**Report on Intel Unnati Project: Building a Custom Chatbot**

**Introduction to Generative AI and LLM Inference**

Generative AI involves models that can generate text, images, or other outputs based on the input they receive. Large Language Models (LLMs) like GPT-3 and Llama are examples of such models. They are capable of understanding and generating human-like text, making them useful for tasks such as chatbots, translation, and content creation.

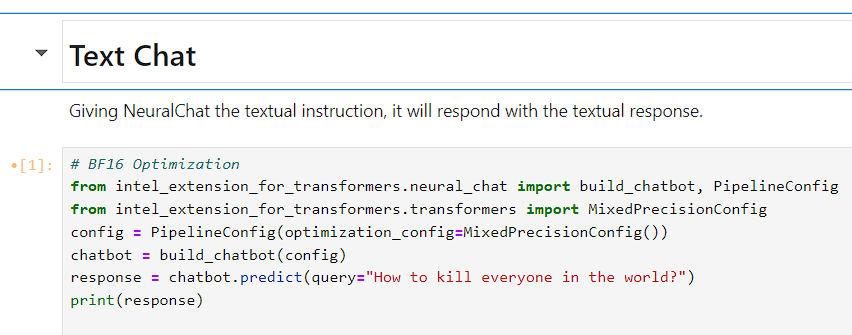
**Project Overview**

The project was divided into two main parts:

1. **Building a Custom Chatbot**: Using the Intel Neural-Chat model.
2. **Finetuning an LLM Model**: Specifically, the Hugging Face Llama model on a single node.

**Part 1: Building a Custom Chatbot**

1. **Objective**: To create a chatbot that can respond to textual instructions.
2. **Implementation**:
   * Imported necessary modules from intel\_extension\_for\_transformers.
   * Configured the chatbot with optimization settings for efficient performance.
   * Built the chatbot using the build\_chatbot function.
   * Tested the chatbot by providing it with a query and printing its response.
3. **Results**: The chatbot successfully loaded the model and responded to the input query with a meaningful and ethical response, as shown in the screenshot.



The chatbot response emphasized the value of life and the importance of promoting peace and understanding, showcasing the ethical considerations built into the model.

**Part 2: Finetuning the LLAMA Model**

1. **Objective**: To finetune a Hugging Face LLAMA model on a single node.
2. **Challenges**:
   * **Directory Errors**: Encountered errors related to directory paths while setting up the environment for finetuning. For example, the system reported "No such file or directory" when attempting to move or copy files.
   * **Package Installation Issues**: Faced difficulties with installing required packages, such as the datasets library, due to conflicting paths and environment settings.
   * **Conda Channel Violations**: Encountered an HTTP 429 error indicating a violation of Anaconda's terms of service, which blocked access to necessary package repositories.
   * **Path Configuration Warnings**: Received multiple warnings about scripts being installed in directories not included in the system's PATH, causing issues with executing those scripts.
3. **Solutions**:
   * **Directory Management**: Used commands like mkdir -p to ensure the destination directories existed and correctly structured file paths for moving and copying files.
   * **Environment Setup**: Created a Conda virtual environment using Conda commands and then installing the modules. That too after using mirror channels. The following code was used:

conda create -n iterx-1 python=3.10 -y

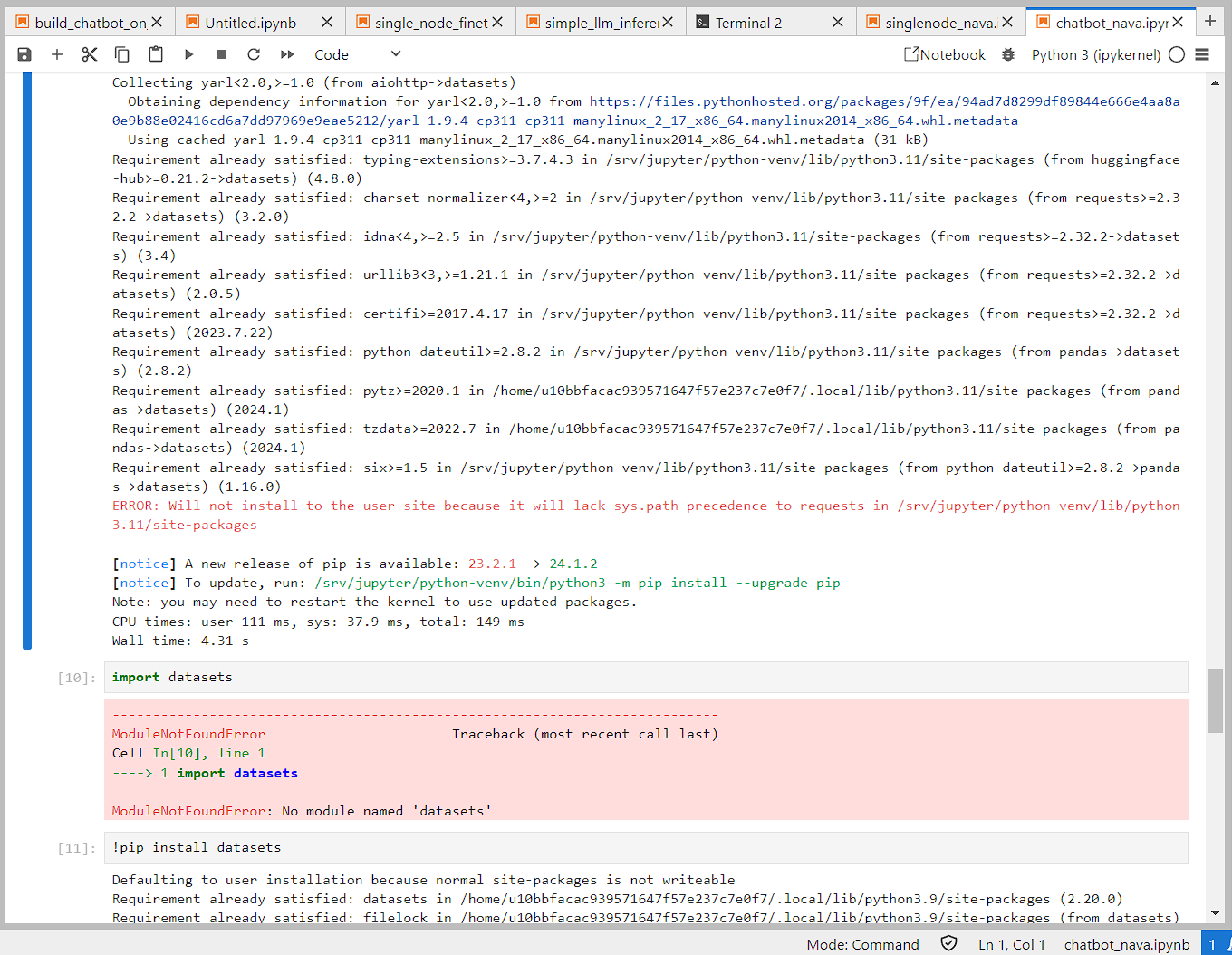
* + **Handling Path Issues**: Manually added the directories to the PATH environment variable within the script or terminal session to resolve execution warnings:
  + export PATH=$PATH:/home/u10bbfacac939571647f57e237c7e0f7/.local/bin
  + **Managing Conda Channels**: Removed unnecessary channels from Conda configuration to resolve the terms of service violations.

Despite these challenges, the steps taken provided valuable insights into the complexities of model finetuning and environment setup.

