INTEL UNNATI INDUSTRIAL TRAINING PROGRAM 2024

Implementing GenAI : CPU-Based LLM Inference and Custom Chatbot Fine-Tuning

Report submitted by students of R V College of Engineering
Mentor: **Ms.Rajatha**
Application ID: **358**
Team Name: **Intellects**

Sl No.	Name	USN	Email id
1.	**Hitesh S P**	1RV22CS070	hiteshp.cs22@rvce.edu.in
2.	**Darshan Kashyap N**	1RV22CS044	darshankn.cs22@rvce.edu.in
3.	**Arahanth M**	1RV22CS027	arahanthm.cs22@rvce.edu.in
4.	**Pramath K P**	1RV22CS142	pramathkp.cs22@rvce.edu.in
5.	**Pranav S Kameshwar**	1RV22CD041	pranavsk.cd22@rvce.edu.in

1. INTRODUCTION:

Problem statement: Introduction to GenAl and Simple LLM Inference on CPU and fine-tuning of LLM Model to create a Custom Chatbot.

Generative AI (GenAI) encompasses technologies that produce new content, ranging from text to images, based on existing data. Large Language Models (LLMs), a cornerstone of GenAI, can generate coherent and contextually relevant text, making them ideal for applications like chatbots. This report explores the basics of LLM inference on CPUs, demonstrating how even without specialized hardware, powerful AI models can function effectively. Additionally, it delves into the fine-tuning process, where pre-trained LLMs are adapted to specific tasks, enabling the creation of custom chatbots tailored to unique requirements. This blend of theoretical insights and practical steps provides a comprehensive guide for leveraging GenAI in real-world applications.

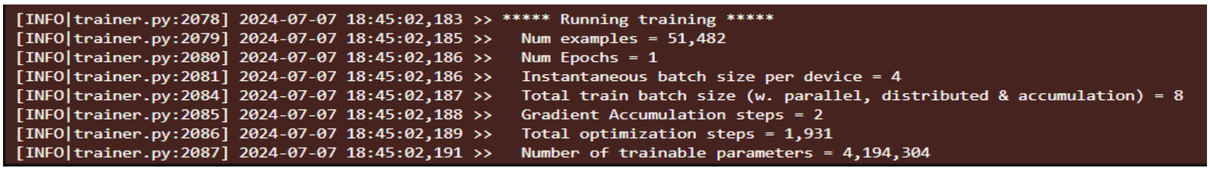
2. UNIQUE IDEA & APPROACH:

While fine-tuning without GPU people face various issues such as long training time and lack of a proper method to conduct the process, so in this project we have customized the "alpaca dataset" before training with our custom data with fewer parameters and then trained significantly **reduced the time of training**, the same process was also performed with other models to obtain the required fine-tuned model and then a chatbot was created to using all the available and created models.
The other unique feature of the project is the exploration of a new tech called **Ollama** using which we can run LLM models on a CPU without an internet connection and even the fine-tuned files could be copied to the existing LLM models in Ollama.

To build a custom chatbot, two major methodologies were taken into consideration. The first method is about using Jupyter Notebook to download LLM, finetune and develop the chatbot on Jupyter Notebook itself without any user interfaces.
The Second method is about using Ollama to build our custom chatbot taking into consideration the user interface and deploying the chatbot fully locally which can run on the CPU.
Both the methods were tried by considering various LLMs such as "Intel/neural-chat-7b-v3-1", "meta-llama/Llama-2-7b-chat-hf" and at last our own fine-tuned LLM.

2.1. Custom Chatbot using Jupyter Notebook

Fine-tuning was conducted on the Intel Developer Cloud's Jupyter notebook environment. Initially attempted with the Alpaca Dataset, which contained approximately **26,000 parameters**, the process required **17 hours**. To expedite, parameter reduction and inclusion of custom data related to **Intel, RVCE, and customized information** were implemented, reducing fine-tuning time to **45-50 minutes**. Subsequently, models such as "Intel/neural-chat-7b-v3-1" and "meta-llama/Llama-2-7b-chat-hf" were downloaded and configured. Outputs from the fine-tuned model were compared with the original, utilized in developing the web interface for the Ollama model which will be explained in detail in the next section of the report. Additionally, a custom IDC chatbot was developed, offering users options to interact with different models, as detailed in the GitHub repository containing all training and chatbot information.



```
[INFO|trainer.py:2078] 2024-07-07 18:45:02,183 >> ***** Running training *****
[INFO|trainer.py:2079] 2024-07-07 18:45:02,185 >>   Num examples = 51,482
[INFO|trainer.py:2080] 2024-07-07 18:45:02,186 >>   Num Epochs = 1
[INFO|trainer.py:2081] 2024-07-07 18:45:02,186 >>   Instantaneous batch size per device = 4
[INFO|trainer.py:2084] 2024-07-07 18:45:02,187 >>   Total train batch size (w. parallel, distributed & accumulation) = 8
[INFO|trainer.py:2085] 2024-07-07 18:45:02,188 >>   Gradient Accumulation steps = 2
[INFO|trainer.py:2086] 2024-07-07 18:45:02,189 >>   Total optimization steps = 1,931
[INFO|trainer.py:2087] 2024-07-07 18:45:02,191 >>   Number of trainable parameters = 4,194,304
```

