Step 3: Model Fine-tuning

In this notebook, you'll fine-tune the Meta Llama 2 7B large language model, deploy the fine-tuned model, and test it's text generation and domain knowledge capabilities.

Fine-tuning refers to the process of taking a pre-trained language model and retraining it for a different but related task using specific data. This approach is also known as transfer learning, which involves transferring the knowledge learned from one task to another. Large language models (LLMs) like Llama 2 7B are trained on massive amounts of unlabeled data and can be fine-tuned on domain domain datasets, making the model perform better on that specific domain.

Input: A train and an optional validation directory. Each directory contains a CSV/JSON/TXT file. For CSV/JSON files, the train or validation data is used from the column called 'text' or the first column if no column called 'text' is found. The number of files under train and validation should equal to one.

- **You'll choose your dataset below based on the domain you've chosen**

Output: A trained model that can be deployed for inference.
After you've fine-tuned the model, you'll evaluate it with the same input you used in project step 2: model evaluation.

Set up

Install and import the necessary packages. Restart the kernel after executing the cell below.

In [1]:
```
!pip install --upgrade sagemaker datasets
```

```
Requirement already satisfied: sagemaker in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (2.237.1)
Requirement already satisfied: datasets in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (3.2.0)
Requirement already satisfied: attrs<24,>=23.1.0 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages
(from sagemaker) (23.2.0)
Requirement already satisfied: boto3<2.0,>=1.35.75 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages
(from sagemaker) (1.35.76)
Requirement already satisfied: cloudpickle==2.2.1 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages
(from sagemaker) (2.2.1)
Requirement already satisfied: docker in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (from sagema
ker) (7.1.0)
Requirement already satisfied: fastapi in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (from sagem
aker) (0.115.6)
Requirement already satisfied: google-pasta in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (from
sagemaker) (0.2.0)
Requirement already satisfied: importlib-metadata<7.0,>=1.4.0 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/si
te-packages (from sagemaker) (6.11.0)
Requirement already satisfied: jsonschema in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (from sa
gemaker) (4.23.0)
Requirement already satisfied: numpy<2.0,>=1.9.0 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages
(from sagemaker) (1.26.4)
Requirement already satisfied: omegaconf<2.3,>=2.2 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages
(from sagemaker) (2.2.3)
Requirement already satisfied: packaging>=20.0 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (fr
om sagemaker) (24.2)
Requirement already satisfied: pandas in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (from sagema
ker) (1.5.3)
Requirement already satisfied: pathos in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (from sagema
ker) (0.3.2)
Requirement already satisfied: platformdirs in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (from
sagemaker) (4.3.6)
Requirement already satisfied: protobuf<6.0,>=3.12 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages
(from sagemaker) (4.25.5)
Requirement already satisfied: psutil in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (from sagema
ker) (6.1.0)
Requirement already satisfied: pyyaml~=6.0 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (from s
agemaker) (6.0.2)
Requirement already satisfied: requests in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (from sage
maker) (2.32.3)
Requirement already satisfied: sagemaker-core<2.0.0,>=1.0.17 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/sit
e-packages (from sagemaker) (1.0.17)
Requirement already satisfied: schema in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (from sagema
```

```
ker) (0.7.7)
Requirement already satisfied: smdebug-rulesconfig==1.0.1 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-p
ackages (from sagemaker) (1.0.1)
Requirement already satisfied: tblib<4,>=1.7.0 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (fr
om sagemaker) (3.0.0)
Requirement already satisfied: tqdm in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (from sagemake
r) (4.67.0)
Requirement already satisfied: urllib3<3.0.0,>=1.26.8 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packa
ges (from sagemaker) (2.2.3)
Requirement already satisfied: uvicorn in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (from sagem
aker) (0.32.1)
Requirement already satisfied: filelock in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (from data
sets) (3.16.1)
Requirement already satisfied: pyarrow>=15.0.0 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (fr
om datasets) (17.0.0)
Requirement already satisfied: dill<0.3.9,>=0.3.0 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages
(from datasets) (0.3.8)
Requirement already satisfied: xxhash in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (from datase
ts) (3.5.0)
Requirement already satisfied: multiprocess<0.70.17 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-package
s (from datasets) (0.70.16)
Requirement already satisfied: fsspec<=2024.9.0,>=2023.1.0 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-
packages (from fsspec[http]<=2024.9.0,>=2023.1.0->datasets) (2024.9.0)
Requirement already satisfied: aiohttp in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (from datas
ets) (3.11.11)
Requirement already satisfied: huggingface-hub>=0.23.0 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-pack
