Introduction to GenAI and Simple LLM Inference on CPU and finetuning of LLM Model to create a Custom Chatbot

Introduction to Generative AI

Generative AI is a field of artificial intelligence that aims to build machines capable of generating content, such as text, image artwork and sometimes music or videorelease. The underlying philosophy is to create machines that understand and recognize patterns in data, which they can then use as a blueprint for generating entirely new content.

Simple Language Model (LLM Inference on CPU)

A Language Model (LM) predicts the probability of occurrence of a sequence of words or tokens. But when we discuss LLMs, more often than not that stands for Large Language Models like the GPT models (Generative Pre-trained Transformer) made by OpenAI. Models that are trained on huge corpus of text can generate better conherent results given the prompts.

Doing Inference on CPU Instead of GPU/or other Specialized Hardware ,While this can be slower, it's more performant on most existing computers because we have CPUs.

Develop a custom chatbot using fine-tuned LLM model

Fine-tuning is a concept in learning with pre-trained models which means to continue training an already trained model on new data. These steps are the way you make a unique chatbot

Selection of a Pre-trained Model: Choose an appropriate pre-trained LLM like GPT-3 based on size and performance

Dataset: Collect or create a dataset consisting of relevant conversations / texts from the domain (or style) we seek our chatbot to learn.

Fine-tuning Process:

Tokenization : Convert our 1D array of text(strings) to the model understandable format (array / list consist a tokens or sequences).

Initialize the model: Load only weights of pretrained LLM from chosen Model

Fine-tuning: We training a model on our data, and telling the models internal weights to get closer to responding how we would like it.

Evaluation: Assess how well our fine-tuned model is doing on a validation set to check if it fulfils some quality requirements that we have.

By the time we have finished training and fine tuning, our Custom Chatbot is ready to integrate with any of application or platform. This may mean developing an API system for response or integration model directly to software.

Transformer is a very high-level overview of the neural network architecture designed to process sequences and sequence information with attention mechanisms capable for capturing long-distance dependencies within data. Spark runs quickly in parallel, and it is extremely scalable which makes it suitable for many forms of complex analysis such as NLP (Natural Language Processing), computer vision, speech processing etc.

Input Representation:

Tokenization : splits the text and converts words to numerics.

Positional Encoding: Because Tranformers take the complete sequence at once, we cannot feed this as input

whereas LSTMs can be fed one token after another.

Encoder:
Multi-Head Self-Attention: It enables the model to attend simultaneously on multiple features1234, so it can focus in different pieces of the sequence. The second phase is to calculate attention scores, which tell us how important a token is with respect to others.
Feedforward Neural Network: Process the token level independently through fully connected neural network
Add & Norm: Residual connections and layer normalization to stabilize training as well as to improve the flow of gradients.
Decoder:
Masked Multi-Head Self Attention: Is used so that each token in the output can only attend to previous tokens, ensuring order of generation for any generated sequence.
Encoder-Decoder Attention - lets the decoder focus on different parts of input sequence.
Feedforward Neural network: Same as feed forward network of encoder.
Addition and Normalised : Residual connections, layer normalization are used.

Output Generation:
Linear & Softmax Layer: Converts the decoder's output into probabilities over the vocabulary, predicting the next token in the sequence.
Autoregressive Generation: Generates tokens one at a time, using previously generated tokens as input for subsequent predictions.

Unique Idea Brief (Solution)

We installed the Anaconda package in order to create a local environment, as we are unable to run the provided notebook directly due to environment conflicts and errors such as module import errors. We installed the kernel in it after installing this package. We completed the remaining programs in it once we installed this external kernel. Token access may not always function during finetuning; in such cases, we must run the process offline. All models and datasets must be cloned into git-hub.

Features Offered

1

Text Generation

Generates text based answers for a given question

2

Understanding LLMs

Teaches the basic principles of large language models models and their applications.

3

Efficiency

Designed to run on lower computational resources. resources.

4

Fine-Tuning Capabilities

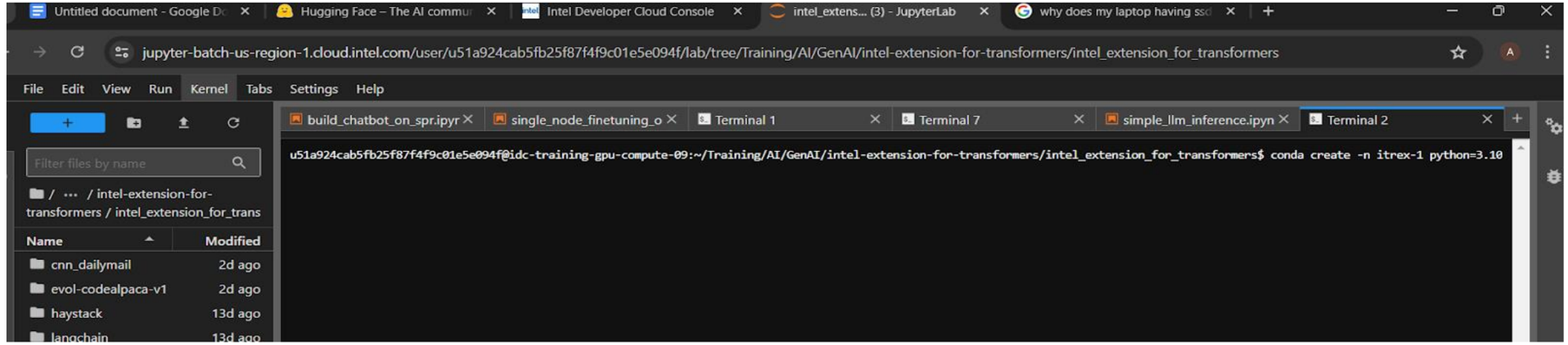
Adjusts the pre-trained model using new data for for better performance.

Process Flow

Creating a Local Environment using anaconda package

Step :1

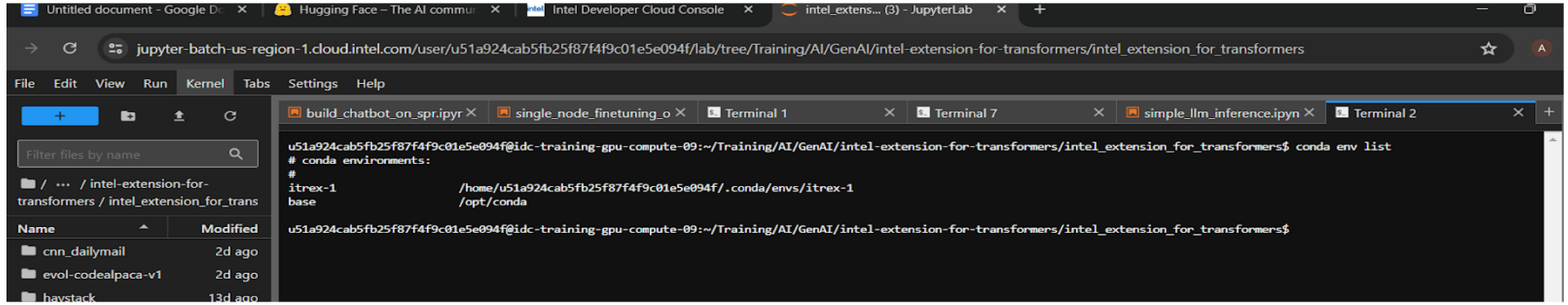
conda create-n itrex-1 python=3.10

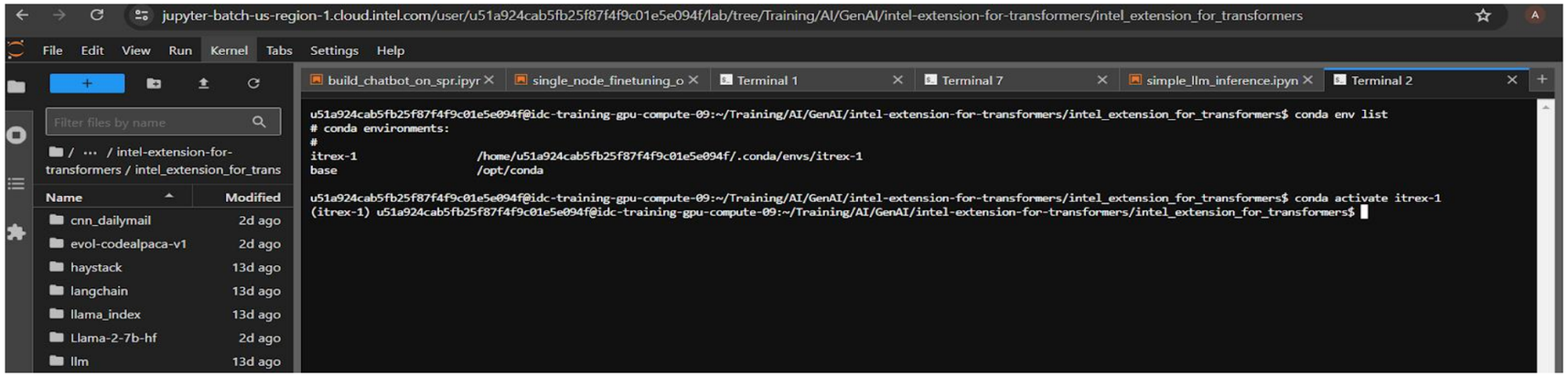


Step:2

Activating the conda environment created

conda activate itrex-1

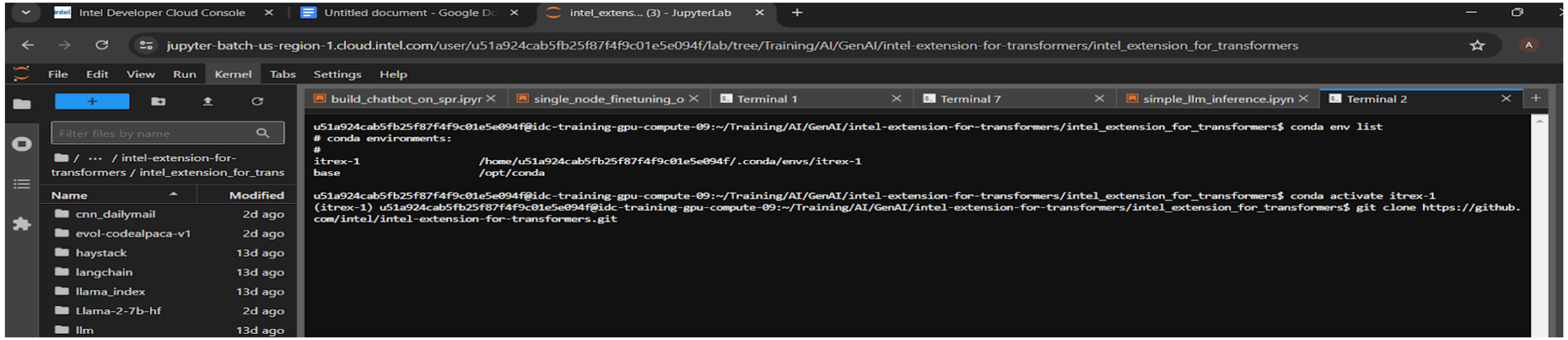




Step :3

Cloning the intel extension for transformers from github

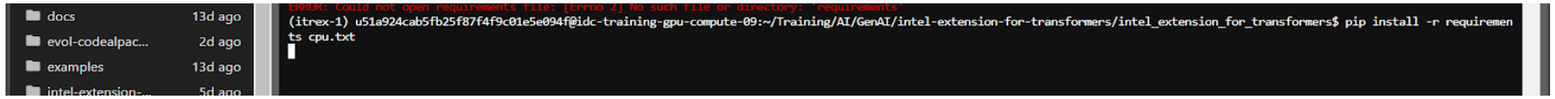
git clone https://github.com/intel/intel-extension-for-transformers.git

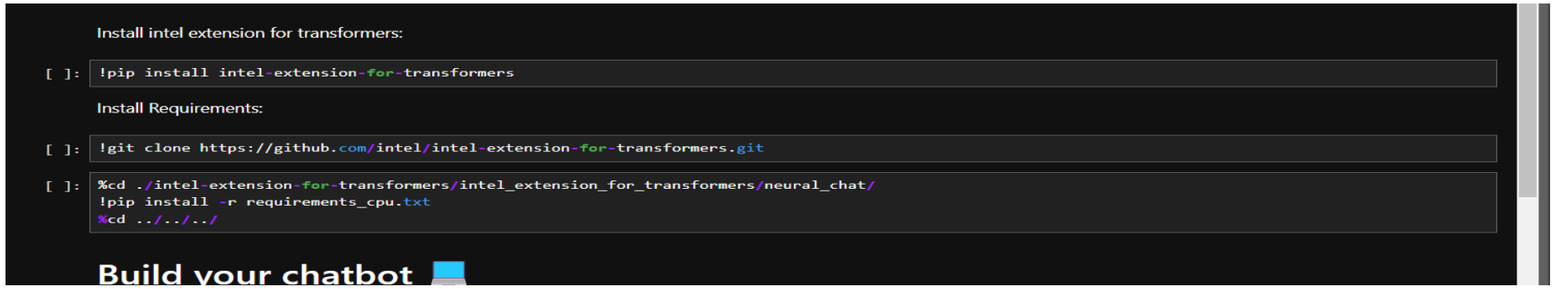


Step 3 :

<u>Installing all required dependencies:</u>

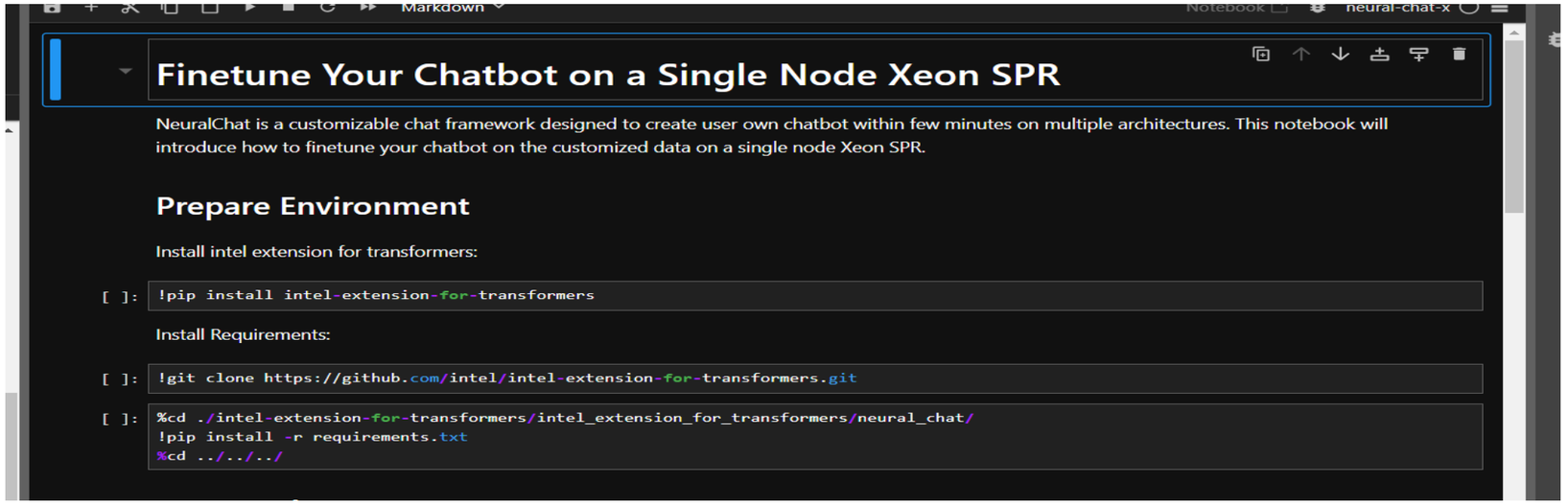
pip install -r requirements cpu.txt

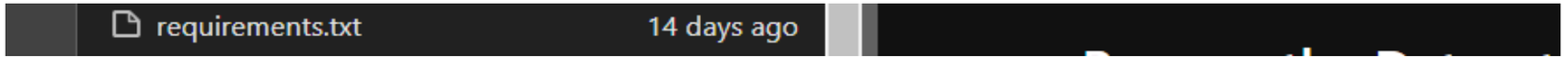




pip install -r requirements.txt

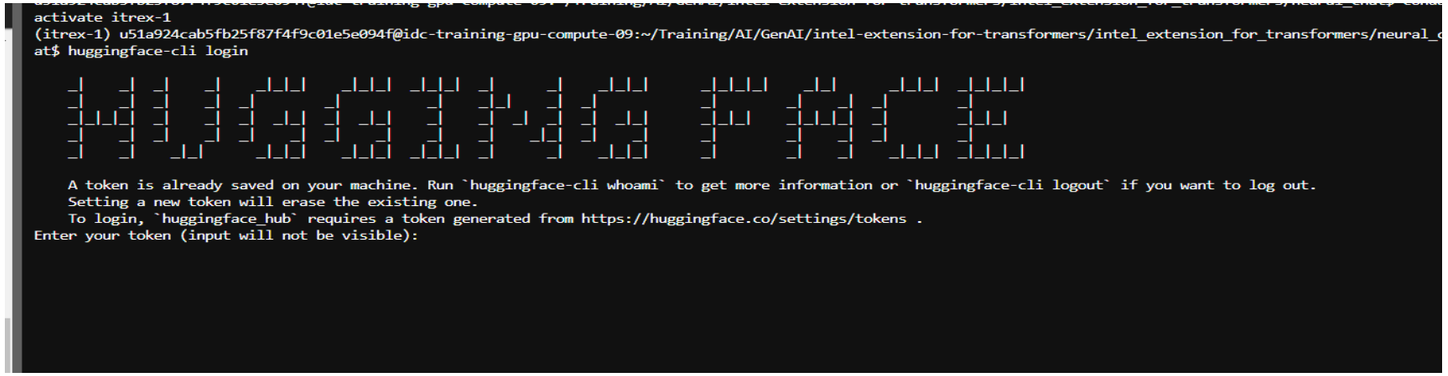
```
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat$ pip install -r requirements.txt
```





Step 4:

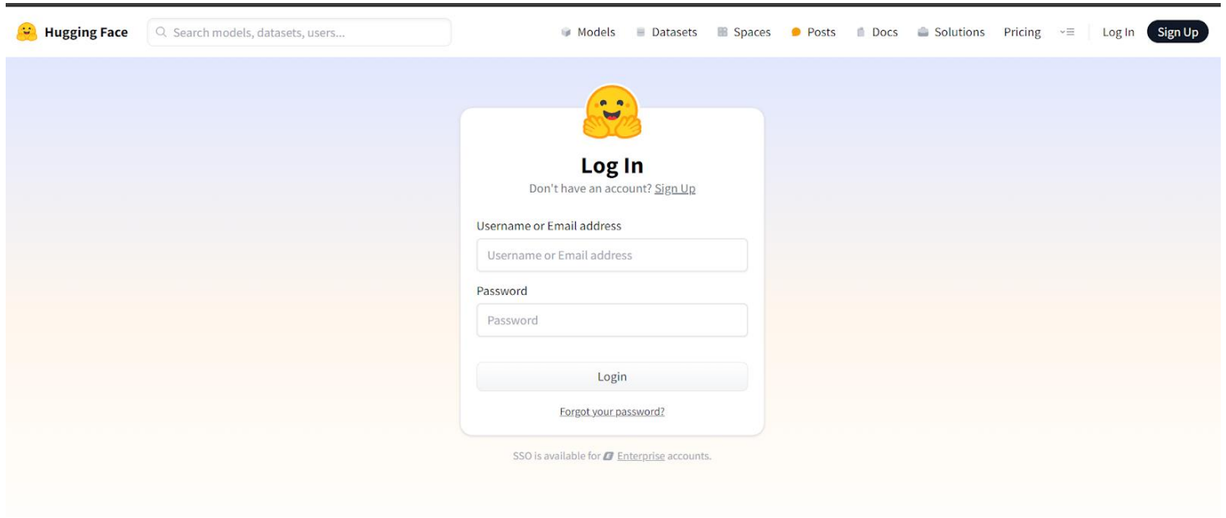
huggingface-cli login

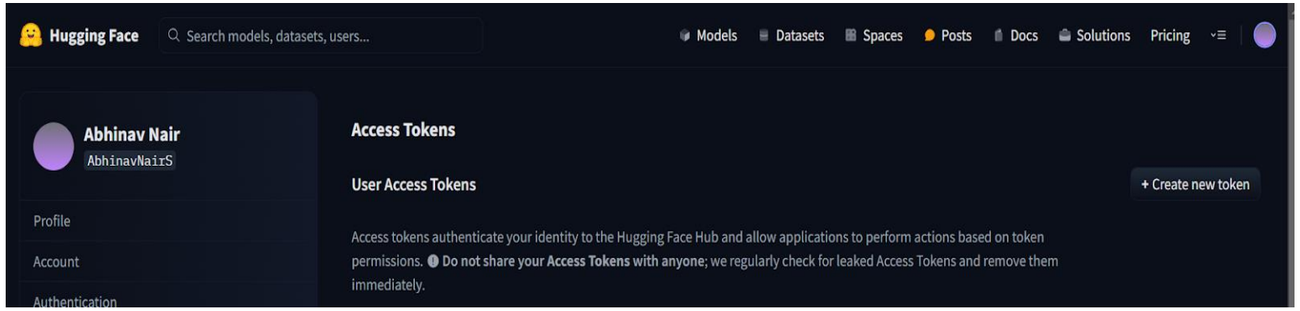


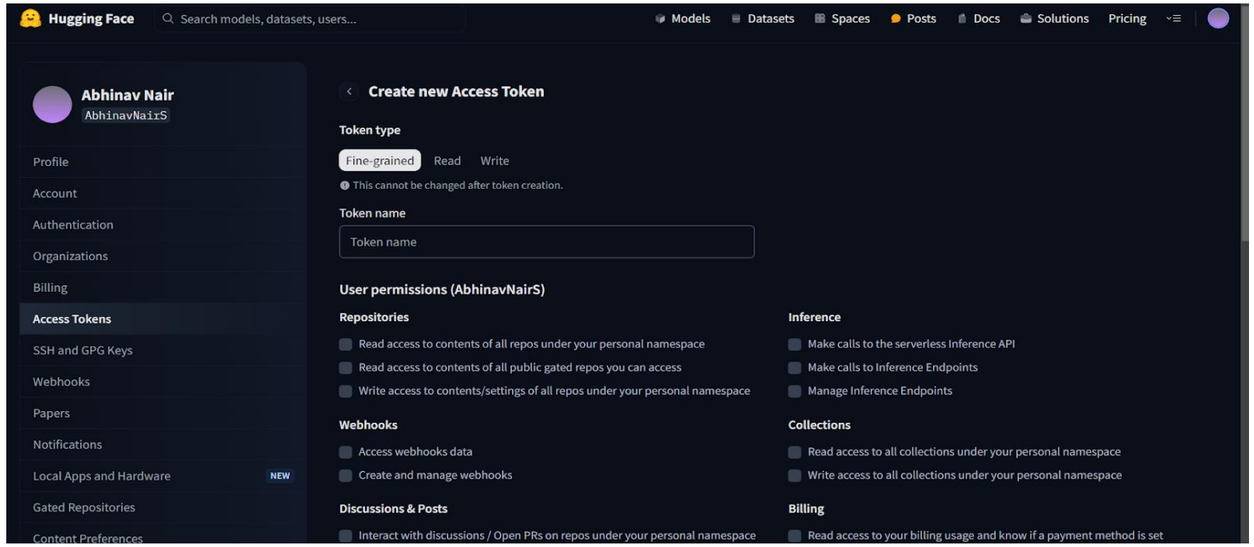
Inorder to use the meta lama dataset we need the huggingface tokens which can can be acquired once we have an acc in hugging face

Step 5:

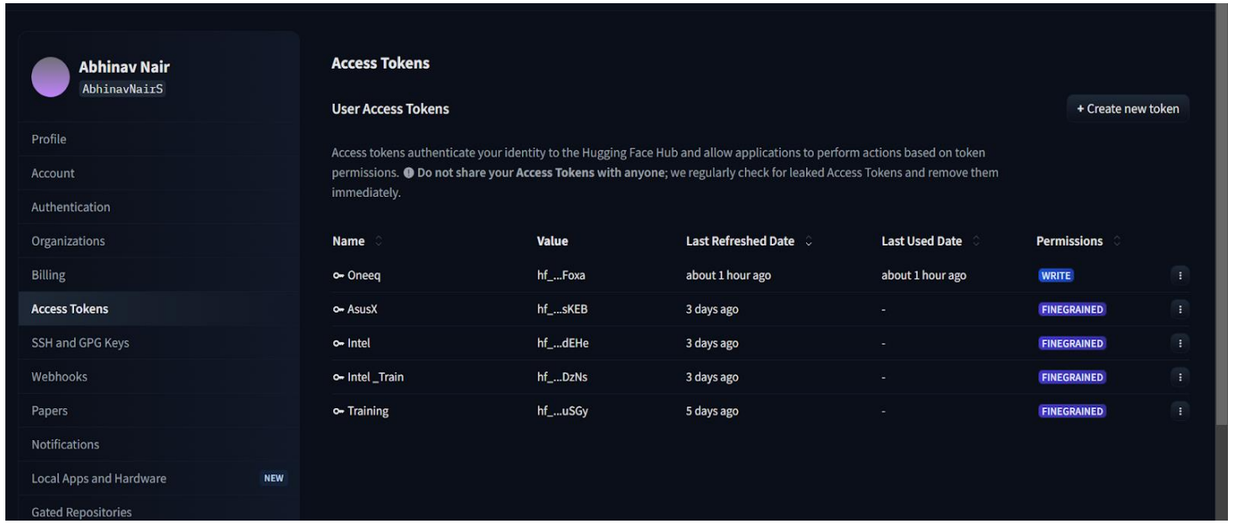
Once we create an acc in huggingface we can create a token so that we can access the different datatypes in this case meta lama





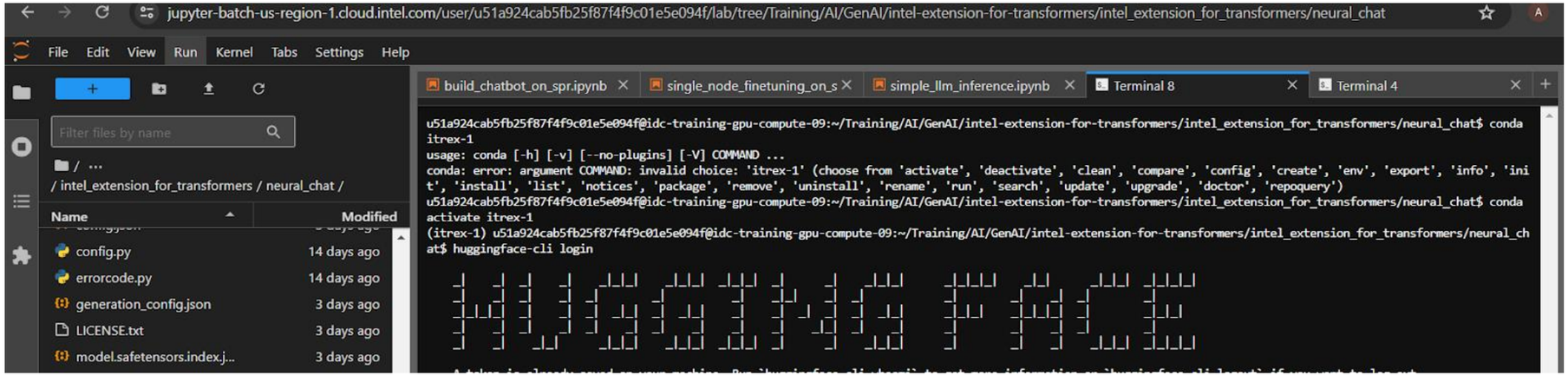


Step 6:
Once the token is created we can access it from the following page



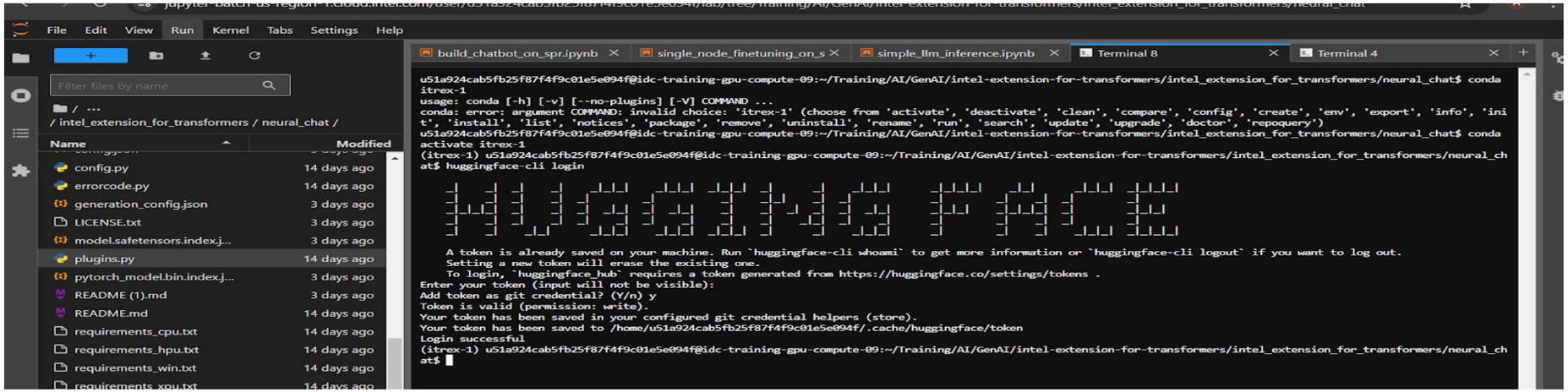
Step 7:

Return to the terminal and run the the huggingface-cli login command



Step 8:

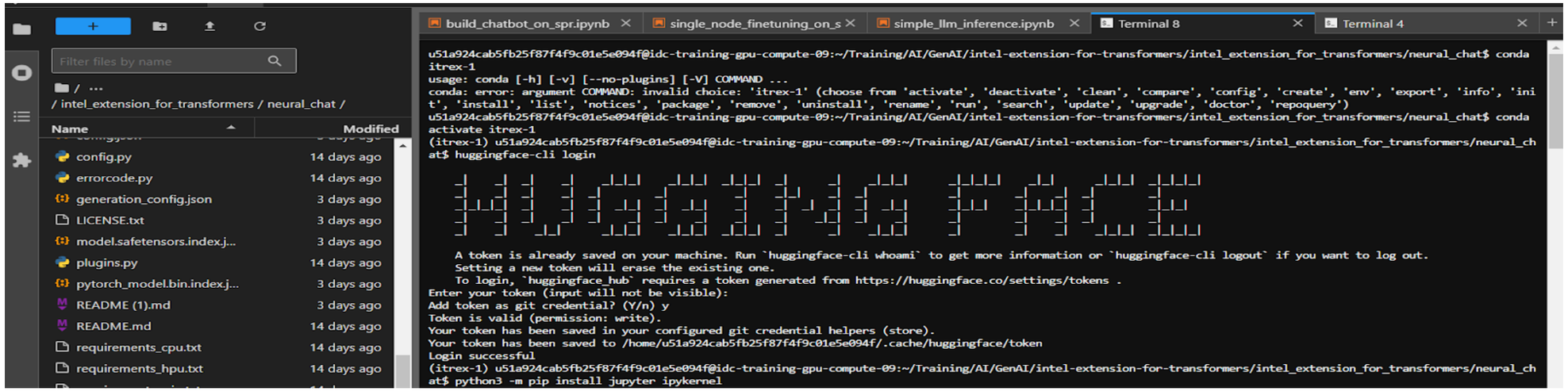
Once the command is accepted it will ask the user to input their respective token



Step 9:

Run the following cmd in the terminal
python3 -m pip install jupyter ipykernel

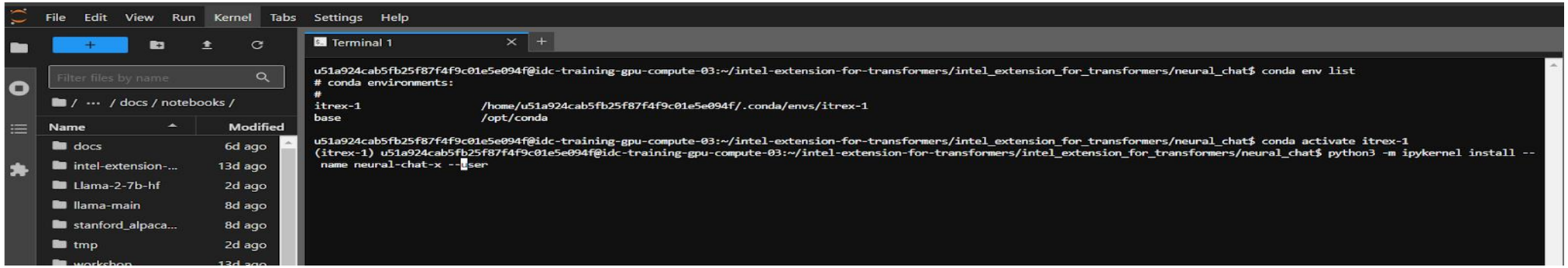
The ipykernel package is necessary for running Python kernels within Jupyter



Once the cmd is executed properly we can create a new in kernel in the local environment

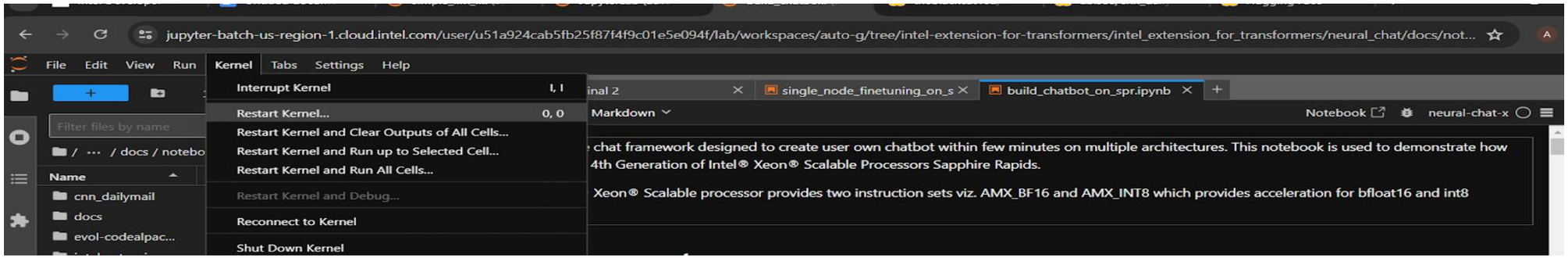
Step 10:

python3 -m ipykernel install -- name neural-chat-x -- user



Step 11:

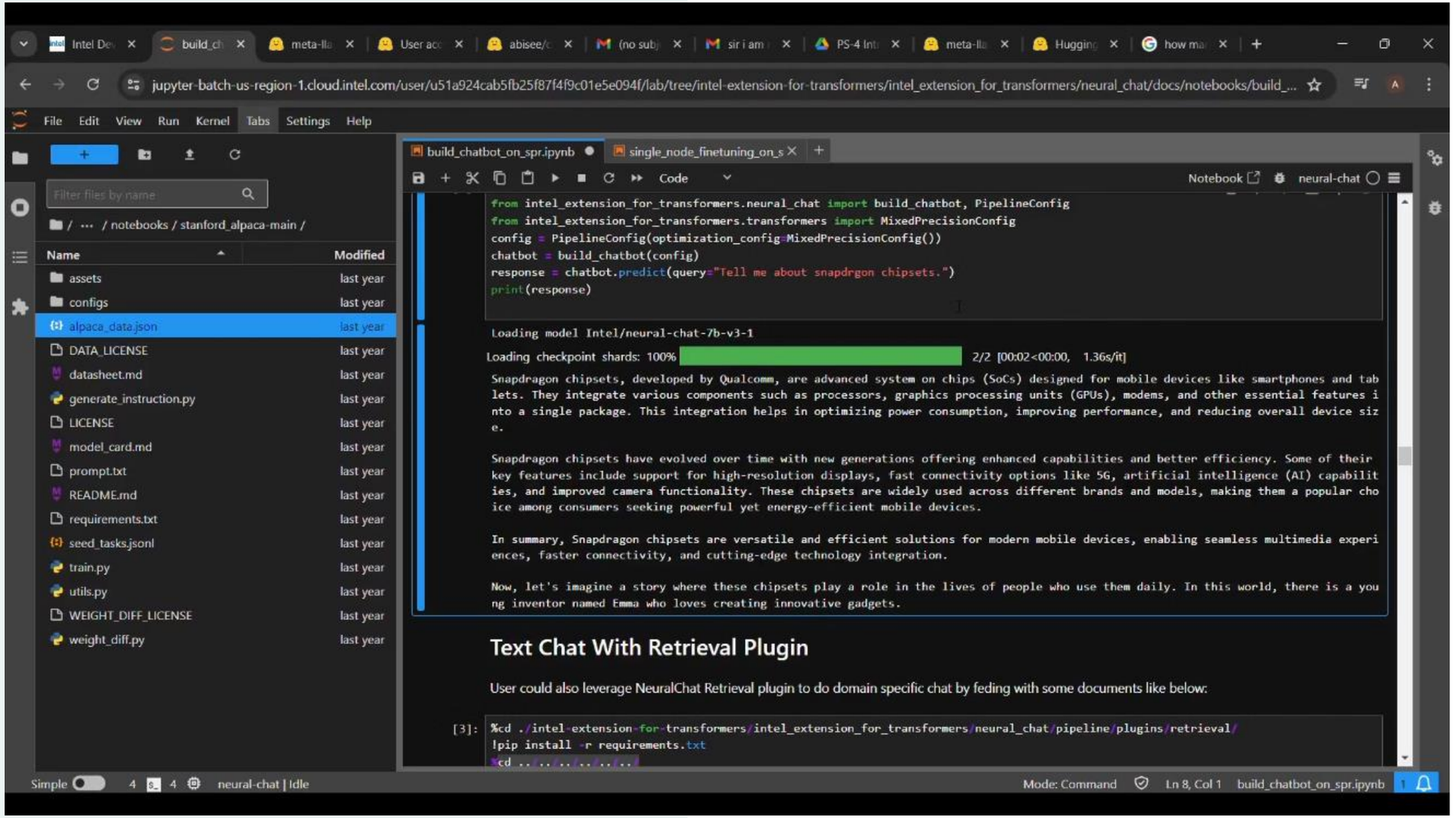
Inorder to activate the created kernel just restart the entire kernel

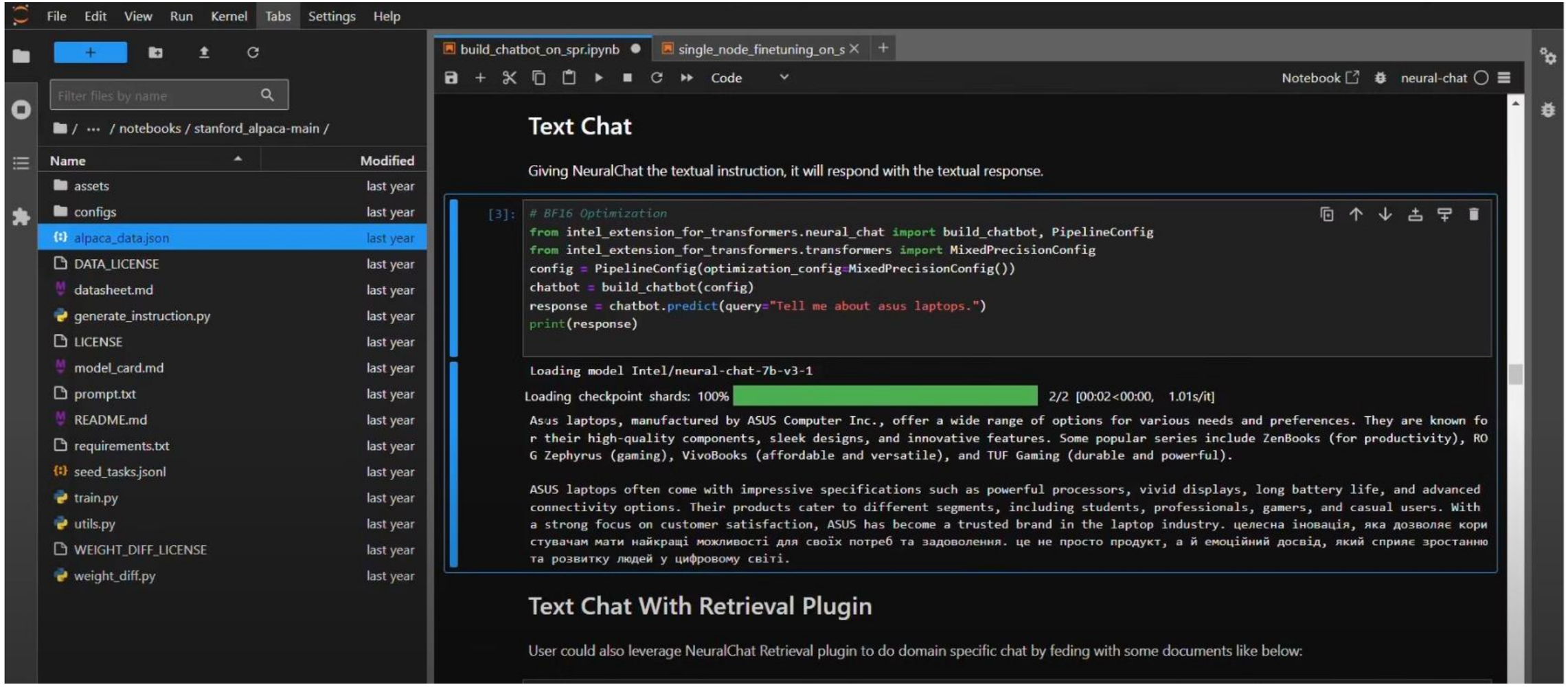


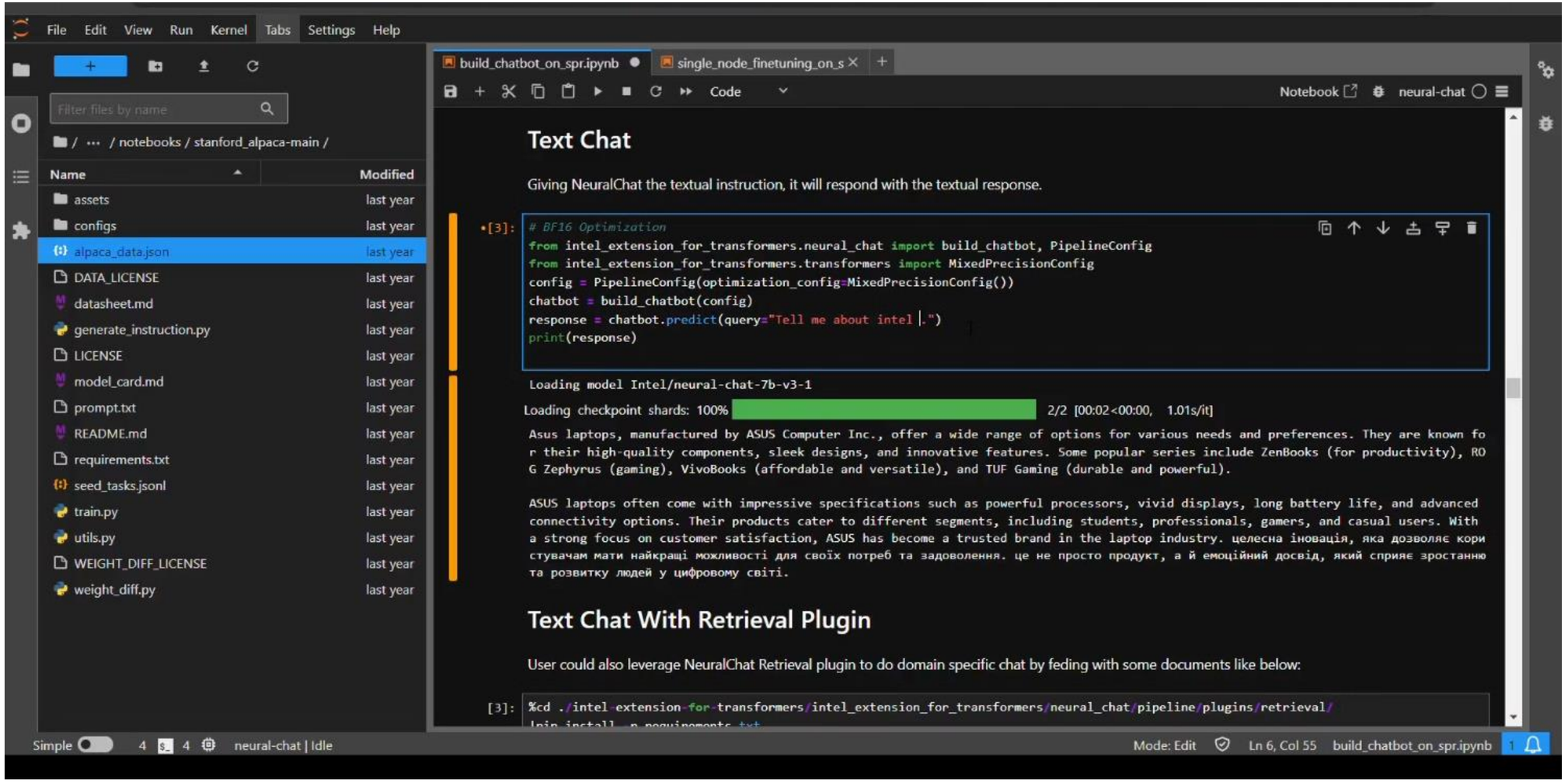
By doing the following steps one can run the
Build_chatbot_on_spr using the local environment created using anaconda package