2.2. Custom chatbot using Ollama:

i. Introduction to Ollama:

Ollama is an **open-source project** that serves as a powerful and user-friendly platform for running LLMs on your local machine. It acts as a bridge between the complexities of LLM technology and the desire for an accessible and customizable AI experience.

At its core, Ollama simplifies the process of downloading, installing, and interacting with a wide range of LLMs, empowering users to explore their capabilities **without the need for extensive technical expertise or reliance on cloud-based platforms**.

ii. Purpose of using Ollama:

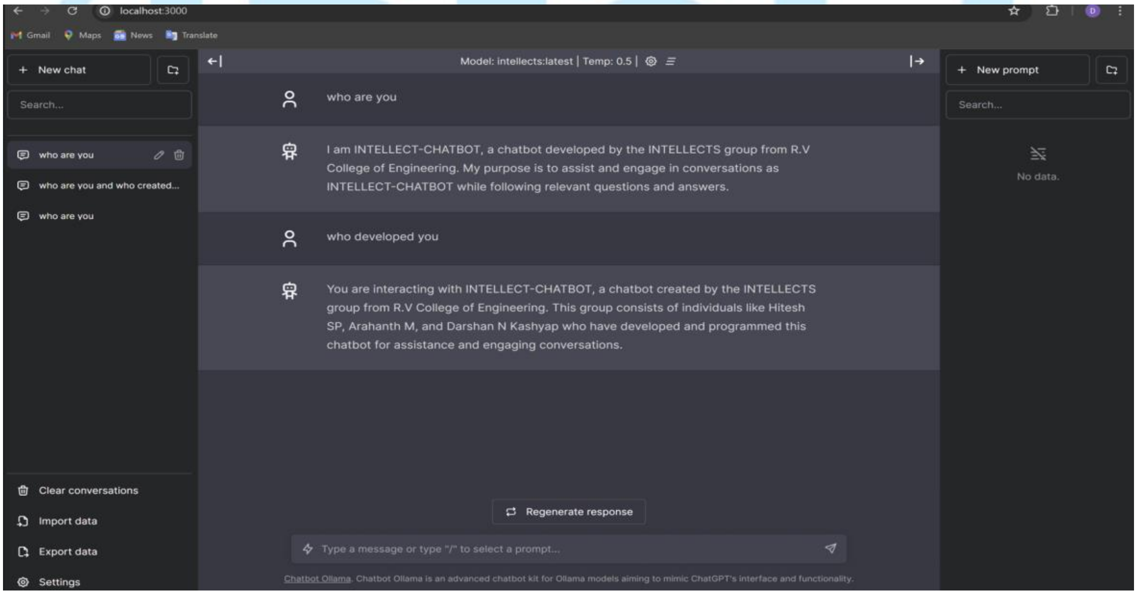
Ollama enhances user experience and maximizes the potential of local LLMs through its comprehensive features. It offers a diverse and expanding library of pre-trained models, ensuring easy download and management. The installation process is user-friendly, and tailored for Windows, macOS, and Linux. Ollama's local API allows seamless integration of LLMs into applications, facilitating efficient communication and customization. The platform supports hardware acceleration, optimizing performance for large-scale models. It provides both command-line and graphical interfaces for intuitive interactions. **Crucially, Ollama can operate entirely offline, ensuring uninterrupted access and addressing privacy concerns by keeping data local**. Additionally, the active open-source community surrounding Ollama fosters continuous development and innovation.

iii. Process flow of developing the Custom Chatbot:

The setup and installation of Ollama are straightforward, with well-documented instructions available for Windows, macOS, and Linux. Following installation, the Ollama library offers an extensive range of pre-trained LLM models, including Llama 2 for versatile text tasks, **Neural-chat by Intel for optimized conversational AI, Mistral for creative writing and Llama model for summarized data**. After downloading the desired model, interaction with the LLM can be conducted using **Ollama's command-line interface** (CLI). The CLI (nothing but the terminal of our system) facilitates engagement in conversations, adjustment of model parameters for customized responses, and utilization of various commands to enhance the experience. Prompts are typed and responses are generated directly in the terminal, ensuring a seamless and interactive chatbot development process. Information regarding the installation, models , prompts and responses are provided in detail in the readme file of the GitHub repository.