ages (from datasets) (0.27.0)
Requirement already satisfied: botocore<1.36.0,>=1.35.76 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-pa
ckages (from boto3<2.0,>=1.35.75->sagemaker) (1.35.76)
Requirement already satisfied: jmespath<2.0.0,>=0.7.1 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packa
ges (from boto3<2.0,>=1.35.75->sagemaker) (1.0.1)
Requirement already satisfied: s3transfer<0.11.0,>=0.10.0 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-p
ackages (from boto3<2.0,>=1.35.75->sagemaker) (0.10.3)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-pack
ages (from aiohttp->datasets) (2.4.4)
Requirement already satisfied: aiosignal>=1.1.2 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (f
rom aiohttp->datasets) (1.3.2)
Requirement already satisfied: async-timeout<6.0,>=4.0 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-pack
ages (from aiohttp->datasets) (5.0.1)
Requirement already satisfied: frozenlist>=1.1.1 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages
(from aiohttp->datasets) (1.5.0)
```

```
Requirement already satisfied: multidict<7.0,>=4.5 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages
(from aiohttp->datasets) (6.1.0)
Requirement already satisfied: propcache>=0.2.0 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (f
rom aiohttp->datasets) (0.2.1)
Requirement already satisfied: yarl<2.0,>=1.17.0 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages
(from aiohttp->datasets) (1.18.3)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-p
ackages (from huggingface-hub>=0.23.0->datasets) (4.12.2)
Requirement already satisfied: zipp>=0.5 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (from imp
ortlib-metadata<7.0,>=1.4.0->sagemaker) (3.21.0)
Requirement already satisfied: antlr4-python3-runtime==4.9.* in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/sit
e-packages (from omegaconf<2.3,>=2.2->sagemaker) (4.9.3)
Requirement already satisfied: charset-normalizer<4,>=2 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-pac
kages (from requests->sagemaker) (3.4.0)
Requirement already satisfied: idna<4,>=2.5 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (from
requests->sagemaker) (3.10)
Requirement already satisfied: certifi>=2017.4.17 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages
(from requests->sagemaker) (2024.8.30)
Requirement already satisfied: pydantic<3.0.0,>=2.0.0 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packa
ges (from sagemaker-core<2.0.0,>=1.0.17->sagemaker) (2.9.2)
Requirement already satisfied: rich<14.0.0,>=13.0.0 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-package
s (from sagemaker-core<2.0.0,>=1.0.17->sagemaker) (13.9.4)
Requirement already satisfied: mock<5.0,>4.0 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (from
sagemaker-core<2.0.0,>=1.0.17->sagemaker) (4.0.3)
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python
3.10/site-packages (from jsonschema->sagemaker) (2024.10.1)
Requirement already satisfied: referencing>=0.28.4 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages
(from jsonschema->sagemaker) (0.35.1)
Requirement already satisfied: rpds-py>=0.7.1 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (fro
m jsonschema->sagemaker) (0.21.0)
Requirement already satisfied: starlette<0.42.0,>=0.40.0 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-pa
ckages (from fastapi->sagemaker) (0.41.3)
Requirement already satisfied: six in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (from google-pa
sta->sagemaker) (1.16.0)
Requirement already satisfied: python-dateutil>=2.8.1 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packa
ges (from pandas->sagemaker) (2.9.0)
Requirement already satisfied: pytz>=2020.1 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (from
pandas->sagemaker) (2024.1)
Requirement already satisfied: ppft>=1.7.6.8 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (from
pathos->sagemaker) (1.7.6.9)
Requirement already satisfied: pox>=0.3.4 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (from pa
```

```
thos->sagemaker) (0.3.5)
Requirement already satisfied: click>=7.0 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (from uv
icorn->sagemaker) (8.1.7)
Requirement already satisfied: h11>=0.8 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (from uvic
orn->sagemaker) (0.14.0)
Requirement already satisfied: annotated-types>=0.6.0 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packa
ges (from pydantic<3.0.0,>=2.0.0->sagemaker-core<2.0.0,>=1.0.17->sagemaker) (0.7.0)
Requirement already satisfied: pydantic-core==2.23.4 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packag
es (from pydantic<3.0.0,>=2.0.0->sagemaker-core<2.0.0,>=1.0.17->sagemaker) (2.23.4)
Requirement already satisfied: markdown-it-py>=2.2.0 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packag
es (from rich<14.0.0,>=13.0.0->sagemaker-core<2.0.0,>=1.0.17->sagemaker) (3.0.0)
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-pack
ages (from rich<14.0.0,>=13.0.0->sagemaker-core<2.0.0,>=1.0.17->sagemaker) (2.18.0)
Requirement already satisfied: anyio<5,>=3.4.0 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (fr
om starlette<0.42.0,>=0.40.0->fastapi->sagemaker) (4.6.2.post1)
Requirement already satisfied: sniffio>=1.1 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (from
anyio<5,>=3.4.0->starlette<0.42.0,>=0.40.0->fastapi->sagemaker) (1.3.1)
Requirement already satisfied: exceptiongroup>=1.0.2 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packag
es (from anyio<5,>=3.4.0->starlette<0.42.0,>=0.40.0->fastapi->sagemaker) (1.2.2)
Requirement already satisfied: mdurl~=0.1 in /home/ec2-user/anaconda3/envs/pytorch_p310/lib/python3.10/site-packages (from ma
rkdown-it-py>=2.2.0->rich<14.0.0,>=13.0.0->sagemaker-core<2.0.0,>=1.0.17->sagemaker) (0.1.2)
```

