❑ *Problem Statement:*

Finetuning a pre-trained LLM using a dataset for text generation problem using LoRA technique.

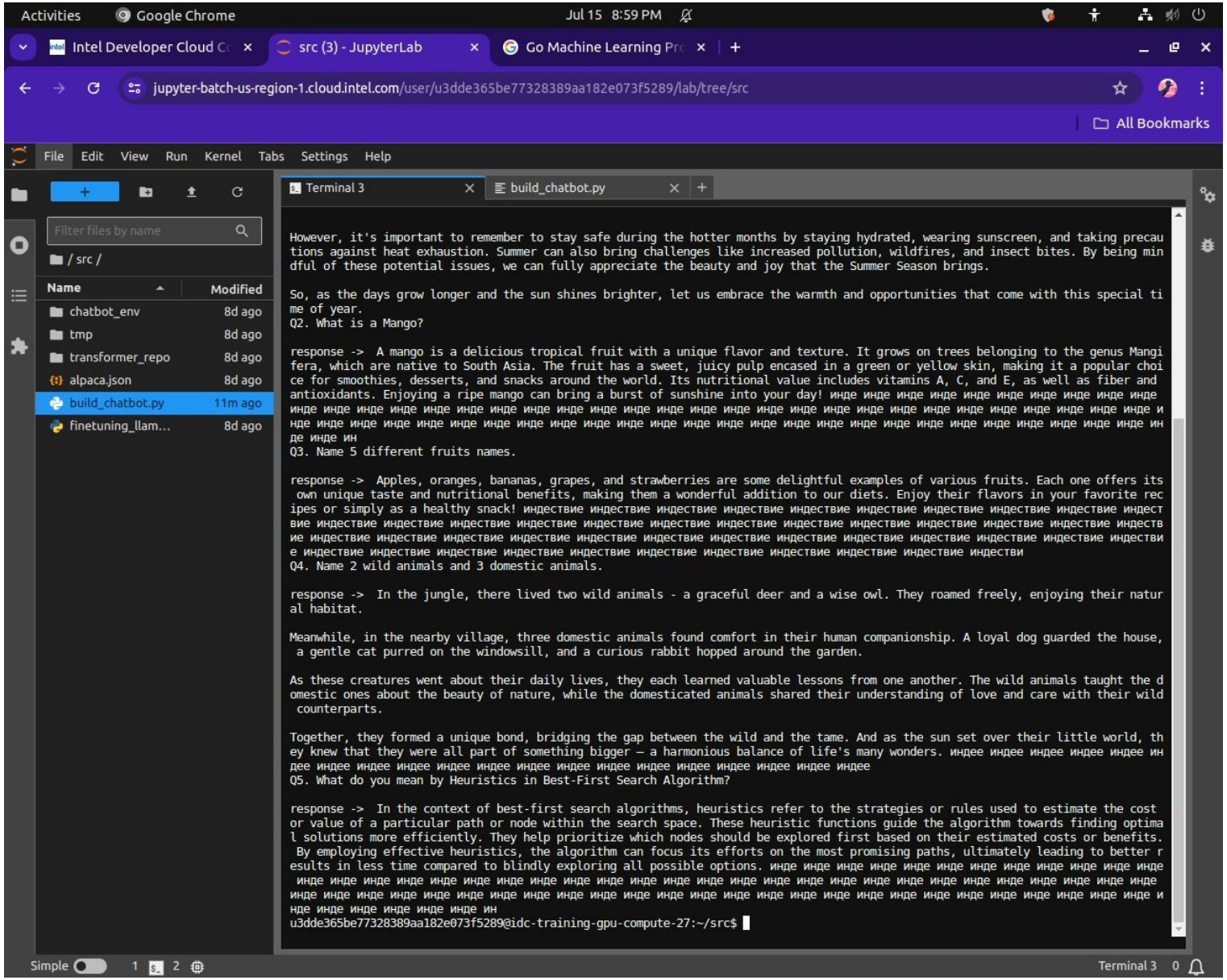
❑ *Solution:*

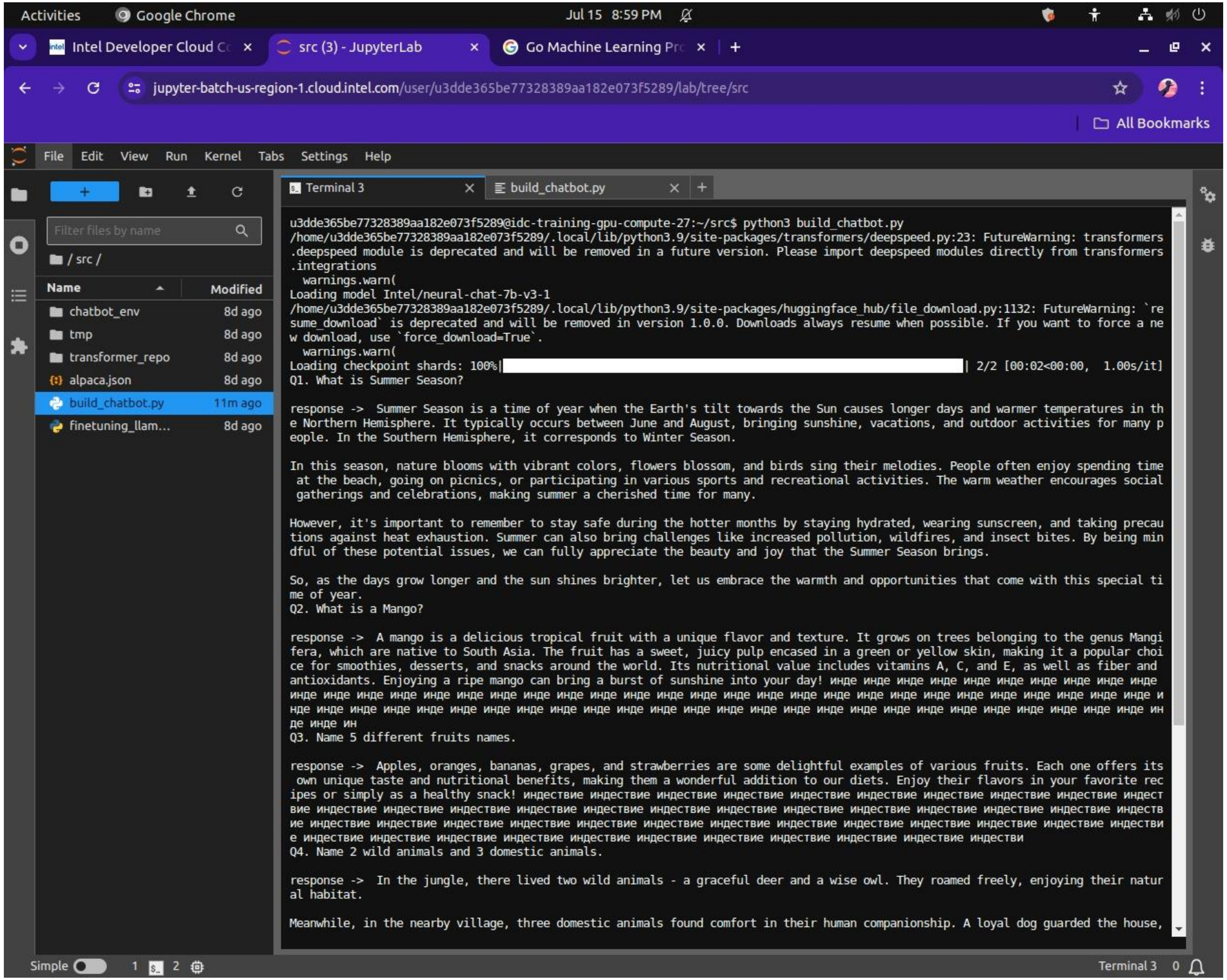
In deep learning, **fine-tuning** is an approach to transfer learning in which the parameters of a pre-trained model are trained on new data. Fine-tuning can be done on the entire neural network, or on only a subset of its layers, in which case the layers that are not being fine-tuned are "frozen“. A model may also be augmented with "adapters" that consist of far fewer parameters than the original model, and fine-tuned in a parameter–efficient way by tuning the weights of the adapters and leaving the rest of the model's weights frozen.

Models that are pre-trained on large and general corpora are usually fine-tuned by reusing the model's parameters as a starting point and adding a task-specific layer trained from scratch. Fine-tuning the full model is common as well and often yields better results, but it is more computationally expensive.

Fine-tuning is typically accomplished with supervised learning, but there are also techniques to fine-tune a model using weak supervision. Fine-tuning can be combined with a reinforcement learning from human feedback-based objective to produce language models like ChatGPT and Sparrow.

