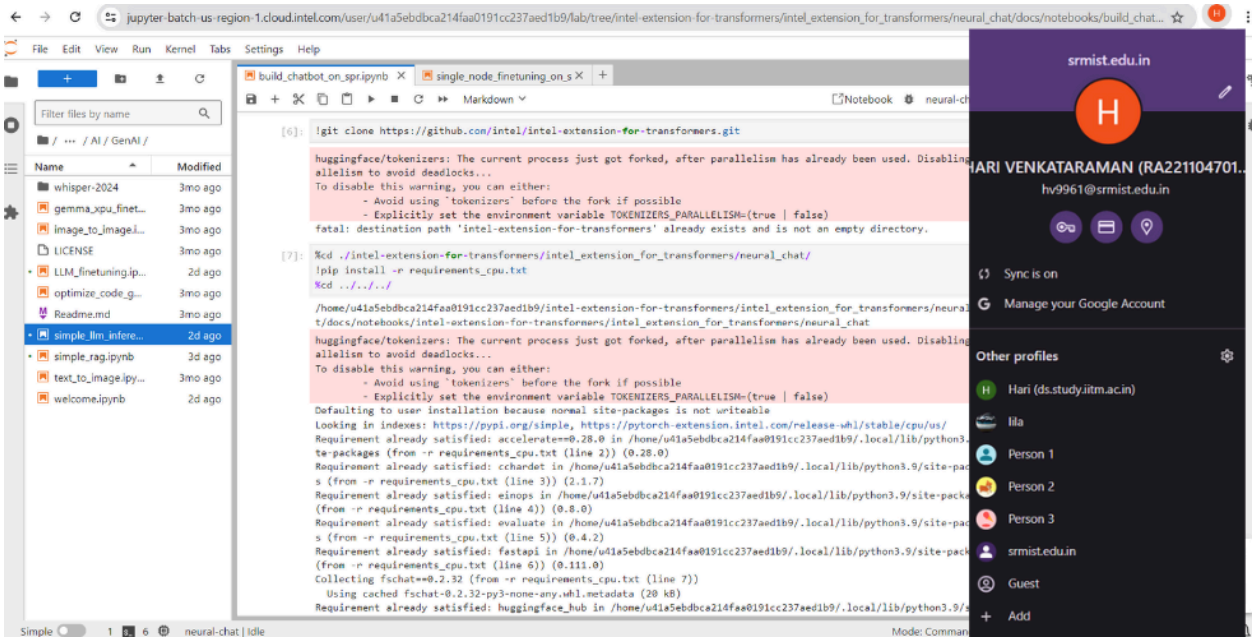

Fig 2: Cloning the Intel Repository

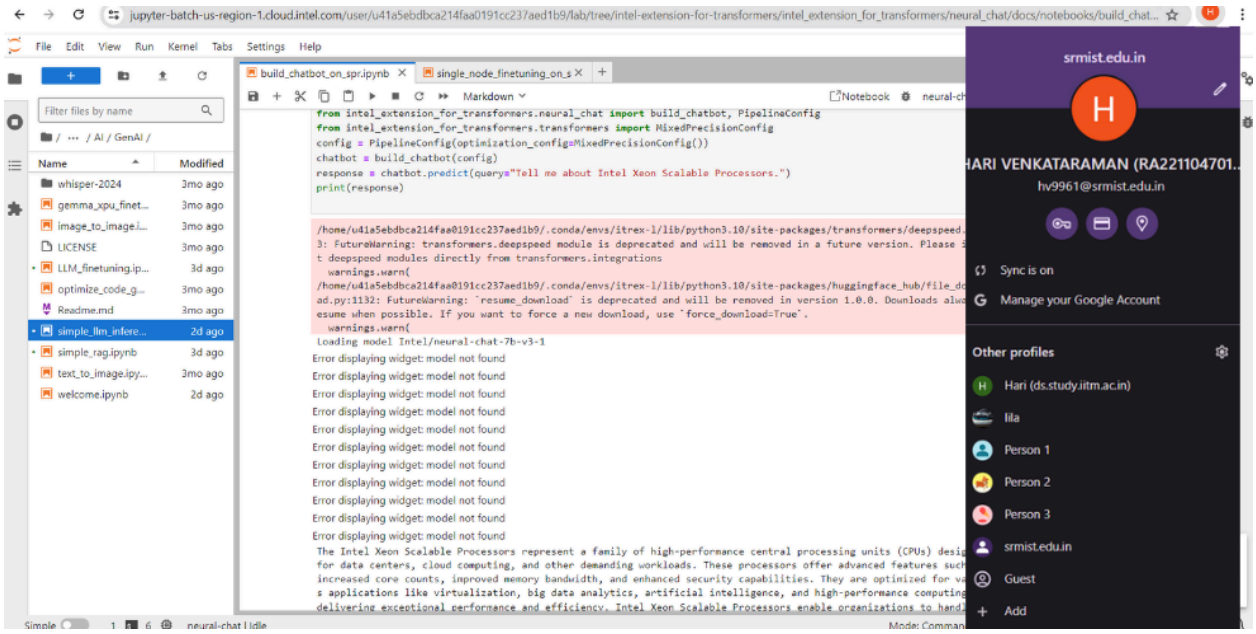

Fig 3: First Query Example to the Chatbot

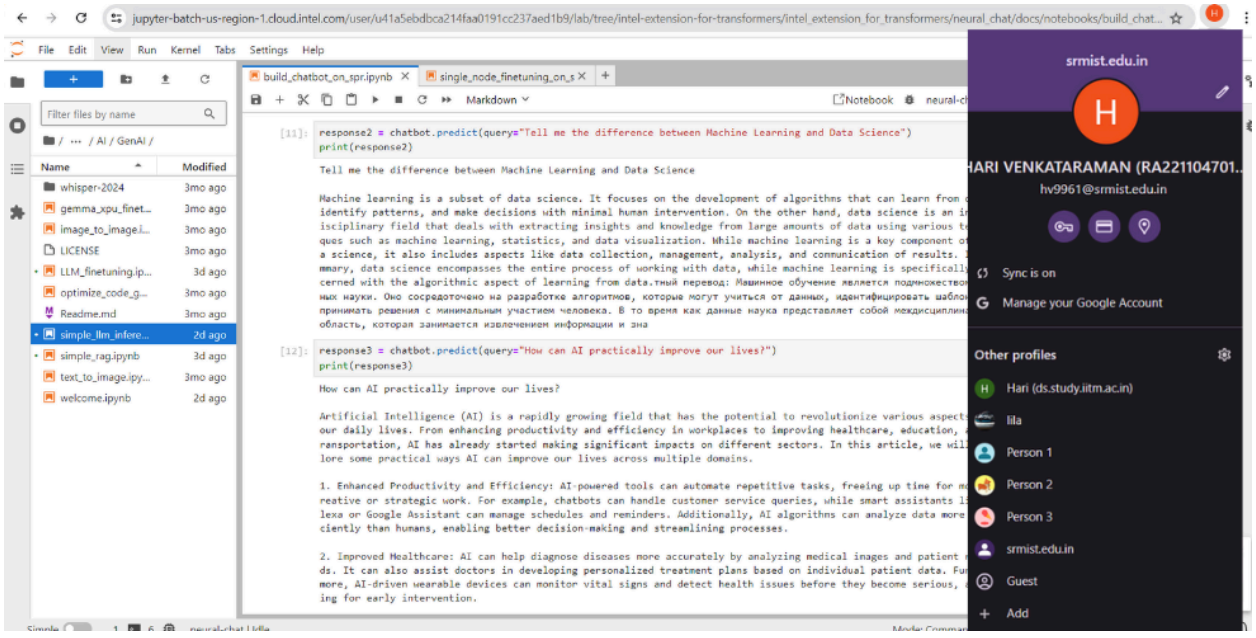

Fig 4: Examples/Responses 2 & 3

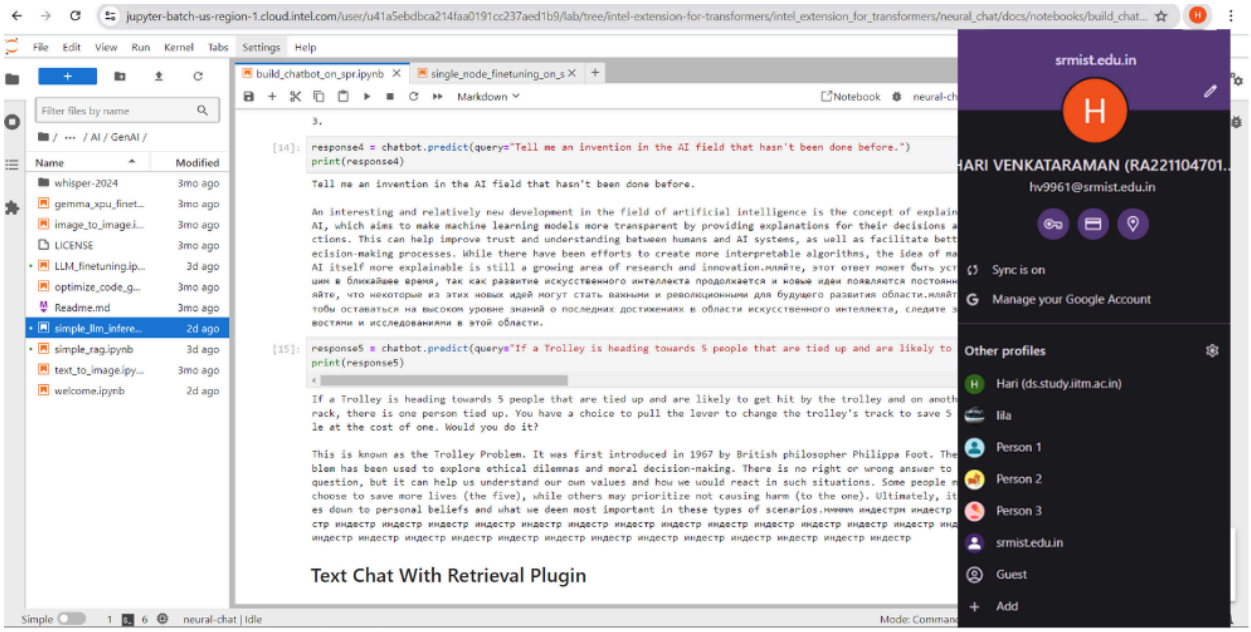

Fig 5: Examples/Responses 4 & 5

The above images show the running of the build_chatbot_on_spr with 5 sample responses and queries to the chatbot. The next part of this implementation is the finetuning of the chatbot (task named single_node_finetuning_on_spr). The fine tuning of the chatbot and the results/metrics as well as sample queries are in the recording below:

https://drive.google.com/file/d/1PTHkHex8fgw1x7xTGv4mBL6jrwYVgXjH/view?usp=sharing

What I learnt

So overall, upon running these Jupyter Notebooks and customizing parameters, I learnt that Large Language Models, like GPT-3, are powerful tools capable of understanding and generating human-like text. Inference refers to the process of using a trained model to make predictions or generate responses. Typically, LLMs require significant computational resources, often utilizing GPUs for efficient processing. However, running inference on CPUs can be beneficial in certain scenarios, such as when GPUs are not available or when deploying models