Select the model to fine-tune

In [7]:
```
model_id, model_version = "meta-textgeneration-llama-2-7b", "2.*"
```

In the cell below, choose the training dataset text for the domain you've chosen and update the code in the cell below:

To create a finance domain expert model:

- `"training": f"s3://genaiwithawsproject2024/training-datasets/finance"`

To create a medical domain expert model:

- `"training": f"s3://genaiwithawsproject2024/training-datasets/medical"`

To create an IT domain expert model:

- `"training": f"s3://genaiwithawsproject2024/training-datasets/it"`

Deploy the fine-tuned model

Next, we deploy the domain fine-tuned model. We will compare the performance of the fine-tuned and pre-trained model.

In [8]:
```
from sagemaker.jumpstart.estimator import JumpStartEstimator
import boto3

estimator = JumpStartEstimator(model_id=model_id,  environment={"accept_eula": "true"},instance_type = "ml.g5.2xlarge")

estimator.set_hyperparameters(instruction_tuned=False, epoch=5)

#Fill in the code below with the dataset you want to use from above
estimator.fit({"training": f"s3://genaiwithawsprojectstarter/financialDataset"})
```

```
[12/24/24 18:54:28] INFO     SageMaker Python SDK will collect telemetry to help us better  telemetry_logging.py:90
                             understand our user's needs, diagnose issues, and deliver
                             additional features.
                             To opt out of telemetry, please disable via TelemetryOptOut
                             parameter in SDK defaults config. For more information, refer
                             to
                             https://sagemaker.readthedocs.io/en/stable/overview.html#confi
                             guring-and-using-defaults-with-the-sagemaker-python-sdk.
```

```
                    INFO     Found credentials from IAM Role:                                credentials.py:1075
                             BaseNotebookInstanceEc2InstanceRole
[12/24/24 18:54:29] INFO     Creating training-job with name:                                    session.py:1042
                             meta-textgeneration-llama-2-7b-2024-12-24-18-54-28-800
```

```
2024-12-24 18:54:29 Starting - Starting the training job
2024-12-24 18:54:29 Pending - Training job waiting for capacity......
2024-12-24 18:55:30 Downloading - Downloading input data.............................
2024-12-24 19:00:23 Training - Training image download completed. Training in progress.bash: cannot set terminal process grou
p (-1): Inappropriate ioctl for device
bash: no job control in this shell
2024-12-24 19:00:26,120 sagemaker-training-toolkit INFO     Imported framework sagemaker_pytorch_container.training
2024-12-24 19:00:26,139 sagemaker-training-toolkit INFO     No Neurons detected (normal if no neurons installed)
2024-12-24 19:00:26,148 sagemaker_pytorch_container.training INFO     Block until all host DNS lookups succeed.
2024-12-24 19:00:26,151 sagemaker_pytorch_container.training INFO     Invoking user training script.
2024-12-24 19:00:35,491 sagemaker-training-toolkit INFO     Installing dependencies from requirements.txt:
/opt/conda/bin/python3.10 -m pip install -r requirements.txt
Processing ./lib/accelerate/accelerate-0.33.0-py3-none-any.whl (from -r requirements.txt (line 1))
Processing ./lib/bitsandbytes/bitsandbytes-0.39.1-py3-none-any.whl (from -r requirements.txt (line 2))
Processing ./lib/black/black-23.7.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (from -r requirements.txt (lin
e 3))
Processing ./lib/brotli/Brotli-1.0.9-cp310-cp310-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_12_x86_64.manylinux2010_x
86_64.whl (from -r requirements.txt (line 4))
Processing ./lib/datasets/datasets-2.14.1-py3-none-any.whl (from -r requirements.txt (line 5))
Processing ./lib/docstring-parser/docstring_parser-0.16-py3-none-any.whl (from -r requirements.txt (line 6))
Processing ./lib/fire/fire-0.5.0.tar.gz
Preparing metadata (setup.py): started