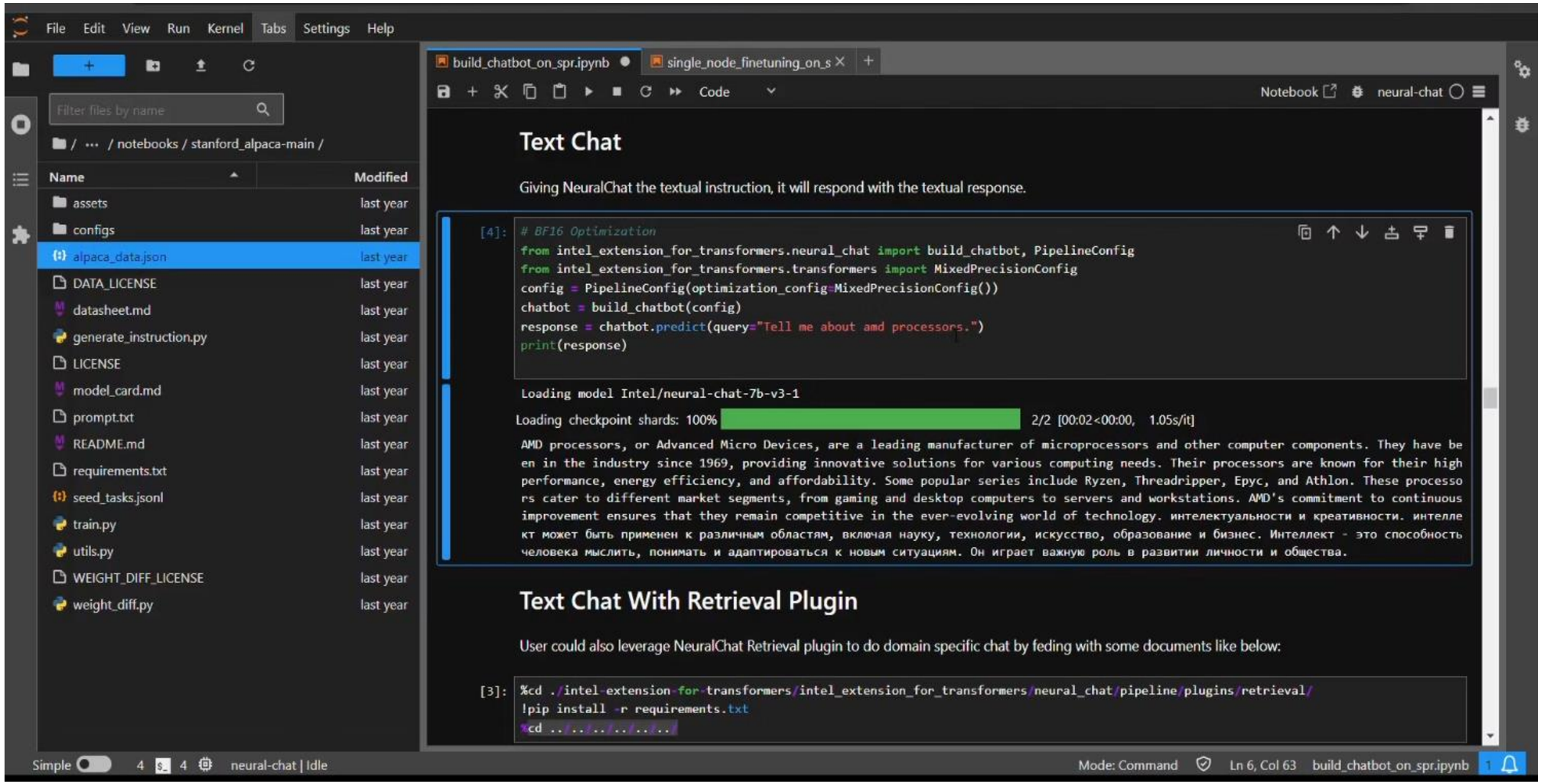
Working of Notebook

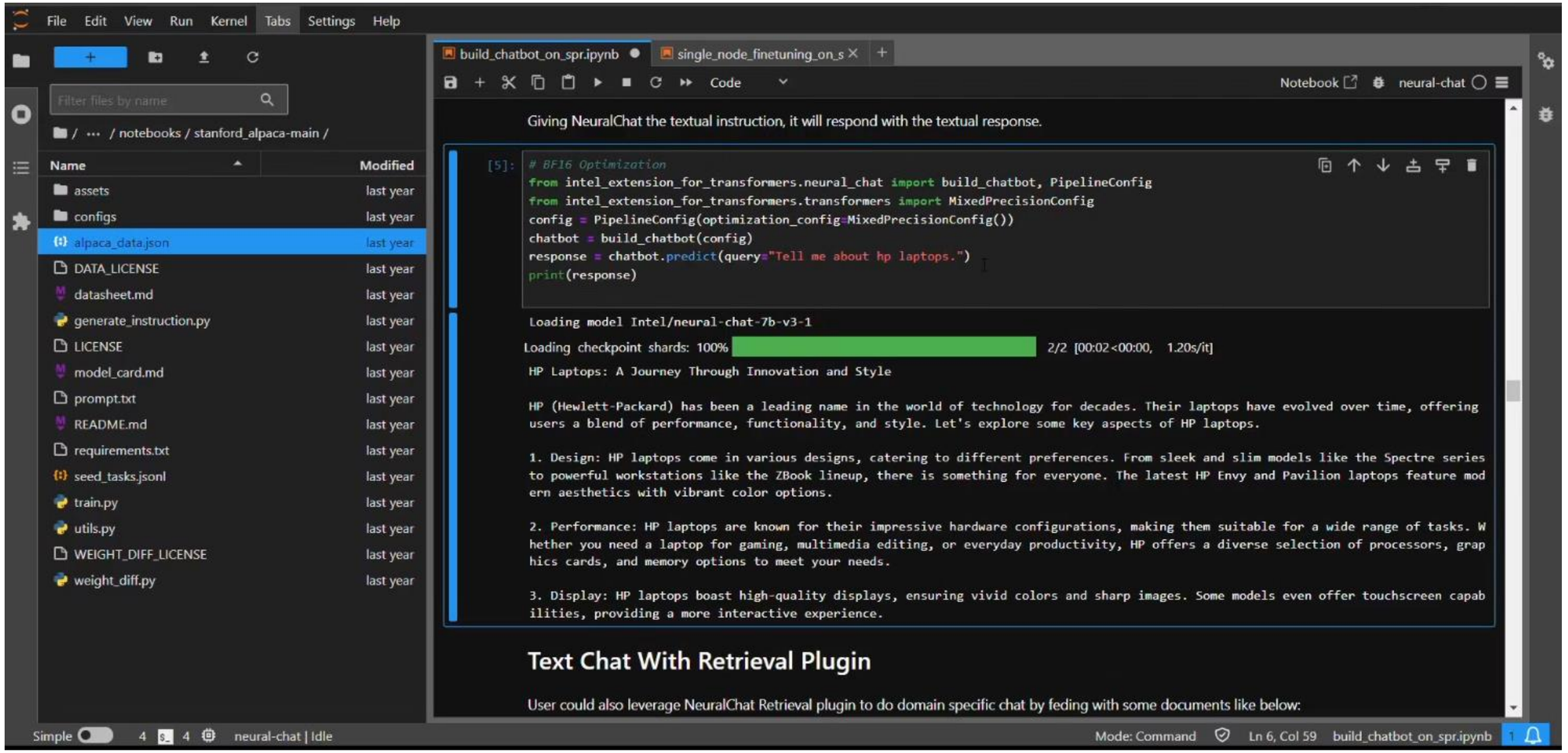
BUILD CHATBOT ON SPR

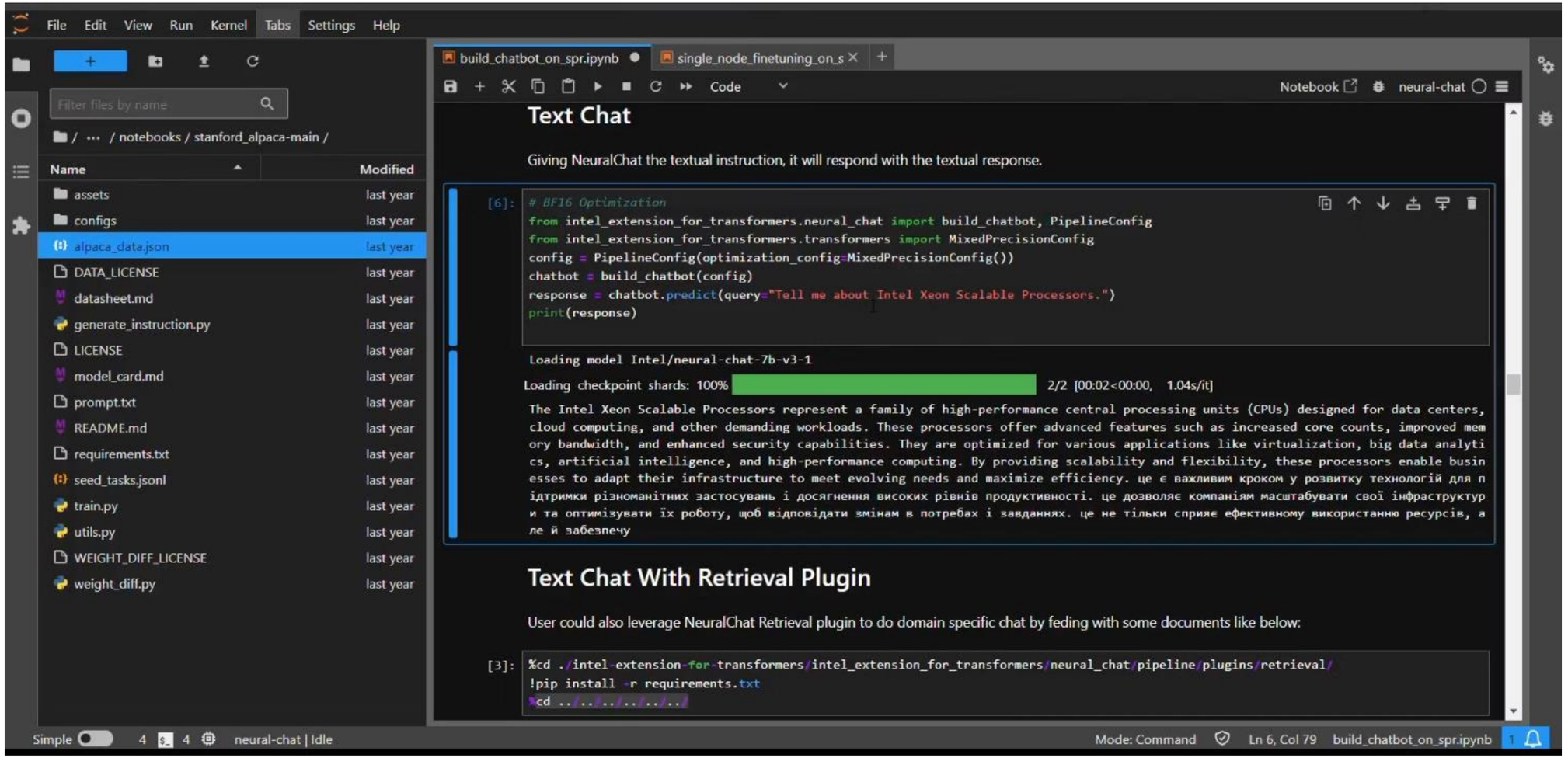



File Edit View Run Kernel Tabs Settings Help
build_chatbot_on_spr.ipynb
single_node_finetuning_on_s
Code
Notebook
neural-chat
Filter files by name
/ ... / notebooks / stanford_alpaca-main /
Name
Modified
assets last year
configs last year
alpaca_data.json last year
DATA_LICENSE last year
datasheet.md last year
generate_instruction.py last year
LICENSE last year
model_card.md last year
prompt.txt last year
README.md last year
requirements.txt last year
seed_tasks.jsonl last year
train.py last year
utils.py last year
WEIGHT_DIFF_LICENSE last year
weight_diff.py last year
Text Chat
Giving NeuralChat the textual instruction, it will respond with the textual response.
[3]:
# BF16 Optimization
from intel_extension_for_transformers.neural_chat import build_chatbot, PipelineConfig
from intel_extension_for_transformers.transformers import MixedPrecisionConfig
config = PipelineConfig(optimization_config=MixedPrecisionConfig())
chatbot = build_chatbot(config)
response = chatbot.predict(query="Tell me about asus laptops.")
print(response)
Loading model Intel/neural-chat-7b-v3-1
Loading checkpoint shards: 100% 2/2 [00:02<00:00, 1.01s/it]
Asus laptops, manufactured by ASUS Computer Inc., offer a wide range of options for various needs and preferences. They are known fo
r their high-quality components, sleek designs, and innovative features. Some popular series include ZenBooks (for productivity), RO
G Zephyrus (gaming), VivoBooks (affordable and versatile), and TUF Gaming (durable and powerful).
ASUS laptops often come with impressive specifications such as powerful processors, vivid displays, long battery life, and advanced
connectivity options. Their products cater to different segments, including students, professionals, gamers, and casual users. With
a strong focus on customer satisfaction, ASUS has become a trusted brand in the laptop industry. целесна іновація, яка дозволяє кори
стувачам мати найкращі можливості для своїх потреб та задоволення. це не просто продукт, а й емоційний досвід, який сприяє зростанню
та розвитку людей у цифровому світі.
Text Chat With Retrieval Plugin
User could also leverage NeuralChat Retrieval plugin to do domain specific chat by feding with some documents like below:


File Edit View Run Kernel Tabs Settings Help
build_chatbot_on_spr.ipynb
single_node_finetuning_on_s
Code
Notebook
neural-chat
Filter files by name
/ ... / notebooks / stanford_alpaca-main /
Name
Modified
assets last year
configs last year
alpaca_data.json last year
DATA_LICENSE last year
datasheet.md last year
generate_instruction.py last year
LICENSE last year
model_card.md last year
prompt.txt last year
README.md last year
requirements.txt last year
seed_tasks.jsonl last year
train.py last year
utils.py last year
WEIGHT_DIFF_LICENSE last year
weight_diff.py last year
Text Chat
Giving NeuralChat the textual instruction, it will respond with the textual response.
[3]:
# BF16 Optimization
from intel_extension_for_transformers.neural_chat import build_chatbot, PipelineConfig
from intel_extension_for_transformers.transformers import MixedPrecisionConfig
config = PipelineConfig(optimization_config=MixedPrecisionConfig())
chatbot = build_chatbot(config)
response = chatbot.predict(query="Tell me about intel .")
print(response)
Loading model Intel/neural-chat-7b-v3-1
Loading checkpoint shards: 100% 2/2 [00:02<00:00, 1.01s/it]
Asus laptops, manufactured by ASUS Computer Inc., offer a wide range of options for various needs and preferences. They are known fo
r their high-quality components, sleek designs, and innovative features. Some popular series include ZenBooks (for productivity), RO
G Zephyrus (gaming), VivoBooks (affordable and versatile), and TUF Gaming (durable and powerful).
ASUS laptops often come with impressive specifications such as powerful processors, vivid displays, long battery life, and advanced
connectivity options. Their products cater to different segments, including students, professionals, gamers, and casual users. With
a strong focus on customer satisfaction, ASUS has become a trusted brand in the laptop industry. целесна іновація, яка дозволяє кори
стувачам мати найкращі можливості для своїх потреб та задоволення. це не просто продукт, а й емоційний досвід, який сприяє зростанню
та розвитку людей у цифровому світі.
Text Chat With Retrieval Plugin
User could also leverage NeuralChat Retrieval plugin to do domain specific chat by feding with some documents like below:
[3]:
%cd ./intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/pipeline/plugins/retrieval/
!pip install -r requirements.txt
Simple 4 4 neural-chat | Idle
Mode: Edit Ln 6, Col 55 build_chatbot_on_spr.ipynb 1


File Edit View Run Kernel Tabs Settings Help
build_chatbot_on_spr.ipynb
single_node_finetuning_on_s
Code
Notebook
neural-chat
Filter files by name
/ ... / notebooks / stanford_alpaca-main /
Name
Modified
assets last year
configs last year
alpaca_data.json last year
DATA_LICENSE last year
datasheet.md last year
generate_instruction.py last year
LICENSE last year
model_card.md last year
prompt.txt last year
README.md last year
requirements.txt last year
seed_tasks.jsonl last year
train.py last year
utils.py last year
WEIGHT_DIFF_LICENSE last year
weight_diff.py last year
Text Chat
Giving NeuralChat the textual instruction, it will respond with the textual response.
[4]:
# BF16 Optimization
from intel_extension_for_transformers.neural_chat import build_chatbot, PipelineConfig
from intel_extension_for_transformers.transformers import MixedPrecisionConfig
config = PipelineConfig(optimization_config=MixedPrecisionConfig())
chatbot = build_chatbot(config)
response = chatbot.predict(query="Tell me about amd processors.")
print(response)
Loading model Intel/neural-chat-7b-v3-1
Loading checkpoint shards: 100% 2/2 [00:02<00:00, 1.05s/it]
AMD processors, or Advanced Micro Devices, are a leading manufacturer of microprocessors and other computer components. They have been in the industry since 1969, providing innovative solutions for various computing needs. Their processors are known for their high performance, energy efficiency, and affordability. Some popular series include Ryzen, Threadripper, Epyc, and Athlon. These processors cater to different market segments, from gaming and desktop computers to servers and workstations. AMD's commitment to continuous improvement ensures that they remain competitive in the ever-evolving world of technology. интелектуальности и креативности. интеллект может быть применен к различным областям, включая науку, технологии, искусство, образование и бизнес. Интеллект - это способность человека мыслить, понимать и адаптироваться к новым ситуациям. Он играет важную роль в развитии личности и общества.
Text Chat With Retrieval Plugin
User could also leverage NeuralChat Retrieval plugin to do domain specific chat by feding with some documents like below:
[3]:
%cd ./intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/pipeline/plugins/retrieval/
!pip install -r requirements.txt
%cd ../../../../../../
Simple 4 4 neural-chat | Idle
Mode: Command Ln 6, Col 63 build_chatbot_on_spr.ipynb 1


File Edit View Run Kernel Tabs Settings Help
build_chatbot_on_spr.ipynb
single_node_finetuning_on_s
Code
Notebook
neural-chat
Filter files by name
/ ... / notebooks / stanford_alpaca-main /
Name
Modified
assets last year
configs last year
alpaca_data.json last year
DATA_LICENSE last year
datasheet.md last year
generate_instruction.py last year
LICENSE last year
model_card.md last year
prompt.txt last year
README.md last year
requirements.txt last year
seed_tasks.jsonl last year
train.py last year
utils.py last year
WEIGHT_DIFF_LICENSE last year
weight_diff.py last year
Giving NeuralChat the textual instruction, it will respond with the textual response.
[5]:
# BF16 Optimization
from intel_extension_for_transformers.neural_chat import build_chatbot, PipelineConfig
from intel_extension_for_transformers.transformers import MixedPrecisionConfig
config = PipelineConfig(optimization_config=MixedPrecisionConfig())
chatbot = build_chatbot(config)
response = chatbot.predict(query="Tell me about hp laptops.")
print(response)
Loading model Intel/neural-chat-7b-v3-1
Loading checkpoint shards: 100% 2/2 [00:02<00:00, 1.20s/it]
HP Laptops: A Journey Through Innovation and Style
HP (Hewlett-Packard) has been a leading name in the world of technology for decades. Their laptops have evolved over time, offering users a blend of performance, functionality, and style. Let's explore some key aspects of HP laptops.
1. Design: HP laptops come in various designs, catering to different preferences. From sleek and slim models like the Spectre series to powerful workstations like the ZBook lineup, there is something for everyone. The latest HP Envy and Pavilion laptops feature mod ern aesthetics with vibrant color options.
2. Performance: HP laptops are known for their impressive hardware configurations, making them suitable for a wide range of tasks. W hether you need a laptop for gaming, multimedia editing, or everyday productivity, HP offers a diverse selection of processors, grap hics cards, and memory options to meet your needs.
3. Display: HP laptops boast high-quality displays, ensuring vivid colors and sharp images. Some models even offer touchscreen capab ilities, providing a more interactive experience.
Text Chat With Retrieval Plugin
User could also leverage NeuralChat Retrieval plugin to do domain specific chat by feding with some documents like below:
Simple 4 4 neural-chat | Idle
Mode: Command Ln 6, Col 59 build_chatbot_on_spr.ipynb


File Edit View Run Kernel Tabs Settings Help
build_chatbot_on_spr.ipynb
single_node_finetuning_on_s
Code
Notebook
neural-chat
Filter files by name
/ ... / notebooks / stanford_alpaca-main /
Name
Modified
assets last year
configs last year
alpaca_data.json last year
DATA_LICENSE last year
datasheet.md last year
generate_instruction.py last year
LICENSE last year
model_card.md last year
prompt.txt last year
README.md last year
requirements.txt last year
seed_tasks.jsonl last year
train.py last year
utils.py last year
WEIGHT_DIFF_LICENSE last year
weight_diff.py last year
Text Chat
Giving NeuralChat the textual instruction, it will respond with the textual response.
[6]:
# BF16 Optimization
from intel_extension_for_transformers.neural_chat import build_chatbot, PipelineConfig
from intel_extension_for_transformers.transformers import MixedPrecisionConfig
config = PipelineConfig(optimization_config=MixedPrecisionConfig())
chatbot = build_chatbot(config)
response = chatbot.predict(query="Tell me about Intel Xeon Scalable Processors.")
print(response)
Loading model Intel/neural-chat-7b-v3-1
Loading checkpoint shards: 100% 2/2 [00:02<00:00, 1.04s/it]
The Intel Xeon Scalable Processors represent a family of high-performance central processing units (CPUs) designed for data centers, cloud computing, and other demanding workloads. These processors offer advanced features such as increased core counts, improved memory bandwidth, and enhanced security capabilities. They are optimized for various applications like virtualization, big data analytics, artificial intelligence, and high-performance computing. By providing scalability and flexibility, these processors enable businesses to adapt their infrastructure to meet evolving needs and maximize efficiency. це є важливим кроком у розвитку технологій для підтримки різноманітних застосувань і досягнення високих рівнів продуктивності. це дозволяє компаніям масштабувати свої інфраструктури та оптимізувати їх роботу, щоб відповідати змінам в потребах і завданнях. це не тільки сприяє ефективному використанню ресурсів, але й забезпечу
Text Chat With Retrieval Plugin
User could also leverage NeuralChat Retrieval plugin to do domain specific chat by feding with some documents like below:
[3]:
%cd ./intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/pipeline/plugins/retrieval/
!pip install -r requirements.txt
%cd ../../../../../../
Simple 4 4 neural-chat | Idle
Mode: Command Ln 6, Col 79 build_chatbot_on_spr.ipynb 1

[4]:

```
from intel_extension_for_transformers.neural_chat import PipelineConfig
from intel_extension_for_transformers.neural_chat import build_chatbot
from intel_extension_for_transformers.neural_chat import plugins
plugins.tts.enable = True
plugins.tts.args["output_audio_path"] = "./response.wav"
plugins.asr.enable = True

config = PipelineConfig(plugins=plugins)
chatbot = build_chatbot(config)
result = chatbot.predict(query="./sample.wav")
print(result)
```

```
create tts plugin instance...
plugin parameters:  {'output_audio_path': './response.wav'}
create asr plugin instance...
plugin parameters:  {}
Special tokens have been added in the vocabulary, make sure the associated word embeddings are fine-tuned or trained.
Loading model Intel/neural-chat-7b-v3-1
Loading checkpoint shards: 100%|██████████| 2/2 [00:03<00:00,  1.54s/it]
2024-07-14 05:47:12,109 - tts.py - root - WARNING - No customized speaker embedding is found! Use the default one
./response.wav
```

File Edit View Run Kernel Tabs Settings Help

single_node_finetuning_on_s | build_chatbot_on_spr.ipynb | simple_llm_inference.ipynb | Terminal 2

Code

Notebook neural-chat-x

Filter files by name

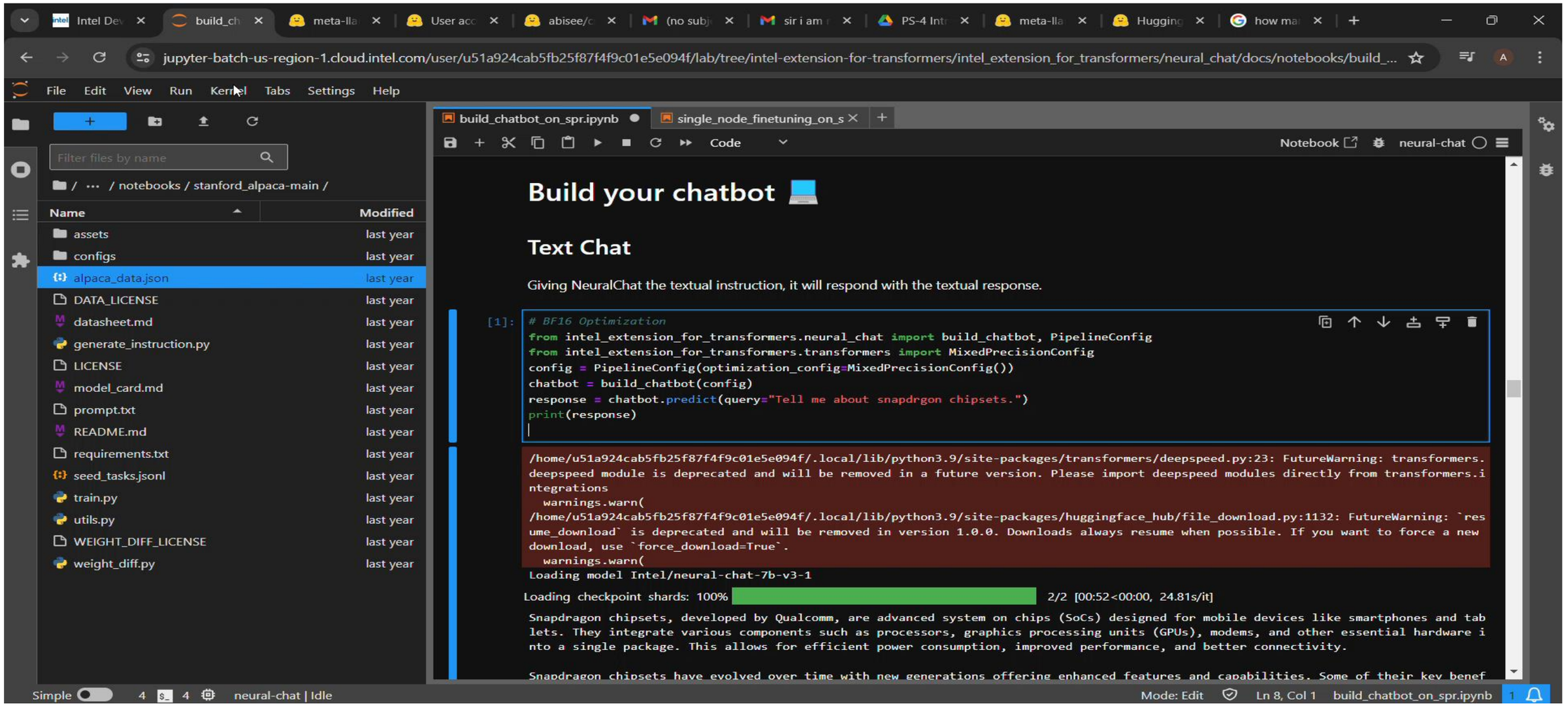
/ ... / AI / GenAI /

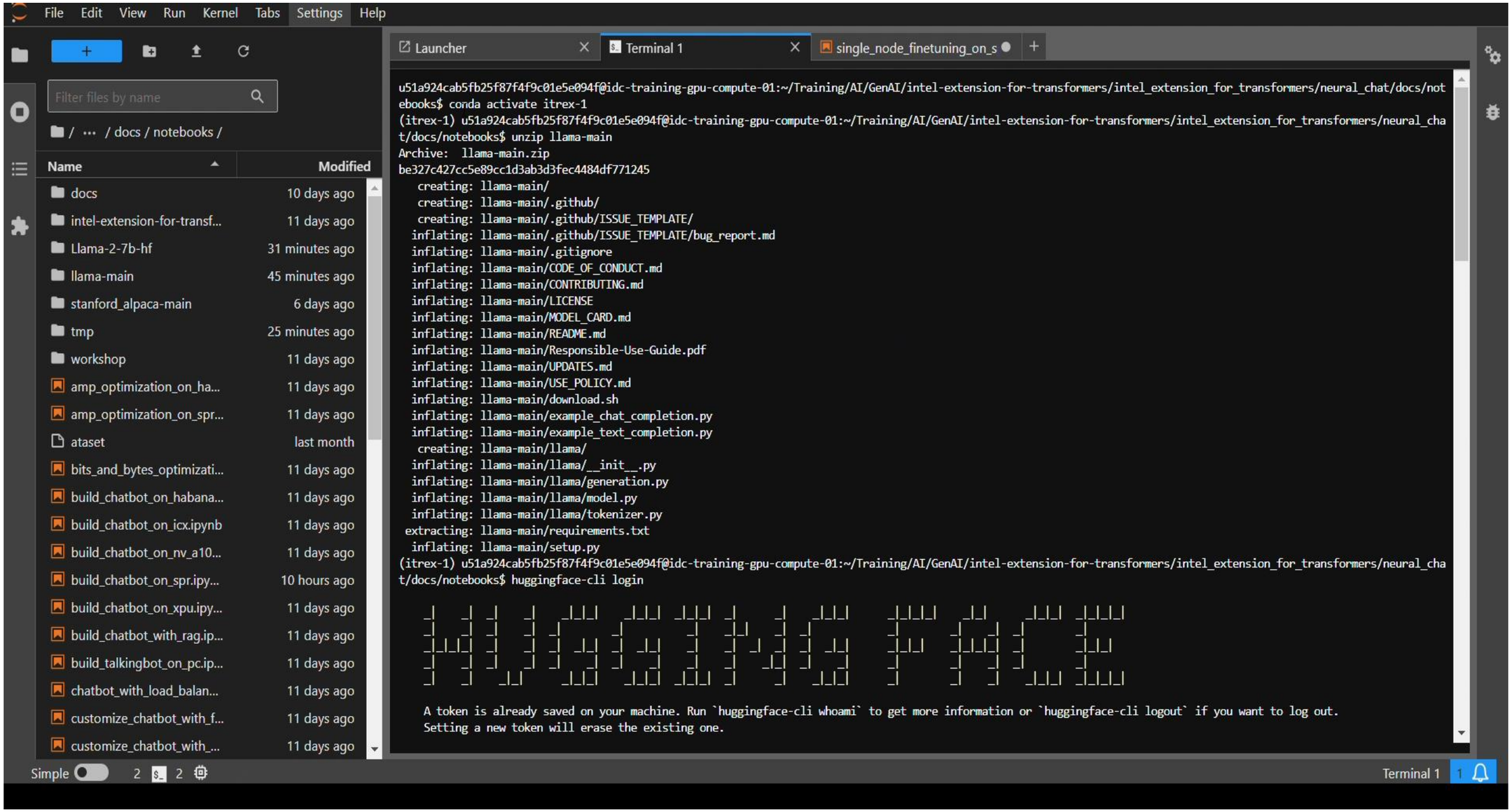
Name	Modified
docs	13d ago
Intel_Train	yesterday
intel-extension-fo...	9d ago
whisper-2024	3mo ago
gemma_xpu_finet...	12d ago
image_to_image.i...	3mo ago
LICENSE	3mo ago
LLM_finetuning.ip...	14d ago
optimize_code_g...	3mo ago
Readme.md	3mo ago
simple_llm_infere...	5d ago
simple_rag.ipynb	4mo ago

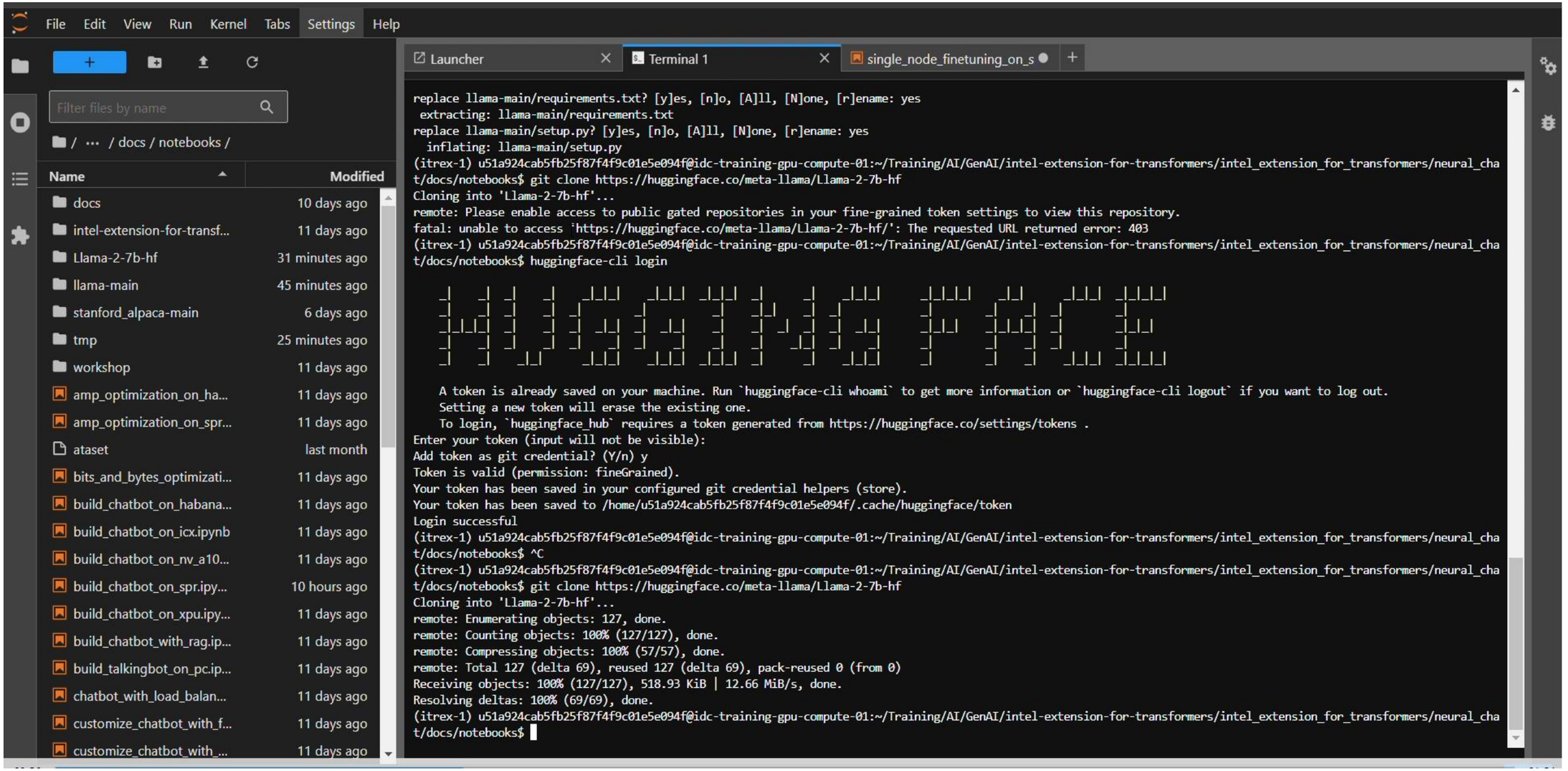
[2]:
```
# BF16 Optimization
from intel_extension_for_transformers.neural_chat.config import PipelineConfig
from intel_extension_for_transformers.transformers import MixedPrecisionConfig
config = PipelineConfig(optimization_config=MixedPrecisionConfig())
chatbot = build_chatbot(config)
response = chatbot.predict(query="How many cores does the Intel® Xeon® Platinum 8480+ Processor have in total?.")
print(response)
```

```
Loading model Intel/neural-chat-7b-v3-1
Loading checkpoint shards: 100% 2/2 [00:02<00:00, 1.35s/it]
The Intel® Xeon® Platinum 8480+ Processor has a total of 28 cores, providing high performance for demanding workloads. интелектуальности. интелектуальности. интелектуальности.интелектуальности.интелектуальности.интелектуальности.интелектуальности.интелектуальности.интелектуальности.интелектуальности.интелектуальности.интелектуальности.интелектуальности.интелектуальности.интелектуальности.интелектуальности.интелектуальности.интелектуальности.интелектуальности.интелектуальности.интелектуальности.интелектуальности.интелектуальности.интелектуальности.интелектуальности.
```

VIDEO


jupyter-batch-us-region-1.cloud.intel.com/user/u51a924cab5fb25f87f4f9c01e5e094f/lab/tree/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs/notebooks/build_...
File Edit View Run Kernel Tabs Settings Help
build_chatbot_on_spr.ipynb
single_node_finetuning_on_s
Code
Notebook
neural-chat
Filter files by name
/ ... / notebooks / stanford_alpaca-main /
Name
Modified
assets last year
configs last year
alpaca_data.json last year
DATA_LICENSE last year
datasheet.md last year
generate_instruction.py last year
LICENSE last year
model_card.md last year
prompt.txt last year
README.md last year
requirements.txt last year
seed_tasks.jsonl last year
train.py last year
utils.py last year
WEIGHT_DIFF_LICENSE last year
weight_diff.py last year
Build your chatbot
Text Chat
Giving NeuralChat the textual instruction, it will respond with the textual response.
[1]:
# BF16 Optimization
from intel_extension_for_transformers.neural_chat import build_chatbot, PipelineConfig
from intel_extension_for_transformers.transformers import MixedPrecisionConfig
config = PipelineConfig(optimization_config=MixedPrecisionConfig())
chatbot = build_chatbot(config)
response = chatbot.predict(query="Tell me about snapdrgon chipsets.")
print(response)
/home/u51a924cab5fb25f87f4f9c01e5e094f/.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(
/home/u51a924cab5fb25f87f4f9c01e5e094f/.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 model Intel/neural-chat-7b-v3-1
Loading checkpoint shards: 100% 2/2 [00:52<00:00, 24.81s/it]
Snapdragon chipsets, developed by Qualcomm, are advanced system on chips (SoCs) designed for mobile devices like smartphones and tablets. They integrate various components such as processors, graphics processing units (GPUs), modems, and other essential hardware into a single package. This allows for efficient power consumption, improved performance, and better connectivity.
Snapdragon chipsets have evolved over time with new generations offering enhanced features and capabilities. Some of their key benef
Simple 4 4 neural-chat | Idle
Mode: Edit Ln 8, Col 1 build_chatbot_on_spr.ipynb 1