Ollama provides a visual and intuitive way to interact with LLMs through seamless **integration with various web-based user interfaces** developed by the community. Activating the web-based UI involves running specific commands and making configuration changes. These web UIs were later modified and customized according to specific requirements using **TypeScript, React, and Tailwind CSS** frameworks. Extensive work was done in developing the UI to enhance its functionality and user experience. Detailed information and step-by-step instructions for setting up the web UI can be found in the **README file** of the GitHub repository.

Fine-tuning involves customizing a base LLM model by modifying its parameters and weights to optimize performance for specific tasks or domains. Neural-chat LLM was considered to be fine-tuned and the new LLM was named as "**intellects**" .The process includes data preparation, model selection, configuration of fine-tuning parameters, training, evaluation, and deployment. Once the Neural-Chat model was fine-tuned in IDC with a custom dataset it was added to the same folder(where the chatbot is present on the local device) and then copied to the intellects model by using the following command(**Ollama cp fine-tuned-model intellects**). Fine-tuning can significantly enhance the accuracy and relevance of the LLM for specialized tasks. In this project, fine-tuning was primarily focused on topics related **to Intel, and the Intel UNNATI program, and included some questions about RVCE and some general details about the project**. This tailored approach ensures the model's responses are well-suited to the specific context and requirements.



who are you

I am INTELLECT-CHATBOT, a chatbot developed by the INTELLECTS group from R.V College of Engineering. My purpose is to assist and engage in conversations as INTELLECT-CHATBOT while following relevant questions and answers.

who developed you

You are interacting with INTELLECT-CHATBOT, a chatbot created by the INTELLECTS group from R.V College of Engineering. This group consists of individuals like Hitesh SP, Arahanth M, and Darshan N Kashyap who have developed and programmed this chatbot for assistance and engaging conversations.

Chatbot Ollama. Chatbot Ollama is an advanced chatbot kit for Ollama models aiming to mimic ChatGPT's interface and functionality.