Preparing metadata (setup.py): finished with status 'done'
Processing ./lib/huggingface-hub/huggingface_hub-0.24.2-py3-none-any.whl (from -r requirements.txt (line 8))
Processing ./lib/inflate64/inflate64-0.3.1-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (from -r requirements.t
xt (line 9))
Processing ./lib/loralib/loralib-0.1.1-py3-none-any.whl (from -r requirements.txt (line 10))
Processing ./lib/multivolumefile/multivolumefile-0.2.3-py3-none-any.whl (from -r requirements.txt (line 11))
Processing ./lib/mypy-extensions/mypy_extensions-1.0.0-py3-none-any.whl (from -r requirements.txt (line 12))
Processing ./lib/nvidia-cublas-cu12/nvidia_cublas_cu12-12.1.3.1-py3-none-manylinux1_x86_64.whl (from -r requirements.txt (lin
e 13))
Processing ./lib/nvidia-cuda-cupti-cu12/nvidia_cuda_cupti_cu12-12.1.105-py3-none-manylinux1_x86_64.whl (from -r requirements.
txt (line 14))
Processing ./lib/nvidia-cuda-nvrtc-cu12/nvidia_cuda_nvrtc_cu12-12.1.105-py3-none-manylinux1_x86_64.whl (from -r requirements.
txt (line 15))
Processing ./lib/nvidia-cuda-runtime-cu12/nvidia_cuda_runtime_cu12-12.1.105-py3-none-manylinux1_x86_64.whl (from -r requireme
nts.txt (line 16))
Processing ./lib/nvidia-cudnn-cu12/nvidia_cudnn_cu12-8.9.2.26-py3-none-manylinux1_x86_64.whl (from -r requirements.txt (line
17))
Processing ./lib/nvidia-cufft-cu12/nvidia_cufft_cu12-11.0.2.54-py3-none-manylinux1_x86_64.whl (from -r requirements.txt (line
18))
```

```
Processing ./lib/nvidia-curand-cu12/nvidia_curand_cu12-10.3.2.106-py3-none-manylinux1_x86_64.whl (from -r requirements.txt (l
ine 19))
Processing ./lib/nvidia-cusolver-cu12/nvidia_cusolver_cu12-11.4.5.107-py3-none-manylinux1_x86_64.whl (from -r requirements.tx
t (line 20))
Processing ./lib/nvidia-cusparse-cu12/nvidia_cusparse_cu12-12.1.0.106-py3-none-manylinux1_x86_64.whl (from -r requirements.tx
t (line 21))
Processing ./lib/nvidia-nccl-cu12/nvidia_nccl_cu12-2.19.3-py3-none-manylinux1_x86_64.whl (from -r requirements.txt (line 22))
Processing ./lib/nvidia-nvjitlink-cu12/nvidia_nvjitlink_cu12-12.3.101-py3-none-manylinux1_x86_64.whl (from -r requirements.tx
t (line 23))
Processing ./lib/nvidia-nvtx-cu12/nvidia_nvtx_cu12-12.1.105-py3-none-manylinux1_x86_64.whl (from -r requirements.txt (line 2
4))
Processing ./lib/pathspec/pathspec-0.11.1-py3-none-any.whl (from -r requirements.txt (line 25))
Processing ./lib/peft/peft-0.4.0-py3-none-any.whl (from -r requirements.txt (line 26))
Processing ./lib/py7zr/py7zr-0.20.5-py3-none-any.whl (from -r requirements.txt (line 27))
Processing ./lib/pybcj/pybcj-1.0.1-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (from -r requirements.txt (line
28))
Processing ./lib/pycryptodomex/pycryptodomex-3.18.0-cp35-abi3-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (from -r require
ments.txt (line 29))
Processing ./lib/pyppmd/pyppmd-1.0.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (from -r requirements.txt (li
ne 30))
Processing ./lib/pyzstd/pyzstd-0.15.9-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (from -r requirements.txt (l
ine 31))
Processing ./lib/safetensors/safetensors-0.4.2-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (from -r requiremen
ts.txt (line 32))
Processing ./lib/scipy/scipy-1.11.1-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (from -r requirements.txt (lin
e 33))
Processing ./lib/shtab/shtab-1.7.1-py3-none-any.whl (from -r requirements.txt (line 34))
Processing ./lib/termcolor/termcolor-2.3.0-py3-none-any.whl (from -r requirements.txt (line 35))
Processing ./lib/texttable/texttable-1.6.7-py2.py3-none-any.whl (from -r requirements.txt (line 36))
Processing ./lib/tokenize-rt/tokenize_rt-5.1.0-py2.py3-none-any.whl (from -r requirements.txt (line 37))
Processing ./lib/tokenizers/tokenizers-0.19.1-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (from -r requirement
s.txt (line 38))
Processing ./lib/torch/torch-2.2.0-cp310-cp310-manylinux1_x86_64.whl (from -r requirements.txt (line 39))
Processing ./lib/transformers/transformers-4.43.1-py3-none-any.whl (from -r requirements.txt (line 40))