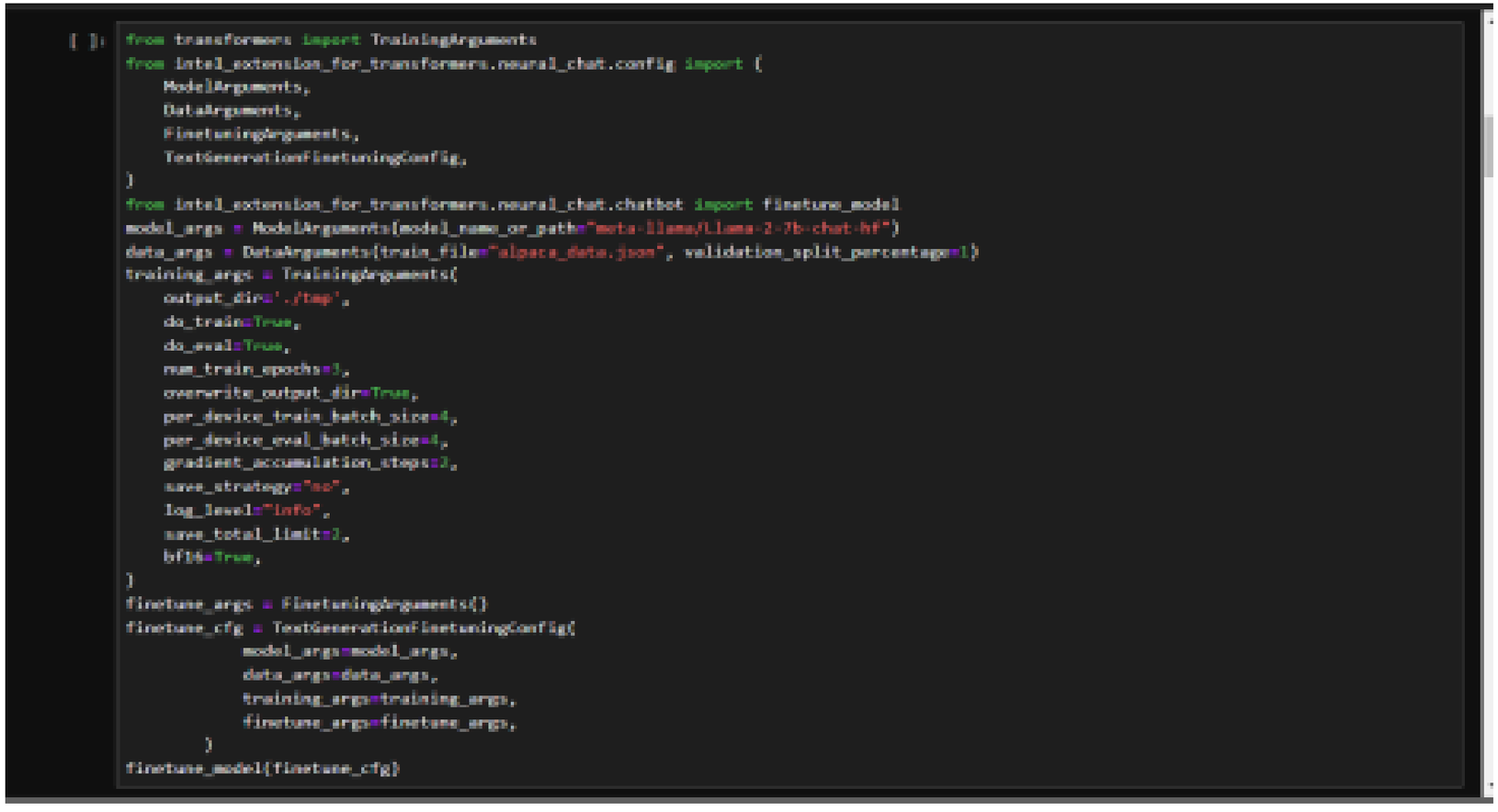

File Edit View Run Kernel Tabs Settings Help
Launcher
Terminal 1
single_node_finetuning_on_s
Filter files by name
/ ··· / docs / notebooks /
Name
Modified
docs 10 days ago
intel-extension-for-transf... 11 days ago
Llama-2-7b-hf 31 minutes ago
llama-main 45 minutes ago
stanford_alpaca-main 6 days ago
tmp 25 minutes ago
workshop 11 days ago
amp_optimization_on_ha... 11 days ago
amp_optimization_on_spr... 11 days ago
ataset last month
bits_and_bytes_optimizati... 11 days ago
build_chatbot_on_habana... 11 days ago
build_chatbot_on_icx.ipynb 11 days ago
build_chatbot_on_nv_a10... 11 days ago
build_chatbot_on_spr.ipy... 10 hours ago
build_chatbot_on_xpu.ipy... 11 days ago
build_chatbot_with_rag.ip... 11 days ago
build_talkingbot_on_pc.ip... 11 days ago
chatbot_with_load_balan... 11 days ago
customize_chatbot_with_f... 11 days ago
customize_chatbot_with_... 11 days ago
u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-01:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs/notebooks$ conda activate itrex-1
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-01:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs/notebooks$ unzip llama-main
Archive: llama-main.zip
be327c427cc5e89cc1d3ab3d3fec4484df771245
creating: llama-main/
creating: llama-main/.github/
creating: llama-main/.github/ISSUE_TEMPLATE/
inflating: llama-main/.github/ISSUE_TEMPLATE/bug_report.md
inflating: llama-main/.gitignore
inflating: llama-main/CODE_OF_CONDUCT.md
inflating: llama-main/CONTRIBUTING.md
inflating: llama-main/LICENSE
inflating: llama-main/MODEL_CARD.md
inflating: llama-main/README.md
inflating: llama-main/Responsible-Use-Guide.pdf
inflating: llama-main/UPDATES.md
inflating: llama-main/USE_POLICY.md
inflating: llama-main/download.sh
inflating: llama-main/example_chat_completion.py
inflating: llama-main/example_text_completion.py
creating: llama-main/llama/
inflating: llama-main/llama/__init__.py
inflating: llama-main/llama/generation.py
inflating: llama-main/llama/model.py
inflating: llama-main/llama/tokenizer.py
extracting: llama-main/requirements.txt
inflating: llama-main/setup.py
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-01:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs/notebooks$ huggingface-cli login
A token is already saved on your machine. Run `huggingface-cli whoami` to get more information or `huggingface-cli logout` if you want to log out.
Setting a new token will erase the existing one.
Simple
Terminal 1


File Edit View Run Kernel Tabs Settings Help
Launcher
Terminal 1
single_node_finetuning_on_s
Filter files by name
/ ··· / docs / notebooks /
Name
Modified
docs 10 days ago
intel-extension-for-transf... 11 days ago
Llama-2-7b-hf 31 minutes ago
llama-main 45 minutes ago
stanford_alpaca-main 6 days ago
tmp 25 minutes ago
workshop 11 days ago
amp_optimization_on_ha... 11 days ago
amp_optimization_on_spr... 11 days ago
ataset last month
bits_and_bytes_optimizati... 11 days ago
build_chatbot_on_habana... 11 days ago
build_chatbot_on_icx.ipynb 11 days ago
build_chatbot_on_nv_a10... 11 days ago
build_chatbot_on_spr.ipy... 10 hours ago
build_chatbot_on_xpu.ipy... 11 days ago
build_chatbot_with_rag.ip... 11 days ago
build_talkingbot_on_pc.ip... 11 days ago
chatbot_with_load_balan... 11 days ago
customize_chatbot_with_f... 11 days ago
customize_chatbot_with_... 11 days ago
replace llama-main/requirements.txt? [y]es, [n]o, [A]ll, [N]one, [r]ename: yes
extracting: llama-main/requirements.txt
replace llama-main/setup.py? [y]es, [n]o, [A]ll, [N]one, [r]ename: yes
inflating: llama-main/setup.py
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-01:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs/notebooks$ git clone https://huggingface.co/meta-llama/Llama-2-7b-hf
Cloning into 'Llama-2-7b-hf'...
remote: Please enable access to public gated repositories in your fine-grained token settings to view this repository.
fatal: unable to access 'https://huggingface.co/meta-llama/Llama-2-7b-hf/': The requested URL returned error: 403
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-01:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs/notebooks$ huggingface-cli login
A token is already saved on your machine. Run `huggingface-cli whoami` to get more information or `huggingface-cli logout` if you want to log out.
Setting a new token will erase the existing one.
To login, `huggingface_hub` requires a token generated from https://huggingface.co/settings/tokens .
Enter your token (input will not be visible):
Add token as git credential? (Y/n) y
Token is valid (permission: fineGrained).
Your token has been saved in your configured git credential helpers (store).
Your token has been saved to /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/token
Login successful
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-01:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs/notebooks$ ^C
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-01:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs/notebooks$ git clone https://huggingface.co/meta-llama/Llama-2-7b-hf
Cloning into 'Llama-2-7b-hf'...
remote: Enumerating objects: 127, done.
remote: Counting objects: 100% (127/127), done.
remote: Compressing objects: 100% (57/57), done.
remote: Total 127 (delta 69), reused 127 (delta 69), pack-reused 0 (from 0)
Receiving objects: 100% (127/127), 518.93 KiB | 12.66 MiB/s, done.
Resolving deltas: 100% (69/69), done.
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-01:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs/notebooks$

Single_node_finetuning_on_spr

What is Finetuning?

Fine-tuning is a process in which a data is fed to the model and tells the models internal weights to get closer to responding how we would like it. For example we can finetune a model for code generation , text generation and summarisation.In simple terms finetuning is a process in
which we can train a model to do specific task by providing the right dataset



```
[ ]: from transformers import TrainingArguments
from intel_extension_for_transformers.neural_chat.config import (
    ModelArguments,
    DataArguments,
    FinetuningArguments,
    TextGenerationFinetuningConfig,
)
from intel_extension_for_transformers.neural_chat.chatbot import finetune_model
model_args = ModelArguments(model_name_or_path="meta-llama/Llama-2-7b-chat-hf")
data_args = DataArguments(train_file="alpaca_data.json", validation_split_percentage=1)
training_args = TrainingArguments(
    output_dir='./tmp',
    do_train=True,
    do_eval=True,
    num_train_epochs=1,
    overwrite_output_dir=True,
    per_device_train_batch_size=4,
    per_device_eval_batch_size=4,
    gradient_accumulation_steps=2,
    save_strategy="no",
    log_level="info",
    save_total_limit=1,
    bf16=True,
)
finetune_args = FinetuningArguments()
finetune_cfg = TextGenerationFinetuningConfig(
            model_args=model_args,
            data_args=data_args,
            training_args=training_args,
            finetune_args=finetune_args,
        )
finetune_model(finetune_cfg)
```

Step 1:
Activate the conda environment

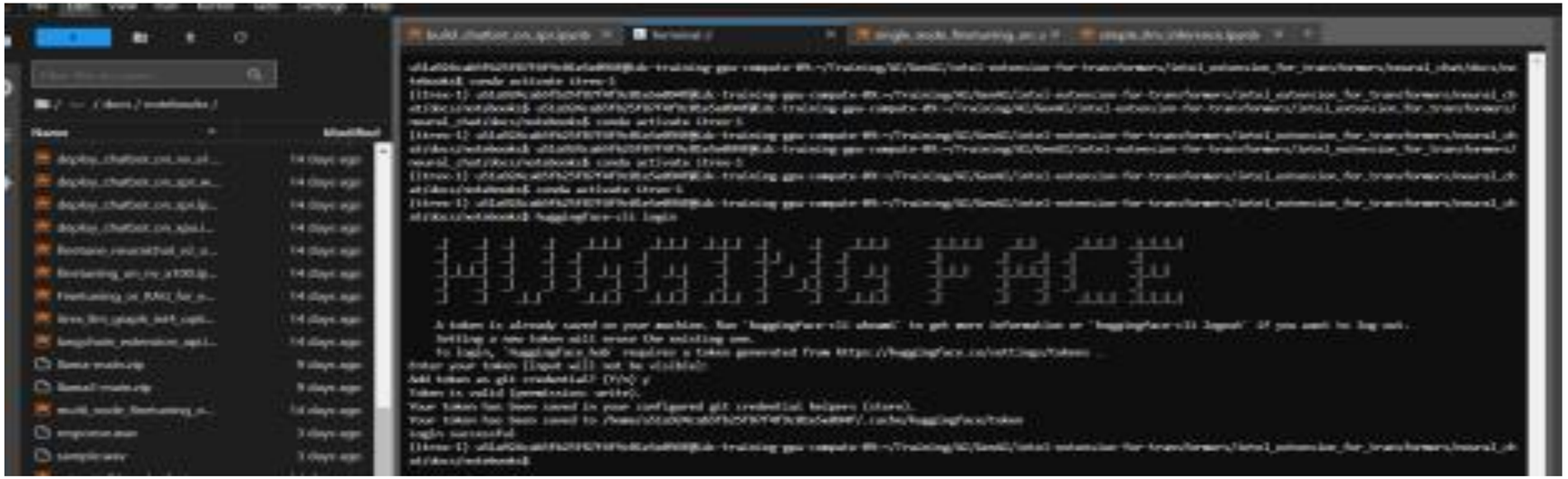
Step2:
Activate the hugging face hub using huggingface-cli login

Step 3:

Input the user token

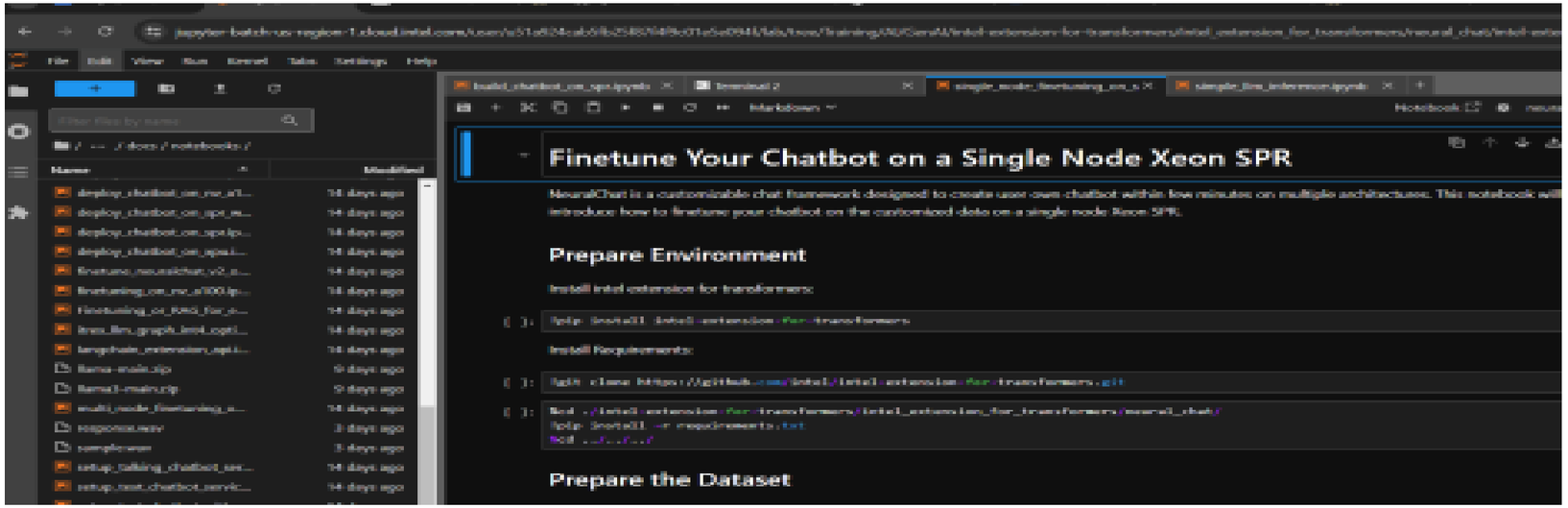
Step 4:

Once the token is accepted the user has the id to use hugging face datasets



Step 5:

Open the single_node_finetuning_on_spr on idc



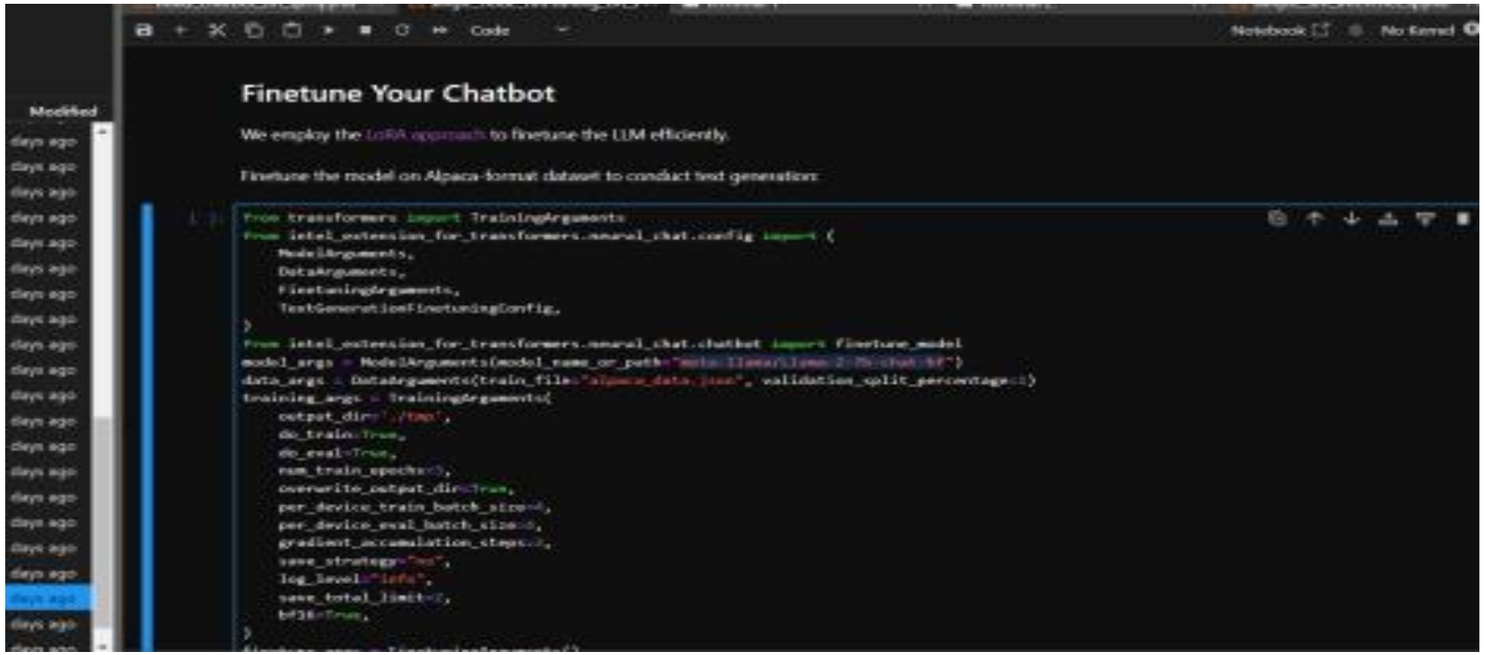
Step 6:

Inorder to finetune the 3 notebooks we need 3 different datasets required in the notebook . In the given notebook 3 different datasets are provided. The model mention here is the meta-llama/Llama-2-7b-chat-hf

The model required to run the given notebook is

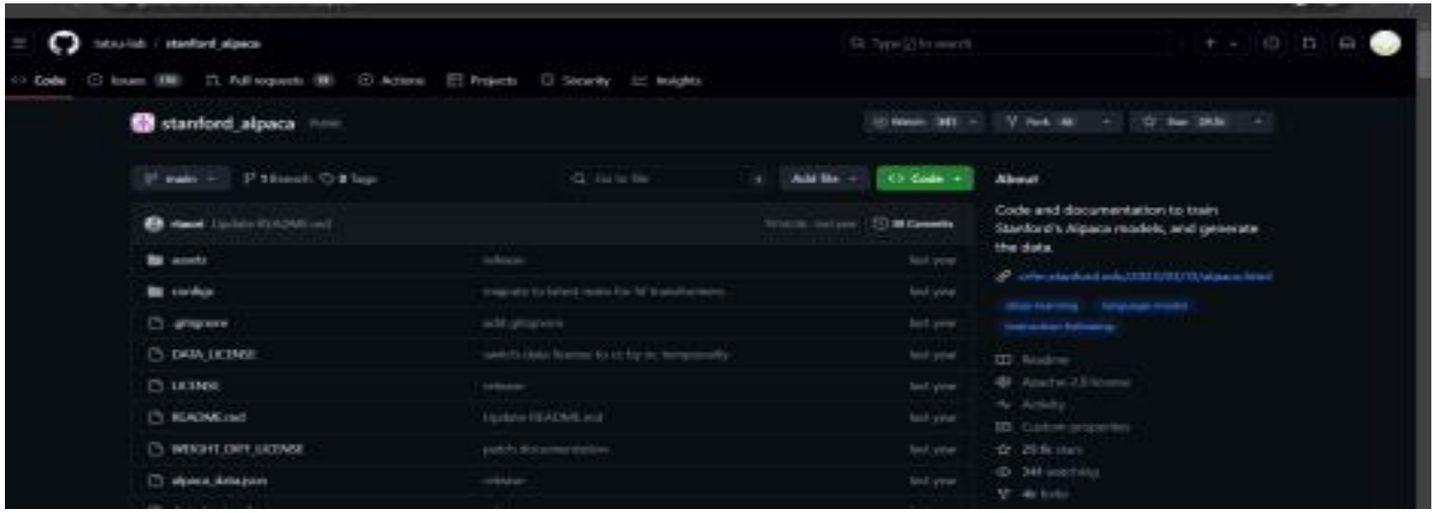
1. Alcapa dataset
2. Cnn_dailymail
3. theblackcat102/evol-codealpaca-v1

Step 7:



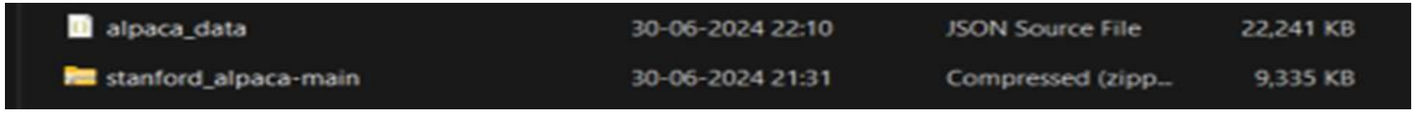
In the given notebook we have specified the model name and path as well as the dataset required for the finetuning The dataset mentioned in the notebook is the alpaca dataset.

Step 8:
Click the alpaca dataset in the notebook . Once its clicked it will redirect us to a github rep having the alpaca dataset



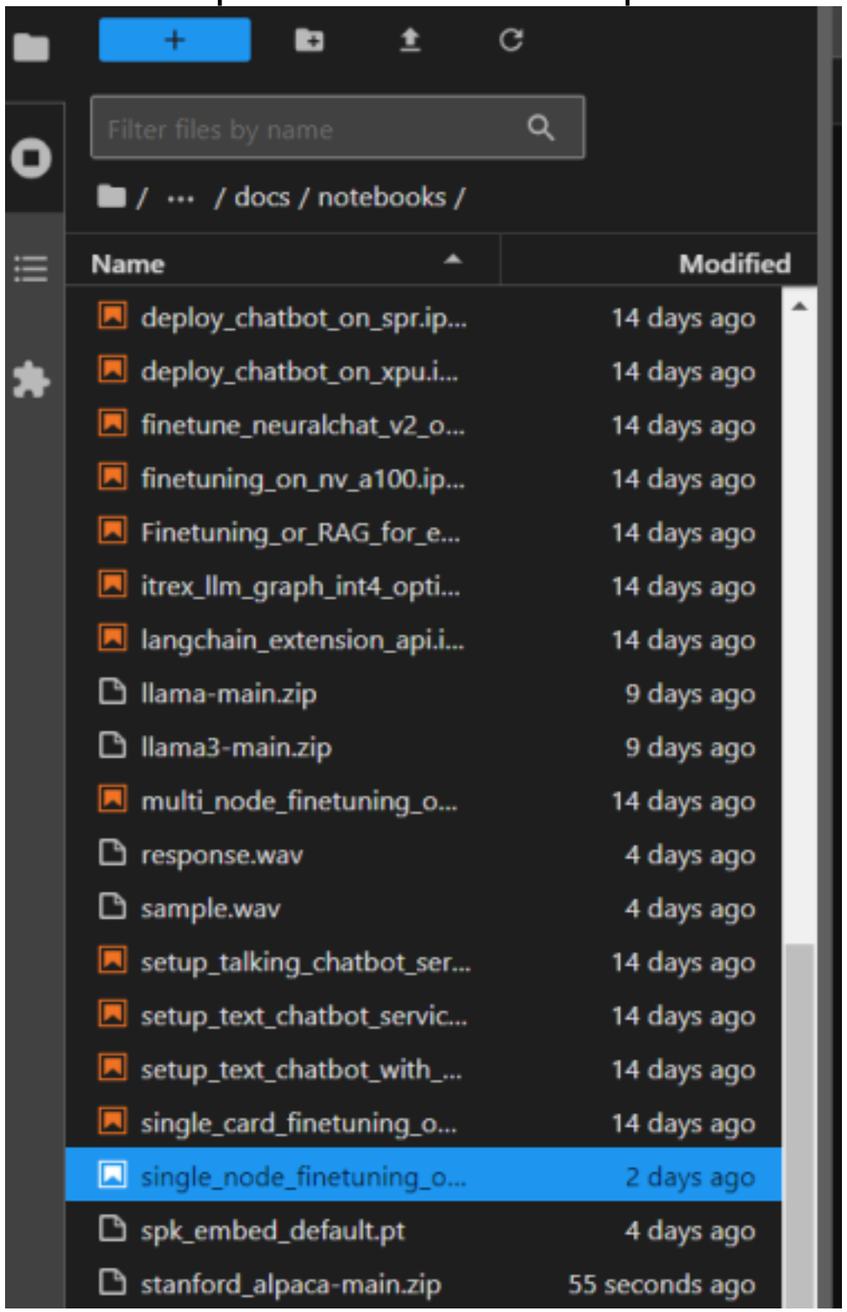
Step 9:
Inorder to use the dataset we need to download it as a zip file

Step 10: Once the file is downloaded it needs to be uploaded in idc or the local terminal for unzipping the zip file .



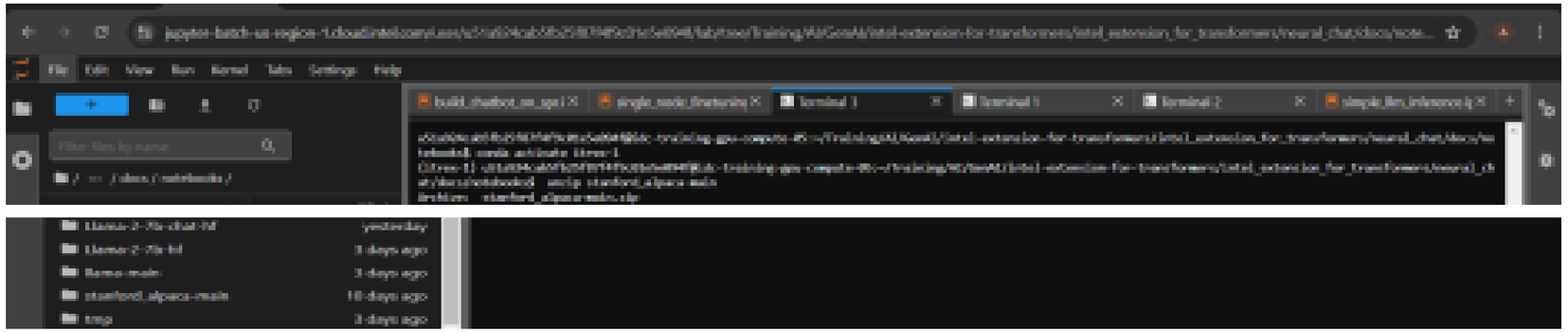
Step 11:

Once the file is uploaded we can unzip it in the terminal in idc server



Step 12:

Unzip it in the idc terminal



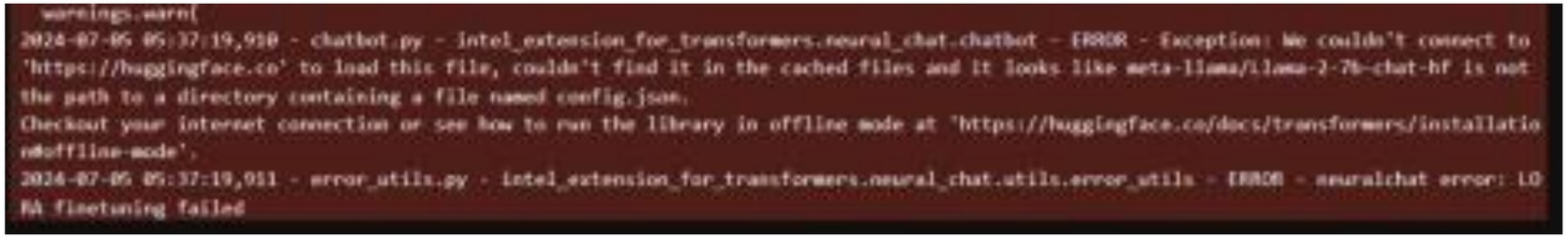
Step 13:

Run the notebook

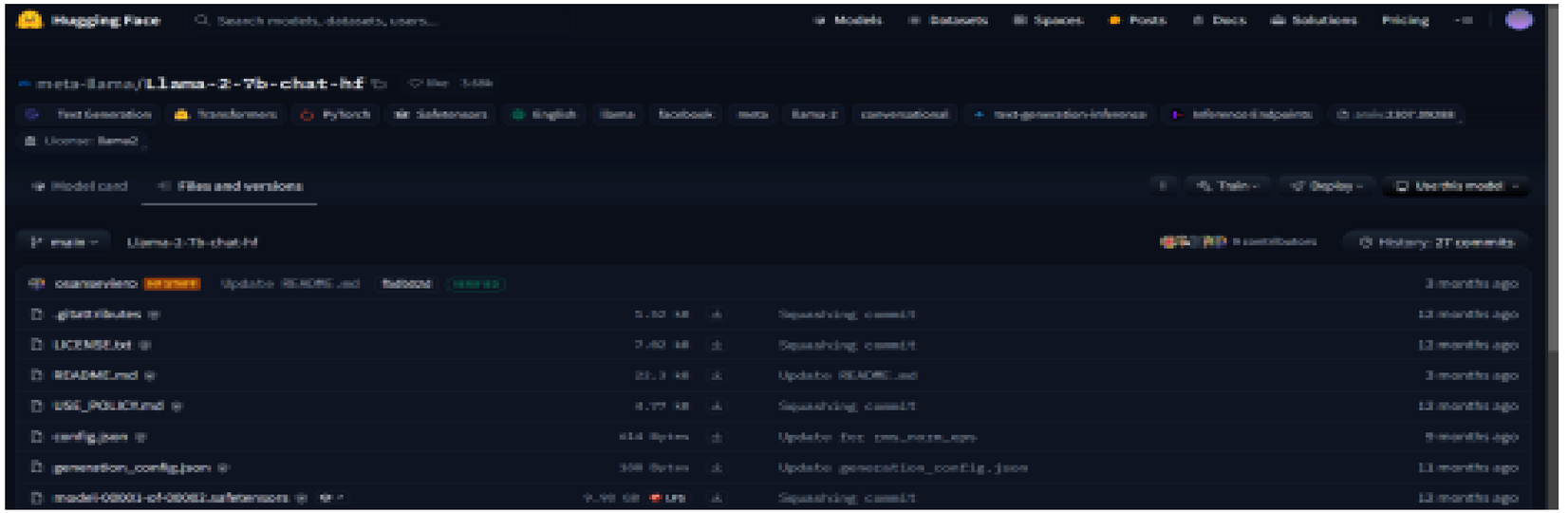
```
[ ]: from transformers import TrainingArguments
     from intel_extension_for_transformers.neural_chat.config import (
         ModelArguments,
         DataArguments,
         FinetuningArguments,
         TextGenerationFinetuningConfig,
     )
     from intel_extension_for_transformers.neural_chat.chatbot import finetune_model
     model_args = ModelArguments(model_name_or_path="meta-llama/Llama-2-7b-chat-hf")
     data_args = DataArguments(train_file="alpaca_data.json", validation_split_percentage=1)
     training_args = TrainingArguments(
         output_dir='./tmp',
         do_train=True,
         do_eval=True,
         num_train_epochs=3,
         overwrite_output_dir=True,
         per_device_train_batch_size=4,
         per_device_eval_batch_size=4,
         gradient_accumulation_steps=2,
         save_strategy="no",
         log_level="info",
         save_total_limit=2,
         bf16=True,
     )
     finetune_args = FinetuningArguments()
     finetune_cfg = TextGenerationFinetuningConfig(
                 model_args=model_args,
                 data_args=data_args,
                 training_args=training_args,
                 finetune_args=finetune_args,
             )
     finetune_model(finetune_cfg)
```

Step 14:

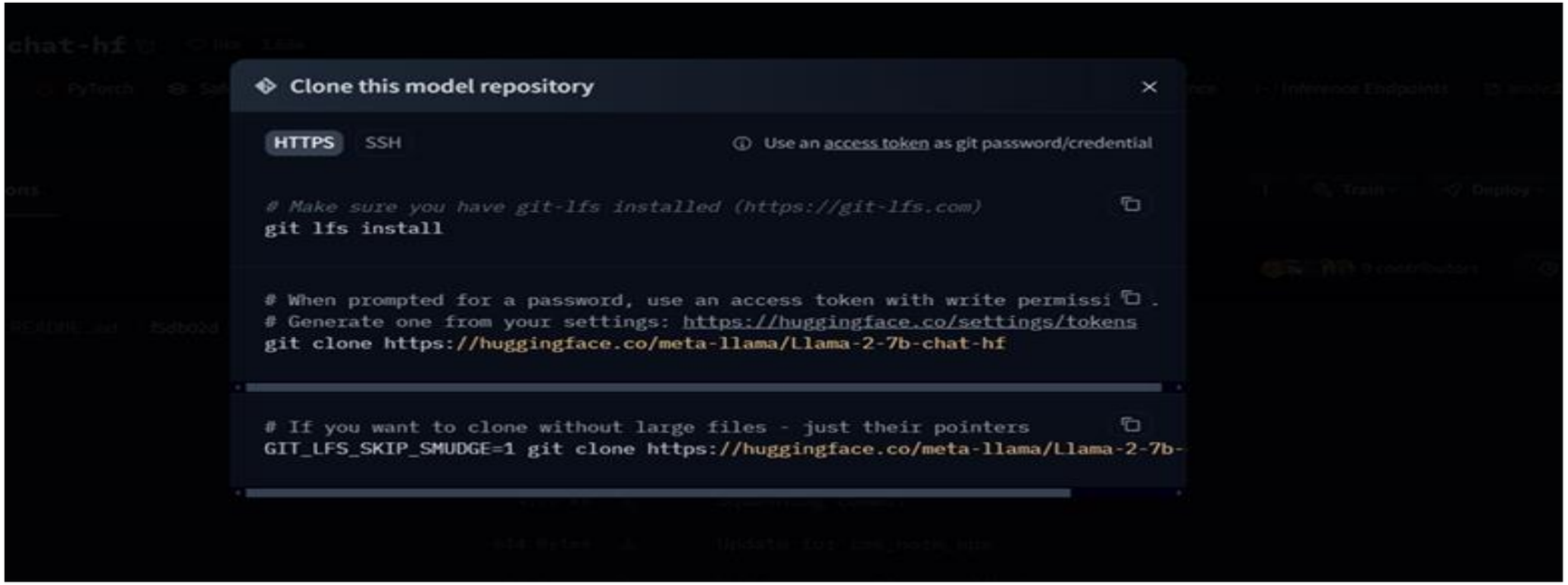
In certain cases the notebook may show an error saying the huggingface token is not able to connect to the meta-llama/Llama-2-7b-chat-hf model The error msg indicates



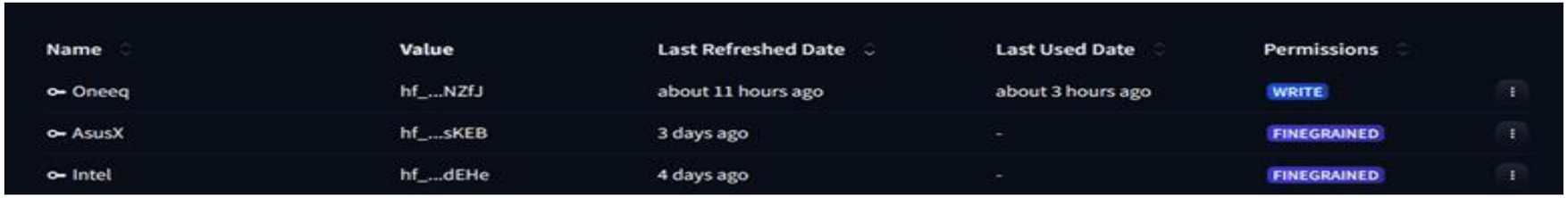
The only way to run the notebook is by loading the meta lama mode in offline mode ; inorder to do so we need to download the entire files of lama model



Or we can clone the entire meta lama repo



In order to clone this repo we need a Hugging face token with write permission



Step 16:

Now we need to login to the hugging face id using the write permitted token

```
u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-18:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs/notebooks$ conda ac
tivate itrex-1
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-18:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs/notebooks
$ huggingface-cli login

    A token is already saved on your machine. Run `huggingface-cli whoami` to get more information or `huggingface-cli logout` if you want to log out.
    Setting a new token will erase the existing one.
    To login, `huggingface_hub` requires a token generated from https://huggingface.co/settings/tokens .
Enter your token (input will not be visible):
Add token as git credential? (Y/n) y
Token is valid (permission: write).
Your token has been saved in your configured git credential helpers (store).
Your token has been saved to /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/token
Login successful
```

Step 17:

Since meta lama model repo huge we cannot clone it directly ; in order to clone it in windows we need to install git lfs :

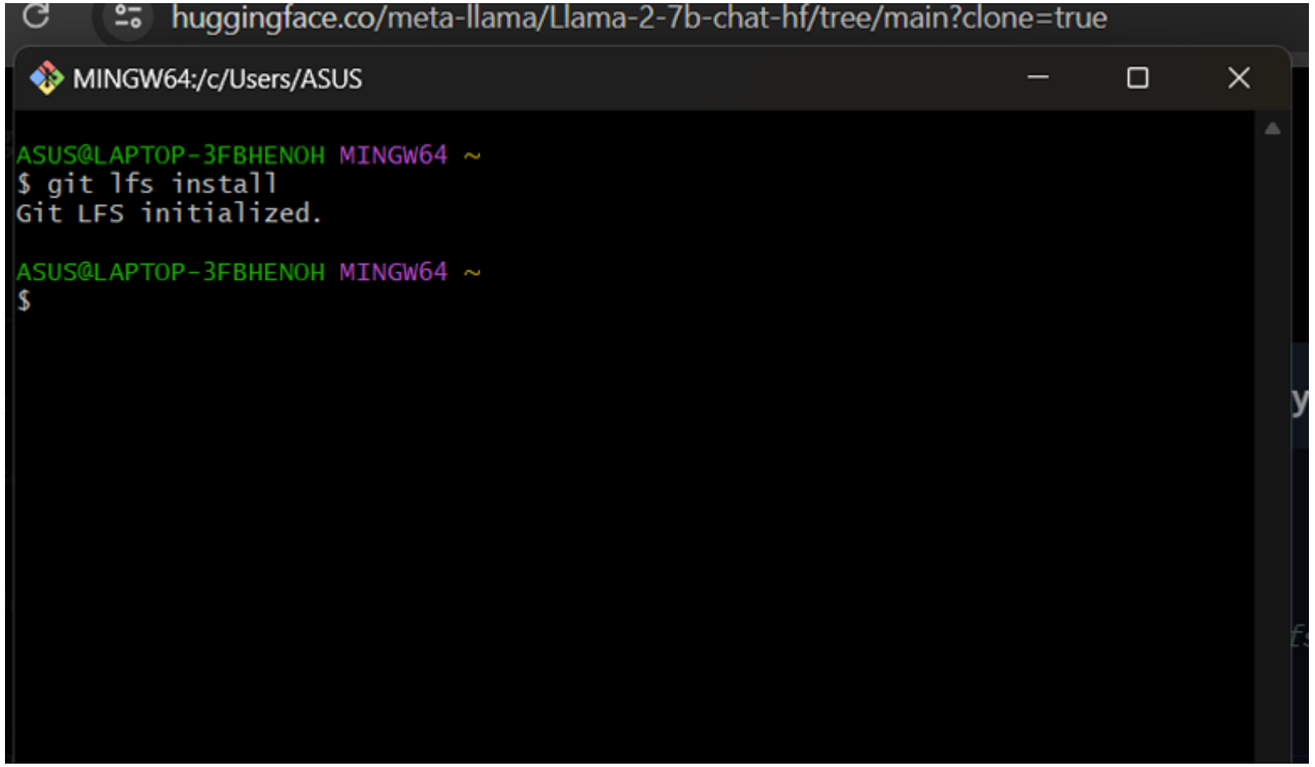
(Git large file storage) which can be done by manually installing it from the git website

Or by clone git lfs repo into the terminal

```
Login successful
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-18:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs/notebooks
$ git clone https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
fatal: destination path 'Llama-2-7b-chat-hf' already exists and is not an empty directory.
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-18:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs/notebooks
$ git clone https://github.com/git-lfs/git-lfs.git
Cloning into 'git-lfs'...
remote: Enumerating objects: 49113, done.
remote: Counting objects: 100% (294/294), done.
remote: Compressing objects: 100% (161/161), done.
remote: Total 49113 (delta 149), reused 243 (delta 133), pack-reused 48819
Receiving objects: 100% (49113/49113), 19.30 MiB | 20.74 MiB/s, done.
Resolving deltas: 100% (33940/33940), done.
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-18:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs/notebooks
$
```

Step 18:

Once the installation or cloning is completely done In case of manual installation we may need to activate lfs in git bash