on edge devices. As far as the Fine-Tuning of the Chatbots, I learnt that it involves adjusting a pre-trained model on a specific dataset to improve its performance on a particular task. In our project, we focused on fine-tuning a chatbot to enhance its response accuracy. The process involved several steps:

- **Data Collection**: We gathered a dataset of conversations relevant to our chatbot's intended use. This included customer service interactions, FAQs, and other domain-specific dialogues.
- **Preprocessing**: The data was cleaned and formatted to ensure consistency and relevance. This step involved removing unnecessary text, correcting errors, and organizing the data into a suitable structure for training.
- **Model Training**: Using Intel's hardware, we fine-tuned the LLM on our dataset. This involved adjusting the model's parameters and optimizing it for our specific use case.
- **Inference on CPU**: To make the chatbot more accessible and cost-effective, we implemented inference on CPUs. This required optimizing the model to ensure it could run efficiently without compromising performance. Techniques such as model quantization and pruning were employed to reduce the computational load.
- **Testing and Evaluation**: The fine-tuned chatbot was rigorously tested to assess its accuracy and responsiveness. We compared its performance against baseline models to ensure significant improvements.

Results and Reflections

Overall, participating in Intel's Industrial Training Program was an invaluable experience. It provided practical skills, industry insights, and the opportunity to work on cutting-edge

technology. The project not only enhanced my understanding of LLMs and NLP but also equipped me with the knowledge to tackle real-world challenges in the tech industry. Intel's Industrial Training Program offers a unique blend of theoretical knowledge and practical experience, preparing students for careers in technology. The skills and knowledge gained through this experience will undoubtedly be beneficial in my future endeavors in the field of technology.