Processing ./lib/triton/triton-2.2.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (from -r requirements.txt (li
ne 41))
Processing ./lib/trl/trl-0.8.1-py3-none-any.whl (from -r requirements.txt (line 42))
Processing ./lib/typing-extensions/typing_extensions-4.8.0-py3-none-any.whl (from -r requirements.txt (line 43))
Processing ./lib/tyro/tyro-0.7.3-py3-none-any.whl (from -r requirements.txt (line 44))
Processing ./lib/sagemaker_jumpstart_script_utilities/sagemaker_jumpstart_script_utilities-1.1.9-py2.py3-none-any.whl (from -
r requirements.txt (line 45))
```

```
Processing ./lib/sagemaker_jumpstart_huggingface_script_utilities/sagemaker_jumpstart_huggingface_script_utilities-1.2.7-py2.
py3-none-any.whl (from -r requirements.txt (line 46))
Requirement already satisfied: numpy<2.0.0,>=1.17 in /opt/conda/lib/python3.10/site-packages (from accelerate==0.33.0->-r req
uirements.txt (line 1)) (1.24.4)
Requirement already satisfied: packaging>=20.0 in /opt/conda/lib/python3.10/site-packages (from accelerate==0.33.0->-r requir
ements.txt (line 1)) (23.1)
Requirement already satisfied: psutil in /opt/conda/lib/python3.10/site-packages (from accelerate==0.33.0->-r requirements.tx
t (line 1)) (5.9.5)
Requirement already satisfied: pyyaml in /opt/conda/lib/python3.10/site-packages (from accelerate==0.33.0->-r requirements.tx
t (line 1)) (6.0)
Requirement already satisfied: click>=8.0.0 in /opt/conda/lib/python3.10/site-packages (from black==23.7.0->-r requirements.t
xt (line 3)) (8.1.4)
Requirement already satisfied: platformdirs>=2 in /opt/conda/lib/python3.10/site-packages (from black==23.7.0->-r requirement
s.txt (line 3)) (3.8.1)
Requirement already satisfied: tomli>=1.1.0 in /opt/conda/lib/python3.10/site-packages (from black==23.7.0->-r requirements.t
xt (line 3)) (2.0.1)
Requirement already satisfied: pyarrow>=8.0.0 in /opt/conda/lib/python3.10/site-packages (from datasets==2.14.1->-r requireme
nts.txt (line 5)) (14.0.2)
Requirement already satisfied: dill<0.3.8,>=0.3.0 in /opt/conda/lib/python3.10/site-packages (from datasets==2.14.1->-r requi
rements.txt (line 5)) (0.3.6)
Requirement already satisfied: pandas in /opt/conda/lib/python3.10/site-packages (from datasets==2.14.1->-r requirements.txt
(line 5)) (2.0.3)
Requirement already satisfied: requests>=2.19.0 in /opt/conda/lib/python3.10/site-packages (from datasets==2.14.1->-r require
ments.txt (line 5)) (2.31.0)
Requirement already satisfied: tqdm>=4.62.1 in /opt/conda/lib/python3.10/site-packages (from datasets==2.14.1->-r requirement
s.txt (line 5)) (4.65.0)
Requirement already satisfied: xxhash in /opt/conda/lib/python3.10/site-packages (from datasets==2.14.1->-r requirements.txt
(line 5)) (3.4.1)
Requirement already satisfied: multiprocess in /opt/conda/lib/python3.10/site-packages (from datasets==2.14.1->-r requirement
s.txt (line 5)) (0.70.14)
Requirement already satisfied: fsspec>=2021.11.1 in /opt/conda/lib/python3.10/site-packages (from fsspec[http]>=2021.11.1->da
tasets==2.14.1->-r requirements.txt (line 5)) (2023.6.0)
Requirement already satisfied: aiohttp in /opt/conda/lib/python3.10/site-packages (from datasets==2.14.1->-r requirements.txt
(line 5)) (3.9.3)
Requirement already satisfied: six in /opt/conda/lib/python3.10/site-packages (from fire==0.5.0->-r requirements.txt (line
7)) (1.16.0)
Requirement already satisfied: filelock in /opt/conda/lib/python3.10/site-packages (from huggingface-hub==0.24.2->-r requirem
ents.txt (line 8)) (3.12.2)
Requirement already satisfied: sympy in /opt/conda/lib/python3.10/site-packages (from torch==2.2.0->-r requirements.txt (line
39)) (1.12)
Requirement already satisfied: networkx in /opt/conda/lib/python3.10/site-packages (from torch==2.2.0->-r requirements.txt (l
```

```
ine 39)) (3.1)
Requirement already satisfied: jinja2 in /opt/conda/lib/python3.10/site-packages (from torch==2.2.0->-r requirements.txt (lin
e 39)) (3.1.2)
Requirement already satisfied: regex!=2019.12.17 in /opt/conda/lib/python3.10/site-packages (from transformers==4.43.1->-r re