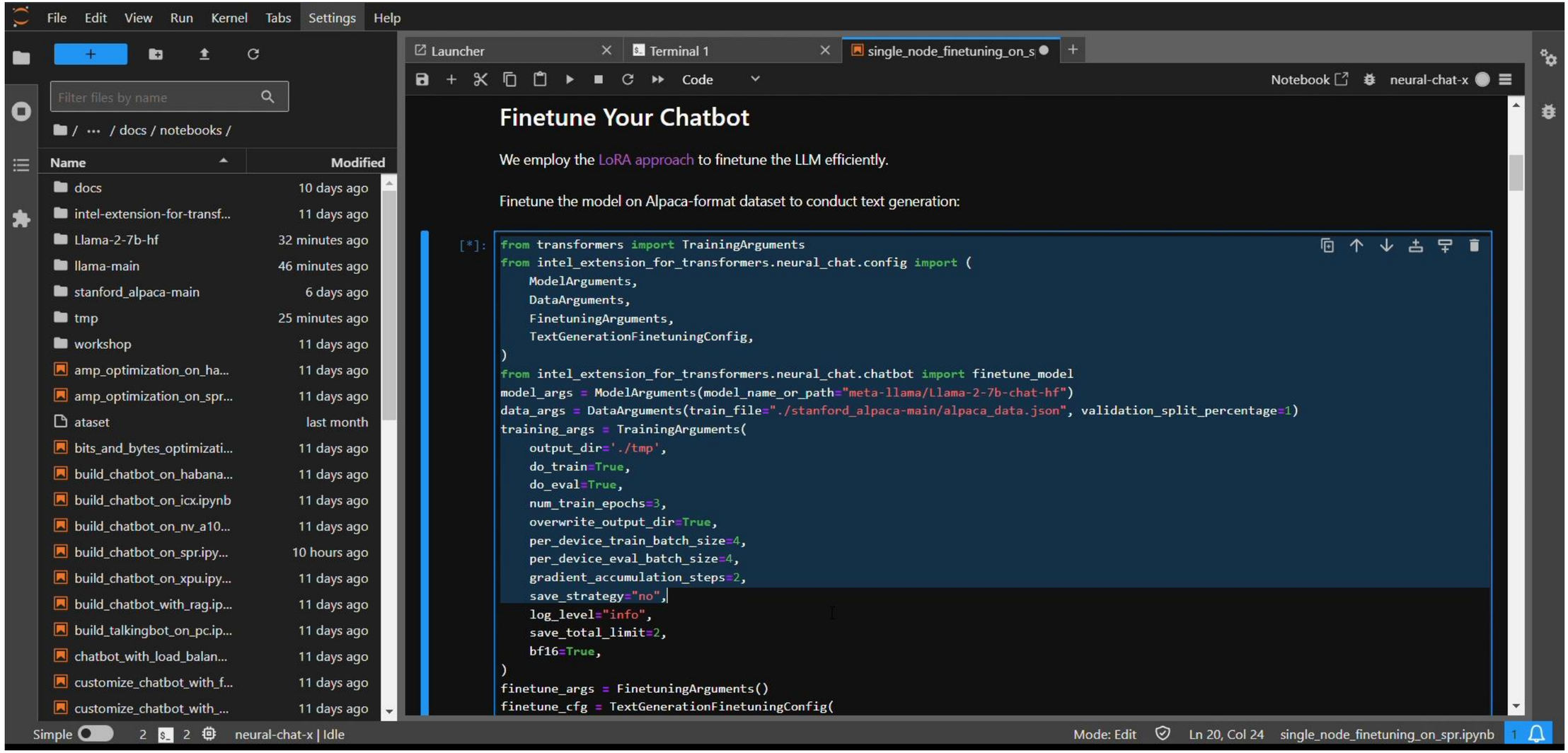


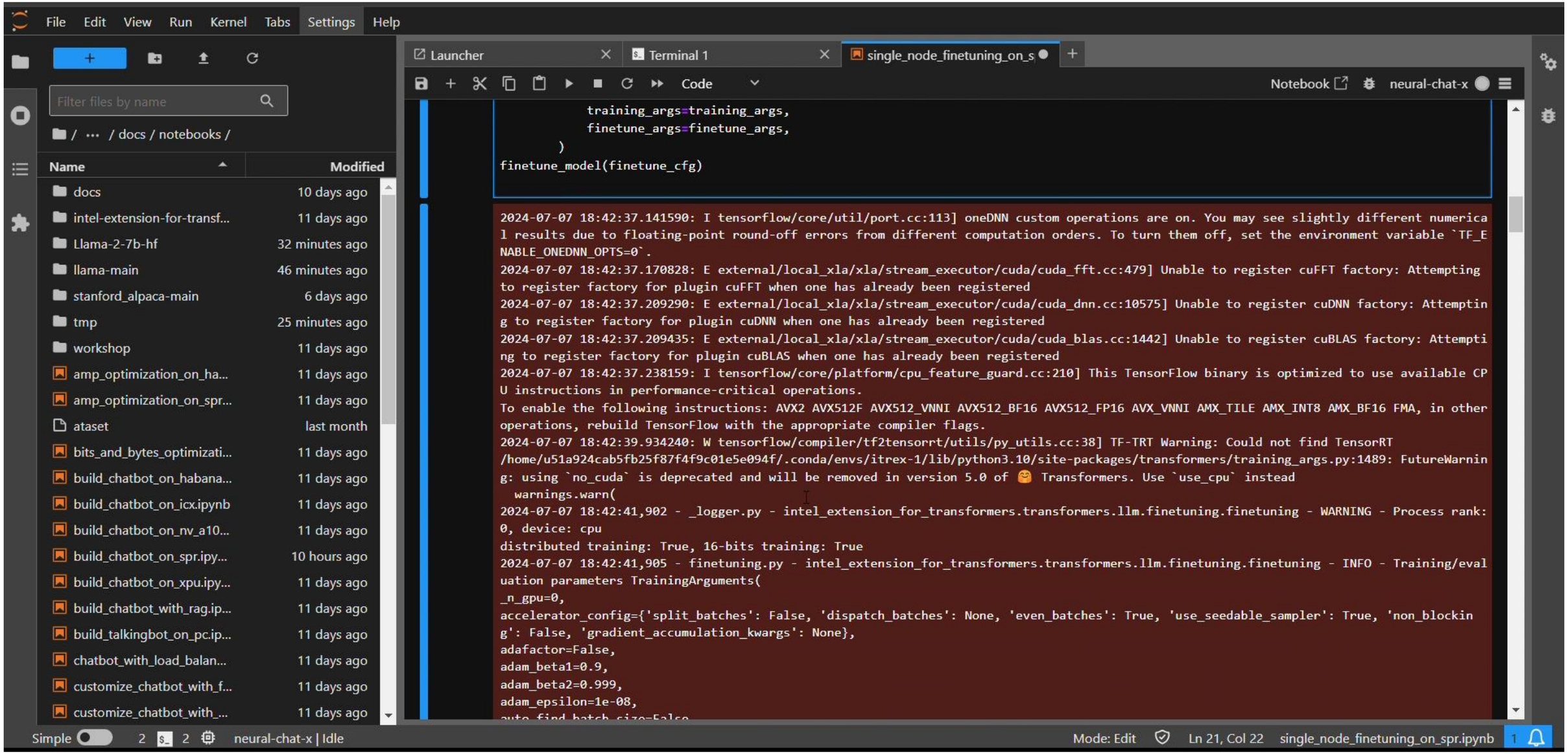
In case of cloning lfs ; we can directly clone the meta lama model

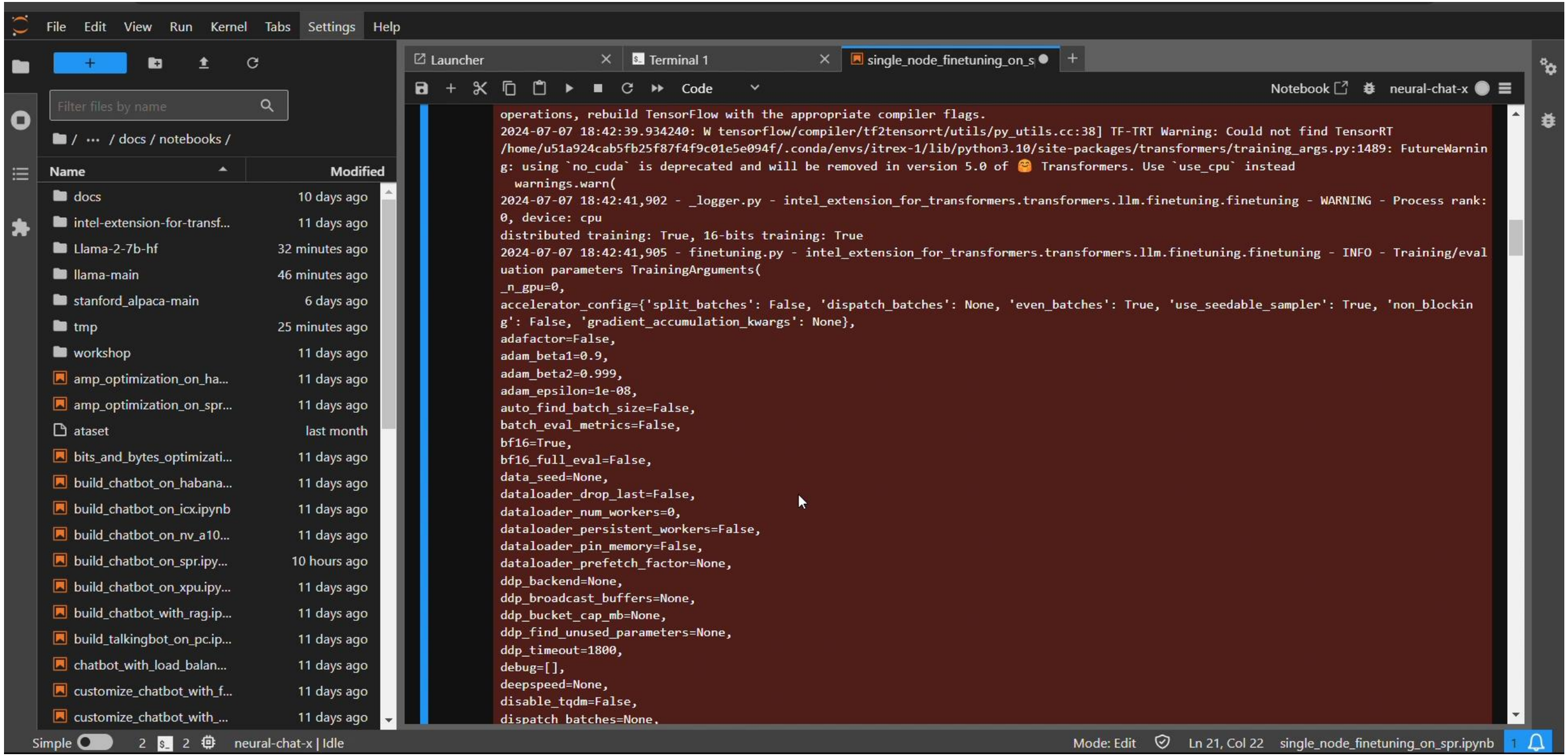
Step 19:

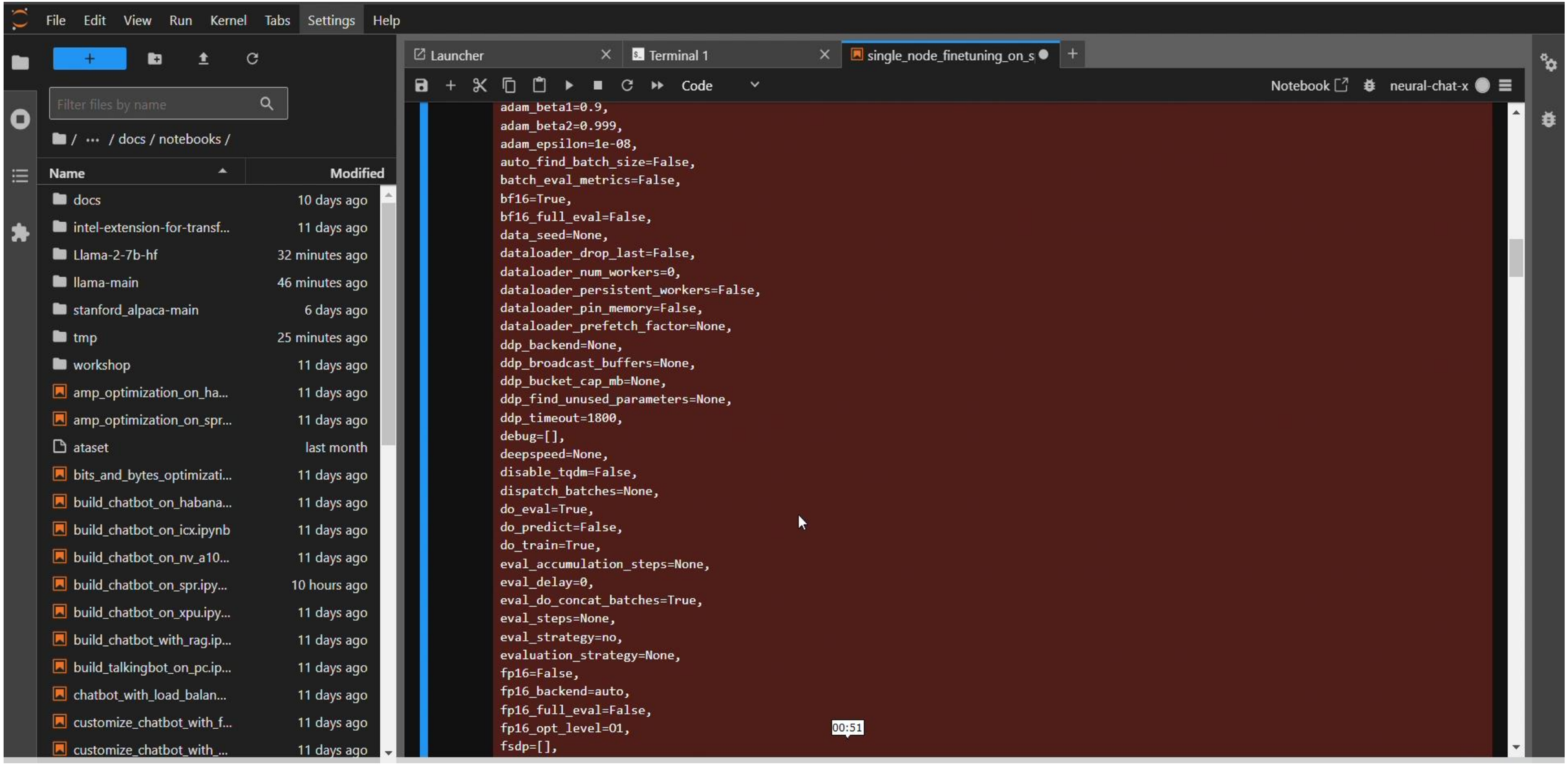
Once the model is completely cloned or installed we can run the notebook with the alpaca dataset in the correct directory . Similarly we can finetune 3 notebooks by accessing the hugging face tokens or by manually cloning or installing the required datasets in the correct directories

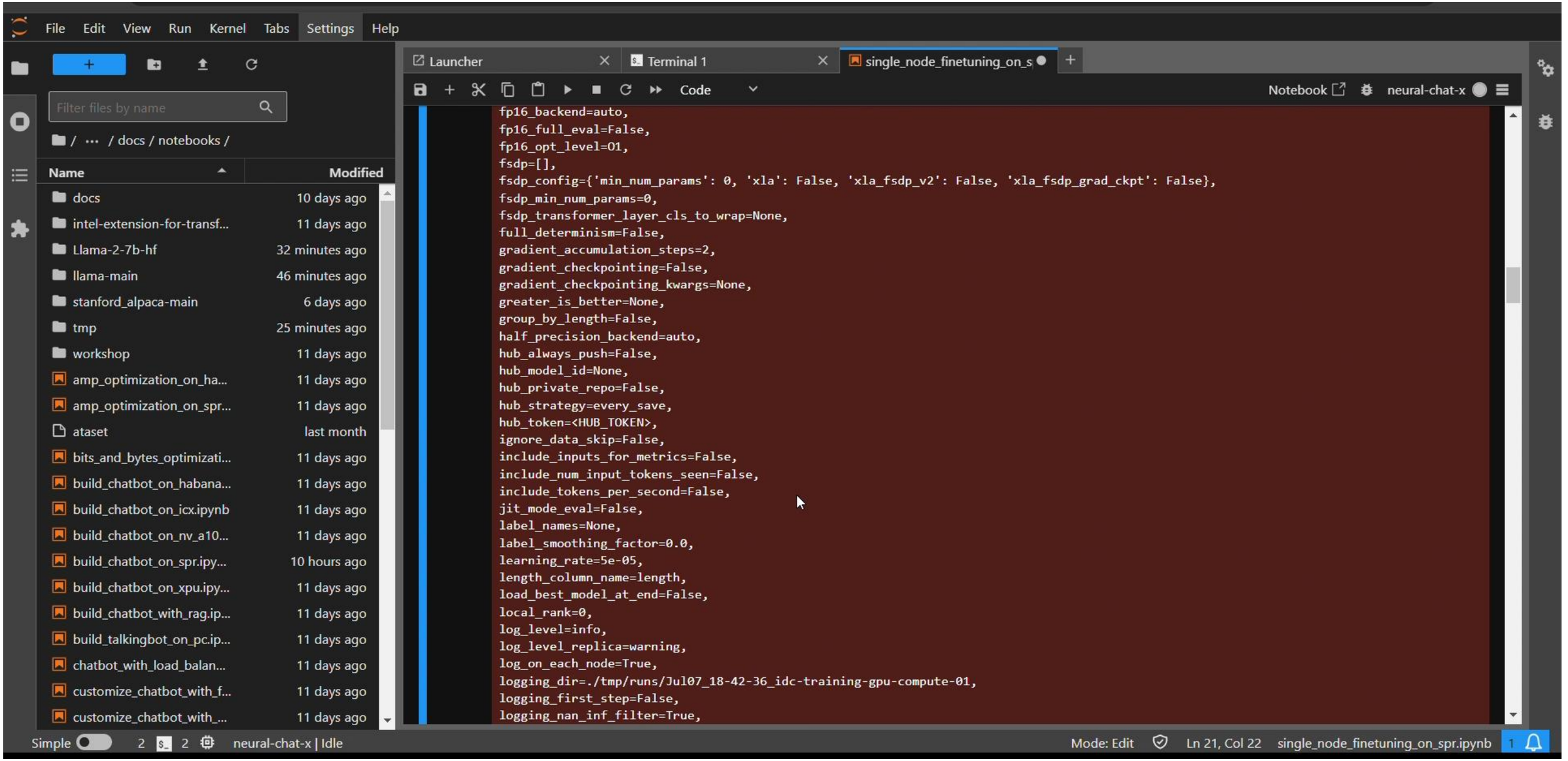
Finetuning chatbot

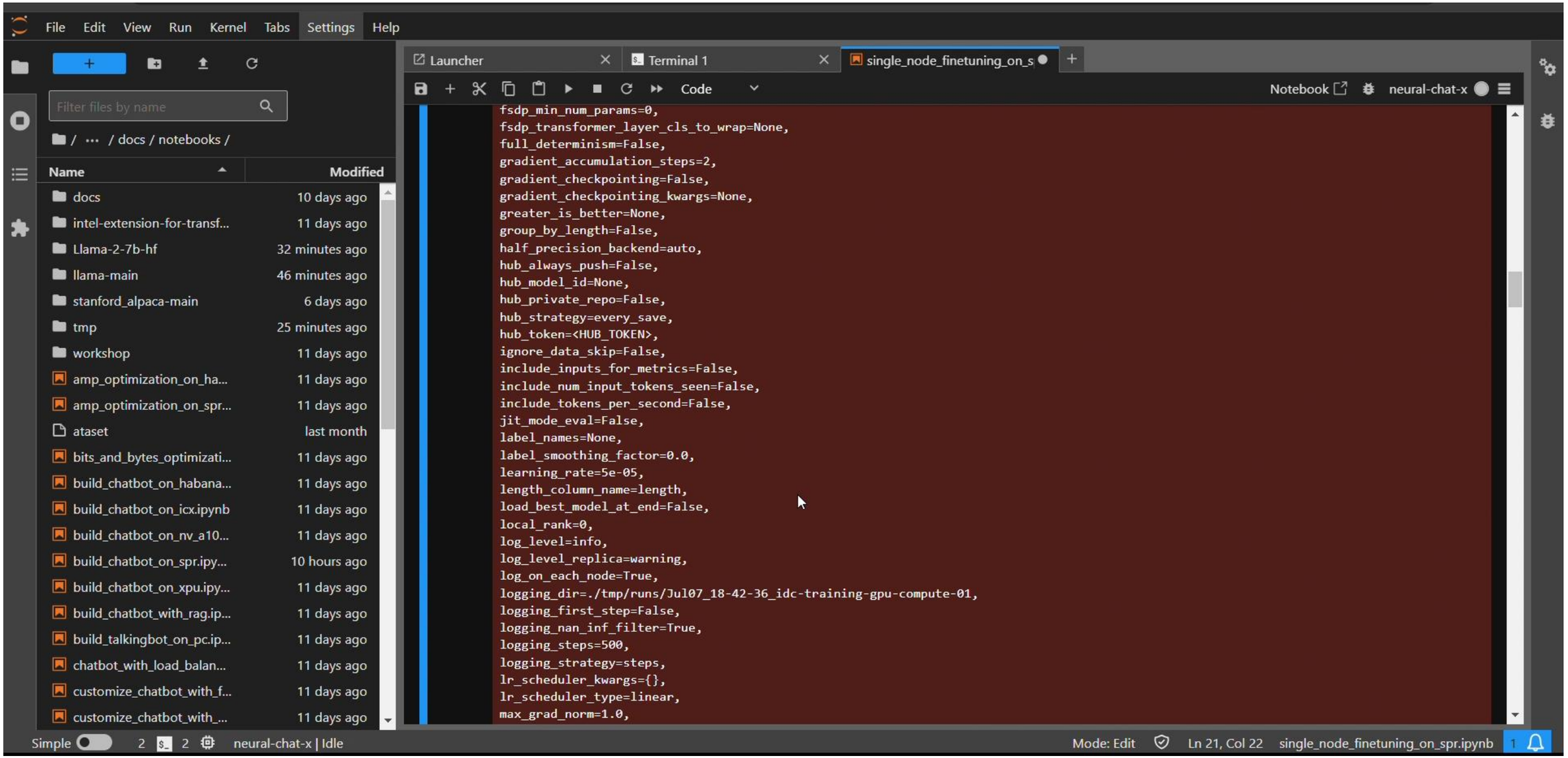


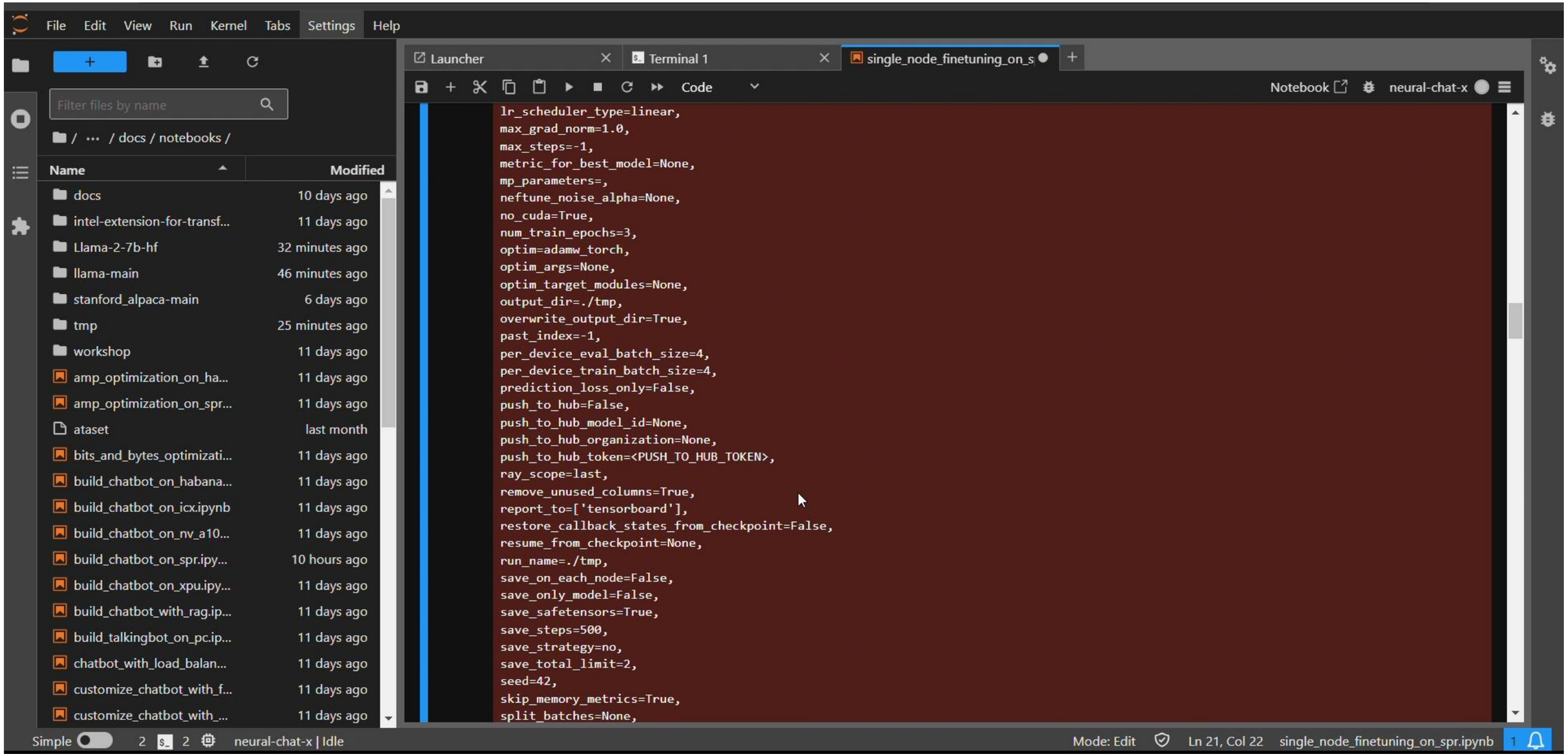

File Edit View Run Kernel Tabs Settings Help
Launcher
Terminal 1
single_node_finetuning_on_s
Code
Notebook
neural-chat-x
Filter files by name
/ ... / docs / notebooks /
Name
Modified
docs 10 days ago
intel-extension-for-transf... 11 days ago
Llama-2-7b-hf 32 minutes ago
llama-main 46 minutes ago
stanford_alpaca-main 6 days ago
tmp 25 minutes ago
workshop 11 days ago
amp_optimization_on_ha... 11 days ago
amp_optimization_on_spr... 11 days ago
ataset last month
bits_and_bytes_optimizati... 11 days ago
build_chatbot_on_habana... 11 days ago
build_chatbot_on_icx.ipynb 11 days ago
build_chatbot_on_nv_a10... 11 days ago
build_chatbot_on_spr.ipy... 10 hours ago
build_chatbot_on_xpu.ipy... 11 days ago
build_chatbot_with_rag.ip... 11 days ago
build_talkingbot_on_pc.ip... 11 days ago
chatbot_with_load_balan... 11 days ago
customize_chatbot_with_f... 11 days ago
customize_chatbot_with_... 11 days ago
training_args=training_args,
finetune_args=finetune_args,
)
finetune_model(finetune_cfg)
2024-07-07 18:42:37.141590: I tensorflow/core/util/port.cc:113] oneDNN custom operations are on. You may see slightly different numerical results due to floating-point round-off errors from different computation orders. To turn them off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2024-07-07 18:42:37.170828: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:479] Unable to register cuFFT factory: Attempting to register factory for plugin cuFFT when one has already been registered
2024-07-07 18:42:37.209290: E external/local_xla/xla/stream_executor/cuda/cuda_dnn.cc:10575] Unable to register cuDNN factory: Attempting to register factory for plugin cuDNN when one has already been registered
2024-07-07 18:42:37.209435: E external/local_xla/xla/stream_executor/cuda/cuda_blas.cc:1442] Unable to register cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has already been registered
2024-07-07 18:42:37.238159: I tensorflow/core/platform/cpu_feature_guard.cc:210] This TensorFlow binary is optimized to use available CPU instructions in performance-critical operations.
To enable the following instructions: AVX2 AVX512F AVX512_VNNI AVX512_BF16 AVX512_FP16 AVX_VNNI AMX_TILE AMX_INT8 AMX_BF16 FMA, in other operations, rebuild TensorFlow with the appropriate compiler flags.
2024-07-07 18:42:39.934240: W tensorflow/compiler/tf2tensorrt/utils/py_utils.cc:38] TF-TRT Warning: Could not find TensorRT
/home/u51a924cab5fb25f87f4f9c01e5e094f/.conda/envs/itrex-1/lib/python3.10/site-packages/transformers/training_args.py:1489: FutureWarning: using `no_cuda` is deprecated and will be removed in version 5.0 of 🤗 Transformers. Use `use_cpu` instead
warnings.warn(
2024-07-07 18:42:41,902 - _logger.py - intel_extension_for_transformers.transformers.llm.finetuning.finetuning - WARNING - Process rank: 0, device: cpu
distributed training: True, 16-bits training: True
2024-07-07 18:42:41,905 - finetuning.py - intel_extension_for_transformers.transformers.llm.finetuning.finetuning - INFO - Training/evaluation parameters TrainingArguments(
_n_gpu=0,
accelerator_config={'split_batches': False, 'dispatch_batches': None, 'even_batches': True, 'use_seedable_sampler': True, 'non_blocking': False, 'gradient_accumulation_kwargs': None},
adafactor=False,
adam_beta1=0.9,
adam_beta2=0.999,
adam_epsilon=1e-08,
Simple
neural-chat-x | Idle
Mode: Edit
Ln 21, Col 22
single_node_finetuning_on_spr.ipynb

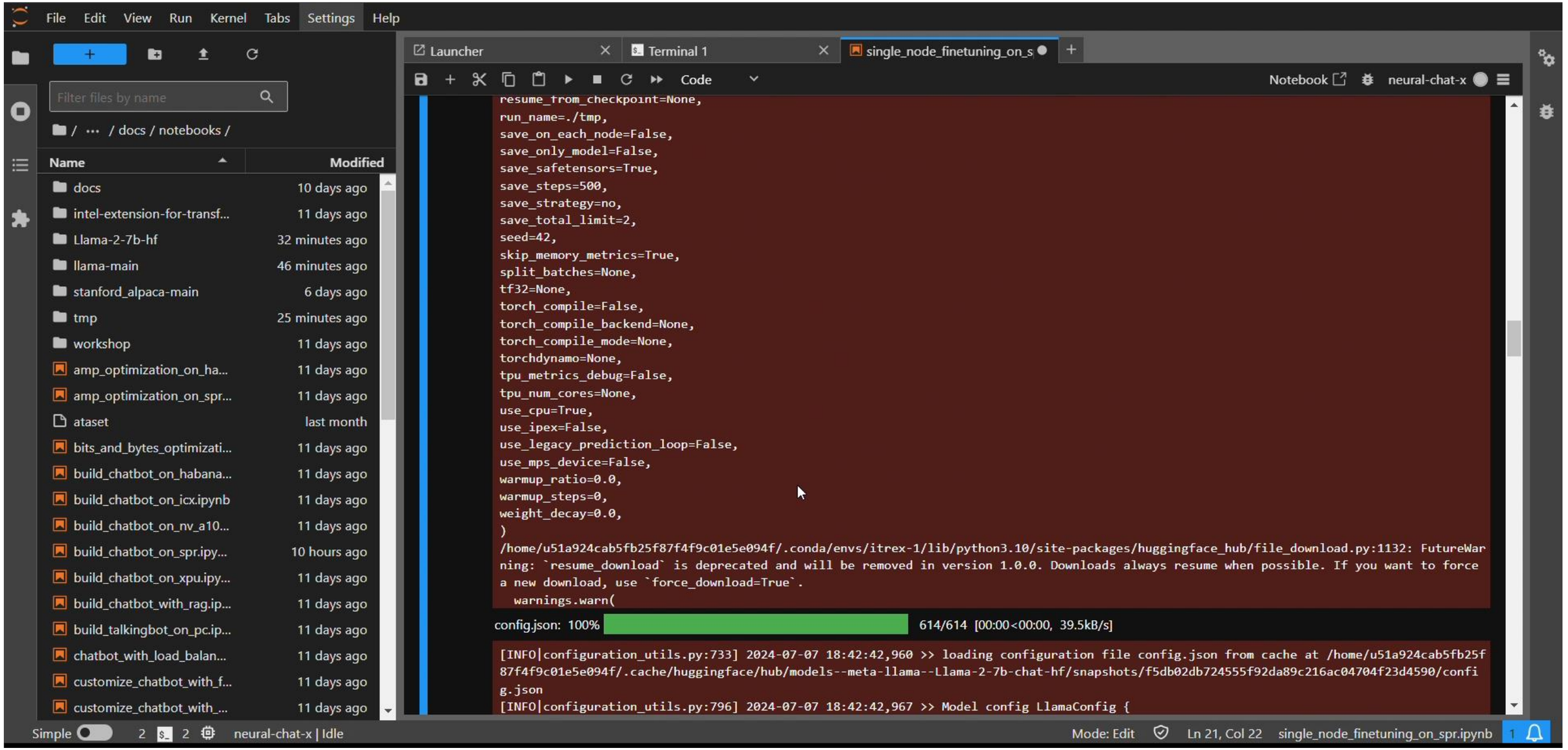

File Edit View Run Kernel Tabs Settings Help
Launcher
Terminal 1
single_node_finetuning_on_s
Code
Notebook
neural-chat-x
Filter files by name
/ ... / docs / notebooks /
Name
Modified
docs 10 days ago
intel-extension-for-transf... 11 days ago
Llama-2-7b-hf 32 minutes ago
llama-main 46 minutes ago
stanford_alpaca-main 6 days ago
tmp 25 minutes ago
workshop 11 days ago
amp_optimization_on_ha... 11 days ago
amp_optimization_on_spr... 11 days ago
ataset last month
bits_and_bytes_optimizati... 11 days ago
build_chatbot_on_habana... 11 days ago
build_chatbot_on_icx.ipynb 11 days ago
build_chatbot_on_nv_a10... 11 days ago
build_chatbot_on_spr.ipy... 10 hours ago
build_chatbot_on_xpu.ipy... 11 days ago
build_chatbot_with_rag.ip... 11 days ago
build_talkingbot_on_pc.ip... 11 days ago
chatbot_with_load_balan... 11 days ago
customize_chatbot_with_f... 11 days ago
customize_chatbot_with_... 11 days ago
operations, rebuild TensorFlow with the appropriate compiler flags.
2024-07-07 18:42:39.934240: W tensorflow/compiler/tf2tensorrt/utils/py_utils.cc:38] TF-TRT Warning: Could not find TensorRT
/home/u51a924cab5fb25f87f4f9c01e5e094f/.conda/envs/itrex-1/lib/python3.10/site-packages/transformers/training_args.py:1489: FutureWarnin
g: using `no_cuda` is deprecated and will be removed in version 5.0 of 🤗 Transformers. Use `use_cpu` instead
warnings.warn(
2024-07-07 18:42:41,902 - _logger.py - intel_extension_for_transformers.transformers.llm.finetuning.finetuning - WARNING - Process rank:
0, device: cpu
distributed training: True, 16-bits training: True
2024-07-07 18:42:41,905 - finetuning.py - intel_extension_for_transformers.transformers.llm.finetuning.finetuning - INFO - Training/eval
uation parameters TrainingArguments(
_n_gpu=0,
accelerator_config={'split_batches': False, 'dispatch_batches': None, 'even_batches': True, 'use_seedable_sampler': True, 'non_blockin
g': False, 'gradient_accumulation_kwargs': None},
adafactor=False,
adam_beta1=0.9,
adam_beta2=0.999,
adam_epsilon=1e-08,
auto_find_batch_size=False,
batch_eval_metrics=False,
bf16=True,
bf16_full_eval=False,
data_seed=None,
dataloader_drop_last=False,
dataloader_num_workers=0,
dataloader_persistent_workers=False,
dataloader_pin_memory=False,
dataloader_prefetch_factor=None,
ddp_backend=None,
ddp_broadcast_buffers=None,
ddp_bucket_cap_mb=None,
ddp_find_unused_parameters=None,
ddp_timeout=1800,
debug=[],
deepspeed=None,
disable_tqdm=False,
dispatch_batches=None,
Simple 2 2 neural-chat-x | Idle
Mode: Edit Ln 21, Col 22 single_node_finetuning_on_spr.ipynb 1

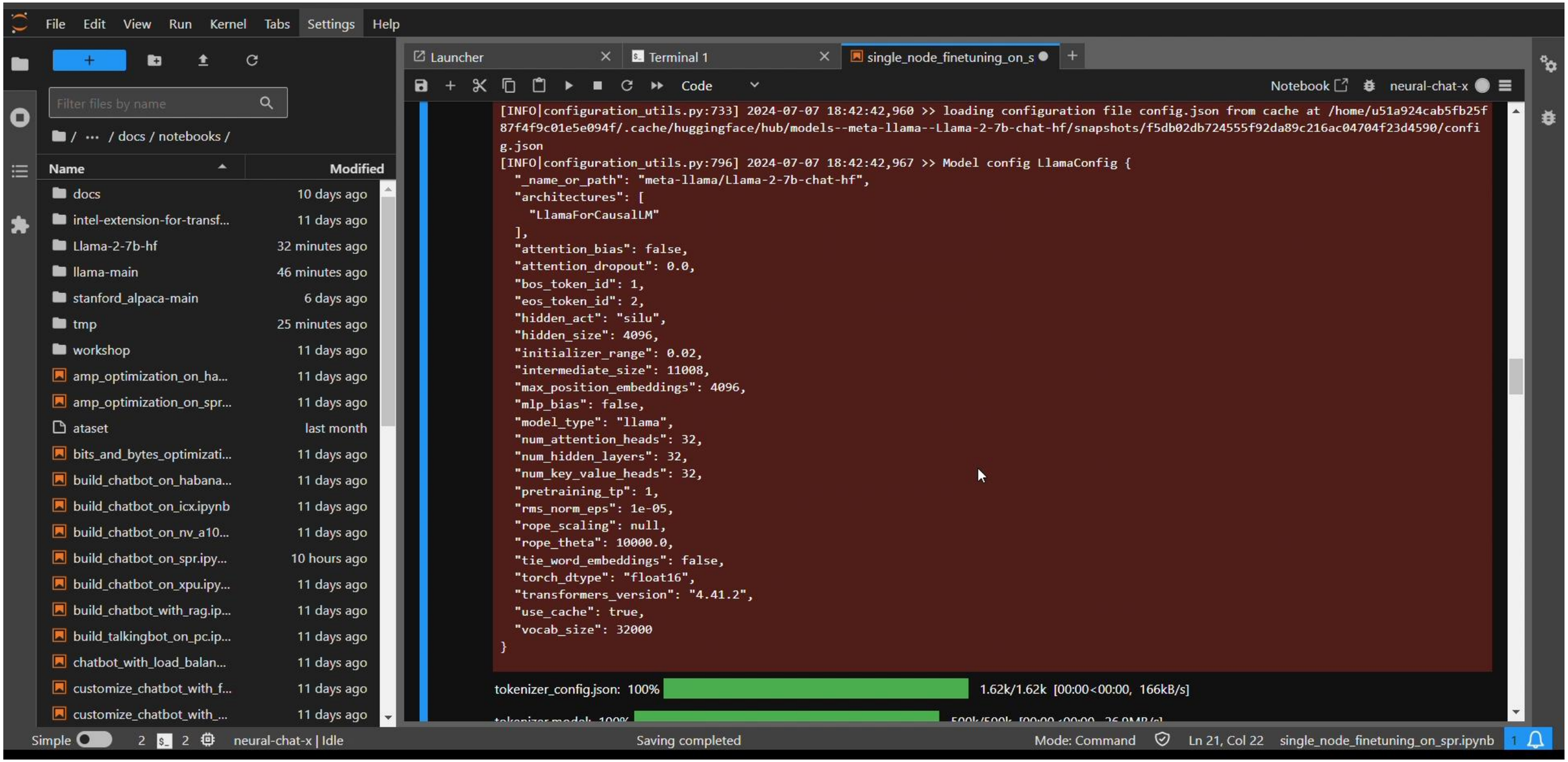

File Edit View Run Kernel Tabs Settings Help
Launcher
Terminal 1
single_node_finetuning_on_s
Code
Notebook
neural-chat-x
Filter files by name
/ ··· / docs / notebooks /
Name
Modified
docs 10 days ago
intel-extension-for-transf... 11 days ago
Llama-2-7b-hf 32 minutes ago
llama-main 46 minutes ago
stanford_alpaca-main 6 days ago
tmp 25 minutes ago
workshop 11 days ago
amp_optimization_on_ha... 11 days ago
amp_optimization_on_spr... 11 days ago
ataset last month
bits_and_bytes_optimizati... 11 days ago
build_chatbot_on_habana... 11 days ago
build_chatbot_on_icx.ipynb 11 days ago
build_chatbot_on_nv_a10... 11 days ago
build_chatbot_on_spr.ipy... 10 hours ago
build_chatbot_on_xpu.ipy... 11 days ago
build_chatbot_with_rag.ip... 11 days ago
build_talkingbot_on_pc.ip... 11 days ago
chatbot_with_load_balan... 11 days ago
customize_chatbot_with_f... 11 days ago
customize_chatbot_with_... 11 days ago
adam_beta1=0.9,
adam_beta2=0.999,
adam_epsilon=1e-08,
auto_find_batch_size=False,
batch_eval_metrics=False,
bf16=True,
bf16_full_eval=False,
data_seed=None,
dataloader_drop_last=False,
dataloader_num_workers=0,
dataloader_persistent_workers=False,
dataloader_pin_memory=False,
dataloader_prefetch_factor=None,
ddp_backend=None,
ddp_broadcast_buffers=None,
ddp_bucket_cap_mb=None,
ddp_find_unused_parameters=None,
ddp_timeout=1800,
debug=[],
deepspeed=None,
disable_tqdm=False,
dispatch_batches=None,
do_eval=True,
do_predict=False,
do_train=True,
eval_accumulation_steps=None,
eval_delay=0,
eval_do_concat_batches=True,
eval_steps=None,
eval_strategy=no,
evaluation_strategy=None,
fp16=False,
fp16_backend=auto,
fp16_full_eval=False,
fp16_opt_level=O1,
fsdp=[],
00:51