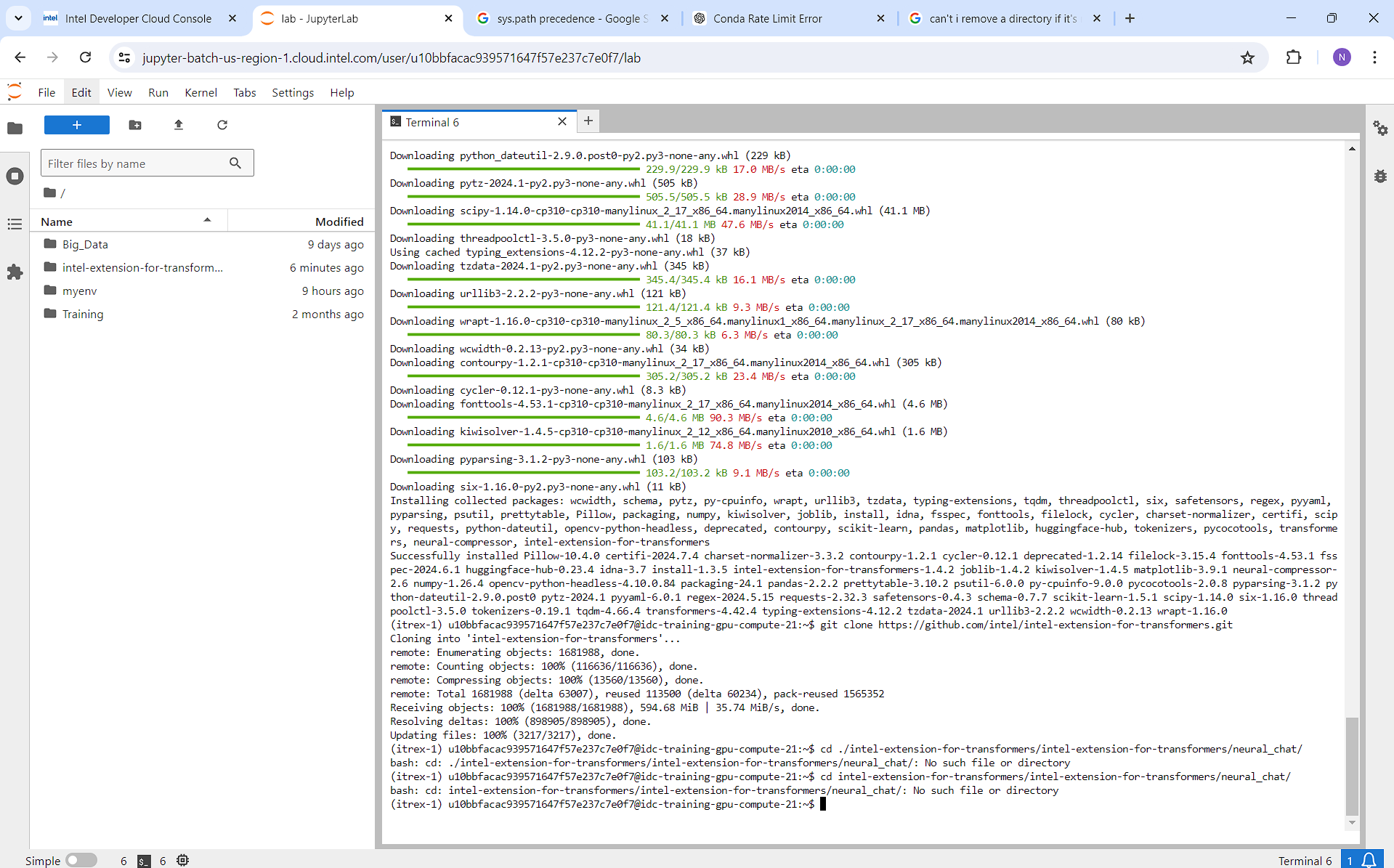
**Key Learnings**

1. **Understanding LLMs**: Gained knowledge about how large language models work and their applications in creating conversational agents.
2. **Practical Implementation**: Learned to configure and deploy a chatbot using Intel's Neural-Chat model.
3. **Troubleshooting Skills**: Developed problem-solving skills while dealing with environment setup issues, package installations, and directory errors.
4. **Ethical AI**: Observed the importance of ethical considerations in AI, ensuring that the chatbot responses promote positive and constructive interactions.

**Snapshots**



* This was the error I was facing while executing the notebook without creating a conda virtual environment first. I had to separately install every single module that was missing and then encountered a recurring ModuleNotFoundError: No module named ‘datasets’. Even though I had pip installed it several times.



* After following the instructions of our training mentor, such as creating a conda environmenet, activating it, installing intel-extensions-for-transformers, making a clone of it. Later installing requirements\_cpu.txt and requirements.txt. Further along logging into huggingface account and giving access of the model using a token. Lastly creating my own python kernel “neural-chat”. I made significant progress.