quirements.txt (line 40)) (2023.12.25)
Requirement already satisfied: rich>=11.1.0 in /opt/conda/lib/python3.10/site-packages (from tyro==0.7.3->-r requirements.txt
(line 44)) (13.4.2)
Requirement already satisfied: aiosignal>=1.1.2 in /opt/conda/lib/python3.10/site-packages (from aiohttp->datasets==2.14.1->-
r requirements.txt (line 5)) (1.3.1)
Requirement already satisfied: attrs>=17.3.0 in /opt/conda/lib/python3.10/site-packages (from aiohttp->datasets==2.14.1->-r r
equirements.txt (line 5)) (23.1.0)
Requirement already satisfied: frozenlist>=1.1.1 in /opt/conda/lib/python3.10/site-packages (from aiohttp->datasets==2.14.1->
-r requirements.txt (line 5)) (1.4.1)
Requirement already satisfied: multidict<7.0,>=4.5 in /opt/conda/lib/python3.10/site-packages (from aiohttp->datasets==2.14.1
->-r requirements.txt (line 5)) (6.0.5)
Requirement already satisfied: yarl<2.0,>=1.0 in /opt/conda/lib/python3.10/site-packages (from aiohttp->datasets==2.14.1->-r
requirements.txt (line 5)) (1.9.4)
Requirement already satisfied: async-timeout<5.0,>=4.0 in /opt/conda/lib/python3.10/site-packages (from aiohttp->datasets==2.
14.1->-r requirements.txt (line 5)) (4.0.3)
Requirement already satisfied: charset-normalizer<4,>=2 in /opt/conda/lib/python3.10/site-packages (from requests>=2.19.0->da
tasets==2.14.1->-r requirements.txt (line 5)) (3.1.0)
Requirement already satisfied: idna<4,>=2.5 in /opt/conda/lib/python3.10/site-packages (from requests>=2.19.0->datasets==2.1
4.1->-r requirements.txt (line 5)) (3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in /opt/conda/lib/python3.10/site-packages (from requests>=2.19.0->datasets
==2.14.1->-r requirements.txt (line 5)) (1.26.15)
Requirement already satisfied: certifi>=2017.4.17 in /opt/conda/lib/python3.10/site-packages (from requests>=2.19.0->datasets
==2.14.1->-r requirements.txt (line 5)) (2024.2.2)
Requirement already satisfied: markdown-it-py>=2.2.0 in /opt/conda/lib/python3.10/site-packages (from rich>=11.1.0->tyro==0.
7.3->-r requirements.txt (line 44)) (3.0.0)
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in /opt/conda/lib/python3.10/site-packages (from rich>=11.1.0->tyro==
0.7.3->-r requirements.txt (line 44)) (2.15.1)
Requirement already satisfied: MarkupSafe>=2.0 in /opt/conda/lib/python3.10/site-packages (from jinja2->torch==2.2.0->-r requ
irements.txt (line 39)) (2.1.3)
Requirement already satisfied: python-dateutil>=2.8.2 in /opt/conda/lib/python3.10/site-packages (from pandas->datasets==2.1
4.1->-r requirements.txt (line 5)) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /opt/conda/lib/python3.10/site-packages (from pandas->datasets==2.14.1->-r req
uirements.txt (line 5)) (2023.3)
Requirement already satisfied: tzdata>=2022.1 in /opt/conda/lib/python3.10/site-packages (from pandas->datasets==2.14.1->-r r
equirements.txt (line 5)) (2023.3)
Requirement already satisfied: mpmath>=0.19 in /opt/conda/lib/python3.10/site-packages (from sympy->torch==2.2.0->-r requirem
ents.txt (line 39)) (1.3.0)
```

```
Requirement already satisfied: mdurl~=0.1 in /opt/conda/lib/python3.10/site-packages (from markdown-it-py>=2.2.0->rich>=11.1.
0->tyro==0.7.3->-r requirements.txt (line 44)) (0.1.0)
scipy is already installed with the same version as the provided wheel. Use --force-reinstall to force an installation of the
wheel.
Building wheels for collected packages: fire
Building wheel for fire (setup.py): started
Building wheel for fire (setup.py): finished with status 'done'
Created wheel for fire: filename=fire-0.5.0-py2.py3-none-any.whl size=116932 sha256=8199004eee9fa68709465a4d51c0596fb8e1b6a39
c8aa617223d2db3a322efcf
Stored in directory: /root/.cache/pip/wheels/db/3d/41/7e69dca5f61e37d109a4457082ffc5c6edb55ab633bafded38
Successfully built fire
Installing collected packages: texttable, Brotli, bitsandbytes, typing-extensions, triton, tokenize-rt, termcolor, shtab, sag
emaker-jumpstart-script-utilities, sagemaker-jumpstart-huggingface-script-utilities, safetensors, pyzstd, pyppmd, pycryptodom