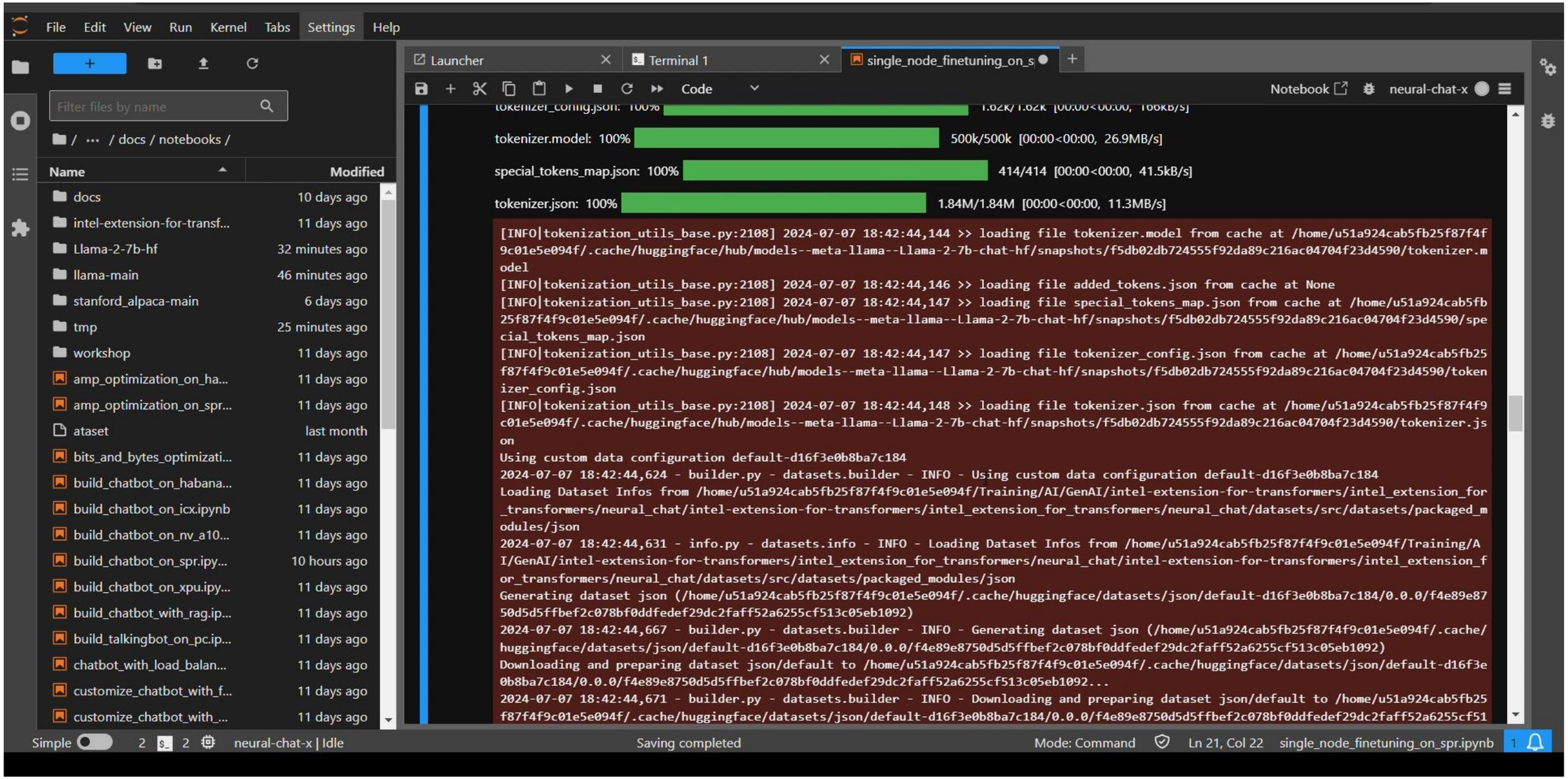

File Edit View Run Kernel Tabs Settings Help
Launcher
Terminal 1
single_node_finetuning_on_s
Code
Notebook
neural-chat-x
Filter files by name
/ ... / docs / notebooks /
Name
Modified
docs 10 days ago
intel-extension-for-transf... 11 days ago
Llama-2-7b-hf 32 minutes ago
llama-main 46 minutes ago
stanford_alpaca-main 6 days ago
tmp 25 minutes ago
workshop 11 days ago
amp_optimization_on_ha... 11 days ago
amp_optimization_on_spr... 11 days ago
ataset last month
bits_and_bytes_optimizati... 11 days ago
build_chatbot_on_habana... 11 days ago
build_chatbot_on_icx.ipynb 11 days ago
build_chatbot_on_nv_a10... 11 days ago
build_chatbot_on_spr.ipy... 10 hours ago
build_chatbot_on_xpu.ipy... 11 days ago
build_chatbot_with_rag.ip... 11 days ago
build_talkingbot_on_pc.ip... 11 days ago
chatbot_with_load_balan... 11 days ago
customize_chatbot_with_f... 11 days ago
customize_chatbot_with_... 11 days ago
fp16_backend=auto,
fp16_full_eval=False,
fp16_opt_level=O1,
fsdp=[],
fsdp_config={'min_num_params': 0, 'xla': False, 'xla_fsdp_v2': False, 'xla_fsdp_grad_ckpt': False},
fsdp_min_num_params=0,
fsdp_transformer_layer_cls_to_wrap=None,
full_determinism=False,
gradient_accumulation_steps=2,
gradient_checkpointing=False,
gradient_checkpointing_kwargs=None,
greater_is_better=None,
group_by_length=False,
half_precision_backend=auto,
hub_always_push=False,
hub_model_id=None,
hub_private_repo=False,
hub_strategy=every_save,
hub_token=<HUB_TOKEN>,
ignore_data_skip=False,
include_inputs_for_metrics=False,
include_num_input_tokens_seen=False,
include_tokens_per_second=False,
jit_mode_eval=False,
label_names=None,
label_smoothing_factor=0.0,
learning_rate=5e-05,
length_column_name=length,
load_best_model_at_end=False,
local_rank=0,
log_level=info,
log_level_replica=warning,
log_on_each_node=True,
logging_dir=./tmp/runs/Jul07_18-42-36_idc-training-gpu-compute-01,
logging_first_step=False,
logging_nan_inf_filter=True,
Simple
2 2
neural-chat-x | Idle
Mode: Edit
Ln 21, Col 22
single_node_finetuning_on_spr.ipynb
1

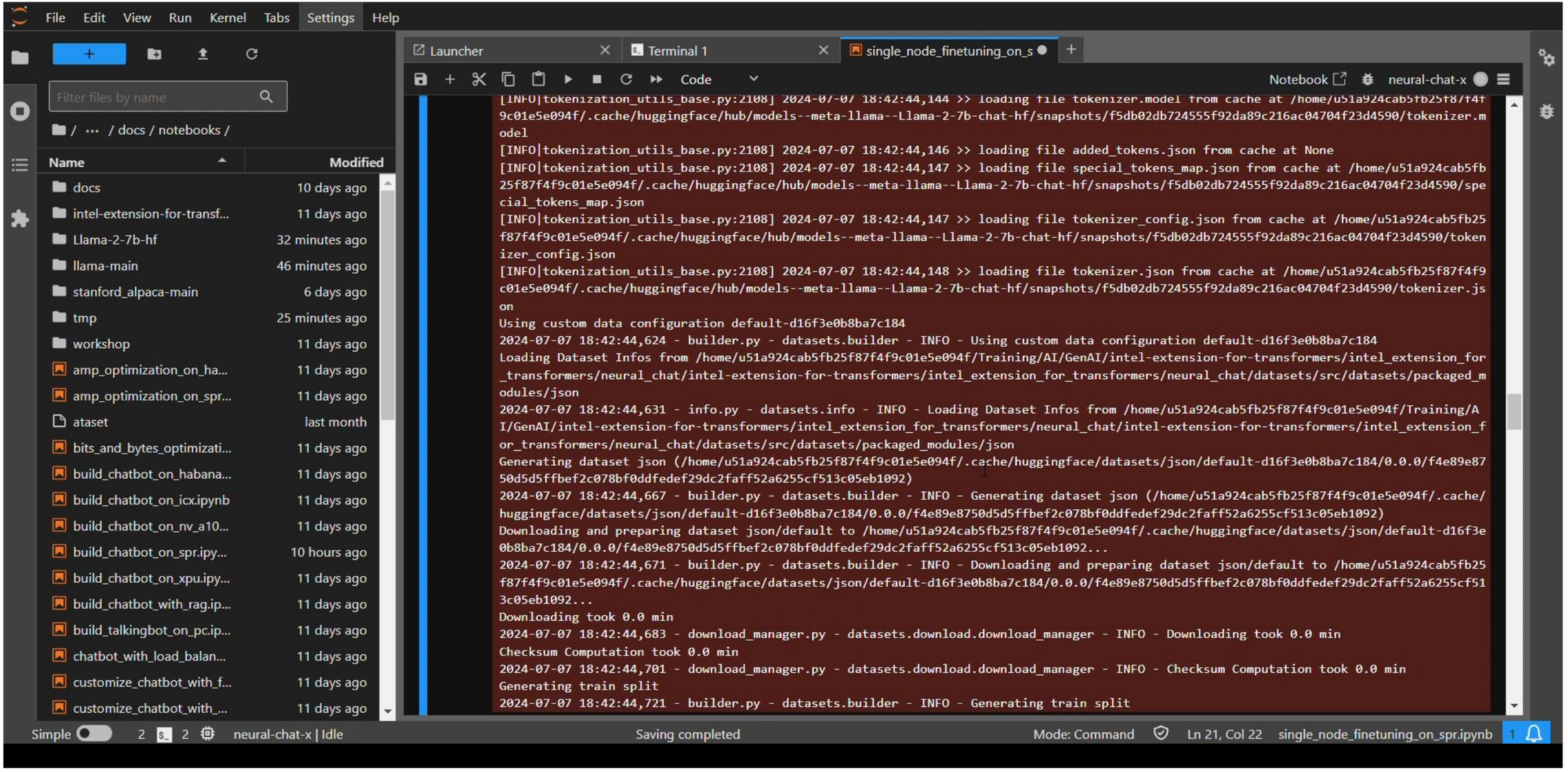

File Edit View Run Kernel Tabs Settings Help
Launcher
Terminal 1
single_node_finetuning_on_s
Code
Notebook
neural-chat-x
Filter files by name
/ ... / docs / notebooks /
Name
Modified
docs 10 days ago
intel-extension-for-transf... 11 days ago
Llama-2-7b-hf 32 minutes ago
llama-main 46 minutes ago
stanford_alpaca-main 6 days ago
tmp 25 minutes ago
workshop 11 days ago
amp_optimization_on_ha... 11 days ago
amp_optimization_on_spr... 11 days ago
ataset last month
bits_and_bytes_optimizati... 11 days ago
build_chatbot_on_habana... 11 days ago
build_chatbot_on_icx.ipynb 11 days ago
build_chatbot_on_nv_a10... 11 days ago
build_chatbot_on_spr.ipy... 10 hours ago
build_chatbot_on_xpu.ipy... 11 days ago
build_chatbot_with_rag.ip... 11 days ago
build_talkingbot_on_pc.ip... 11 days ago
chatbot_with_load_balan... 11 days ago
customize_chatbot_with_f... 11 days ago
customize_chatbot_with_... 11 days ago
fsdp_min_num_params=0,
fsdp_transformer_layer_cls_to_wrap=None,
full_determinism=False,
gradient_accumulation_steps=2,
gradient_checkpointing=False,
gradient_checkpointing_kwargs=None,
greater_is_better=None,
group_by_length=False,
half_precision_backend=auto,
hub_always_push=False,
hub_model_id=None,
hub_private_repo=False,
hub_strategy=every_save,
hub_token=<HUB_TOKEN>,
ignore_data_skip=False,
include_inputs_for_metrics=False,
include_num_input_tokens_seen=False,
include_tokens_per_second=False,
jit_mode_eval=False,
label_names=None,
label_smoothing_factor=0.0,
learning_rate=5e-05,
length_column_name=length,
load_best_model_at_end=False,
local_rank=0,
log_level=info,
log_level_replica=warning,
log_on_each_node=True,
logging_dir=./tmp/runs/Jul07_18-42-36_idc-training-gpu-compute-01,
logging_first_step=False,
logging_nan_inf_filter=True,
logging_steps=500,
logging_strategy=steps,
lr_scheduler_kwargs={},
lr_scheduler_type=linear,
max_grad_norm=1.0,
Simple
2 2
neural-chat-x | Idle
Mode: Edit
Ln 21, Col 22
single_node_finetuning_on_spr.ipynb
1

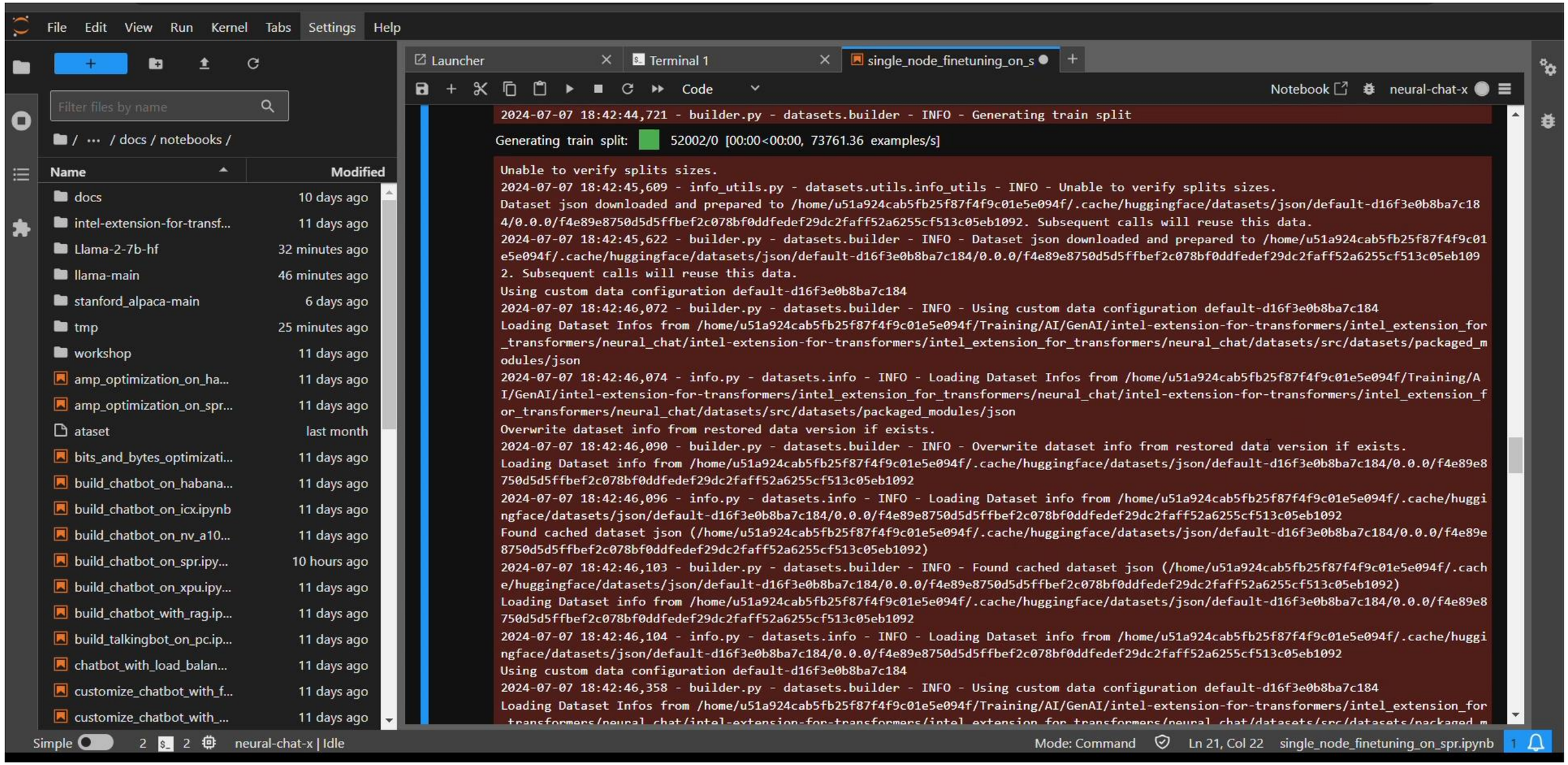

File Edit View Run Kernel Tabs Settings Help
Launcher
Terminal 1
single_node_finetuning_on_s
Code
Notebook
neural-chat-x
Filter files by name
/ ... / docs / notebooks /
Name
Modified
docs 10 days ago
intel-extension-for-transf... 11 days ago
Llama-2-7b-hf 32 minutes ago
llama-main 46 minutes ago
stanford_alpaca-main 6 days ago
tmp 25 minutes ago
workshop 11 days ago
amp_optimization_on_ha... 11 days ago
amp_optimization_on_spr... 11 days ago
ataset last month
bits_and_bytes_optimizati... 11 days ago
build_chatbot_on_habana... 11 days ago
build_chatbot_on_icx.ipynb 11 days ago
build_chatbot_on_nv_a10... 11 days ago
build_chatbot_on_spr.ipy... 10 hours ago
build_chatbot_on_xpu.ipy... 11 days ago
build_chatbot_with_rag.ip... 11 days ago
build_talkingbot_on_pc.ip... 11 days ago
chatbot_with_load_balan... 11 days ago
customize_chatbot_with_f... 11 days ago
customize_chatbot_with_... 11 days ago
lr_scheduler_type=linear,
max_grad_norm=1.0,
max_steps=-1,
metric_for_best_model=None,
mp_parameters=,
neftune_noise_alpha=None,
no_cuda=True,
num_train_epochs=3,
optim=adamw_torch,
optim_args=None,
optim_target_modules=None,
output_dir=./tmp,
overwrite_output_dir=True,
past_index=-1,
per_device_eval_batch_size=4,
per_device_train_batch_size=4,
prediction_loss_only=False,
push_to_hub=False,
push_to_hub_model_id=None,
push_to_hub_organization=None,
push_to_hub_token=<PUSH_TO_HUB_TOKEN>,
ray_scope=last,
remove_unused_columns=True,
report_to=['tensorboard'],
restore_callback_states_from_checkpoint=False,
resume_from_checkpoint=None,
run_name=./tmp,
save_on_each_node=False,
save_only_model=False,
save_safetensors=True,
save_steps=500,
save_strategy=no,
save_total_limit=2,
seed=42,
skip_memory_metrics=True,
split_batches=None,
Simple 2 2 neural-chat-x | Idle
Mode: Edit Ln 21, Col 22 single_node_finetuning_on_spr.ipynb 1

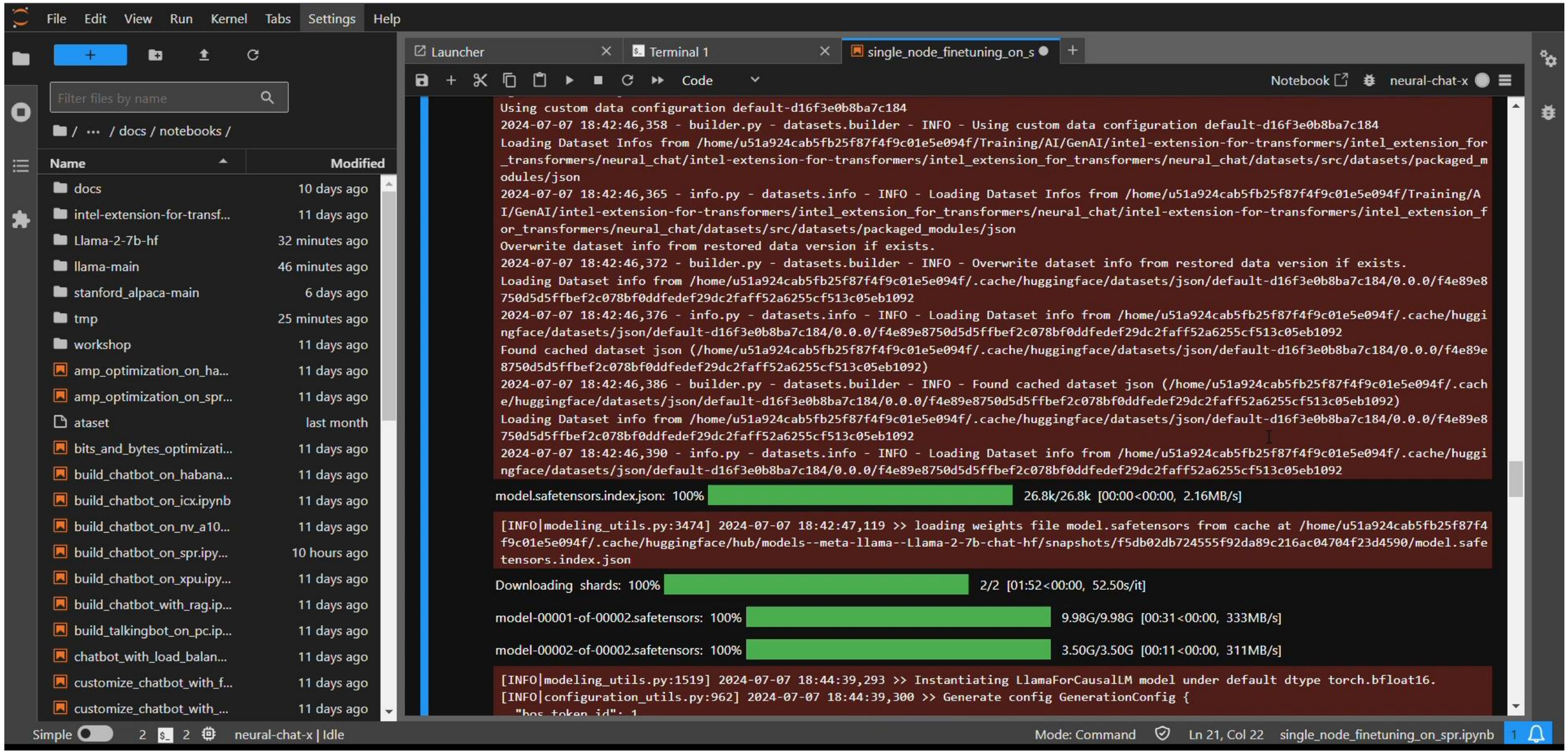

File Edit View Run Kernel Tabs Settings Help
Launcher
Terminal 1
single_node_finetuning_on_s
Code
Notebook
neural-chat-x
Filter files by name
/ ... / docs / notebooks /
Name
Modified
docs 10 days ago
intel-extension-for-transf... 11 days ago
Llama-2-7b-hf 32 minutes ago
llama-main 46 minutes ago
stanford_alpaca-main 6 days ago
tmp 25 minutes ago
workshop 11 days ago
amp_optimization_on_ha... 11 days ago
amp_optimization_on_spr... 11 days ago
ataset last month
bits_and_bytes_optimizati... 11 days ago
build_chatbot_on_habana... 11 days ago
build_chatbot_on_icx.ipynb 11 days ago
build_chatbot_on_nv_a10... 11 days ago
build_chatbot_on_spr.ipy... 10 hours ago
build_chatbot_on_xpu.ipy... 11 days ago
build_chatbot_with_rag.ip... 11 days ago
build_talkingbot_on_pc.ip... 11 days ago
chatbot_with_load_balan... 11 days ago
customize_chatbot_with_f... 11 days ago
customize_chatbot_with_... 11 days ago
resume_from_checkpoint=None,
run_name=./tmp,
save_on_each_node=False,
save_only_model=False,
save_safetensors=True,
save_steps=500,
save_strategy=no,
save_total_limit=2,
seed=42,
skip_memory_metrics=True,
split_batches=None,
tf32=None,
torch_compile=False,
torch_compile_backend=None,
torch_compile_mode=None,
torchdynamo=None,
tpu_metrics_debug=False,
tpu_num_cores=None,
use_cpu=True,
use_ipex=False,
use_legacy_prediction_loop=False,
use_mps_device=False,
warmup_ratio=0.0,
warmup_steps=0,
weight_decay=0.0,
)
/home/u51a924cab5fb25f87f4f9c01e5e094f/.conda/envs/itrex-1/lib/python3.10/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(
config.json: 100% 614/614 [00:00<00:00, 39.5kB/s]
[INFO|configuration_utils.py:733] 2024-07-07 18:42:42,960 >> loading configuration file config.json from cache at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/hub/models--meta-llama--Llama-2-7b-chat-hf/snapshots/f5db02db724555f92da89c216ac04704f23d4590/config.json
[INFO|configuration_utils.py:796] 2024-07-07 18:42:42,967 >> Model config LlamaConfig {
Simple 2 2 neural-chat-x | Idle
Mode: Edit Ln 21, Col 22 single_node_finetuning_on_spr.ipynb 1

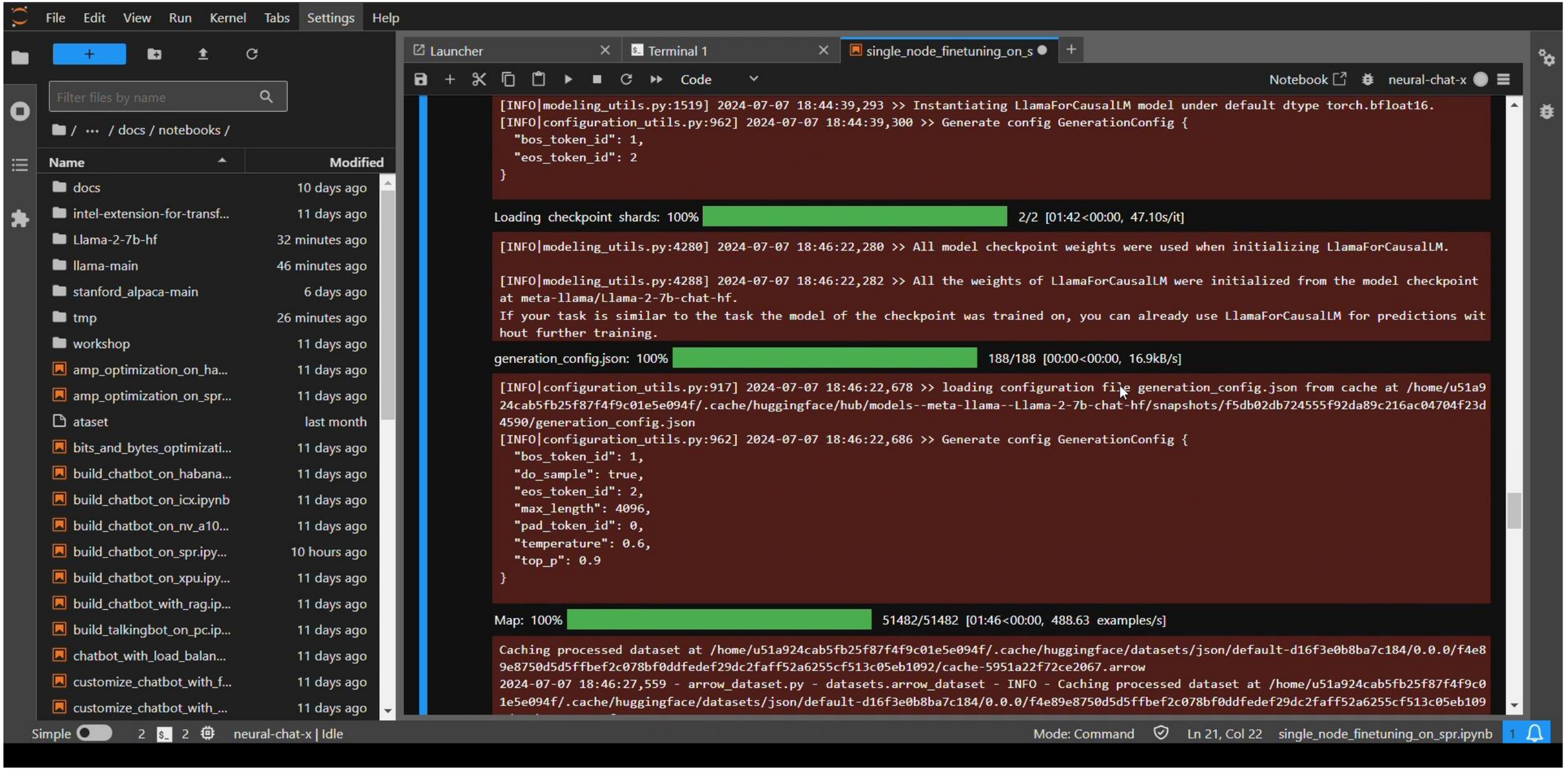

File Edit View Run Kernel Tabs Settings Help
Launcher
Terminal 1
single_node_finetuning_on_s
Code
Notebook
neural-chat-x
Filter files by name
/ ... / docs / notebooks /
Name
Modified
docs 10 days ago
intel-extension-for-transf... 11 days ago
Llama-2-7b-hf 32 minutes ago
llama-main 46 minutes ago
stanford_alpaca-main 6 days ago
tmp 25 minutes ago
workshop 11 days ago
amp_optimization_on_ha... 11 days ago
amp_optimization_on_spr... 11 days ago
ataset last month
bits_and_bytes_optimizati... 11 days ago
build_chatbot_on_habana... 11 days ago
build_chatbot_on_icx.ipynb 11 days ago
build_chatbot_on_nv_a10... 11 days ago
build_chatbot_on_spr.ipy... 10 hours ago
build_chatbot_on_xpu.ipy... 11 days ago
build_chatbot_with_rag.ip... 11 days ago
build_talkingbot_on_pc.ip... 11 days ago
chatbot_with_load_balan... 11 days ago
customize_chatbot_with_f... 11 days ago
customize_chatbot_with_... 11 days ago
[INFO|configuration_utils.py:733] 2024-07-07 18:42:42,960 >> loading configuration file config.json from cache at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/hub/models--meta-llama--Llama-2-7b-chat-hf/snapshots/f5db02db724555f92da89c216ac04704f23d4590/config.json
[INFO|configuration_utils.py:796] 2024-07-07 18:42:42,967 >> Model config LlamaConfig {
"_name_or_path": "meta-llama/Llama-2-7b-chat-hf",
"architectures": [
"LlamaForCausalLM"
],
"attention_bias": false,
"attention_dropout": 0.0,
"bos_token_id": 1,
"eos_token_id": 2,
"hidden_act": "silu",
"hidden_size": 4096,
"initializer_range": 0.02,
"intermediate_size": 11008,
"max_position_embeddings": 4096,
"mlp_bias": false,
"model_type": "llama",
"num_attention_heads": 32,
"num_hidden_layers": 32,
"num_key_value_heads": 32,
"pretraining_tp": 1,
"rms_norm_eps": 1e-05,
"rope_scaling": null,
"rope_theta": 10000.0,
"tie_word_embeddings": false,
"torch_dtype": "float16",
"transformers_version": "4.41.2",
"use_cache": true,
"vocab_size": 32000
}
tokenizer_config.json: 100% 1.62k/1.62k [00:00<00:00, 166kB/s]
Simple
neural-chat-x | Idle
Saving completed
Mode: Command
Ln 21, Col 22
single_node_finetuning_on_spr.ipynb

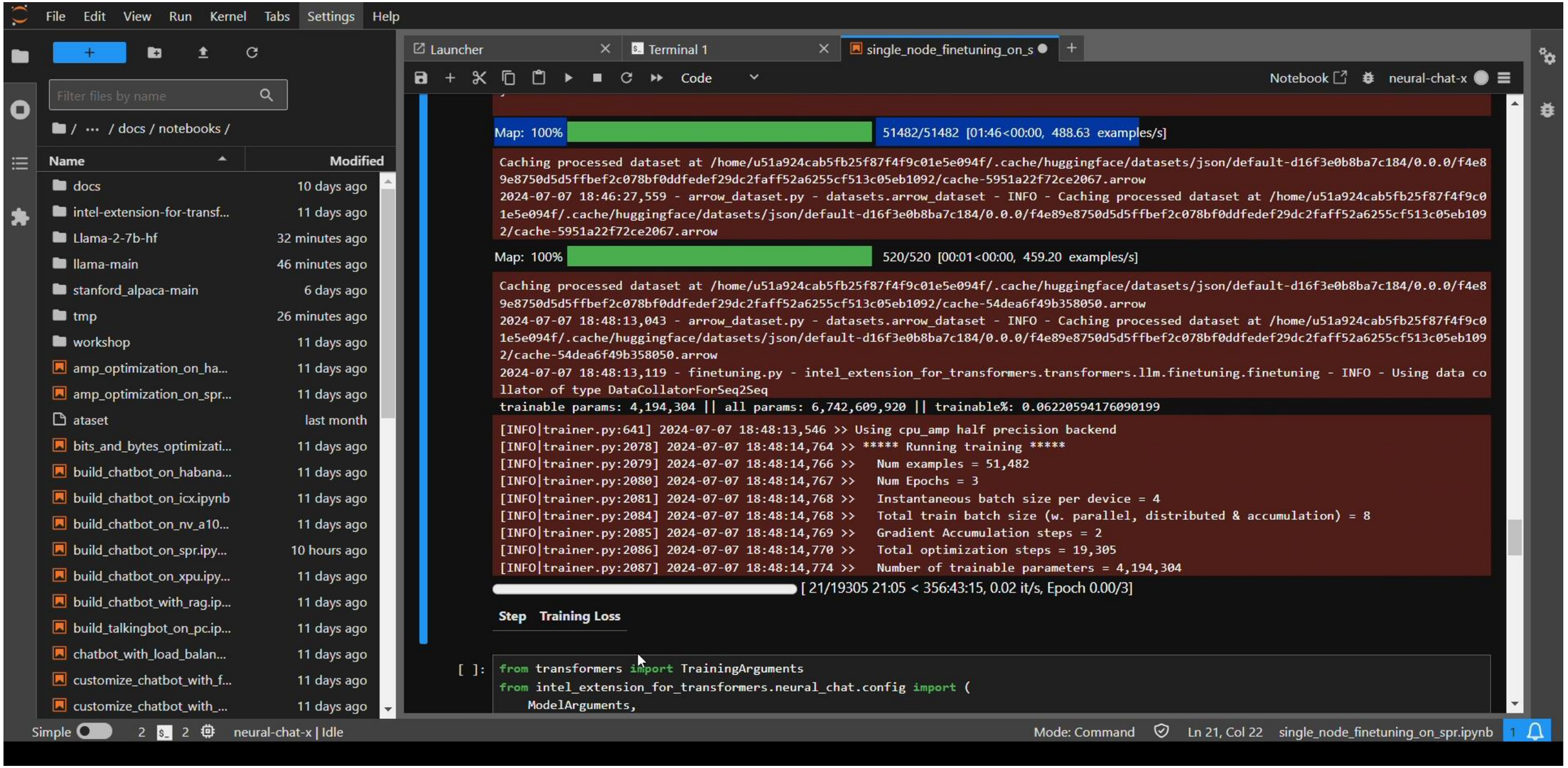

File Edit View Run Kernel Tabs Settings Help
Launcher
Terminal 1
single_node_finetuning_on_s
Code
Notebook
neural-chat-x
Filter files by name
/ ... / docs / notebooks /
Name
Modified
docs 10 days ago
intel-extension-for-transf... 11 days ago
Llama-2-7b-hf 32 minutes ago
llama-main 46 minutes ago
stanford_alpaca-main 6 days ago
tmp 25 minutes ago
workshop 11 days ago
amp_optimization_on_ha... 11 days ago
amp_optimization_on_spr... 11 days ago
ataset last month
bits_and_bytes_optimizati... 11 days ago
build_chatbot_on_habana... 11 days ago
build_chatbot_on_icx.ipynb 11 days ago
build_chatbot_on_nv_a10... 11 days ago
build_chatbot_on_spr.ipy... 10 hours ago
build_chatbot_on_xpu.ipy... 11 days ago
build_chatbot_with_rag.ip... 11 days ago
build_talkingbot_on_pc.ip... 11 days ago
chatbot_with_load_balan... 11 days ago
customize_chatbot_with_f... 11 days ago
customize_chatbot_with_... 11 days ago
tokenizer.model: 100% 500k/500k [00:00<00:00, 26.9MB/s]
special_tokens_map.json: 100% 414/414 [00:00<00:00, 41.5kB/s]
tokenizer.json: 100% 1.84M/1.84M [00:00<00:00, 11.3MB/s]
[INFO|tokenization_utils_base.py:2108] 2024-07-07 18:42:44,144 >> loading file tokenizer.model from cache at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/hub/models--meta-llama--Llama-2-7b-chat-hf/snapshots/f5db02db724555f92da89c216ac04704f23d4590/tokenizer.model
[INFO|tokenization_utils_base.py:2108] 2024-07-07 18:42:44,146 >> loading file added_tokens.json from cache at None
[INFO|tokenization_utils_base.py:2108] 2024-07-07 18:42:44,147 >> loading file special_tokens_map.json from cache at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/hub/models--meta-llama--Llama-2-7b-chat-hf/snapshots/f5db02db724555f92da89c216ac04704f23d4590/special_tokens_map.json
[INFO|tokenization_utils_base.py:2108] 2024-07-07 18:42:44,147 >> loading file tokenizer_config.json from cache at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/hub/models--meta-llama--Llama-2-7b-chat-hf/snapshots/f5db02db724555f92da89c216ac04704f23d4590/tokenizer_config.json
[INFO|tokenization_utils_base.py:2108] 2024-07-07 18:42:44,148 >> loading file tokenizer.json from cache at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/hub/models--meta-llama--Llama-2-7b-chat-hf/snapshots/f5db02db724555f92da89c216ac04704f23d4590/tokenizer.json
Using custom data configuration default-d16f3e0b8ba7c184
2024-07-07 18:42:44,624 - builder.py - datasets.builder - INFO - Using custom data configuration default-d16f3e0b8ba7c184
Loading Dataset Infos from /home/u51a924cab5fb25f87f4f9c01e5e094f/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/datasets/src/datasets/packaged_modules/json
2024-07-07 18:42:44,631 - info.py - datasets.info - INFO - Loading Dataset Infos from /home/u51a924cab5fb25f87f4f9c01e5e094f/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/datasets/src/datasets/packaged_modules/json
Generating dataset json (/home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/json/default-d16f3e0b8ba7c184/0.0.0/f4e89e8750d5d5ffbef2c078bf0ddfedef29dc2faff52a6255cf513c05eb1092)
2024-07-07 18:42:44,667 - builder.py - datasets.builder - INFO - Generating dataset json (/home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/json/default-d16f3e0b8ba7c184/0.0.0/f4e89e8750d5d5ffbef2c078bf0ddfedef29dc2faff52a6255cf513c05eb1092)
Downloading and preparing dataset json/default to /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/json/default-d16f3e0b8ba7c184/0.0.0/f4e89e8750d5d5ffbef2c078bf0ddfedef29dc2faff52a6255cf513c05eb1092...
2024-07-07 18:42:44,671 - builder.py - datasets.builder - INFO - Downloading and preparing dataset json/default to /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/json/default-d16f3e0b8ba7c184/0.0.0/f4e89e8750d5d5ffbef2c078bf0ddfedef29dc2faff52a6255cf51
Simple 2 2 neural-chat-x | Idle
Saving completed
Mode: Command Ln 21, Col 22 single_node_finetuning_on_spr.ipynb 1


File Edit View Run Kernel Tabs Settings Help
Launcher
Terminal 1
single_node_finetuning_on_s
Code
Notebook
neural-chat-x
Filter files by name
/ ... / docs / notebooks /
Name
Modified
docs 10 days ago
intel-extension-for-transf... 11 days ago
Llama-2-7b-hf 32 minutes ago
llama-main 46 minutes ago
stanford_alpaca-main 6 days ago
tmp 25 minutes ago
workshop 11 days ago
amp_optimization_on_ha... 11 days ago
amp_optimization_on_spr... 11 days ago
ataset last month
bits_and_bytes_optimizati... 11 days ago
build_chatbot_on_habana... 11 days ago
build_chatbot_on_icx.ipynb 11 days ago
build_chatbot_on_nv_a10... 11 days ago
build_chatbot_on_spr.ipy... 10 hours ago
build_chatbot_on_xpu.ipy... 11 days ago
build_chatbot_with_rag.ip... 11 days ago
build_talkingbot_on_pc.ip... 11 days ago
chatbot_with_load_balan... 11 days ago
customize_chatbot_with_f... 11 days ago
customize_chatbot_with_... 11 days ago
[INFO|tokenization_utils_base.py:2108] 2024-07-07 18:42:44,144 >> loading file tokenizer.model from cache at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/hub/models--meta-llama--Llama-2-7b-chat-hf/snapshots/f5db02db724555f92da89c216ac04704f23d4590/tokenizer.model
[INFO|tokenization_utils_base.py:2108] 2024-07-07 18:42:44,146 >> loading file added_tokens.json from cache at None
[INFO|tokenization_utils_base.py:2108] 2024-07-07 18:42:44,147 >> loading file special_tokens_map.json from cache at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/hub/models--meta-llama--Llama-2-7b-chat-hf/snapshots/f5db02db724555f92da89c216ac04704f23d4590/special_tokens_map.json
[INFO|tokenization_utils_base.py:2108] 2024-07-07 18:42:44,147 >> loading file tokenizer_config.json from cache at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/hub/models--meta-llama--Llama-2-7b-chat-hf/snapshots/f5db02db724555f92da89c216ac04704f23d4590/tokenizer_config.json
[INFO|tokenization_utils_base.py:2108] 2024-07-07 18:42:44,148 >> loading file tokenizer.json from cache at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/hub/models--meta-llama--Llama-2-7b-chat-hf/snapshots/f5db02db724555f92da89c216ac04704f23d4590/tokenizer.json
Using custom data configuration default-d16f3e0b8ba7c184
2024-07-07 18:42:44,624 - builder.py - datasets.builder - INFO - Using custom data configuration default-d16f3e0b8ba7c184
Loading Dataset Infos from /home/u51a924cab5fb25f87f4f9c01e5e094f/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/datasets/src/datasets/packaged_modules/json
2024-07-07 18:42:44,631 - info.py - datasets.info - INFO - Loading Dataset Infos from /home/u51a924cab5fb25f87f4f9c01e5e094f/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/datasets/src/datasets/packaged_modules/json
Generating dataset json (/home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/json/default-d16f3e0b8ba7c184/0.0.0/f4e89e8750d5d5ffbef2c078bf0ddfedef29dc2faff52a6255cf513c05eb1092)
2024-07-07 18:42:44,667 - builder.py - datasets.builder - INFO - Generating dataset json (/home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/json/default-d16f3e0b8ba7c184/0.0.0/f4e89e8750d5d5ffbef2c078bf0ddfedef29dc2faff52a6255cf513c05eb1092)
Downloading and preparing dataset json/default to /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/json/default-d16f3e0b8ba7c184/0.0.0/f4e89e8750d5d5ffbef2c078bf0ddfedef29dc2faff52a6255cf513c05eb1092...
2024-07-07 18:42:44,671 - builder.py - datasets.builder - INFO - Downloading and preparing dataset json/default to /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/json/default-d16f3e0b8ba7c184/0.0.0/f4e89e8750d5d5ffbef2c078bf0ddfedef29dc2faff52a6255cf513c05eb1092...
Downloading took 0.0 min
2024-07-07 18:42:44,683 - download_manager.py - datasets.download.download_manager - INFO - Downloading took 0.0 min
Checksum Computation took 0.0 min
2024-07-07 18:42:44,701 - download_manager.py - datasets.download.download_manager - INFO - Checksum Computation took 0.0 min
Generating train split
2024-07-07 18:42:44,721 - builder.py - datasets.builder - INFO - Generating train split
Simple
2 2 neural-chat-x | Idle
Saving completed
Mode: Command
Ln 21, Col 22
single_node_finetuning_on_spr.ipynb
1