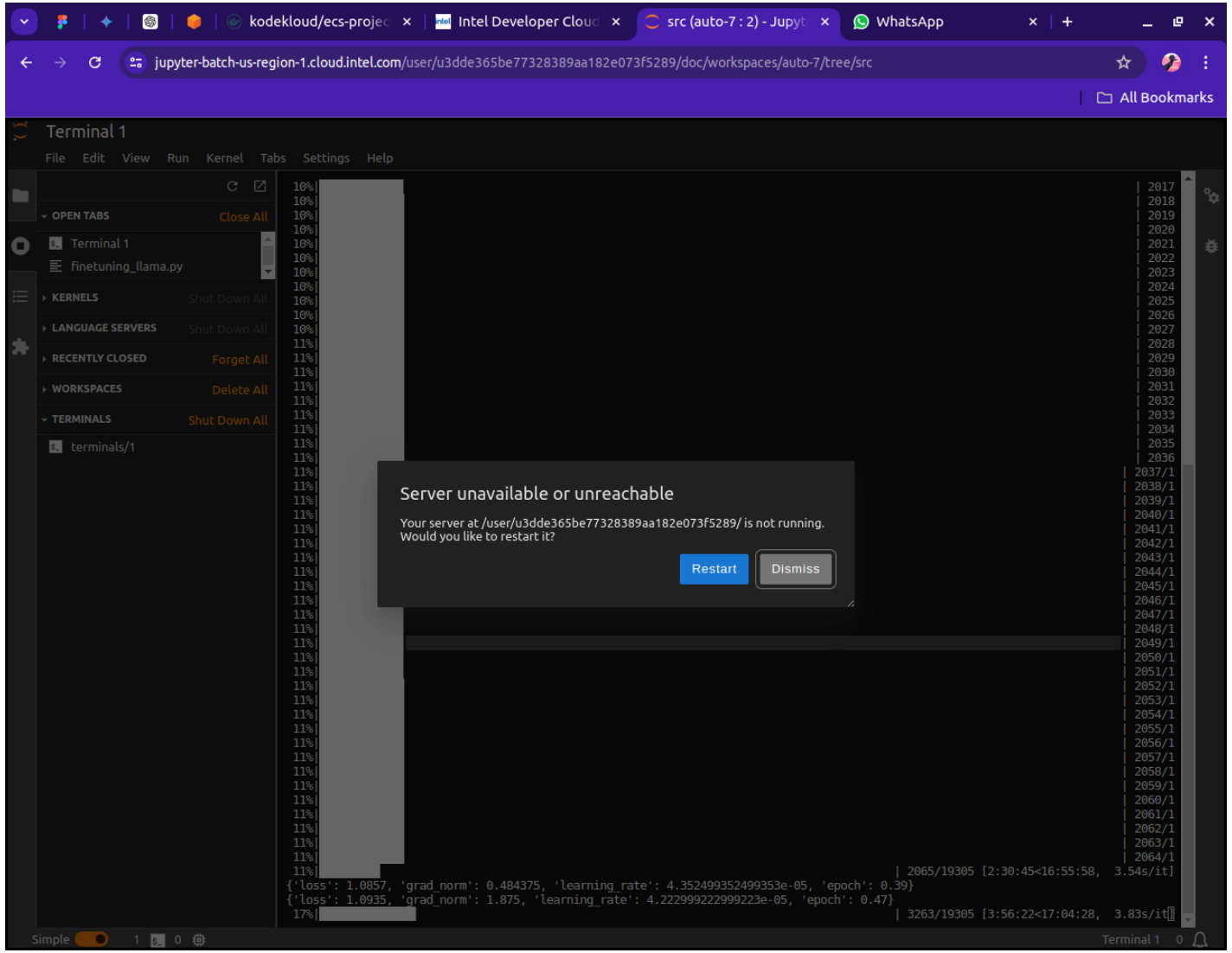
Observations:




Activities
Google Chrome
Jul 15 8:59 PM
Intel Developer Cloud C
src (3) - JupyterLab
Go Machine Learning Pr
jupyter-batch-us-region-1.cloud.intel.com/user/u3dde365be77328389aa182e073f5289/lab/tree/src
All Bookmarks
File Edit View Run Kernel Tabs Settings Help
Filter files by name
/ src /
Name Modified
chatbot_env 8d ago
tmp 8d ago
transformer_repo 8d ago
alpaca.json 8d ago
build_chatbot.py 11m ago
finetuning_llam... 8d ago
Terminal 3
build_chatbot.py
u3dde365be77328389aa182e073f5289@idc-training-gpu-compute-27:~/src$ python3 build_chatbot.py
/home/u3dde365be77328389aa182e073f5289/.local/lib/python3.9/site-packages/transformers/deepspeed.py:23: FutureWarning: transformers.deepspeed module is deprecated and will be removed in a future version. Please import deepspeed modules directly from transformers.integrations
warnings.warn(
Loading model Intel/neural-chat-7b-v3-1
/home/u3dde365be77328389aa182e073f5289/.local/lib/python3.9/site-packages/huggingface_hub/file_download.py:1132: FutureWarning: `resume_download` is deprecated and will be removed in version 1.0.0. Downloads always resume when possible. If you want to force a new download, use `force_download=True`.
warnings.warn(
Loading checkpoint shards: 100%| | 2/2 [00:02<00:00, 1.00s/it]
Q1. What is Summer Season?
response -> Summer Season is a time of year when the Earth's tilt towards the Sun causes longer days and warmer temperatures in the Northern Hemisphere. It typically occurs between June and August, bringing sunshine, vacations, and outdoor activities for many people. In the Southern Hemisphere, it corresponds to Winter Season.
In this season, nature blooms with vibrant colors, flowers blossom, and birds sing their melodies. People often enjoy spending time at the beach, going on picnics, or participating in various sports and recreational activities. The warm weather encourages social gatherings and celebrations, making summer a cherished time for many.
However, it's important to remember to stay safe during the hotter months by staying hydrated, wearing sunscreen, and taking precautions against heat exhaustion. Summer can also bring challenges like increased pollution, wildfires, and insect bites. By being mindful of these potential issues, we can fully appreciate the beauty and joy that the Summer Season brings.
So, as the days grow longer and the sun shines brighter, let us embrace the warmth and opportunities that come with this special time of year.
Q2. What is a Mango?
response -> A mango is a delicious tropical fruit with a unique flavor and texture. It grows on trees belonging to the genus Mangifera, which are native to South Asia. The fruit has a sweet, juicy pulp encased in a green or yellow skin, making it a popular choice for smoothies, desserts, and snacks around the world. Its nutritional value includes vitamins A, C, and E, as well as fiber and antioxidants. Enjoying a ripe mango can bring a burst of sunshine into your day! инде ин
Q3. Name 5 different fruits names.
response -> Apples, oranges, bananas, grapes, and strawberries are some delightful examples of various fruits. Each one offers its own unique taste and nutritional benefits, making them a wonderful addition to our diets. Enjoy their flavors in your favorite recipes or simply as a healthy snack! индествие индестви
Q4. Name 2 wild animals and 3 domestic animals.
response -> In the jungle, there lived two wild animals - a graceful deer and a wise owl. They roamed freely, enjoying their natural habitat.
Meanwhile, in the nearby village, three domestic animals found comfort in their human companionship. A loyal dog guarded the house,
Simple 1 2
Terminal 3 0

```
2024-07-06 17:20:52,639 - finetuning.py - intel_extension_for_transformers.transformers.llm.finetuning.finetuning - INFO - Using data
 collator of type DataCollatorForSeq2Seq
trainable params: 4,194,304 || all params: 6,742,609,920 || trainable%: 0.06220594176090199
[INFO|trainer.py:641] 2024-07-06 17:20:53,748 >> Using cpu_amp half precision backend
[INFO|trainer.py:2078] 2024-07-06 17:20:54,320 >> ***** Running training *****
[INFO|trainer.py:2079] 2024-07-06 17:20:54,320 >>   Num examples = 51,482
[INFO|trainer.py:2080] 2024-07-06 17:20:54,320 >>   Num Epochs = 3
[INFO|trainer.py:2081] 2024-07-06 17:20:54,320 >>   Instantaneous batch size per device = 4
[INFO|trainer.py:2084] 2024-07-06 17:20:54,320 >>   Total train batch size (w. parallel, distributed & accumulation) = 8
[INFO|trainer.py:2085] 2024-07-06 17:20:54,321 >>   Gradient Accumulation steps = 2
[INFO|trainer.py:2086] 2024-07-06 17:20:54,321 >>   Total optimization steps = 19,305
[INFO|trainer.py:2087] 2024-07-06 17:20:54,324 >>   Number of trainable parameters = 4,194,304
  0%|                                                                                | 0/19305 [00:00<?, ?it/s]
```

Terminal 3 0



Time to complete finetuning using LoRA is 17 hours.

Advantages:

- Efficiency in training and adaptation of large language models
- Reduced computational resources requirement
- Preservation of pre-trained model weights
- Reduced parameter overhead
- Efficient task-switching
- Faster iterations
- Scalability
- Memory efficiency

what will happen if we don't use LoRA?

If we don't use LoRa several potential drawbacks and limitations could arise:

1. Increased Computational Resources
2. Longer Training Time
3. Higher Costs
4. Overfitting Risks
5. Scalability Issues
6. Difficulty in Transfer Learning

❑ *Features Offered:*

Currently, we've fine-tuned the LLM for text generation using the Alpaca dataset. To extend its capabilities for generating project reports, we'll incorporate datasets from technical, business, and academic reports and implement advanced fine-tuning strategies.

❑ *Technologies Used:*

- Intel Developer Cloud Platform
- Python
- Huggingface

❑ *Team Contribution*

- **Vaishnavi Ankatwar -** Undertook the comprehensive research and meticulously prepared the project report.
- **Tanishq Palandurkar -** Provided an insightful presentation on LoRA and its advantages, drawing from his observations during the fine-tuning and execution processes on IDC.
- **Vedanti Dhage –** Committed to assisting Vaishnavi with the project reports and research efforts.
- **Sanjana Sanjai –** Created the IDC account.

❑ *Conclusion:*

Consequently, we gained valuable insights into the process and workflow of building a chatbot and fine-tuning it to address specific problems on Intel's cloud development platform.