ex, pybcj, pathspec, nvidia-nvtx-cu12, nvidia-nvjitlink-cu12, nvidia-nccl-cu12, nvidia-curand-cu12, nvidia-cufft-cu12, nvidia
-cuda-runtime-cu12, nvidia-cuda-nvrtc-cu12, nvidia-cuda-cupti-cu12, nvidia-cublas-cu12, mypy-extensions, multivolumefile, lor
alib, inflate64, docstring-parser, py7zr, nvidia-cusparse-cu12, nvidia-cudnn-cu12, huggingface-hub, fire, black, tyro, tokeni
zers, nvidia-cusolver-cu12, transformers, torch, datasets, accelerate, trl, peft
Attempting uninstall: typing-extensions
Found existing installation: typing_extensions 4.7.1
Uninstalling typing_extensions-4.7.1:
Successfully uninstalled typing_extensions-4.7.1
Attempting uninstall: triton
Found existing installation: triton 2.0.0.dev20221202
Uninstalling triton-2.0.0.dev20221202:
Successfully uninstalled triton-2.0.0.dev20221202
Attempting uninstall: huggingface-hub
Found existing installation: huggingface-hub 0.20.3
Uninstalling huggingface-hub-0.20.3:
Successfully uninstalled huggingface-hub-0.20.3
Attempting uninstall: tokenizers
Found existing installation: tokenizers 0.13.3
Uninstalling tokenizers-0.13.3:
Successfully uninstalled tokenizers-0.13.3
Attempting uninstall: transformers
Found existing installation: transformers 4.28.1
Uninstalling transformers-4.28.1:
Successfully uninstalled transformers-4.28.1
Attempting uninstall: torch
Found existing installation: torch 2.0.0
Uninstalling torch-2.0.0:
Successfully uninstalled torch-2.0.0
```

```
Attempting uninstall: datasets
Found existing installation: datasets 2.16.1
Uninstalling datasets-2.16.1:
Successfully uninstalled datasets-2.16.1
Attempting uninstall: accelerate
Found existing installation: accelerate 0.19.0
Uninstalling accelerate-0.19.0:
Successfully uninstalled accelerate-0.19.0
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is
the source of the following dependency conflicts.
fastai 2.7.12 requires torch<2.1,>=1.7, but you have torch 2.2.0 which is incompatible.
Successfully installed Brotli-1.0.9 accelerate-0.33.0 bitsandbytes-0.39.1 black-23.7.0 datasets-2.14.1 docstring-parser-0.16
fire-0.5.0 huggingface-hub-0.24.2 inflate64-0.3.1 loralib-0.1.1 multivolumefile-0.2.3 mypy-extensions-1.0.0 nvidia-cublas-cu1
2-12.1.3.1 nvidia-cuda-cupti-cu12-12.1.105 nvidia-cuda-nvrtc-cu12-12.1.105 nvidia-cuda-runtime-cu12-12.1.105 nvidia-cudnn-cu1
2-8.9.2.26 nvidia-cufft-cu12-11.0.2.54 nvidia-curand-cu12-10.3.2.106 nvidia-cusolver-cu12-11.4.5.107 nvidia-cusparse-cu12-12.
1.0.106 nvidia-nccl-cu12-2.19.3 nvidia-nvjitlink-cu12-12.3.101 nvidia-nvtx-cu12-12.1.105 pathspec-0.11.1 peft-0.4.0 py7zr-0.2
0.5 pybcj-1.0.1 pycryptodomex-3.18.0 pyppmd-1.0.0 pyzstd-0.15.9 safetensors-0.4.2 sagemaker-jumpstart-huggingface-script-util
ities-1.2.7 sagemaker-jumpstart-script-utilities-1.1.9 shtab-1.7.1 termcolor-2.3.0 texttable-1.6.7 tokenize-rt-5.1.0 tokenize
rs-0.19.1 torch-2.2.0 transformers-4.43.1 triton-2.2.0 trl-0.8.1 typing-extensions-4.8.0 tyro-0.7.3
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package ma
nager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv
2024-12-24 19:01:51,412 sagemaker-training-toolkit INFO     Waiting for the process to finish and give a return code.
2024-12-24 19:01:51,412 sagemaker-training-toolkit INFO     Done waiting for a return code. Received 0 from exiting process.
2024-12-24 19:01:51,451 sagemaker-training-toolkit INFO     No Neurons detected (normal if no neurons installed)
2024-12-24 19:01:51,481 sagemaker-training-toolkit INFO     No Neurons detected (normal if no neurons installed)
2024-12-24 19:01:51,510 sagemaker-training-toolkit INFO     No Neurons detected (normal if no neurons installed)
2024-12-24 19:01:51,520 sagemaker-training-toolkit INFO     Invoking user script
Training Env:
{
    "additional_framework_parameters": {},
    "channel_input_dirs": {
        "code": "/opt/ml/input/data/code",
        "training": "/opt/ml/input/data/training"
    },
    "current_host": "algo-1",
    "current_instance_group": "homogeneousCluster",
    "current_instance_group_hosts": [
        "algo-1"