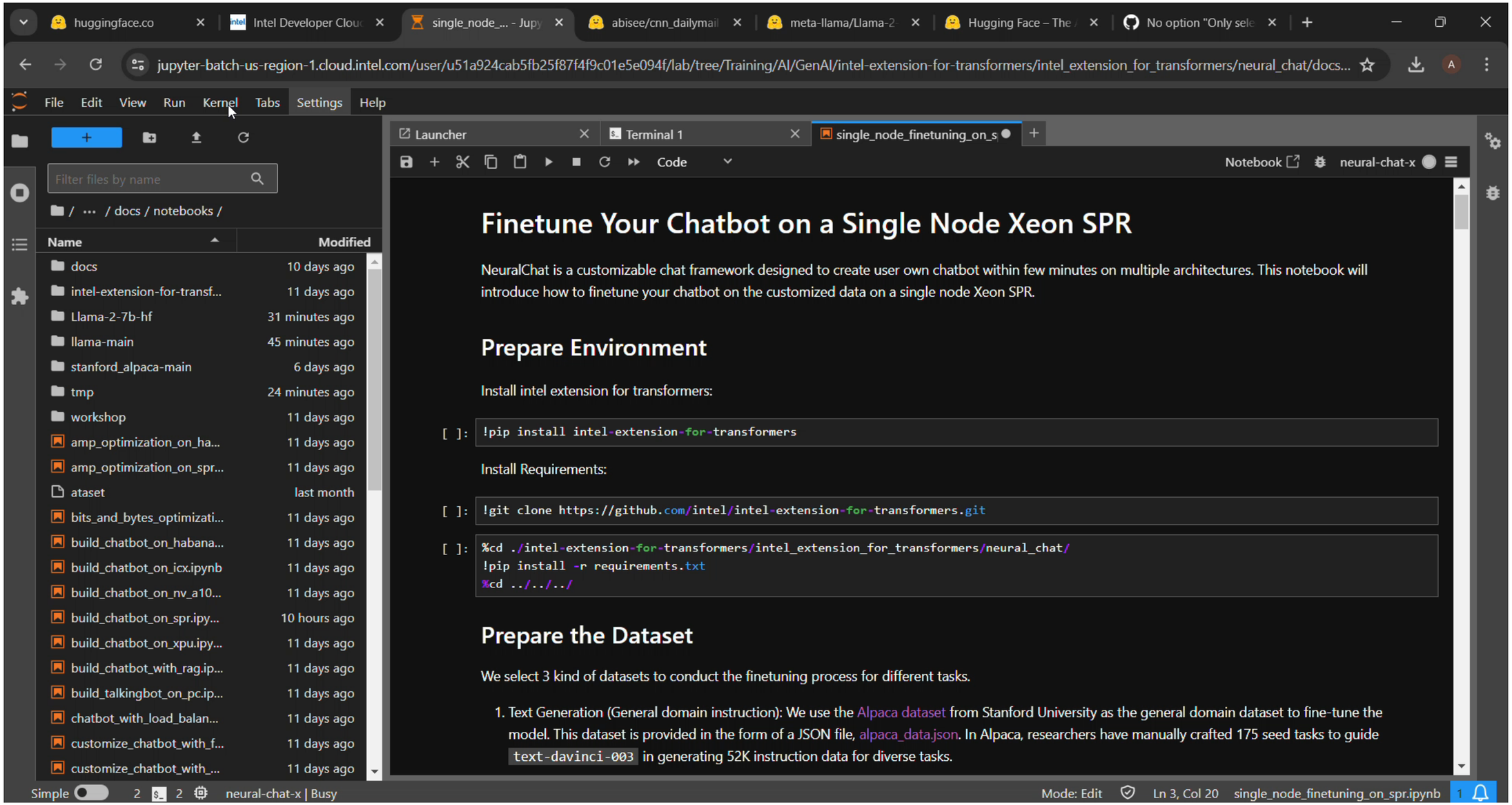

File Edit View Run Kernel Tabs Settings Help
Launcher
Terminal 1
single_node_finetuning_on_s
Code
Notebook
neural-chat-x
Filter files by name
/ ... / docs / notebooks /
Name
Modified
docs 10 days ago
intel-extension-for-transf... 11 days ago
Llama-2-7b-hf 32 minutes ago
llama-main 46 minutes ago
stanford_alpaca-main 6 days ago
tmp 25 minutes ago
workshop 11 days ago
amp_optimization_on_ha... 11 days ago
amp_optimization_on_spr... 11 days ago
ataset last month
bits_and_bytes_optimizati... 11 days ago
build_chatbot_on_habana... 11 days ago
build_chatbot_on_icx.ipynb 11 days ago
build_chatbot_on_nv_a10... 11 days ago
build_chatbot_on_spr.ipy... 10 hours ago
build_chatbot_on_xpu.ipy... 11 days ago
build_chatbot_with_rag.ip... 11 days ago
build_talkingbot_on_pc.ip... 11 days ago
chatbot_with_load_balan... 11 days ago
customize_chatbot_with_f... 11 days ago
customize_chatbot_with_... 11 days ago
2024-07-07 18:42:44,721 - builder.py - datasets.builder - INFO - Generating train split
Generating train split: 52002/0 [00:00<00:00, 73761.36 examples/s]
Unable to verify splits sizes.
2024-07-07 18:42:45,609 - info_utils.py - datasets.utils.info_utils - INFO - Unable to verify splits sizes.
Dataset json downloaded and prepared to /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/json/default-d16f3e0b8ba7c184/0.0.0/f4e89e8750d5d5ffbef2c078bf0ddfedef29dc2faff52a6255cf513c05eb1092. Subsequent calls will reuse this data.
2024-07-07 18:42:45,622 - builder.py - datasets.builder - INFO - Dataset json downloaded and prepared to /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/json/default-d16f3e0b8ba7c184/0.0.0/f4e89e8750d5d5ffbef2c078bf0ddfedef29dc2faff52a6255cf513c05eb1092. Subsequent calls will reuse this data.
Using custom data configuration default-d16f3e0b8ba7c184
2024-07-07 18:42:46,072 - builder.py - datasets.builder - INFO - Using custom data configuration default-d16f3e0b8ba7c184
Loading Dataset Infos from /home/u51a924cab5fb25f87f4f9c01e5e094f/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/datasets/src/datasets/packaged_modules/json
2024-07-07 18:42:46,074 - info.py - datasets.info - INFO - Loading Dataset Infos from /home/u51a924cab5fb25f87f4f9c01e5e094f/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/datasets/src/datasets/packaged_modules/json
Overwrite dataset info from restored data version if exists.
2024-07-07 18:42:46,090 - builder.py - datasets.builder - INFO - Overwrite dataset info from restored data version if exists.
Loading Dataset info from /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/json/default-d16f3e0b8ba7c184/0.0.0/f4e89e8750d5d5ffbef2c078bf0ddfedef29dc2faff52a6255cf513c05eb1092
2024-07-07 18:42:46,096 - info.py - datasets.info - INFO - Loading Dataset info from /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/json/default-d16f3e0b8ba7c184/0.0.0/f4e89e8750d5d5ffbef2c078bf0ddfedef29dc2faff52a6255cf513c05eb1092
Found cached dataset json (/home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/json/default-d16f3e0b8ba7c184/0.0.0/f4e89e8750d5d5ffbef2c078bf0ddfedef29dc2faff52a6255cf513c05eb1092)
2024-07-07 18:42:46,103 - builder.py - datasets.builder - INFO - Found cached dataset json (/home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/json/default-d16f3e0b8ba7c184/0.0.0/f4e89e8750d5d5ffbef2c078bf0ddfedef29dc2faff52a6255cf513c05eb1092)
Loading Dataset info from /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/json/default-d16f3e0b8ba7c184/0.0.0/f4e89e8750d5d5ffbef2c078bf0ddfedef29dc2faff52a6255cf513c05eb1092
2024-07-07 18:42:46,104 - info.py - datasets.info - INFO - Loading Dataset info from /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/json/default-d16f3e0b8ba7c184/0.0.0/f4e89e8750d5d5ffbef2c078bf0ddfedef29dc2faff52a6255cf513c05eb1092
Using custom data configuration default-d16f3e0b8ba7c184
2024-07-07 18:42:46,358 - builder.py - datasets.builder - INFO - Using custom data configuration default-d16f3e0b8ba7c184
Loading Dataset Infos from /home/u51a924cab5fb25f87f4f9c01e5e094f/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for
Simple 2 2 neural-chat-x | Idle
Mode: Command Ln 21, Col 22 single_node_finetuning_on_spr.ipynb 1


File Edit View Run Kernel Tabs Settings Help
Launcher
Terminal 1
single_node_finetuning_on_s
Code
Notebook
neural-chat-x
Filter files by name
/ ... / docs / notebooks /
Name
Modified
docs 10 days ago
intel-extension-for-transf... 11 days ago
Llama-2-7b-hf 32 minutes ago
llama-main 46 minutes ago
stanford_alpaca-main 6 days ago
tmp 25 minutes ago
workshop 11 days ago
amp_optimization_on_ha... 11 days ago
amp_optimization_on_spr... 11 days ago
ataset last month
bits_and_bytes_optimizati... 11 days ago
build_chatbot_on_habana... 11 days ago
build_chatbot_on_icx.ipynb 11 days ago
build_chatbot_on_nv_a10... 11 days ago
build_chatbot_on_spr.ipy... 10 hours ago
build_chatbot_on_xpu.ipy... 11 days ago
build_chatbot_with_rag.ip... 11 days ago
build_talkingbot_on_pc.ip... 11 days ago
chatbot_with_load_balan... 11 days ago
customize_chatbot_with_f... 11 days ago
customize_chatbot_with_... 11 days ago
Using custom data configuration default-d16f3e0b8ba7c184
2024-07-07 18:42:46,358 - builder.py - datasets.builder - INFO - Using custom data configuration default-d16f3e0b8ba7c184
Loading Dataset Infos from /home/u51a924cab5fb25f87f4f9c01e5e094f/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/datasets/src/datasets/packaged_modules/json
2024-07-07 18:42:46,365 - info.py - datasets.info - INFO - Loading Dataset Infos from /home/u51a924cab5fb25f87f4f9c01e5e094f/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/datasets/src/datasets/packaged_modules/json
Overwrite dataset info from restored data version if exists.
2024-07-07 18:42:46,372 - builder.py - datasets.builder - INFO - Overwrite dataset info from restored data version if exists.
Loading Dataset info from /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/json/default-d16f3e0b8ba7c184/0.0.0/f4e89e8750d5d5ffbef2c078bf0ddfedef29dc2faff52a6255cf513c05eb1092
2024-07-07 18:42:46,376 - info.py - datasets.info - INFO - Loading Dataset info from /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/json/default-d16f3e0b8ba7c184/0.0.0/f4e89e8750d5d5ffbef2c078bf0ddfedef29dc2faff52a6255cf513c05eb1092
Found cached dataset json (/home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/json/default-d16f3e0b8ba7c184/0.0.0/f4e89e8750d5d5ffbef2c078bf0ddfedef29dc2faff52a6255cf513c05eb1092)
2024-07-07 18:42:46,386 - builder.py - datasets.builder - INFO - Found cached dataset json (/home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/json/default-d16f3e0b8ba7c184/0.0.0/f4e89e8750d5d5ffbef2c078bf0ddfedef29dc2faff52a6255cf513c05eb1092)
Loading Dataset info from /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/json/default-d16f3e0b8ba7c184/0.0.0/f4e89e8750d5d5ffbef2c078bf0ddfedef29dc2faff52a6255cf513c05eb1092
2024-07-07 18:42:46,390 - info.py - datasets.info - INFO - Loading Dataset info from /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/json/default-d16f3e0b8ba7c184/0.0.0/f4e89e8750d5d5ffbef2c078bf0ddfedef29dc2faff52a6255cf513c05eb1092
model.safetensors.index.json: 100% 26.8k/26.8k [00:00<00:00, 2.16MB/s]
[INFO|modeling_utils.py:3474] 2024-07-07 18:42:47,119 >> loading weights file model.safetensors from cache at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/hub/models--meta-llama--Llama-2-7b-chat-hf/snapshots/f5db02db724555f92da89c216ac04704f23d4590/model.safetensors.index.json
Downloading shards: 100% 2/2 [01:52<00:00, 52.50s/it]
model-00001-of-00002.safetensors: 100% 9.98G/9.98G [00:31<00:00, 333MB/s]
model-00002-of-00002.safetensors: 100% 3.50G/3.50G [00:11<00:00, 311MB/s]
[INFO|modeling_utils.py:1519] 2024-07-07 18:44:39,293 >> Instantiating LlamaForCausalLM model under default dtype torch.bfloat16.
[INFO|configuration_utils.py:962] 2024-07-07 18:44:39,300 >> Generate config GenerationConfig {
"bos_token_id": 1
Simple 2 2 neural-chat-x | Idle
Mode: Command Ln 21, Col 22 single_node_finetuning_on_spr.ipynb 1

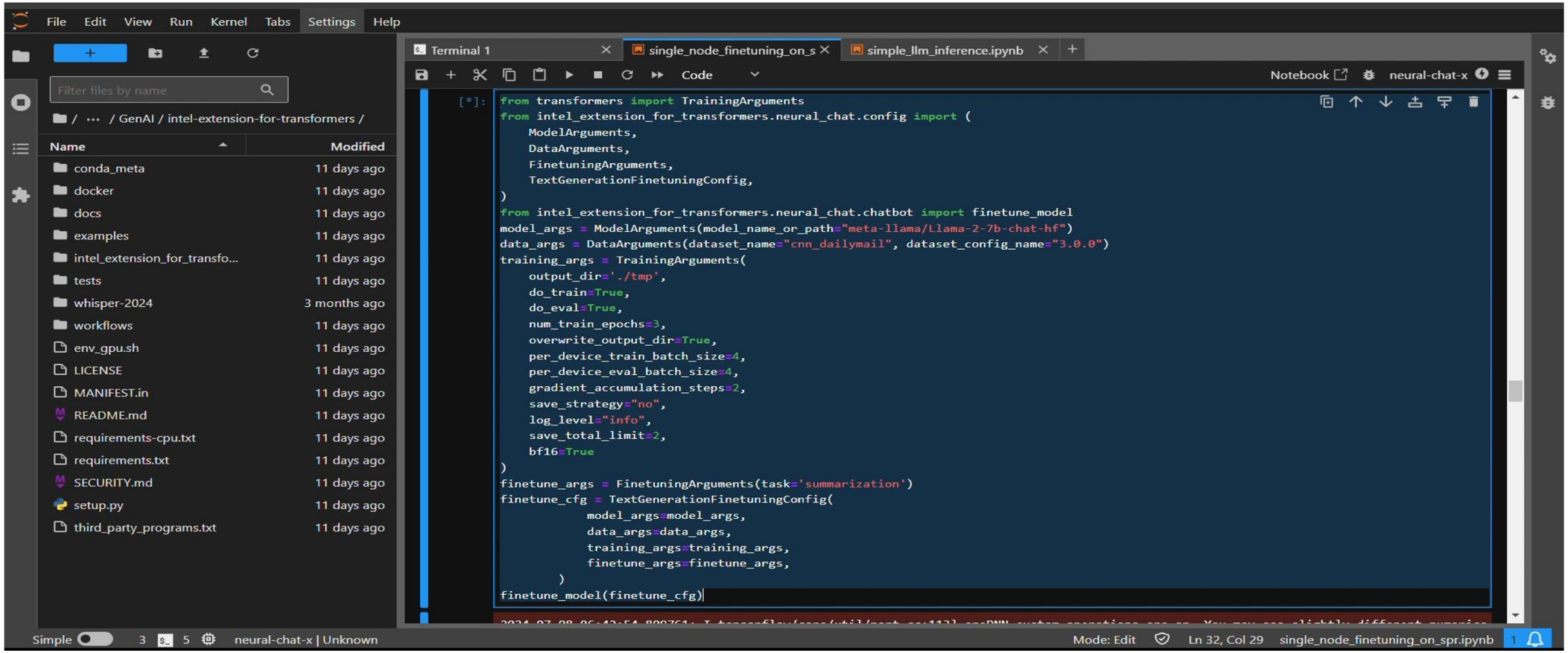

File Edit View Run Kernel Tabs Settings Help
Launcher
Terminal 1
single_node_finetuning_on_s
Code
Notebook
neural-chat-x
Filter files by name
/ ··· / docs / notebooks /
Name
Modified
docs 10 days ago
intel-extension-for-transf... 11 days ago
Llama-2-7b-hf 32 minutes ago
llama-main 46 minutes ago
stanford_alpaca-main 6 days ago
tmp 26 minutes ago
workshop 11 days ago
amp_optimization_on_ha... 11 days ago
amp_optimization_on_spr... 11 days ago
ataset last month
bits_and_bytes_optimizati... 11 days ago
build_chatbot_on_habana... 11 days ago
build_chatbot_on_icx.ipynb 11 days ago
build_chatbot_on_nv_a10... 11 days ago
build_chatbot_on_spr.ipy... 10 hours ago
build_chatbot_on_xpu.ipy... 11 days ago
build_chatbot_with_rag.ip... 11 days ago
build_talkingbot_on_pc.ip... 11 days ago
chatbot_with_load_balan... 11 days ago
customize_chatbot_with_f... 11 days ago
customize_chatbot_with_... 11 days ago
[INFO|modeling_utils.py:1519] 2024-07-07 18:44:39,293 >> Instantiating LlamaForCausalLM model under default dtype torch.bfloat16.
[INFO|configuration_utils.py:962] 2024-07-07 18:44:39,300 >> Generate config GenerationConfig {
"bos_token_id": 1,
"eos_token_id": 2
}
Loading checkpoint shards: 100% 2/2 [01:42<00:00, 47.10s/it]
[INFO|modeling_utils.py:4280] 2024-07-07 18:46:22,280 >> All model checkpoint weights were used when initializing LlamaForCausalLM.
[INFO|modeling_utils.py:4288] 2024-07-07 18:46:22,282 >> All the weights of LlamaForCausalLM were initialized from the model checkpoint at meta-llama/Llama-2-7b-chat-hf.
If your task is similar to the task the model of the checkpoint was trained on, you can already use LlamaForCausalLM for predictions without further training.
generation_config.json: 100% 188/188 [00:00<00:00, 16.9kB/s]
[INFO|configuration_utils.py:917] 2024-07-07 18:46:22,678 >> loading configuration file generation_config.json from cache at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/hub/models--meta-llama--Llama-2-7b-chat-hf/snapshots/f5db02db724555f92da89c216ac04704f23d4590/generation_config.json
[INFO|configuration_utils.py:962] 2024-07-07 18:46:22,686 >> Generate config GenerationConfig {
"bos_token_id": 1,
"do_sample": true,
"eos_token_id": 2,
"max_length": 4096,
"pad_token_id": 0,
"temperature": 0.6,
"top_p": 0.9
}
Map: 100% 51482/51482 [01:46<00:00, 488.63 examples/s]
Caching processed dataset at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/json/default-d16f3e0b8ba7c184/0.0.0/f4e89e8750d5d5ffbef2c078bf0ddfedef29dc2faff52a6255cf513c05eb1092/cache-5951a22f72ce2067.arrow
2024-07-07 18:46:27,559 - arrow_dataset.py - datasets.arrow_dataset - INFO - Caching processed dataset at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/json/default-d16f3e0b8ba7c184/0.0.0/f4e89e8750d5d5ffbef2c078bf0ddfedef29dc2faff52a6255cf513c05eb109
Simple 2 2 neural-chat-x | Idle
Mode: Command Ln 21, Col 22 single_node_finetuning_on_spr.ipynb 1

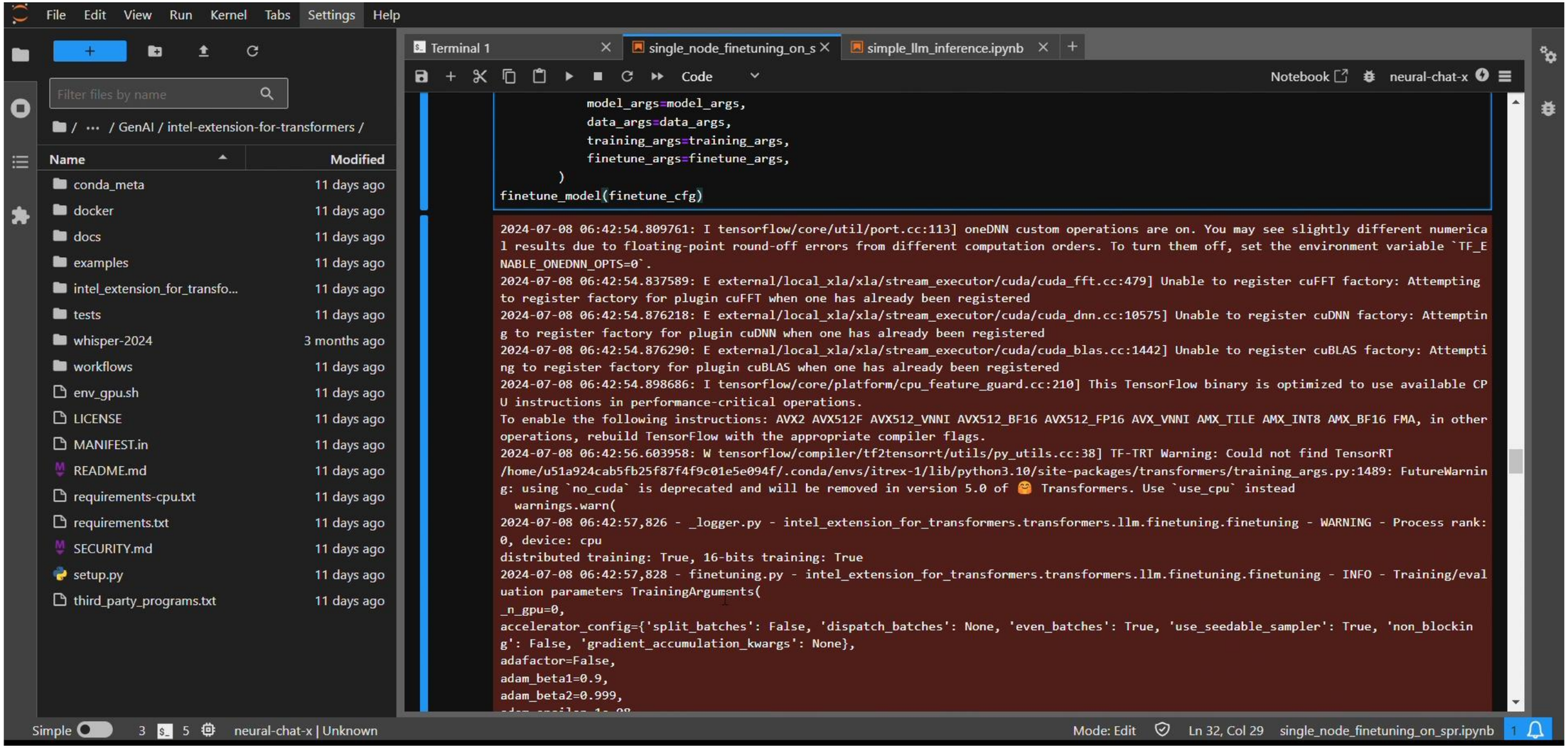

File Edit View Run Kernel Tabs Settings Help
Launcher
Terminal 1
single_node_finetuning_on_s
Code
Notebook
neural-chat-x
Filter files by name
/ ··· / docs / notebooks /
Name
Modified
docs 10 days ago
intel-extension-for-transf... 11 days ago
Llama-2-7b-hf 32 minutes ago
llama-main 46 minutes ago
stanford_alpaca-main 6 days ago
tmp 26 minutes ago
workshop 11 days ago
amp_optimization_on_ha... 11 days ago
amp_optimization_on_spr... 11 days ago
ataset last month
bits_and_bytes_optimizati... 11 days ago
build_chatbot_on_habana... 11 days ago
build_chatbot_on_icx.ipynb 11 days ago
build_chatbot_on_nv_a10... 11 days ago
build_chatbot_on_spr.ipy... 10 hours ago
build_chatbot_on_xpu.ipy... 11 days ago
build_chatbot_with_rag.ip... 11 days ago
build_talkingbot_on_pc.ip... 11 days ago
chatbot_with_load_balan... 11 days ago
customize_chatbot_with_f... 11 days ago
customize_chatbot_with_... 11 days ago
Map: 100% 51482/51482 [01:46<00:00, 488.63 examples/s]
Caching processed dataset at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/json/default-d16f3e0b8ba7c184/0.0.0/f4e89e8750d5d5ffbef2c078bf0ddfedef29dc2faff52a6255cf513c05eb1092/cache-5951a22f72ce2067.arrow
2024-07-07 18:46:27,559 - arrow_dataset.py - datasets.arrow_dataset - INFO - Caching processed dataset at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/json/default-d16f3e0b8ba7c184/0.0.0/f4e89e8750d5d5ffbef2c078bf0ddfedef29dc2faff52a6255cf513c05eb1092/cache-5951a22f72ce2067.arrow
Map: 100% 520/520 [00:01<00:00, 459.20 examples/s]
Caching processed dataset at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/json/default-d16f3e0b8ba7c184/0.0.0/f4e89e8750d5d5ffbef2c078bf0ddfedef29dc2faff52a6255cf513c05eb1092/cache-54dea6f49b358050.arrow
2024-07-07 18:48:13,043 - arrow_dataset.py - datasets.arrow_dataset - INFO - Caching processed dataset at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/json/default-d16f3e0b8ba7c184/0.0.0/f4e89e8750d5d5ffbef2c078bf0ddfedef29dc2faff52a6255cf513c05eb1092/cache-54dea6f49b358050.arrow
2024-07-07 18:48:13,119 - 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] 2024-07-07 18:48:13,546 >> Using cpu_amp half precision backend
[INFO|trainer.py:2078] 2024-07-07 18:48:14,764 >> ***** Running training *****
[INFO|trainer.py:2079] 2024-07-07 18:48:14,766 >> Num examples = 51,482
[INFO|trainer.py:2080] 2024-07-07 18:48:14,767 >> Num Epochs = 3
[INFO|trainer.py:2081] 2024-07-07 18:48:14,768 >> Instantaneous batch size per device = 4
[INFO|trainer.py:2084] 2024-07-07 18:48:14,768 >> Total train batch size (w. parallel, distributed & accumulation) = 8
[INFO|trainer.py:2085] 2024-07-07 18:48:14,769 >> Gradient Accumulation steps = 2
[INFO|trainer.py:2086] 2024-07-07 18:48:14,770 >> Total optimization steps = 19,305
[INFO|trainer.py:2087] 2024-07-07 18:48:14,774 >> Number of trainable parameters = 4,194,304
[21/19305 21:05 < 356:43:15, 0.02 it/s, Epoch 0.00/3]
Step Training Loss
[ ]:
from transformers import TrainingArguments
from intel_extension_for_transformers.neural_chat.config import (
ModelArguments,
Simple 2 2 neural-chat-x | Idle
Mode: Command Ln 21, Col 22 single_node_finetuning_on_spr.ipynb 1

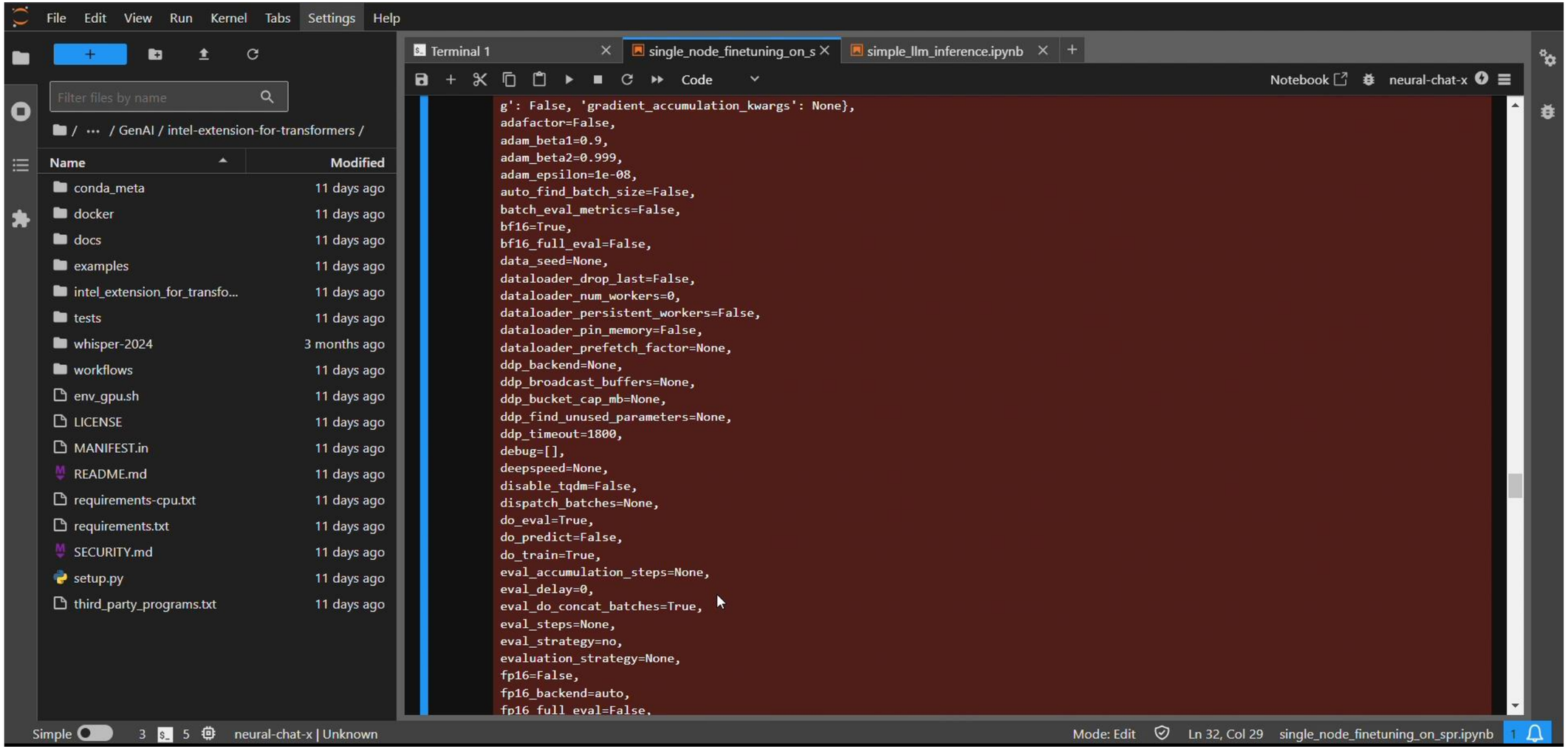
VIDEO

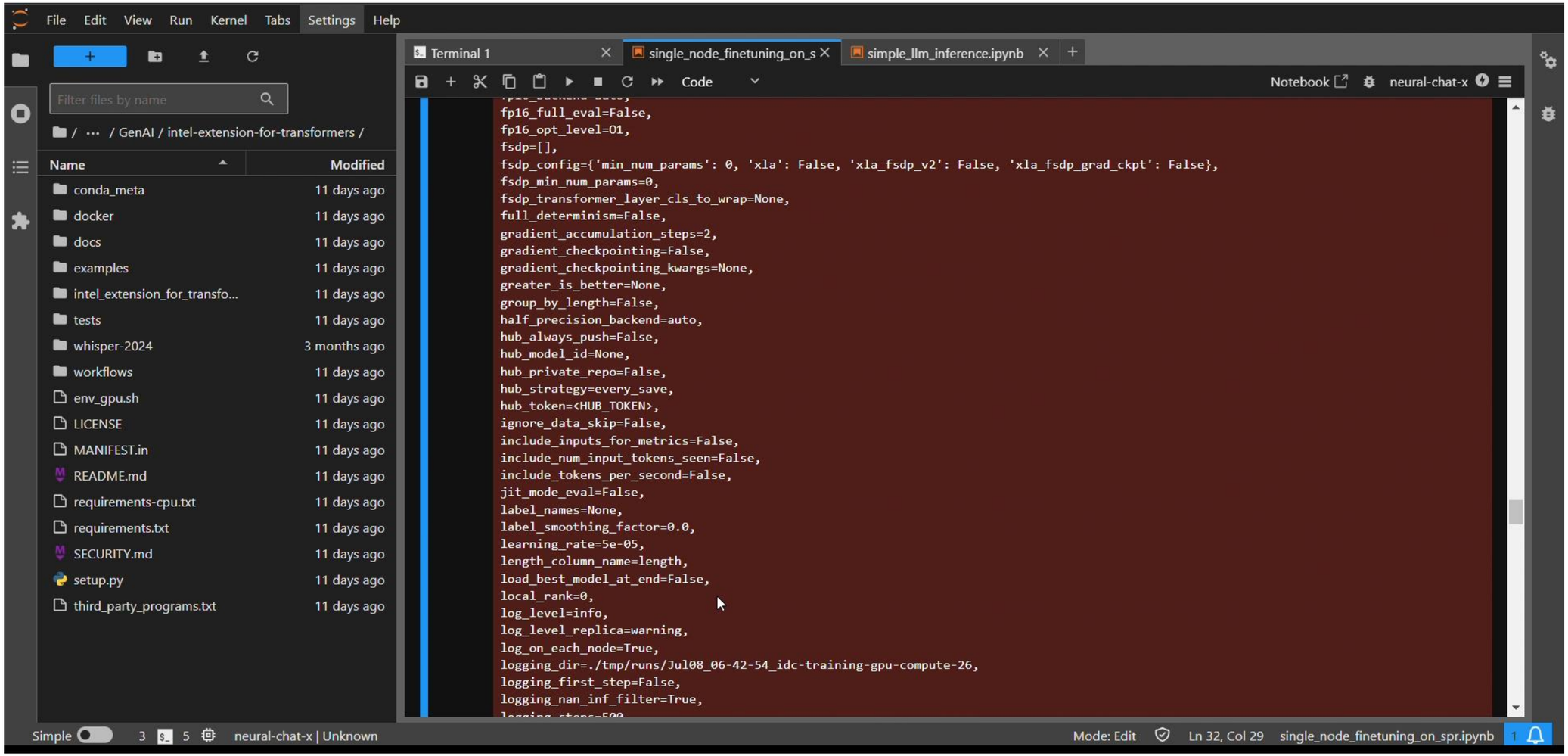

huggingface.co
Intel Developer Clou
single_node_... - Jupy
abisee/cnn_dailymail
meta-llama/Llama-2-
Hugging Face – The
No option "Only sele
jupyter-batch-us-region-1.cloud.intel.com/user/u51a924cab5fb25f87f4f9c01e5e094f/lab/tree/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs...
File Edit View Run Kernel Tabs Settings Help
Filter files by name
/ ... / docs / notebooks /
Name
Modified
docs 10 days ago
intel-extension-for-transf... 11 days ago
Llama-2-7b-hf 31 minutes ago
llama-main 45 minutes ago
stanford_alpaca-main 6 days ago
tmp 24 minutes ago
workshop 11 days ago
amp_optimization_on_ha... 11 days ago
amp_optimization_on_spr... 11 days ago
ataset last month
bits_and_bytes_optimizati... 11 days ago
build_chatbot_on_habana... 11 days ago
build_chatbot_on_icx.ipynb 11 days ago
build_chatbot_on_nv_a10... 11 days ago
build_chatbot_on_spr.ipy... 10 hours ago
build_chatbot_on_xpu.ipy... 11 days ago
build_chatbot_with_rag.ip... 11 days ago
build_talkingbot_on_pc.ip... 11 days ago
chatbot_with_load_balan... 11 days ago
customize_chatbot_with_f... 11 days ago
customize_chatbot_with_... 11 days ago
Launcher
Terminal 1
single_node_finetuning_on_s
Code
Notebook
neural-chat-x
Finetune Your Chatbot on a Single Node Xeon SPR
NeuralChat is a customizable chat framework designed to create user own chatbot within few minutes on multiple architectures. This notebook will introduce how to finetune your chatbot on the customized data on a single node Xeon SPR.
Prepare Environment
Install intel extension for transformers:
[ ]: !pip install intel-extension-for-transformers
Install Requirements:
[ ]: !git clone https://github.com/intel/intel-extension-for-transformers.git
[ ]: %cd ./intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/
!pip install -r requirements.txt
%cd ../../../
Prepare the Dataset
We select 3 kind of datasets to conduct the finetuning process for different tasks.
1. Text Generation (General domain instruction): We use the Alpaca dataset from Stanford University as the general domain dataset to fine-tune the model. This dataset is provided in the form of a JSON file, alpaca_data.json. In Alpaca, researchers have manually crafted 175 seed tasks to guide text-davinci-003 in generating 52K instruction data for diverse tasks.
Simple 2 2 neural-chat-x | Busy
Mode: Edit Ln 3, Col 20 single_node_finetuning_on_spr.ipynb 1