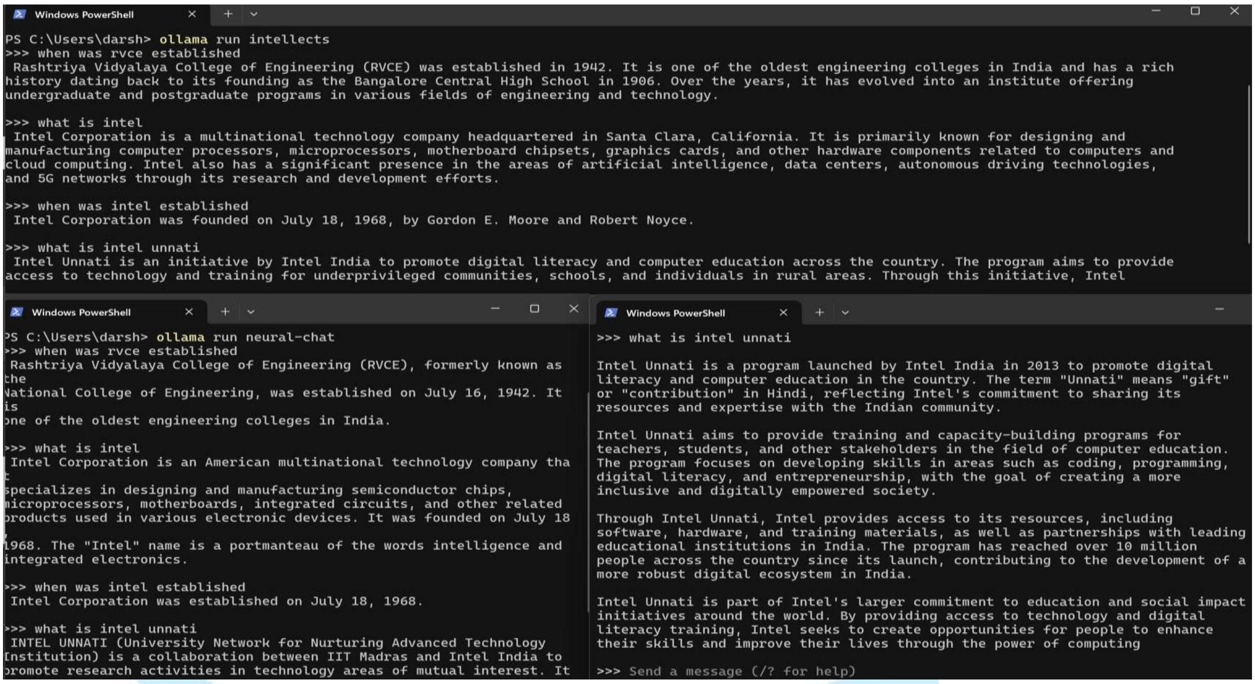
iv. Summary and Future Scope

Ollama demonstrates its versatility across various domains, enabling applications such as creative writing, code generation, language translation, and personalized AI assistants. It supports research with literature reviews and data analysis, enhances education through personalized learning tools, and improves customer service with intelligent chatbots. Ethical considerations are crucial, focusing on bias mitigation, privacy protection, transparency, responsible content generation, and human oversight. Looking forward, Ollama and local LLMs are assured of advancing AI accessibility, innovation, and ethical deployment, driving the future of AI development and empowering diverse applications across industries.

3. COMPARATIVE STUDY:

As mentioned before, our chatbot integrates various models such as "neural-chat," "llama2," and the fine-tuned "intellects" model. Users have the flexibility to select their preferred model for interaction. After conducting a comparative analysis of the aforementioned LLMs by posing identical questions to each and evaluating efficiency, response time, and accuracy, the following conclusions were drawn:

- neural chat works the fastest in the comparison.
- next the intellects model(with answers to the fine-tuned questions also).
- llama is slower compared to the other two models.