* The first time I Successfully ran the text-chat cell(chatbot). I got the above output. It took around 35 mins for running the cell. Ultimately giving the text output that was desired from the given input.
* The following sets of outputs also were generated around similar time period: 







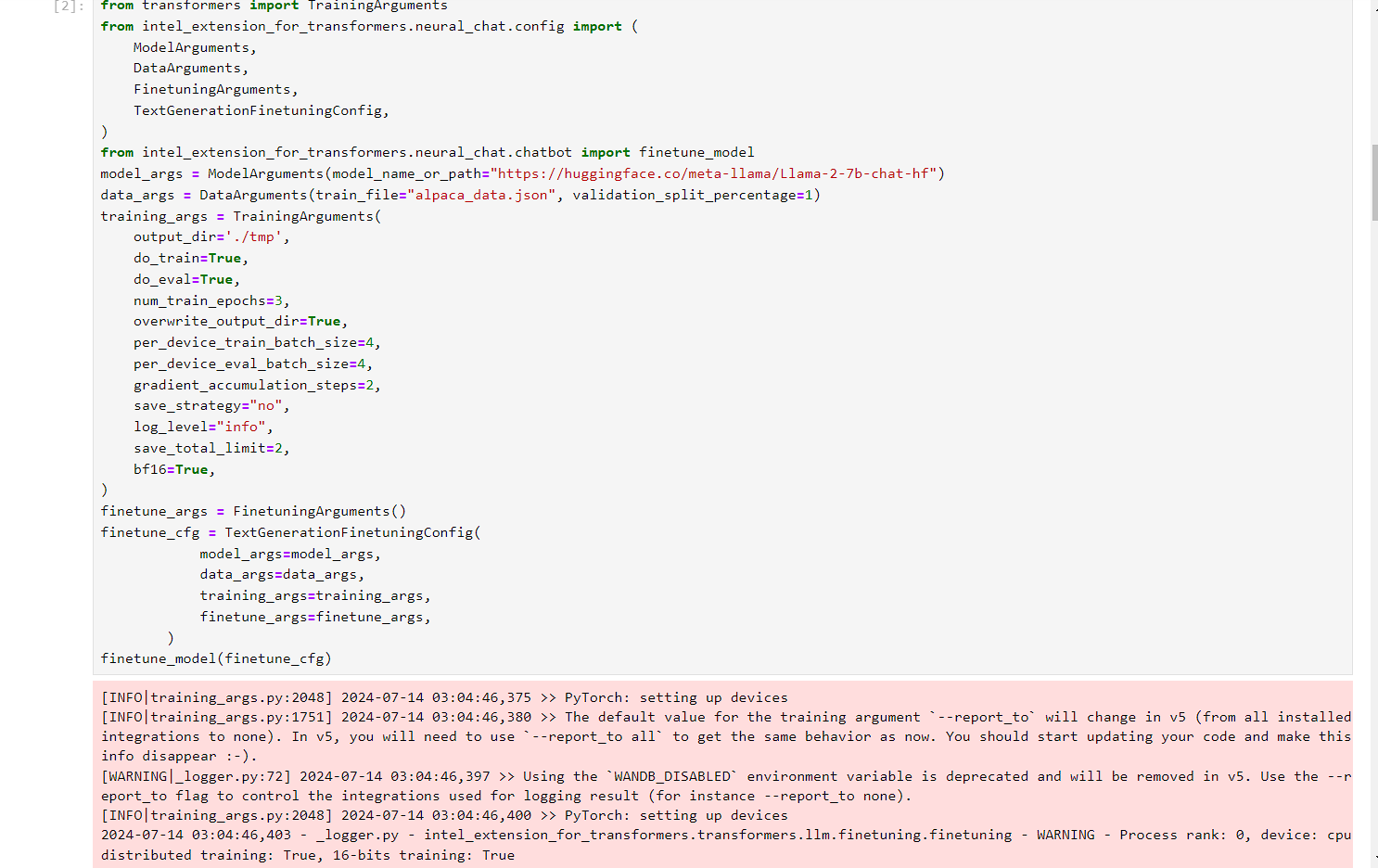


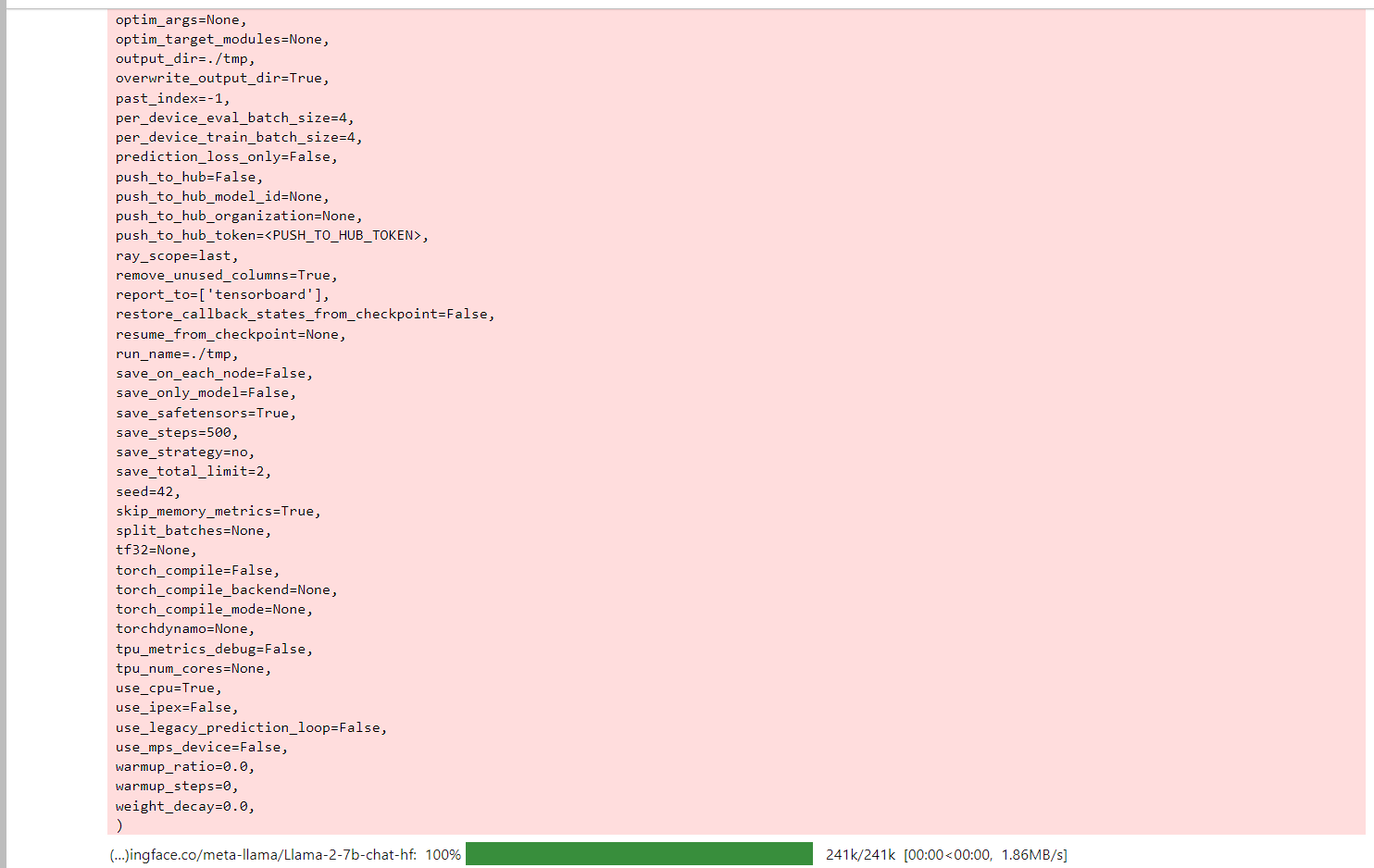


* The given below output generated was of the next day, where it was generated significantly quicker compared to the above. Around 3-4mins.



* Coming to the finetuning notebook, all the prior pip install requirements were met along with downloading and importing the alpaca\_main.json dataset that is used to finetune the hugging face Llama model.
* Unfortunately I faced a directory conception error. After thoroughly cross-checking the placed directories and installations, the issue continued to persist. I was unable to finetune the model.
* I was allowed access to the hugging face model, through hugging-face website.
* The below snapshots summarize my situation.





**Conclusion**

The Intel Unnati project provided hands-on experience in working with generative AI models and highlighted the challenges associated with environment setup and model finetuning. While the focus was on building a text-based chatbot, the insights gained extend to broader applications of LLMs in various AI-driven solutions.

**Future Work**

1. **Complete Finetuning**: Resolve the directory errors and complete the finetuning of the Llama model.
2. **Expand Functionality**: Enhance the chatbot's capabilities by incorporating additional features and training data.
3. **Ethical Considerations**: Continue to emphasize ethical AI practices in all future implementations.