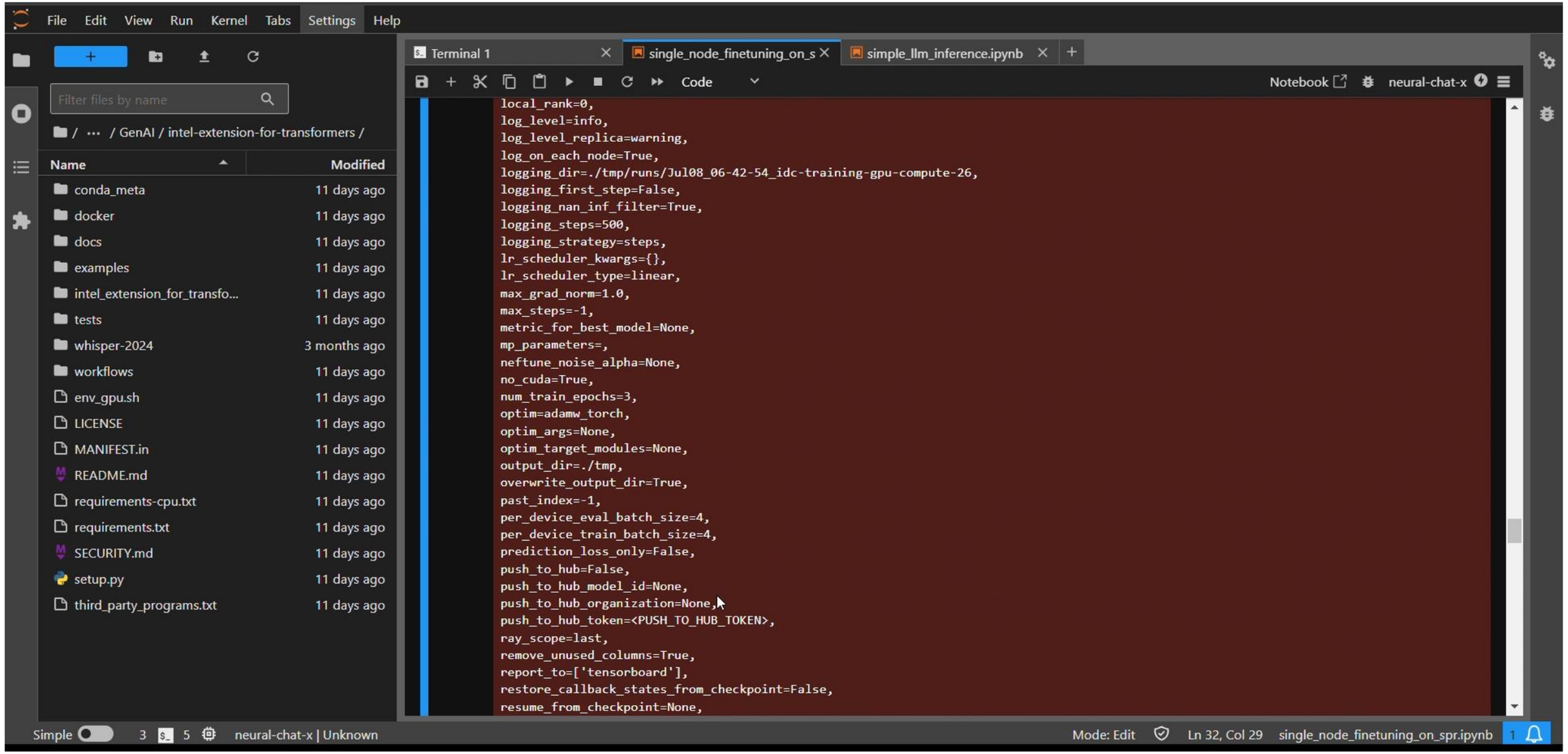
SINGLE NODE FINETUNING SUMMARISATION

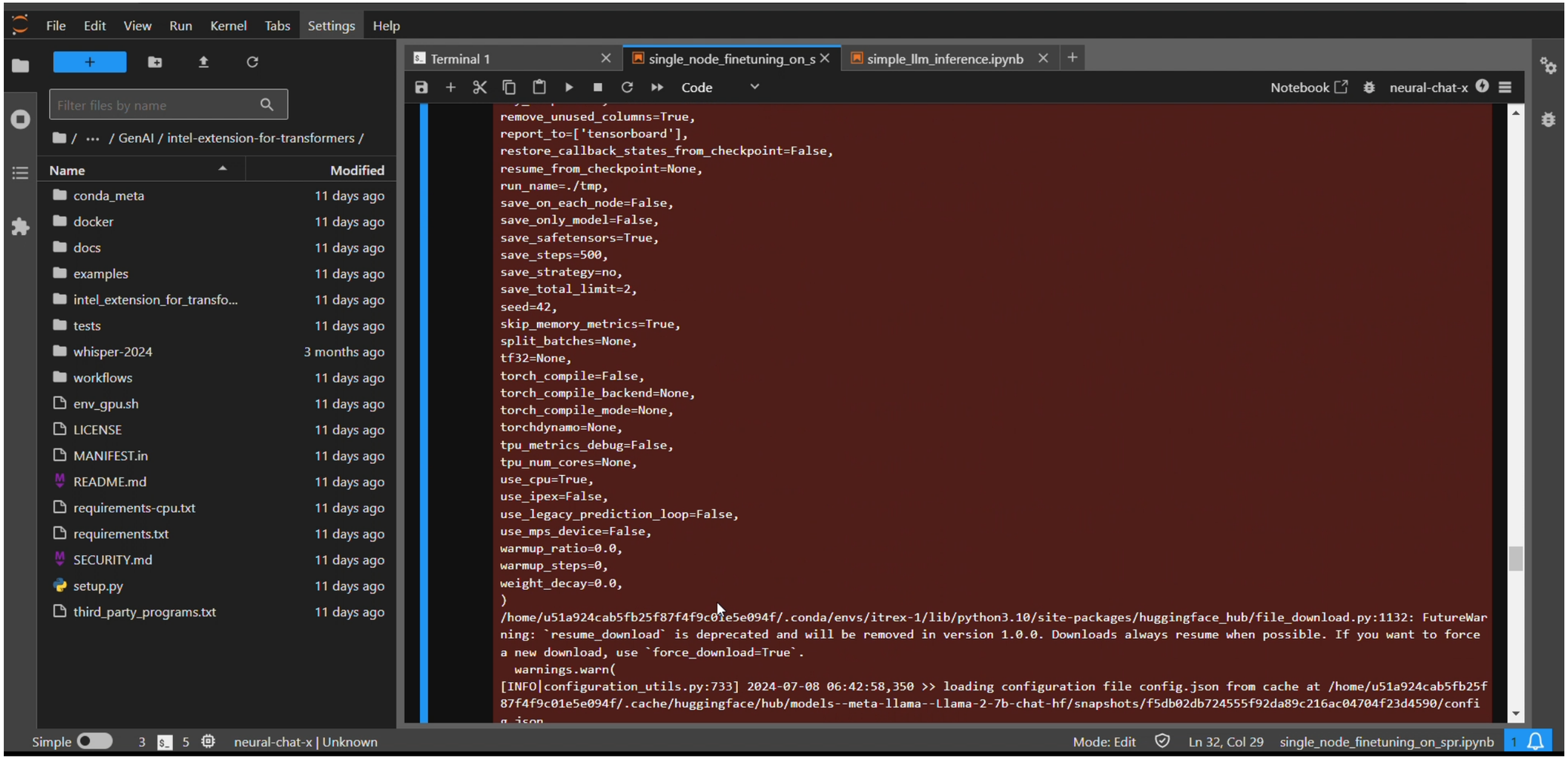


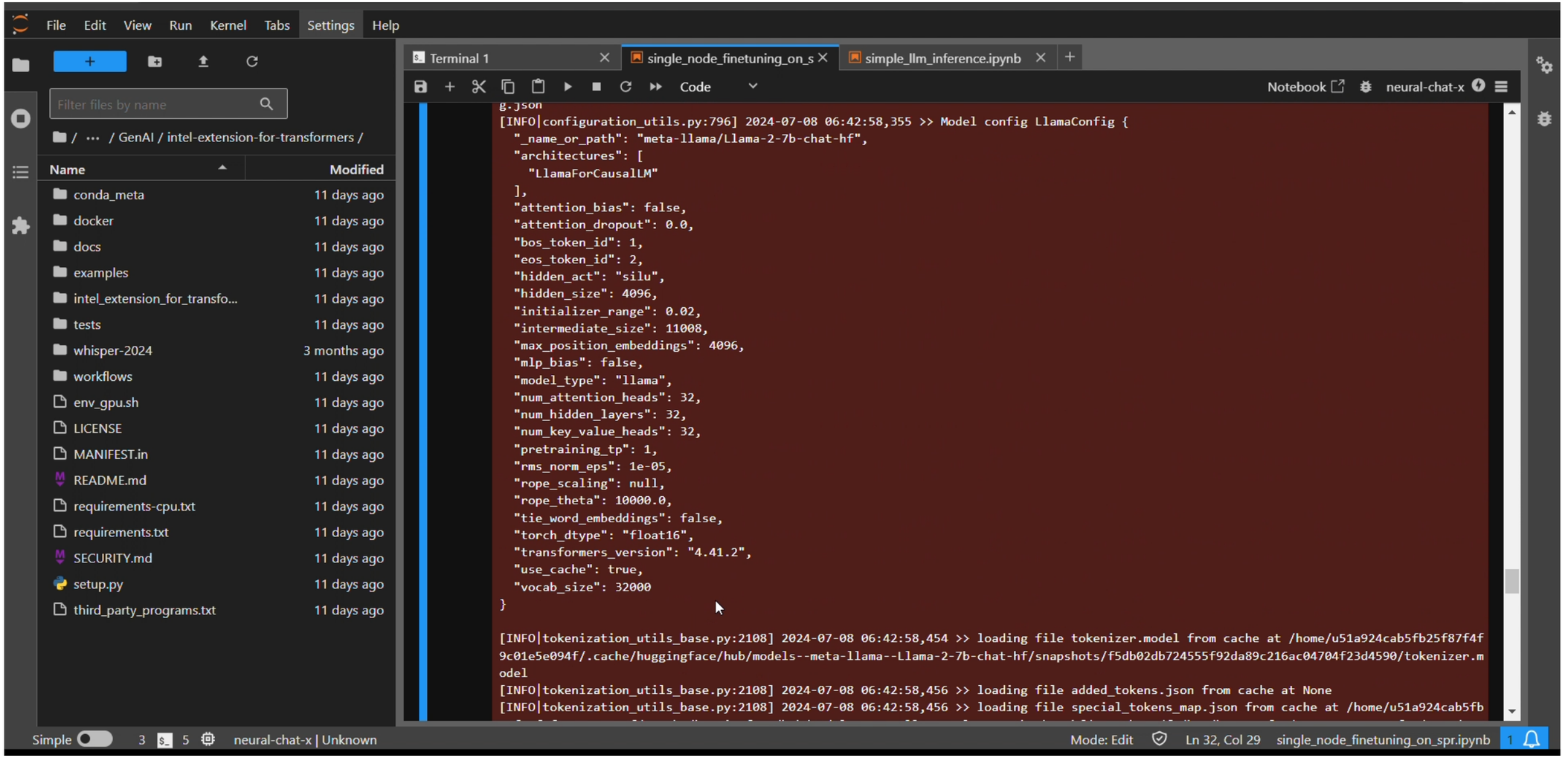

File Edit View Run Kernel Tabs Settings Help
Filter files by name
/ ... / GenAI / intel-extension-for-transformers /
Name
Modified
conda_meta 11 days ago
docker 11 days ago
docs 11 days ago
examples 11 days ago
intel_extension_for_transfo... 11 days ago
tests 11 days ago
whisper-2024 3 months ago
workflows 11 days ago
env_gpu.sh 11 days ago
LICENSE 11 days ago
MANIFEST.in 11 days ago
README.md 11 days ago
requirements-cpu.txt 11 days ago
requirements.txt 11 days ago
SECURITY.md 11 days ago
setup.py 11 days ago
third_party_programs.txt 11 days ago
Terminal 1
single_node_finetuning_on_s
simple_llm_inference.ipynb
Code
Notebook
neural-chat-x
model_args=model_args,
data_args=data_args,
training_args=training_args,
finetune_args=finetune_args,
)
finetune_model(finetune_cfg)
2024-07-08 06:42:54.809761: I tensorflow/core/util/port.cc:113] oneDNN custom operations are on. You may see slightly different numerical results due to floating-point round-off errors from different computation orders. To turn them off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2024-07-08 06:42:54.837589: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:479] Unable to register cuFFT factory: Attempting to register factory for plugin cuFFT when one has already been registered
2024-07-08 06:42:54.876218: E external/local_xla/xla/stream_executor/cuda/cuda_dnn.cc:10575] Unable to register cuDNN factory: Attempting to register factory for plugin cuDNN when one has already been registered
2024-07-08 06:42:54.876290: E external/local_xla/xla/stream_executor/cuda/cuda_blas.cc:1442] Unable to register cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has already been registered
2024-07-08 06:42:54.898686: I tensorflow/core/platform/cpu_feature_guard.cc:210] This TensorFlow binary is optimized to use available CPU instructions in performance-critical operations.
To enable the following instructions: AVX2 AVX512F AVX512_VNNI AVX512_BF16 AVX512_FP16 AVX_VNNI AMX_TILE AMX_INT8 AMX_BF16 FMA, in other operations, rebuild TensorFlow with the appropriate compiler flags.
2024-07-08 06:42:56.603958: W tensorflow/compiler/tf2tensorrt/utils/py_utils.cc:38] TF-TRT Warning: Could not find TensorRT
/home/u51a924cab5fb25f87f4f9c01e5e094f/.conda/envs/itrex-1/lib/python3.10/site-packages/transformers/training_args.py:1489: FutureWarning: using `no_cuda` is deprecated and will be removed in version 5.0 of 🤗 Transformers. Use `use_cpu` instead
warnings.warn(
2024-07-08 06:42:57,826 - _logger.py - intel_extension_for_transformers.transformers.llm.finetuning.finetuning - WARNING - Process rank: 0, device: cpu
distributed training: True, 16-bits training: True
2024-07-08 06:42:57,828 - finetuning.py - intel_extension_for_transformers.transformers.llm.finetuning.finetuning - INFO - Training/evaluation parameters TrainingArguments(
_n_gpu=0,
accelerator_config={'split_batches': False, 'dispatch_batches': None, 'even_batches': True, 'use_seedable_sampler': True, 'non_blocking': False, 'gradient_accumulation_kwargs': None},
adafactor=False,
adam_beta1=0.9,
adam_beta2=0.999,
Simple
neural-chat-x | Unknown
Mode: Edit
Ln 32, Col 29
single_node_finetuning_on_spr.ipynb

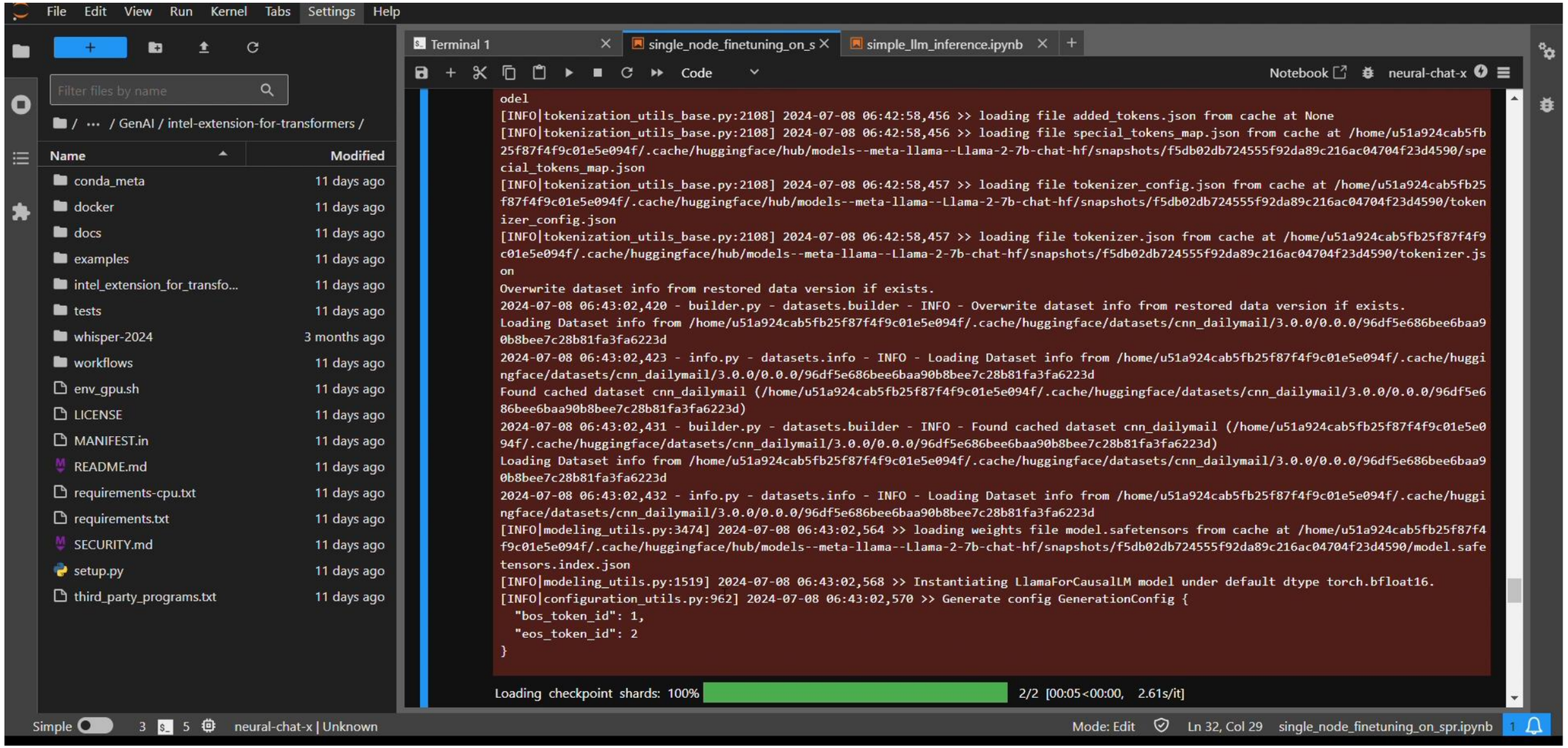

File Edit View Run Kernel Tabs Settings Help
Terminal 1
single_node_finetuning_on_s
simple_llm_inference.ipynb
Code
Notebook
neural-chat-x
Filter files by name
/ ... / GenAI / intel-extension-for-transformers /
Name
Modified
conda_meta 11 days ago
docker 11 days ago
docs 11 days ago
examples 11 days ago
intel_extension_for_transfo... 11 days ago
tests 11 days ago
whisper-2024 3 months ago
workflows 11 days ago
env_gpu.sh 11 days ago
LICENSE 11 days ago
MANIFEST.in 11 days ago
README.md 11 days ago
requirements-cpu.txt 11 days ago
requirements.txt 11 days ago
SECURITY.md 11 days ago
setup.py 11 days ago
third_party_programs.txt 11 days ago
g': False, 'gradient_accumulation_kwargs': None},
adafactor=False,
adam_beta1=0.9,
adam_beta2=0.999,
adam_epsilon=1e-08,
auto_find_batch_size=False,
batch_eval_metrics=False,
bf16=True,
bf16_full_eval=False,
data_seed=None,
dataloader_drop_last=False,
dataloader_num_workers=0,
dataloader_persistent_workers=False,
dataloader_pin_memory=False,
dataloader_prefetch_factor=None,
ddp_backend=None,
ddp_broadcast_buffers=None,
ddp_bucket_cap_mb=None,
ddp_find_unused_parameters=None,
ddp_timeout=1800,
debug=[],
deepspeed=None,
disable_tqdm=False,
dispatch_batches=None,
do_eval=True,
do_predict=False,
do_train=True,
eval_accumulation_steps=None,
eval_delay=0,
eval_do_concat_batches=True,
eval_steps=None,
eval_strategy=no,
evaluation_strategy=None,
fp16=False,
fp16_backend=auto,
fp16_full_eval=False,
Simple
3 5
neural-chat-x | Unknown
Mode: Edit
Ln 32, Col 29
single_node_finetuning_on_spr.ipynb
1

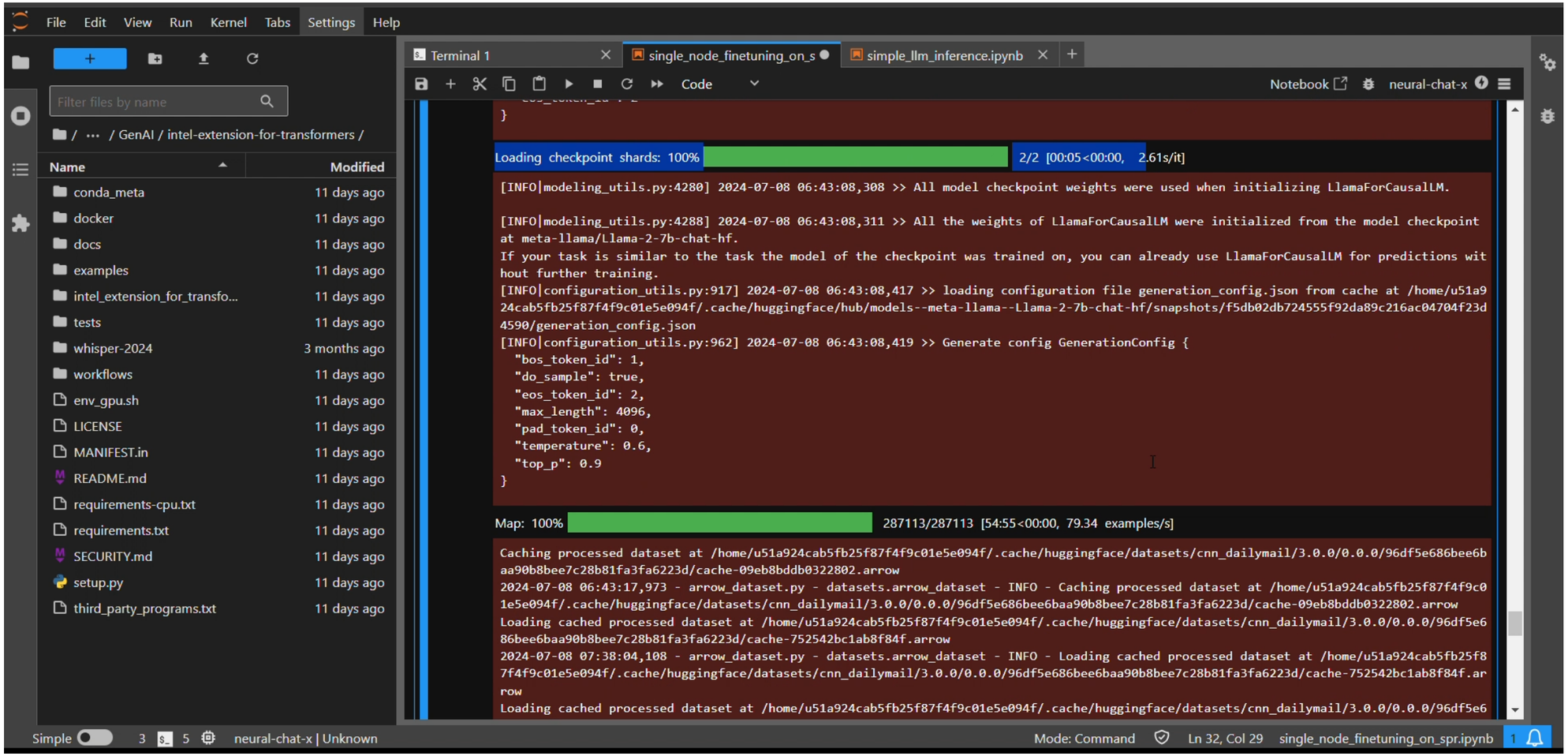

File Edit View Run Kernel Tabs Settings Help
Terminal 1
single_node_finetuning_on_s
simple_llm_inference.ipynb
Code
Notebook
neural-chat-x
Filter files by name
/ ... / GenAI / intel-extension-for-transformers /
Name
Modified
conda_meta 11 days ago
docker 11 days ago
docs 11 days ago
examples 11 days ago
intel_extension_for_transfo... 11 days ago
tests 11 days ago
whisper-2024 3 months ago
workflows 11 days ago
env_gpu.sh 11 days ago
LICENSE 11 days ago
MANIFEST.in 11 days ago
README.md 11 days ago
requirements-cpu.txt 11 days ago
requirements.txt 11 days ago
SECURITY.md 11 days ago
setup.py 11 days ago
third_party_programs.txt 11 days ago
fp16_full_eval=False,
fp16_opt_level=O1,
fsdp=[],
fsdp_config={'min_num_params': 0, 'xla': False, 'xla_fsdp_v2': False, 'xla_fsdp_grad_ckpt': False},
fsdp_min_num_params=0,
fsdp_transformer_layer_cls_to_wrap=None,
full_determinism=False,
gradient_accumulation_steps=2,
gradient_checkpointing=False,
gradient_checkpointing_kwargs=None,
greater_is_better=None,
group_by_length=False,
half_precision_backend=auto,
hub_always_push=False,
hub_model_id=None,
hub_private_repo=False,
hub_strategy=every_save,
hub_token=<HUB_TOKEN>,
ignore_data_skip=False,
include_inputs_for_metrics=False,
include_num_input_tokens_seen=False,
include_tokens_per_second=False,
jit_mode_eval=False,
label_names=None,
label_smoothing_factor=0.0,
learning_rate=5e-05,
length_column_name=length,
load_best_model_at_end=False,
local_rank=0,
log_level=info,
log_level_replica=warning,
log_on_each_node=True,
logging_dir=./tmp/runs/Jul08_06-42-54_idc-training-gpu-compute-26,
logging_first_step=False,
logging_nan_inf_filter=True,
Simple 3 5 neural-chat-x | Unknown
Mode: Edit Ln 32, Col 29 single_node_finetuning_on_spr.ipynb 1

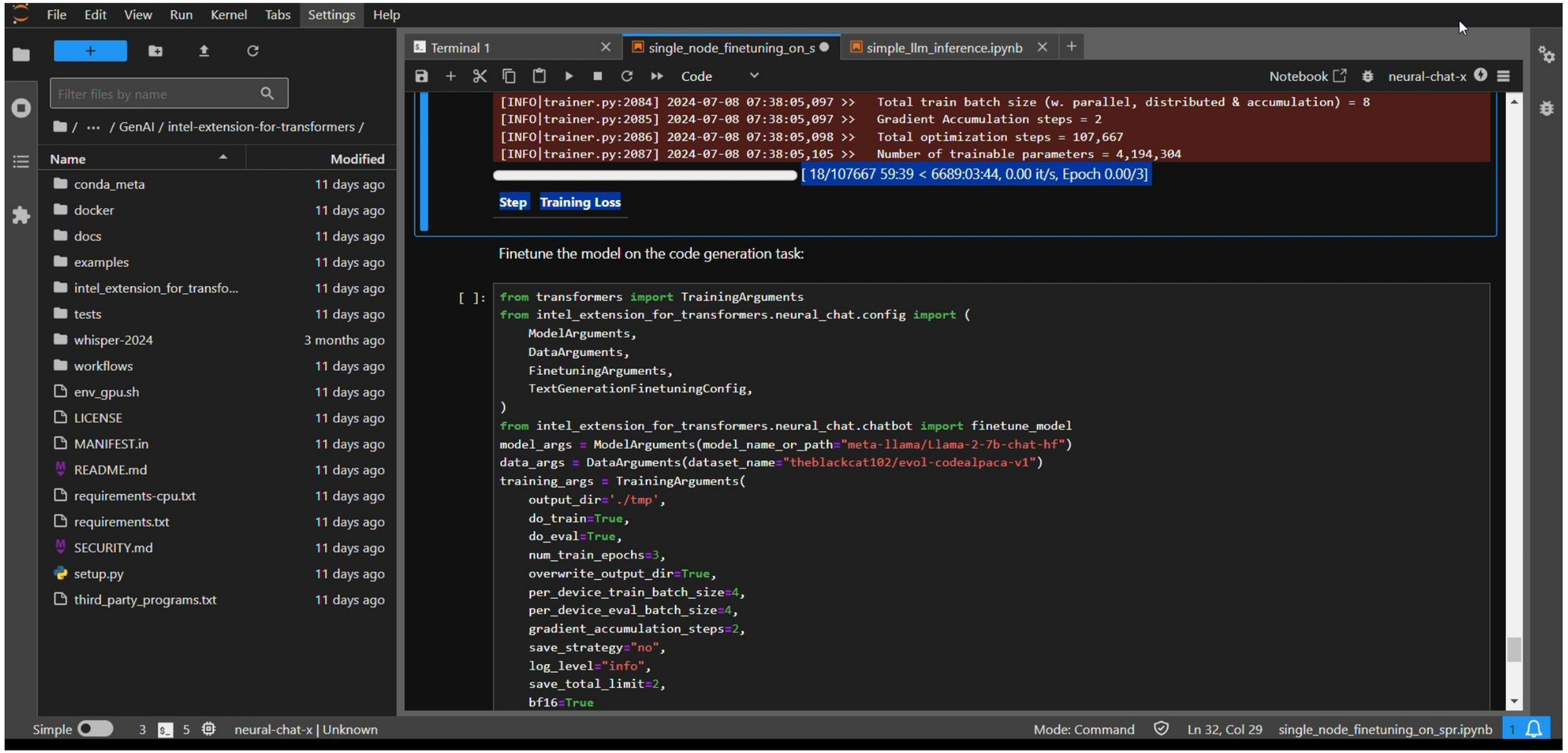

File Edit View Run Kernel Tabs Settings Help
Terminal 1
single_node_finetuning_on_s
simple_llm_inference.ipynb
Code
Notebook
neural-chat-x
Filter files by name
/ ... / GenAI / intel-extension-for-transformers /
Name
Modified
conda_meta 11 days ago
docker 11 days ago
docs 11 days ago
examples 11 days ago
intel_extension_for_transfo... 11 days ago
tests 11 days ago
whisper-2024 3 months ago
workflows 11 days ago
env_gpu.sh 11 days ago
LICENSE 11 days ago
MANIFEST.in 11 days ago
README.md 11 days ago
requirements-cpu.txt 11 days ago
requirements.txt 11 days ago
SECURITY.md 11 days ago
setup.py 11 days ago
third_party_programs.txt 11 days ago
local_rank=0,
log_level=info,
log_level_replica=warning,
log_on_each_node=True,
logging_dir=./tmp/runs/Jul08_06-42-54_idc-training-gpu-compute-26,
logging_first_step=False,
logging_nan_inf_filter=True,
logging_steps=500,
logging_strategy=steps,
lr_scheduler_kwargs={},
lr_scheduler_type=linear,
max_grad_norm=1.0,
max_steps=-1,
metric_for_best_model=None,
mp_parameters=,
neftune_noise_alpha=None,
no_cuda=True,
num_train_epochs=3,
optim=adamw_torch,
optim_args=None,
optim_target_modules=None,
output_dir=./tmp,
overwrite_output_dir=True,
past_index=-1,
per_device_eval_batch_size=4,
per_device_train_batch_size=4,
prediction_loss_only=False,
push_to_hub=False,
push_to_hub_model_id=None,
push_to_hub_organization=None,
push_to_hub_token=<PUSH_TO_HUB_TOKEN>,
ray_scope=last,
remove_unused_columns=True,
report_to=['tensorboard'],
restore_callback_states_from_checkpoint=False,
resume_from_checkpoint=None,
Simple
3 5
neural-chat-x | Unknown
Mode: Edit
Ln 32, Col 29
single_node_finetuning_on_spr.ipynb
1

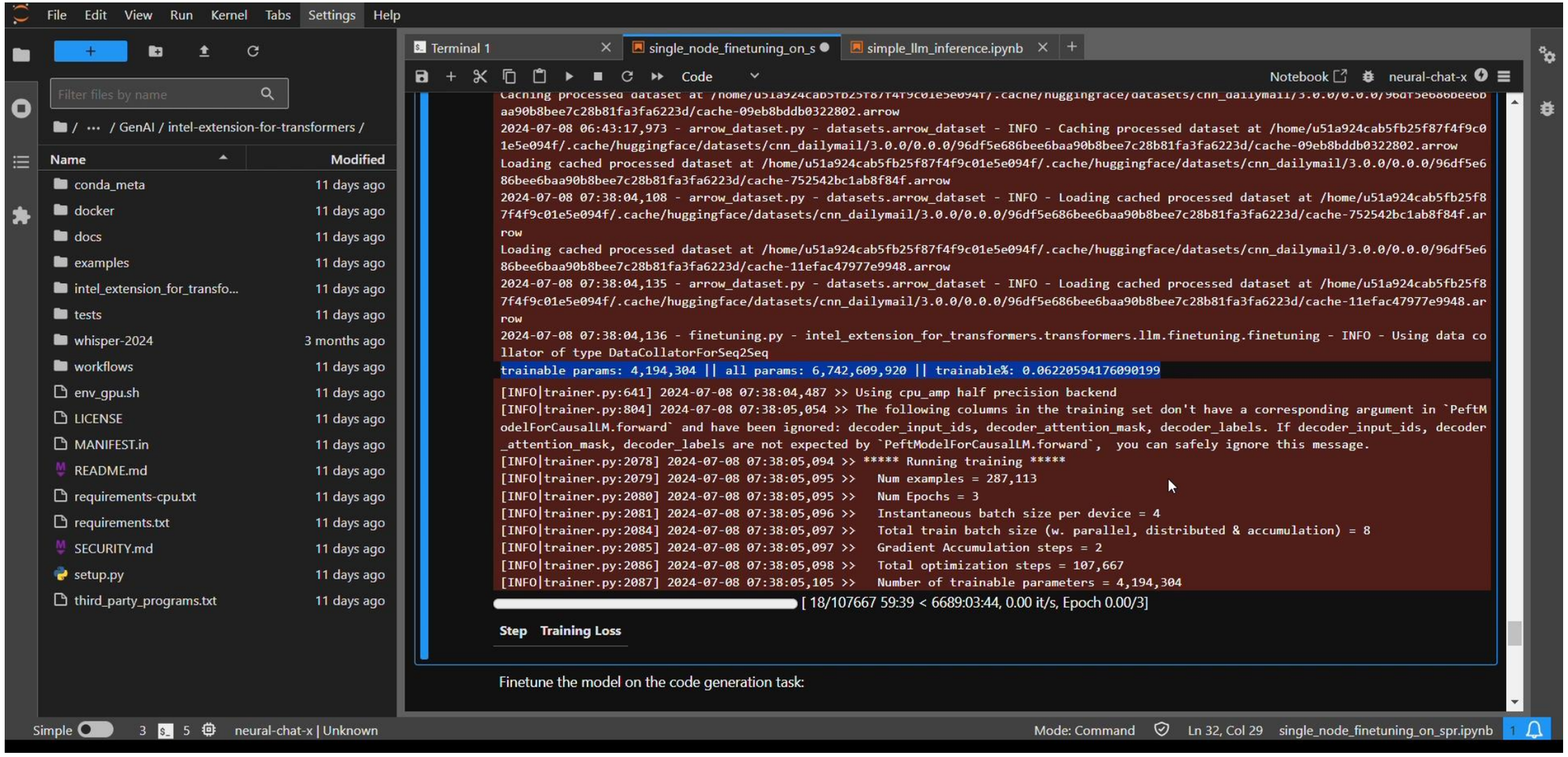

File Edit View Run Kernel Tabs Settings Help
Terminal 1
single_node_finetuning_on_s
simple_llm_inference.ipynb
Code
Notebook
neural-chat-x
Filter files by name
/ ... / GenAI / intel-extension-for-transformers /
Name
Modified
conda_meta 11 days ago
docker 11 days ago
docs 11 days ago
examples 11 days ago
intel_extension_for_transfo... 11 days ago
tests 11 days ago
whisper-2024 3 months ago
workflows 11 days ago
env_gpu.sh 11 days ago
LICENSE 11 days ago
MANIFEST.in 11 days ago
README.md 11 days ago
requirements-cpu.txt 11 days ago
requirements.txt 11 days ago
SECURITY.md 11 days ago
setup.py 11 days ago
third_party_programs.txt 11 days ago
remove_unused_columns=True,
report_to=['tensorboard'],
restore_callback_states_from_checkpoint=False,
resume_from_checkpoint=None,
run_name=./tmp,
save_on_each_node=False,
save_only_model=False,
save_safetensors=True,
save_steps=500,
save_strategy=no,
save_total_limit=2,
seed=42,
skip_memory_metrics=True,
split_batches=None,
tf32=None,
torch_compile=False,
torch_compile_backend=None,
torch_compile_mode=None,
torchdynamo=None,
tpu_metrics_debug=False,
tpu_num_cores=None,
use_cpu=True,
use_ipex=False,
use_legacy_prediction_loop=False,
use_mps_device=False,
warmup_ratio=0.0,
warmup_steps=0,
weight_decay=0.0,
)
/home/u51a924cab5fb25f87f4f9c01e5e094f/.conda/envs/itrex-1/lib/python3.10/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(
[INFO|configuration_utils.py:733] 2024-07-08 06:42:58,350 >> loading configuration file config.json from cache at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/hub/models--meta-llama--Llama-2-7b-chat-hf/snapshots/f5db02db724555f92da89c216ac04704f23d4590/config.json
Simple
3 5
neural-chat-x | Unknown
Mode: Edit
Ln 32, Col 29
single_node_finetuning_on_spr.ipynb
1


File Edit View Run Kernel Tabs Settings Help
Filter files by name
/ ... / GenAI / intel-extension-for-transformers /
Name Modified
conda_meta 11 days ago
docker 11 days ago
docs 11 days ago
examples 11 days ago
intel_extension_for_transfo... 11 days ago
tests 11 days ago
whisper-2024 3 months ago
workflows 11 days ago
env_gpu.sh 11 days ago
LICENSE 11 days ago
MANIFEST.in 11 days ago
README.md 11 days ago
requirements-cpu.txt 11 days ago
requirements.txt 11 days ago
SECURITY.md 11 days ago
setup.py 11 days ago
third_party_programs.txt 11 days ago
Terminal 1
single_node_finetuning_on_s
simple_llm_inference.ipynb
Code
Notebook
neural-chat-x
g.json
[INFO|configuration_utils.py:796] 2024-07-08 06:42:58,355 >> Model config LlamaConfig {
"_name_or_path": "meta-llama/Llama-2-7b-chat-hf",
"architectures": [
"LlamaForCausalLM"
],
"attention_bias": false,
"attention_dropout": 0.0,
"bos_token_id": 1,
"eos_token_id": 2,
"hidden_act": "silu",
"hidden_size": 4096,
"initializer_range": 0.02,
"intermediate_size": 11008,
"max_position_embeddings": 4096,
"mlp_bias": false,
"model_type": "llama",
"num_attention_heads": 32,
"num_hidden_layers": 32,
"num_key_value_heads": 32,
"pretraining_tp": 1,
"rms_norm_eps": 1e-05,
"rope_scaling": null,
"rope_theta": 10000.0,
"tie_word_embeddings": false,
"torch_dtype": "float16",
"transformers_version": "4.41.2",
"use_cache": true,
"vocab_size": 32000
}
[INFO|tokenization_utils_base.py:2108] 2024-07-08 06:42:58,454 >> loading file tokenizer.model from cache at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/hub/models--meta-llama--Llama-2-7b-chat-hf/snapshots/f5db02db724555f92da89c216ac04704f23d4590/tokenizer.model
[INFO|tokenization_utils_base.py:2108] 2024-07-08 06:42:58,456 >> loading file added_tokens.json from cache at None
[INFO|tokenization_utils_base.py:2108] 2024-07-08 06:42:58,456 >> loading file special_tokens_map.json from cache at /home/u51a924cab5fb
Simple 3 5 neural-chat-x | Unknown
Mode: Edit Ln 32, Col 29 single_node_finetuning_on_spr.ipynb 1


File Edit View Run Kernel Tabs Settings Help
Filter files by name
/ ... / GenAI / intel-extension-for-transformers /
Name
Modified
conda_meta 11 days ago
docker 11 days ago
docs 11 days ago
examples 11 days ago
intel_extension_for_transfo... 11 days ago
tests 11 days ago
whisper-2024 3 months ago
workflows 11 days ago
env_gpu.sh 11 days ago
LICENSE 11 days ago
MANIFEST.in 11 days ago
README.md 11 days ago
requirements-cpu.txt 11 days ago
requirements.txt 11 days ago
SECURITY.md 11 days ago
setup.py 11 days ago
third_party_programs.txt 11 days ago
Terminal 1
single_node_finetuning_on_s
simple_llm_inference.ipynb
Code
Notebook
neural-chat-x
odel
[INFO|tokenization_utils_base.py:2108] 2024-07-08 06:42:58,456 >> loading file added_tokens.json from cache at None
[INFO|tokenization_utils_base.py:2108] 2024-07-08 06:42:58,456 >> loading file special_tokens_map.json from cache at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/hub/models--meta-llama--Llama-2-7b-chat-hf/snapshots/f5db02db724555f92da89c216ac04704f23d4590/special_tokens_map.json
[INFO|tokenization_utils_base.py:2108] 2024-07-08 06:42:58,457 >> loading file tokenizer_config.json from cache at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/hub/models--meta-llama--Llama-2-7b-chat-hf/snapshots/f5db02db724555f92da89c216ac04704f23d4590/tokenizer_config.json
[INFO|tokenization_utils_base.py:2108] 2024-07-08 06:42:58,457 >> loading file tokenizer.json from cache at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/hub/models--meta-llama--Llama-2-7b-chat-hf/snapshots/f5db02db724555f92da89c216ac04704f23d4590/tokenizer.json
Overwrite dataset info from restored data version if exists.
2024-07-08 06:43:02,420 - builder.py - datasets.builder - INFO - Overwrite dataset info from restored data version if exists.
Loading Dataset info from /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/cnn_dailymail/3.0.0/0.0.0/96df5e686bee6baa90b8bee7c28b81fa3fa6223d
2024-07-08 06:43:02,423 - info.py - datasets.info - INFO - Loading Dataset info from /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/cnn_dailymail/3.0.0/0.0.0/96df5e686bee6baa90b8bee7c28b81fa3fa6223d
Found cached dataset cnn_dailymail (/home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/cnn_dailymail/3.0.0/0.0.0/96df5e686bee6baa90b8bee7c28b81fa3fa6223d)
2024-07-08 06:43:02,431 - builder.py - datasets.builder - INFO - Found cached dataset cnn_dailymail (/home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/cnn_dailymail/3.0.0/0.0.0/96df5e686bee6baa90b8bee7c28b81fa3fa6223d)
Loading Dataset info from /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/cnn_dailymail/3.0.0/0.0.0/96df5e686bee6baa90b8bee7c28b81fa3fa6223d
2024-07-08 06:43:02,432 - info.py - datasets.info - INFO - Loading Dataset info from /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/cnn_dailymail/3.0.0/0.0.0/96df5e686bee6baa90b8bee7c28b81fa3fa6223d
[INFO|modeling_utils.py:3474] 2024-07-08 06:43:02,564 >> loading weights file model.safetensors from cache at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/hub/models--meta-llama--Llama-2-7b-chat-hf/snapshots/f5db02db724555f92da89c216ac04704f23d4590/model.safetensors.index.json
[INFO|modeling_utils.py:1519] 2024-07-08 06:43:02,568 >> Instantiating LlamaForCausalLM model under default dtype torch.bfloat16.
[INFO|configuration_utils.py:962] 2024-07-08 06:43:02,570 >> Generate config GenerationConfig {
"bos_token_id": 1,
"eos_token_id": 2
}
Loading checkpoint shards: 100%
2/2 [00:05<00:00, 2.61s/it]
Simple
3 5
neural-chat-x | Unknown
Mode: Edit
Ln 32, Col 29
single_node_finetuning_on_spr.ipynb
1


File Edit View Run Kernel Tabs Settings Help
Filter files by name
/ ... / GenAI / intel-extension-for-transformers /
Name
Modified
conda_meta 11 days ago
docker 11 days ago
docs 11 days ago
examples 11 days ago
intel_extension_for_transfo... 11 days ago
tests 11 days ago
whisper-2024 3 months ago
workflows 11 days ago
env_gpu.sh 11 days ago
LICENSE 11 days ago
MANIFEST.in 11 days ago
README.md 11 days ago
requirements-cpu.txt 11 days ago
requirements.txt 11 days ago
SECURITY.md 11 days ago
setup.py 11 days ago
third_party_programs.txt 11 days ago
Terminal 1
single_node_finetuning_on_s
simple_llm_inference.ipynb
Code
Notebook
neural-chat-x
}
Loading checkpoint shards: 100%
2/2 [00:05<00:00, 2.61s/it]
[INFO|modeling_utils.py:4280] 2024-07-08 06:43:08,308 >> All model checkpoint weights were used when initializing LlamaForCausalLM.
[INFO|modeling_utils.py:4288] 2024-07-08 06:43:08,311 >> All the weights of LlamaForCausalLM were initialized from the model checkpoint at meta-llama/Llama-2-7b-chat-hf.
If your task is similar to the task the model of the checkpoint was trained on, you can already use LlamaForCausalLM for predictions without further training.
[INFO|configuration_utils.py:917] 2024-07-08 06:43:08,417 >> loading configuration file generation_config.json from cache at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/hub/models--meta-llama--Llama-2-7b-chat-hf/snapshots/f5db02db724555f92da89c216ac04704f23d4590/generation_config.json
[INFO|configuration_utils.py:962] 2024-07-08 06:43:08,419 >> Generate config GenerationConfig {
"bos_token_id": 1,
"do_sample": true,
"eos_token_id": 2,
"max_length": 4096,
"pad_token_id": 0,
"temperature": 0.6,
"top_p": 0.9
}
Map: 100%
287113/287113 [54:55<00:00, 79.34 examples/s]
Caching processed dataset at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/cnn_dailymail/3.0.0/0.0.0/96df5e686bee6baa90b8bee7c28b81fa3fa6223d/cache-09eb8bddb0322802.arrow
2024-07-08 06:43:17,973 - arrow_dataset.py - datasets.arrow_dataset - INFO - Caching processed dataset at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/cnn_dailymail/3.0.0/0.0.0/96df5e686bee6baa90b8bee7c28b81fa3fa6223d/cache-09eb8bddb0322802.arrow
Loading cached processed dataset at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/cnn_dailymail/3.0.0/0.0.0/96df5e686bee6baa90b8bee7c28b81fa3fa6223d/cache-752542bc1ab8f84f.arrow
2024-07-08 07:38:04,108 - arrow_dataset.py - datasets.arrow_dataset - INFO - Loading cached processed dataset at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/cnn_dailymail/3.0.0/0.0.0/96df5e686bee6baa90b8bee7c28b81fa3fa6223d/cache-752542bc1ab8f84f.arrow
Loading cached processed dataset at /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/datasets/cnn_dailymail/3.0.0/0.0.0/96df5e6
Simple
3 5
neural-chat-x | Unknown
Mode: Command
Ln 32, Col 29
single_node_finetuning_on_spr.ipynb
1


File Edit View Run Kernel Tabs Settings Help
Filter files by name
/ ... / GenAI / intel-extension-for-transformers /
Name
Modified
conda_meta 11 days ago
docker 11 days ago
docs 11 days ago
examples 11 days ago
intel_extension_for_transfo... 11 days ago
tests 11 days ago
whisper-2024 3 months ago
workflows 11 days ago
env_gpu.sh 11 days ago
LICENSE 11 days ago
MANIFEST.in 11 days ago
README.md 11 days ago
requirements-cpu.txt 11 days ago
requirements.txt 11 days ago
SECURITY.md 11 days ago
setup.py 11 days ago
third_party_programs.txt 11 days ago
Terminal 1
single_node_finetuning_on_s
simple_llm_inference.ipynb
Code
Notebook
neural-chat-x
[INFO|trainer.py:2084] 2024-07-08 07:38:05,097 >> Total train batch size (w. parallel, distributed & accumulation) = 8
[INFO|trainer.py:2085] 2024-07-08 07:38:05,097 >> Gradient Accumulation steps = 2
[INFO|trainer.py:2086] 2024-07-08 07:38:05,098 >> Total optimization steps = 107,667
[INFO|trainer.py:2087] 2024-07-08 07:38:05,105 >> Number of trainable parameters = 4,194,304
[18/107667 59:39 < 6689:03:44, 0.00 it/s, Epoch 0.00/3]
Step Training Loss
Finetune the model on the code generation task:
[ ]:
from transformers import TrainingArguments
from intel_extension_for_transformers.neural_chat.config import (
ModelArguments,
DataArguments,
FinetuningArguments,
TextGenerationFinetuningConfig,
)
from intel_extension_for_transformers.neural_chat.chatbot import finetune_model
model_args = ModelArguments(model_name_or_path="meta-llama/Llama-2-7b-chat-hf")
data_args = DataArguments(dataset_name="theblackcat102/evol-codealpaca-v1")
training_args = TrainingArguments(
output_dir='./tmp',
do_train=True,
do_eval=True,
num_train_epochs=3,
overwrite_output_dir=True,
per_device_train_batch_size=4,
per_device_eval_batch_size=4,
gradient_accumulation_steps=2,
save_strategy="no",
log_level="info",
save_total_limit=2,
bf16=True
Simple 3 5 neural-chat-x | Unknown
Mode: Command Ln 32, Col 29 single_node_finetuning_on_spr.ipynb 1



https://jupyter-batch-us-region-1.cloud.intel.com/hub/user-redirect/lab/tree/Training/AI/GenAI/intel-extension-for-transformers

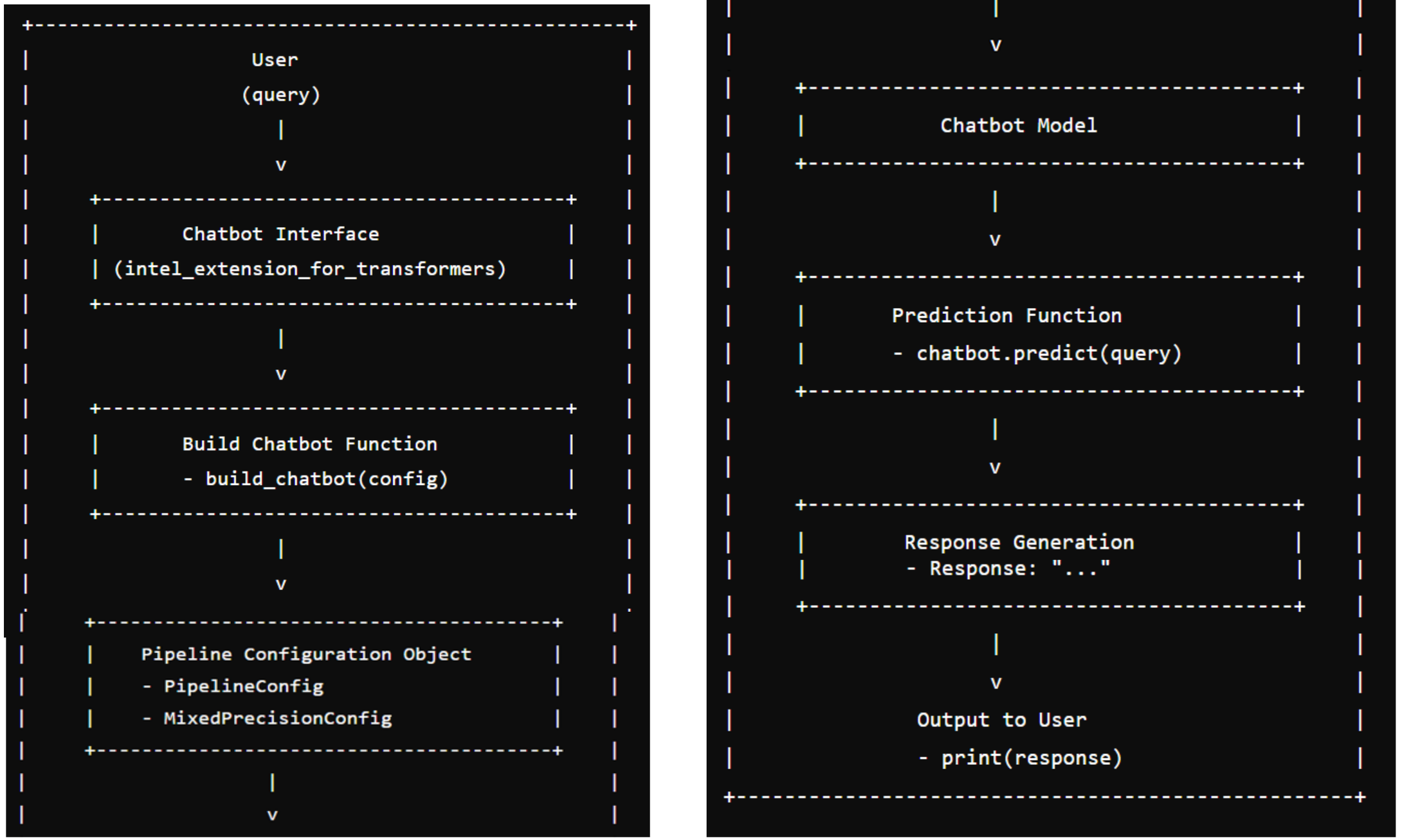
Architecture Diagram

Build your chatbot 💻

Text Chat

Giving NeuralChat the textual instruction, it will respond with the textual response.