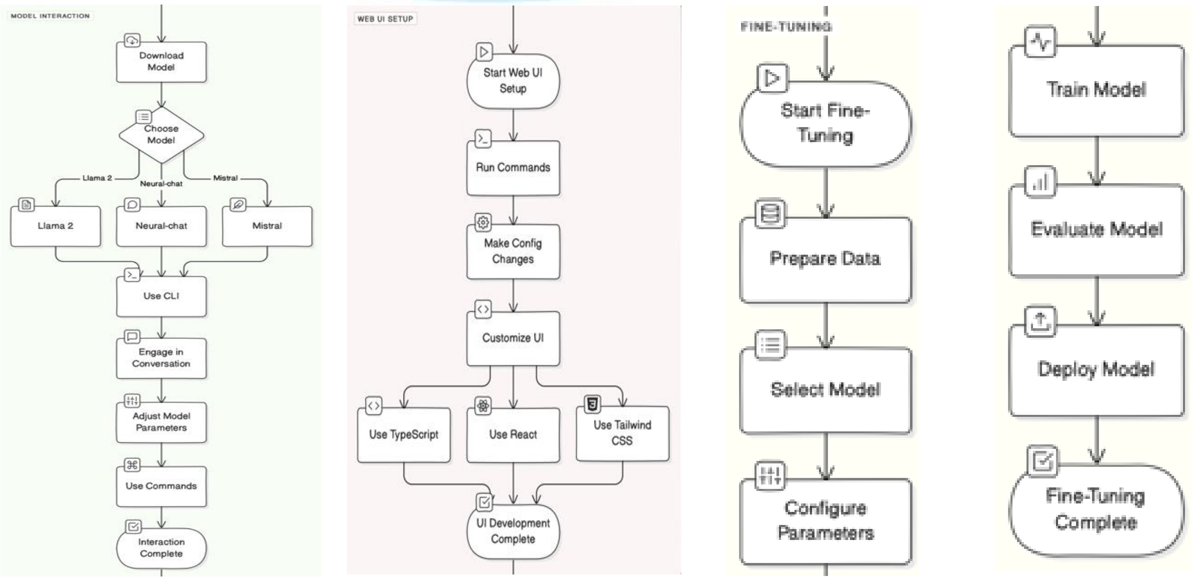


```
PS C:\Users\darsh> ollama run intellects
>>> when was rvce established
 Rashtriya Vidyalaya College of Engineering (RVCE) was established in 1942. It is one of the oldest engineering colleges in India and has a rich
history dating back to its founding as the Bangalore Central High School in 1906. Over the years, it has evolved into an institute offering
undergraduate and postgraduate programs in various fields of engineering and technology.

>>> what is intel
 Intel Corporation is a multinational technology company headquartered in Santa Clara, California. It is primarily known for designing and
manufacturing computer processors, microprocessors, motherboard chipsets, graphics cards, and other hardware components related to computers and
cloud computing. Intel also has a significant presence in the areas of artificial intelligence, data centers, autonomous driving technologies,
and 5G networks through its research and development efforts.

>>> when was intel established
 Intel Corporation was founded on July 18, 1968, by Gordon E. Moore and Robert Noyce.

>>> what is intel unnati
 Intel Unnati is an initiative by Intel India to promote digital literacy and computer education across the country. The program aims to provide
access to technology and training for underprivileged communities, schools, and individuals in rural areas. Through this initiative, Intel
```

```
PS C:\Users\darsh> ollama run neural-chat
>>> when was rvce established
 Rashtriya Vidyalaya College of Engineering (RVCE), formerly known as
the
National College of Engineering, was established on July 16, 1942. It
is
one of the oldest engineering colleges in India.

>>> what is intel
 Intel Corporation is an American multinational technology company tha
t
specializes in designing and manufacturing semiconductor chips,
microprocessors, motherboards, integrated circuits, and other related
products used in various electronic devices. It was founded on July 18
,
1968. The "Intel" name is a portmanteau of the words intelligence and
integrated electronics.

>>> when was intel established
 Intel Corporation was established on July 18, 1968.

>>> what is intel unnati
 INTEL UNNATI (University Network for Nurturing Advanced Technology
Institution) is a collaboration between IIT Madras and Intel India to
promote research activities in technology areas of mutual interest. It
```

```
>>> what is intel unnati

Intel Unnati is a program launched by Intel India in 2013 to promote digital
literacy and computer education in the country. The term "Unnati" means "gift"
or "contribution" in Hindi, reflecting Intel's commitment to sharing its
resources and expertise with the Indian community.

Intel Unnati aims to provide training and capacity-building programs for
teachers, students, and other stakeholders in the field of computer education.
The program focuses on developing skills in areas such as coding, programming,
digital literacy, and entrepreneurship, with the goal of creating a more
inclusive and digitally empowered society.

Through Intel Unnati, Intel provides access to its resources, including
software, hardware, and training materials, as well as partnerships with leading
educational institutions in India. The program has reached over 10 million
people across the country since its launch, contributing to the development of a
more robust digital ecosystem in India.

Intel Unnati is part of Intel's larger commitment to education and social impact
initiatives around the world. By providing access to technology and digital
literacy training, Intel seeks to create opportunities for people to enhance
their skills and improve their lives through the power of computing

>>> Send a message (/? for help)
```

4. ARCHITECTURE DIAGRAM:



5. FEATURES OFFERED:

The various features of our custom chatbot are :

1. **Offline**: The chatbot operates entirely offline, ensuring accurate and reliable performance without requiring an internet connection.

2. **Dark Mode and Light Mode**: Users can toggle between dark and light modes to suit their preference for interface appearance.

3. **Logo**: Our interface prominently displays the Intel logo, reflecting our brand identity and commitment to quality.

4. **Import and Export Documents**: The custom chatbot has the ability to import any file present in the local machine and use it get responses to queries related to that particular file and export the chat in json format.

5. **Copy Option**: Facilitates easy copying of chatbot responses, allowing users to conveniently save or share information provided.

6**. Saving History**: The platform maintains a comprehensive history of users' interactions and searches, enabling convenient access to past queries for reference or review.

6. CONCLUSION:

In conclusion, the process of building a custom chatbot using Generative AI (GenAI) has provided valuable insights into its fundamental principles and practical applications. The use of **Jupyter notebook on the Intel Developer Cloud made it easier to fine-tune** the LLM, greatly improving its performance. This process was key in smoothly integrating the LLM with Ollama, showing how **Ollama can be used for creative writing, code generation, language translation, and personalized AI** applications. It’s important to consider ethical issues like bias and privacy to make sure these advancements are beneficial. Looking ahead, the collaboration between Ollama and local LLM technologies will boost AI accessibility and innovation, ensuring responsible use across various applications. This experience has significantly enhanced proficiency in leveraging advanced AI technologies to develop and optimize tailored chatbot solutions, setting a solid foundation for future advancements in natural language processing and interactive AI systems.

7. REFERENCES:

https://github.com/ivanfioravanti/chatbot-ollama.git
https://github.com/intel/intel-extension-for-transformers.git