[8]:
```
# BF16 Optimization
from intel_extension_for_transformers.neural_chat import build_chatbot, PipelineConf
from intel_extension_for_transformers.transformers import MixedPrecisionConfig
config = PipelineConfig(optimization_config=MixedPrecisionConfig())
chatbot = build_chatbot(config)
print(chatbot)
response = chatbot.predict(query="Tell me about Intel Xeon Scalable Processors.")
print(response)
```


User
(query)
Chatbot Interface
(intel_extension_for_transformers)
Build Chatbot Function
- build_chatbot(config)
Pipeline Configuration Object
- PipelineConfig
- MixedPrecisionConfig
Chatbot Model
Prediction Function
- chatbot.predict(query)
Response Generation
- Response: "..."
Output to User
- print(response)

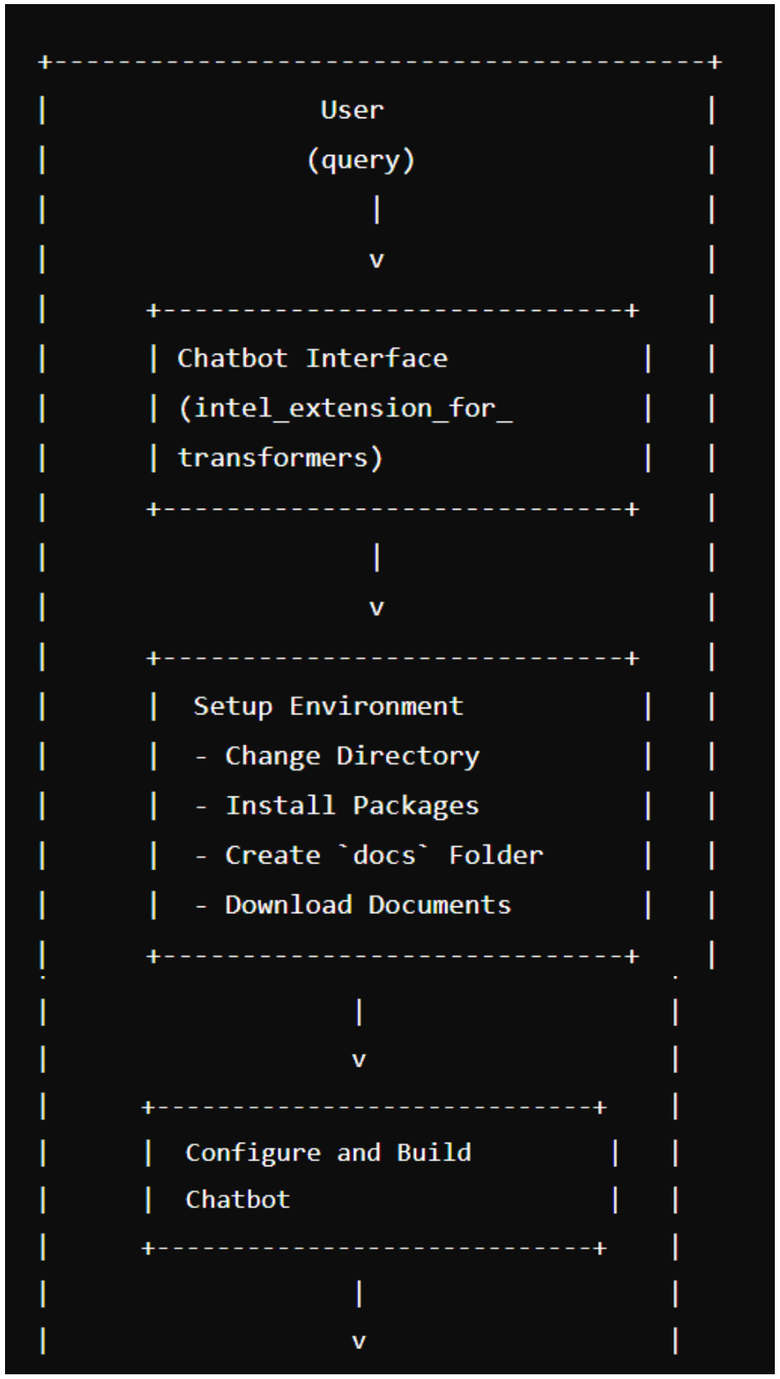
Text Chat With Retrieval Plugin

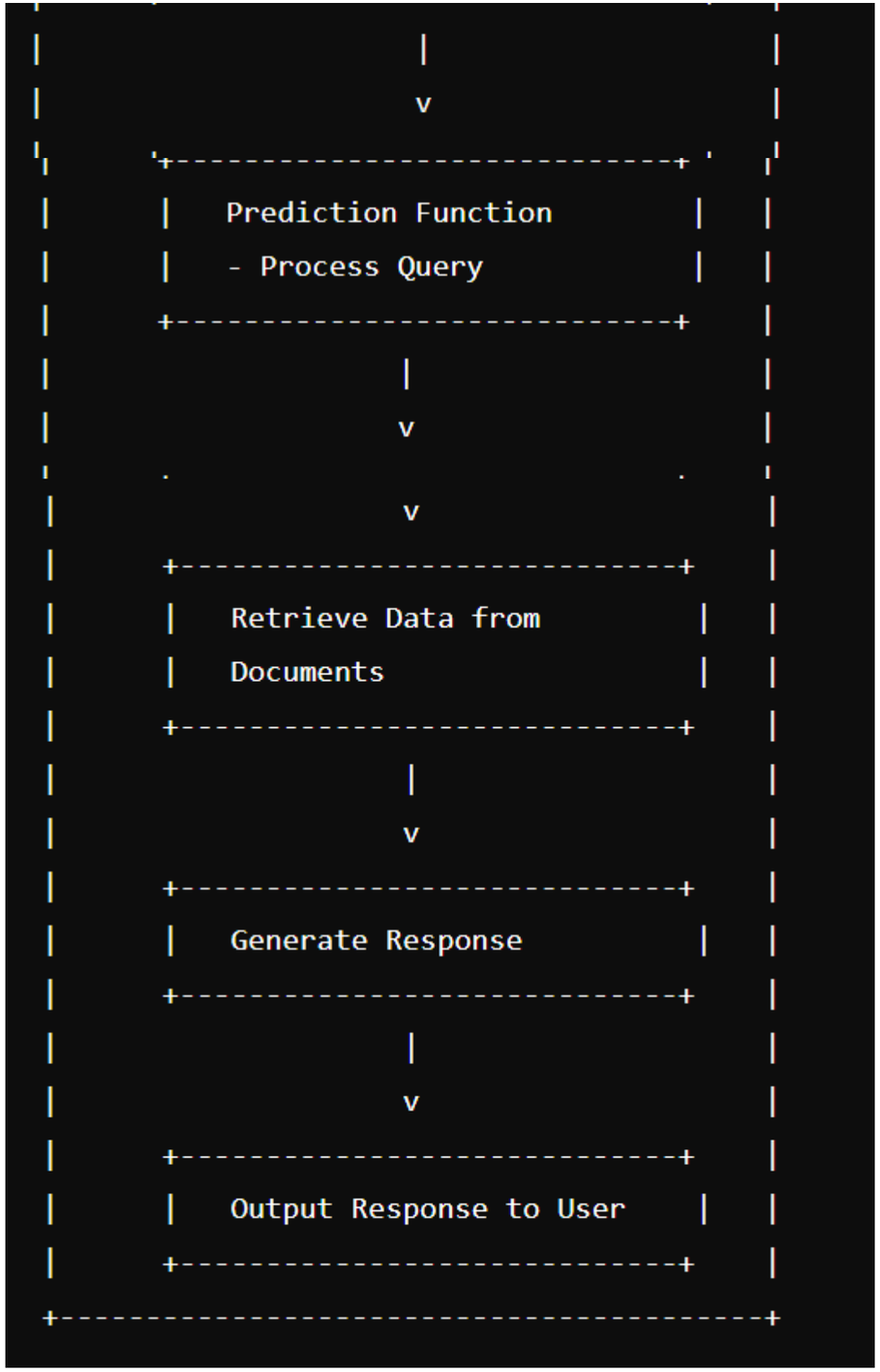
User could also leverage NeuralChat Retrieval plugin to do domain specific chat by feding with some documents like below:

```
%cd ./intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/pipeline/plugins/retrieval/
!pip install -r requirements.txt
%cd ../../../../../../
```

```
!mkdir docs
%cd docs
!curl -OL https://raw.githubusercontent.com/intel/intel-extension-for-transformers/main/intel_extension_for_transformers/neural_chat/assets/docs/
!curl -OL https://raw.githubusercontent.com/intel/intel-extension-for-transformers/main/intel_extension_for_transformers/neural_chat/assets/docs/
!curl -OL https://raw.githubusercontent.com/intel/intel-extension-for-transformers/main/intel_extension_for_transformers/neural_chat/assets/docs/
%cd ..
```

```
from intel_extension_for_transformers.neural_chat import PipelineConfig
from intel_extension_for_transformers.neural_chat import build_chatbot
from intel_extension_for_transformers.neural_chat import plugins
plugins.retrieval.enable=True
plugins.retrieval.args["input_path"]="./docs/"
config = PipelineConfig(plugins=plugins)
chatbot = build_chatbot(config)
response = chatbot.predict("How many cores does the Intel® Xeon® Platinum 8480+ Processor have in total?")
print(response)
```


User
(query)
Chatbot Interface
(intel_extension_for_
transformers)
Setup Environment
- Change Directory
- Install Packages
- Create `docs` Folder
- Download Documents
Configure and Build
Chatbot


Prediction Function
- Process Query
Retrieve Data from
Documents
Generate Response
Output Response to User

Voice Chat with ASR & TTS Plugin

In the context of voice chat, users have the option to engage in various modes: utilizing input audio and receiving output audio, employing input audio and receiving textual output, or providing input in textual form and receiving audio output.

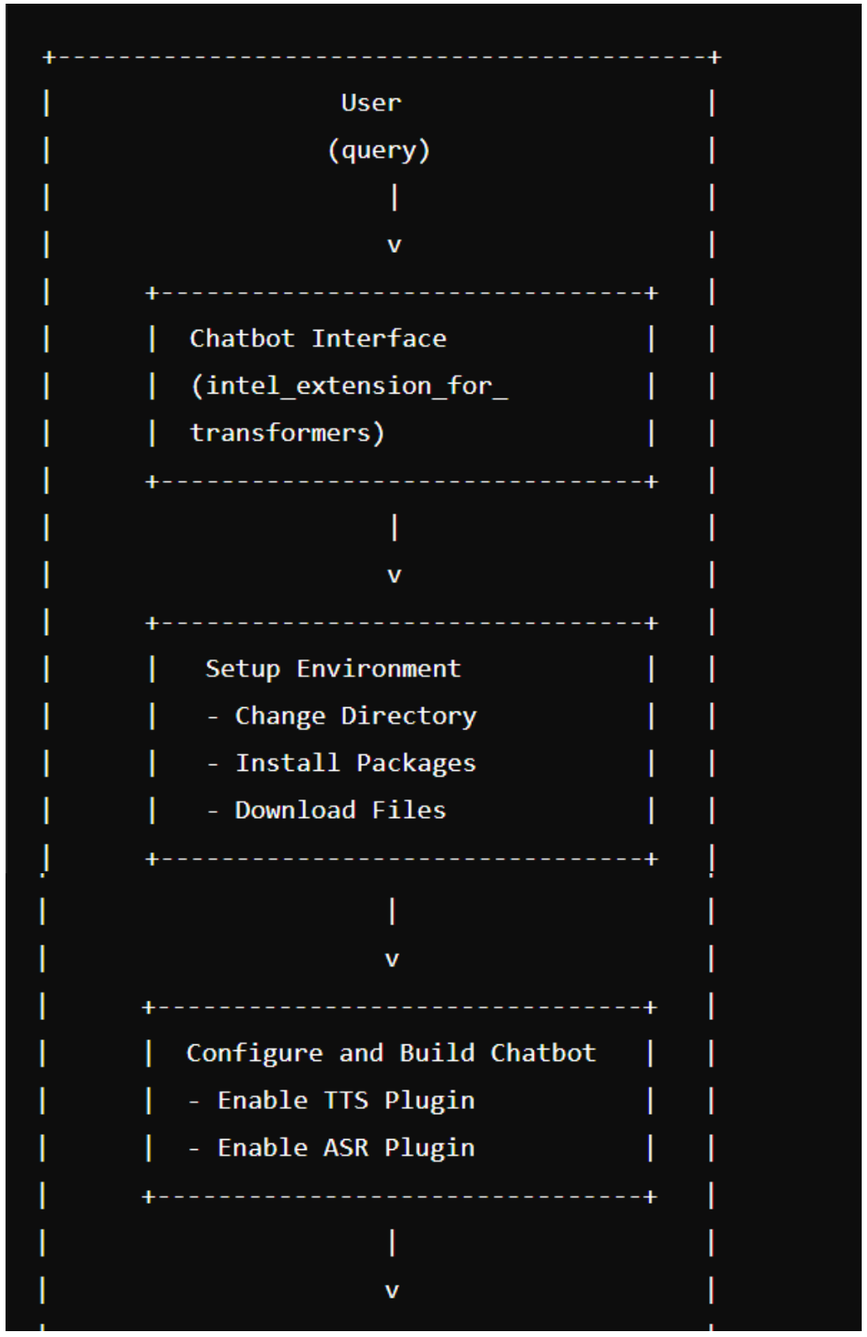
For the Python API code, users have the option to enable different voice chat modes by setting ASR and TTS plugins enable or disable.

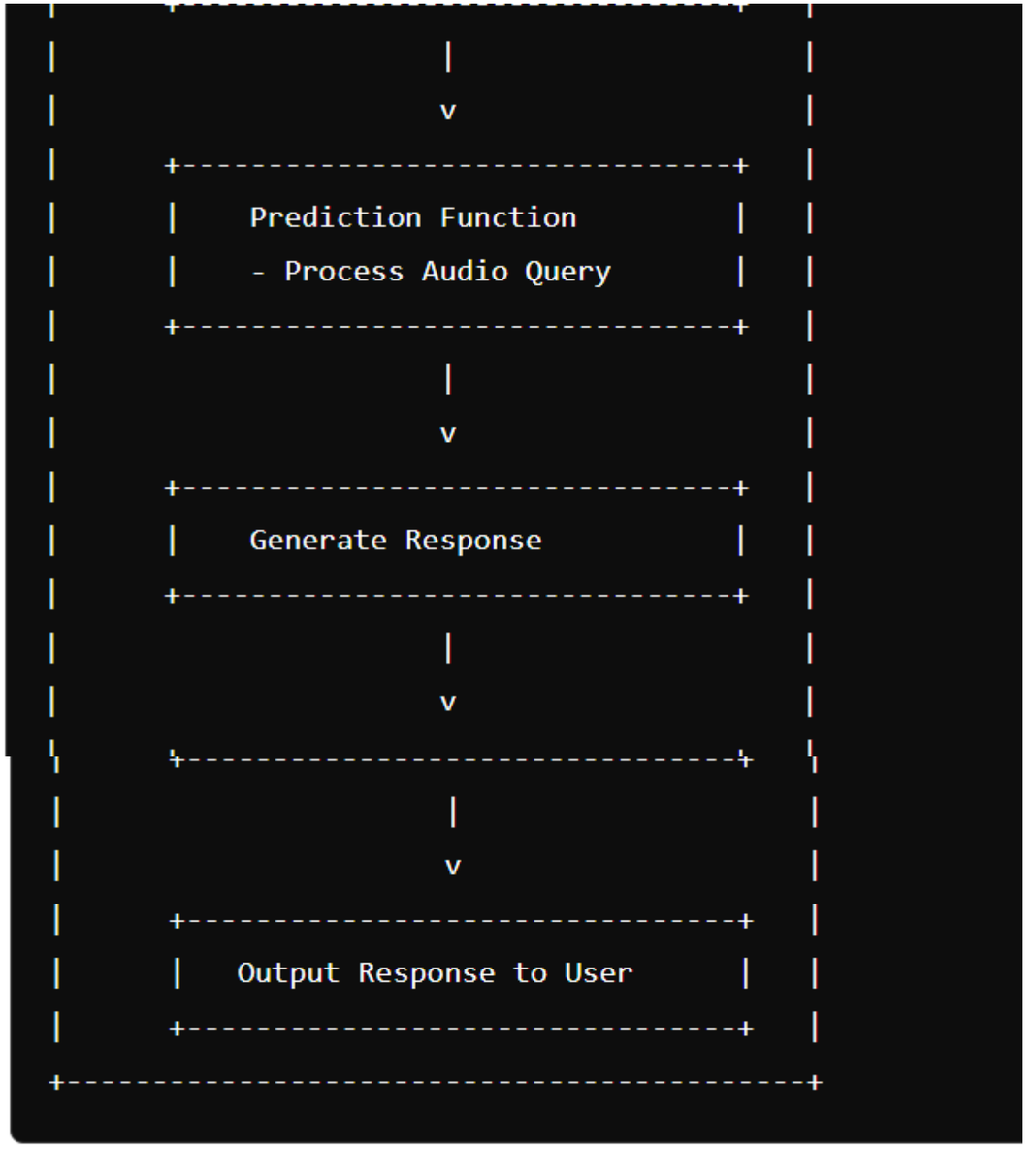
[ ]:
```
%cd ./intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/pipeline/plugins/audio/
!pip install -r requirements.txt
%cd ../../../../../../
```

[ ]:
```
!curl -OL https://raw.githubusercontent.com/intel/intel-extension-for-transformers/main/intel_extension_for_transformers/neural_chat/assets/speak
!curl -OL https://raw.githubusercontent.com/intel/intel-extension-for-transformers/main/intel_extension_for_transformers/neural_chat/assets/audio
```

[ ]:
```
from intel_extension_for_transformers.neural_chat import PipelineConfig
from intel_extension_for_transformers.neural_chat import build_chatbot
from intel_extension_for_transformers.neural_chat import plugins
plugins.tts.enable = True
plugins.tts.args["output_audio_path"] = "./response.wav"
plugins.asr.enable = True

config = PipelineConfig(plugins=plugins)
chatbot = build_chatbot(config)
result = chatbot.predict(query="./sample.wav")
print(result)
```


User
(query)
Chatbot Interface
(intel_extension_for_
transformers)
Setup Environment
- Change Directory
- Install Packages
- Download Files
Configure and Build Chatbot
- Enable TTS Plugin
- Enable ASR Plugin

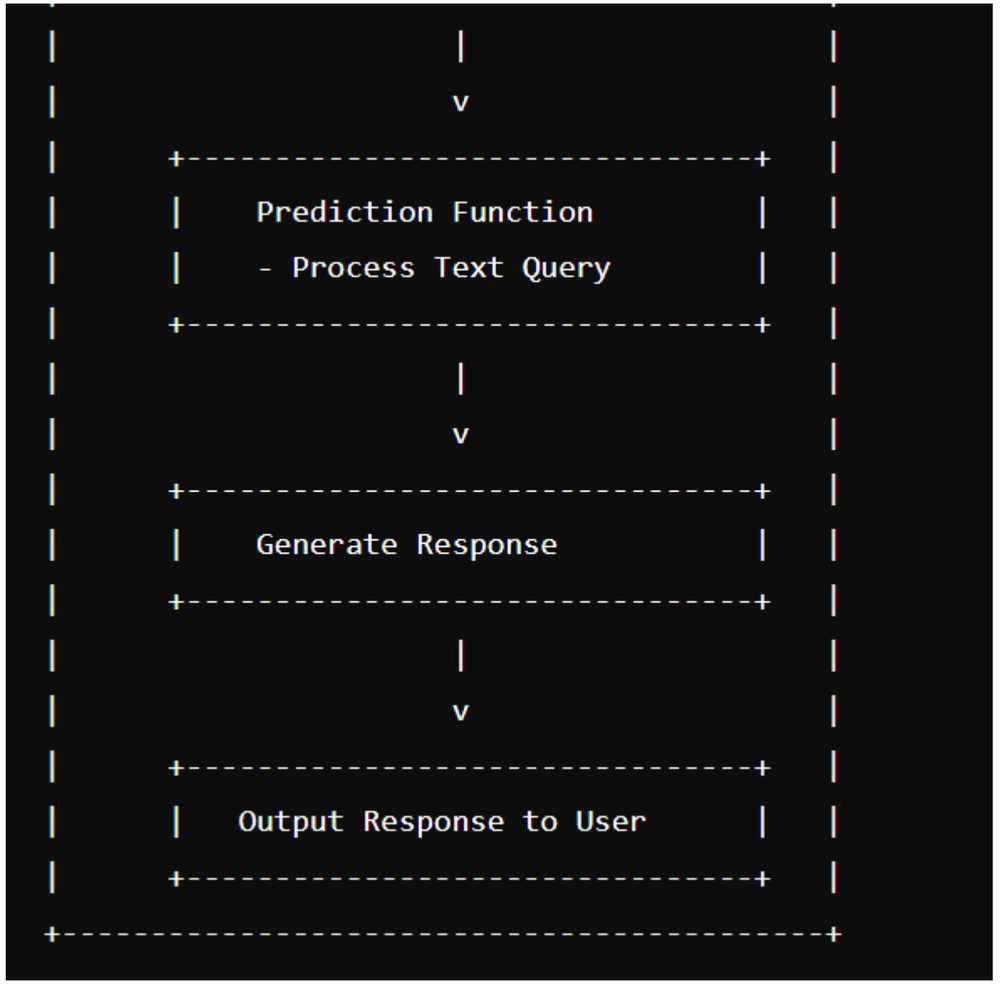

Prediction Function
- Process Audio Query
Generate Response
Output Response to User

Low Precision Optimization

BF16

[ ]:
```
# BF16 Optimization
from intel_extension_for_transformers.neural_chat.config import PipelineConfig
from intel_extension_for_transformers.transformers import MixedPrecisionConfig
config = PipelineConfig(optimization_config=MixedPrecisionConfig())
chatbot = build_chatbot(config)
response = chatbot.predict(query="Tell me about Intel Xeon Scalable Processors.")
print(response)
```


User
(query)
Chatbot Interface
(intel_extension_for_
transformers)
Configure and Build Chatbot
- Mixed Precision Config


Prediction Function
- Process Text Query
Generate Response
Output Response to User

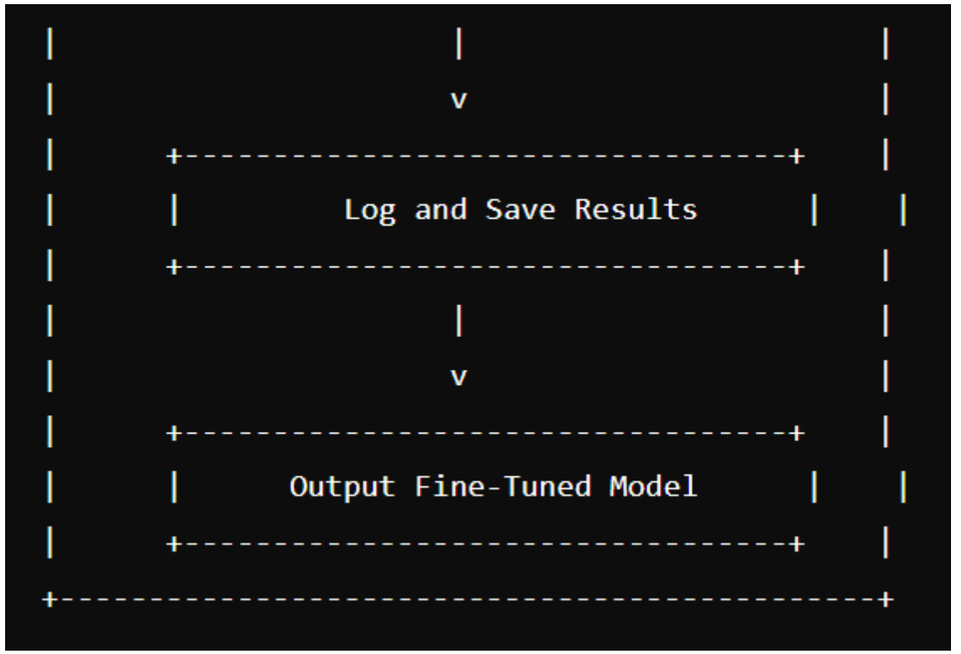
Finetune Your Chatbot

We employ the LoRA approach to finetune the LLM efficiently.

Finetune the model on Alpaca-format dataset to conduct text generation:

[ ]:
```
from transformers import TrainingArguments
from intel_extension_for_transformers.neural_chat.config import (
    ModelArguments,
    DataArguments,
    FinetuningArguments,
    TextGenerationFinetuningConfig,
)
from intel_extension_for_transformers.neural_chat.chatbot import finetune_model
model_args = ModelArguments(model_name_or_path="meta-llama/Llama-2-7b-chat-hf")
data_args = DataArguments(train_file="alpaca_data.json", validation_split_percentage=1)
training_args = TrainingArguments(
    output_dir='./tmp',
    do_train=True,
    do_eval=True,
    num_train_epochs=3,
    overwrite_output_dir=True,
    per_device_train_batch_size=4,
    per_device_eval_batch_size=4,
    gradient_accumulation_steps=2,
    save_strategy="no",
    log_level="info",
    save_total_limit=2,
    bf16=True,
)
finetune_args = FinetuningArguments()
finetune_cfg = TextGenerationFinetuningConfig(
            model_args=model_args,
            data_args=data_args,
            training_args=training_args,
            finetune_args=finetune_args,
        )
finetune_model(finetune_cfg)
```

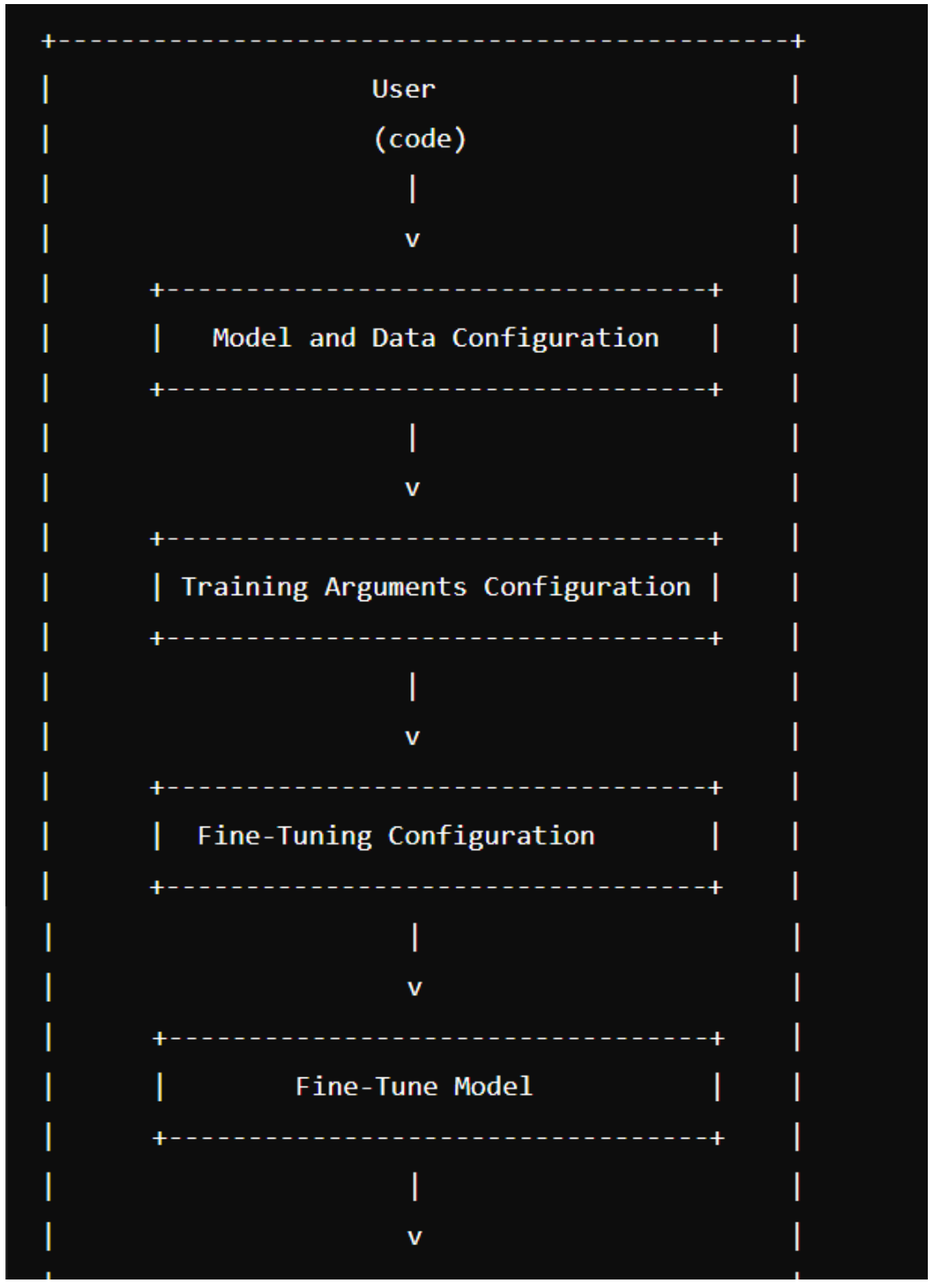

User
(code)
Model and Data Configuration
Training Arguments Configuration
Fine-Tuning Configuration
Fine-Tune Model

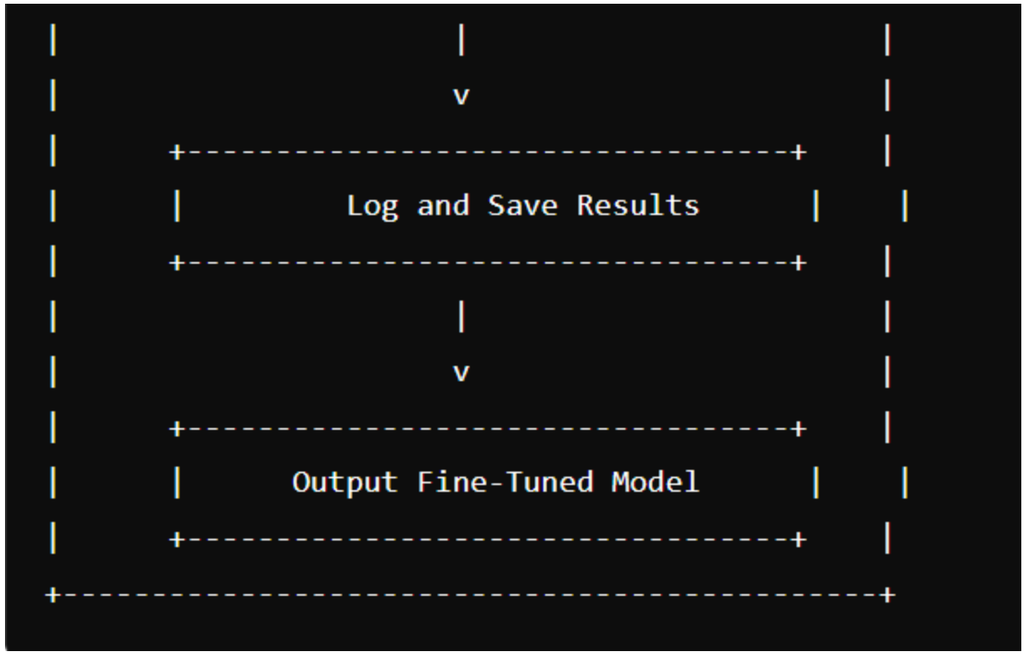

Log and Save Results
Output Fine-Tuned Model

Finetune the model on the summarization task:

[ ]:
```
from transformers import TrainingArguments
from intel_extension_for_transformers.neural_chat.config import (
    ModelArguments,
    DataArguments,
    FinetuningArguments,
    TextGenerationFinetuningConfig,
)
from intel_extension_for_transformers.neural_chat.chatbot import finetune_model
model_args = ModelArguments(model_name_or_path="meta-llama/Llama-2-7b-chat-hf")
data_args = DataArguments(dataset_name="cnn_dailymail", dataset_config_name="3.0.0")
training_args = TrainingArguments(
    output_dir='./tmp',
    do_train=True,
    do_eval=True,
    num_train_epochs=3,
    overwrite_output_dir=True,
    per_device_train_batch_size=4,
    per_device_eval_batch_size=4,
    gradient_accumulation_steps=2,
    save_strategy="no",
    log_level="info",
    save_total_limit=2,
    bf16=True
)
finetune_args = FinetuningArguments(task='summarization')
finetune_cfg = TextGenerationFinetuningConfig(
            model_args=model_args,
            data_args=data_args,
            training_args=training_args,
            finetune_args=finetune_args,
        )
```

```
finetune_model(finetune_cfg)
```

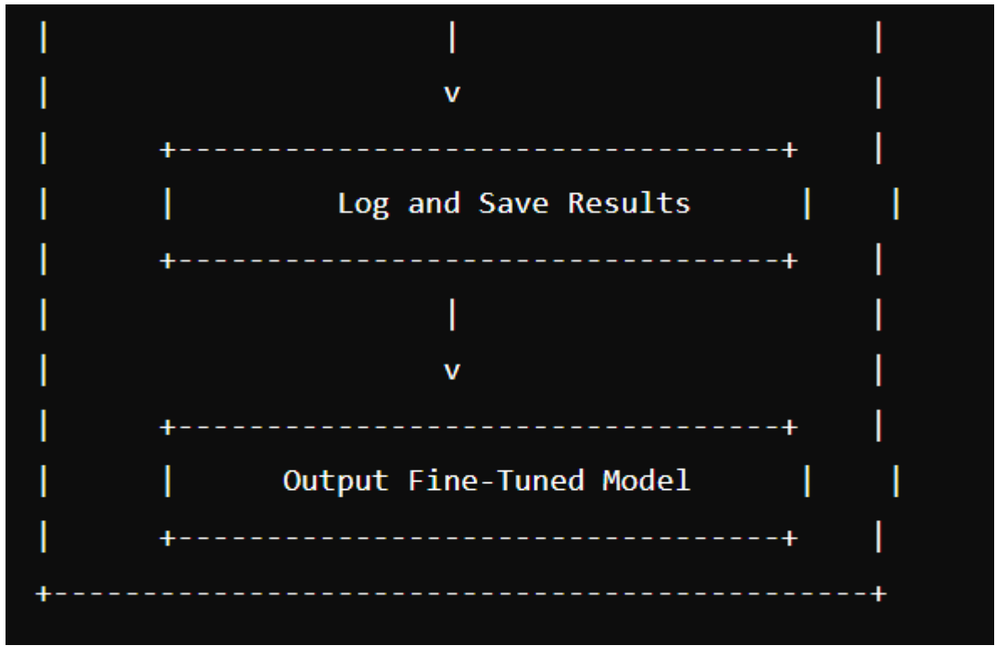

User
(code)
Model and Data Configuration
Training Arguments Configuration
Fine-Tuning Configuration
Fine-Tune Model


Log and Save Results
Output Fine-Tuned Model

Finetune the model on the code generation task:

[ ]:
```
from transformers import TrainingArguments
from intel_extension_for_transformers.neural_chat.config import (
    ModelArguments,
    DataArguments,
    FinetuningArguments,
    TextGenerationFinetuningConfig,
)
from intel_extension_for_transformers.neural_chat.chatbot import finetune_model
model_args = ModelArguments(model_name_or_path="meta-llama/Llama-2-7b-chat-hf")
data_args = DataArguments(dataset_name="theblackcat102/evol-codealpaca-v1")
training_args = TrainingArguments(
    output_dir='./tmp',
    do_train=True,
    do_eval=True,
    num_train_epochs=3,
    overwrite_output_dir=True,
    per_device_train_batch_size=4,
    per_device_eval_batch_size=4,
    gradient_accumulation_steps=2,
    save_strategy="no",
    log_level="info",
    save_total_limit=2,
    bf16=True
)
finetune_args = FinetuningArguments(task='code-generation')
finetune_cfg = TextGenerationFinetuningConfig(
            model_args=model_args,
            data_args=data_args,
            training_args=training_args,
            finetune_args=finetune_args,
        )
finetune_model(finetune_cfg)
```


User
(code)
Model and Data Configuration
Training Arguments Configuration
Fine-Tuning Configuration
Fine-Tune Model


Log and Save Results
Output Fine-Tuned Model

Technologies Used

Intel® Developer Cloud (IDC)

Intel extensions for Transformers
Git-Hub

Intel Server in college

Intel unnati Lab in college

Team Members and Contribution

ABHINAV NAIR S	Working of entire notebook and Report writing.
ARDRA A NAIR	Working of notebook and Report writing
ESA MARIA SHYJU	Report writing
CHAITHANYA P SUNIL	Report writing
SREERAM R NAIR	Report writing

Conclusion

In this report, we explored the fundamental concepts of Generative AI, focusing on the use of simple language model (LLM) interfaces on CPUs and the process of fine-tuning LLMs to create custom chatbots . By understanding the basics of how LLMs function and their applications, we recognized the potential these models hold in transforming various domains through intelligent automation and personalized user interactions. Initially, we delved into the mechanics of LLMs, discussing their architecture, capabilities, and limitations, especially when operating on CPU-based systems. This foundation set the stage for practical applications, highlighting how even with hardware constraints, these models can deliver valuable results. The core of our study involved the fine-tuning process, where we adapted a pre-trained LLM to cater to specific needs. This customization not only demonstrated the flexibility of LLMs but also emphasized the importance of domain-specific data and fine-tuning techniques to enhance the model's performance in targeted applications. By creating a custom chatbot , we illustrated a tangible outcome of this process, showcasing how tailored interactions can significantly improve user experiences