    ],
    "current_instance_type": "ml.g5.2xlarge",
    "distribution_hosts": [],
```

```
"distribution_instance_groups": [],
"framework_module": "sagemaker_pytorch_container.training:main",
"hosts": [
    "algo-1"
],
"hyperparameters": {
    "add_input_output_demarcation_key": "True",
    "chat_dataset": "False",
    "enable_fsdp": "True",
    "epoch": 5,
    "instruction_tuned": false,
    "int8_quantization": "False",
    "learning_rate": "0.0001",
    "lora_alpha": "32",
    "lora_dropout": "0.05",
    "lora_r": "8",
    "max_input_length": "-1",
    "max_train_samples": "-1",
    "max_val_samples": "-1",
    "per_device_eval_batch_size": "1",
    "per_device_train_batch_size": "4",
    "preprocessing_num_workers": "None",
    "seed": "10",
    "target_modules": "q_proj,v_proj",
    "train_data_split_seed": "0",
    "validation_split_ratio": "0.2"
},
"input_config_dir": "/opt/ml/input/config",
"input_data_config": {
    "code": {
        "TrainingInputMode": "File",
        "S3DistributionType": "FullyReplicated",
        "RecordWrapperType": "None"
    },
    "training": {
        "TrainingInputMode": "File",
        "S3DistributionType": "FullyReplicated",
        "RecordWrapperType": "None"
    }
},
"input_dir": "/opt/ml/input",
```

```
"instance_groups": [
    "homogeneousCluster"
],
"instance_groups_dict": {
    "homogeneousCluster": {
        "instance_group_name": "homogeneousCluster",
        "instance_type": "ml.g5.2xlarge",
        "hosts": [
            "algo-1"
        ]
    }
},
"is_hetero": false,
"is_master": true,
"is_modelparallel_enabled": null,
"is_smddpmprun_installed": true,
"job_name": "meta-textgeneration-llama-2-7b-2024-12-24-18-54-28-800",
"log_level": 20,
"master_hostname": "algo-1",
"model_dir": "/opt/ml/model",
"module_dir": "/opt/ml/input/data/code/sourcedir.tar.gz",
"module_name": "transfer_learning",
"network_interface_name": "eth0",
"num_cpus": 8,
"num_gpus": 1,
"num_neurons": 0,
"output_data_dir": "/opt/ml/output/data",
"output_dir": "/opt/ml/output",
"output_intermediate_dir": "/opt/ml/output/intermediate",
"resource_config": {
    "current_host": "algo-1",
    "current_instance_type": "ml.g5.2xlarge",
    "current_group_name": "homogeneousCluster",
    "hosts": [
        "algo-1"
    ],
    "instance_groups": [
        {
            "instance_group_name": "homogeneousCluster",
            "instance_type": "ml.g5.2xlarge",
            "hosts": [
```

```
                    "algo-1"
                ]
            }
        ],
        "network_interface_name": "eth0"
    },
    "user_entry_point": "transfer_learning.py"
}
Environment variables:
SM_HOSTS=["algo-1"]
SM_NETWORK_INTERFACE_NAME=eth0
SM_HPS={"add_input_output_demarcation_key":"True","chat_dataset":"False","enable_fsdp":"True","epoch":5,"instruction_tuned":f
alse,"int8_quantization":"False","learning_rate":"0.0001","lora_alpha":"32","lora_dropout":"0.05","lora_r":"8","max_input_len
gth":"-1","max_train_samples":"-1","max_val_samples":"-1","per_device_eval_batch_size":"1","per_device_train_batch_siz
e":"4","preprocessing_num_workers":"None","seed":"10","target_modules":"q_proj,v_proj","train_data_split_seed":"0","validatio
n_split_ratio":"0.2"}
SM_USER_ENTRY_POINT=transfer_learning.py
SM_FRAMEWORK_PARAMS={}
SM_RESOURCE_CONFIG={"current_group_name":"homogeneousCluster","current_host":"algo-1","current_instance_type":"ml.g5.2xlarg
e","hosts":["algo-1"],"instance_groups":[{"hosts":["algo-1"],"instance_group_name":"homogeneousCluster","instance_type":"ml.g
5.2xlarge"}],"network_interface_name":"eth0"}
SM_INPUT_DATA_CONFIG={"code":{"RecordWrapperType":"None","S3DistributionType":"FullyReplicated","TrainingInputMode":"Fil
e"},"training":{"RecordWrapperType":"None","S3DistributionType":"FullyReplicated","TrainingInputMode":"File"}}
SM_OUTPUT_DATA_DIR=/opt/ml/output/data
SM_CHANNELS=["code","training"]
SM_CURRENT_HOST=algo-1
SM_CURRENT_INSTANCE_TYPE=ml.g5.2xlarge
SM_CURRENT_INSTANCE_GROUP=homogeneousCluster
SM_CURRENT_INSTANCE_GROUP_HOSTS=["algo-1"]
SM_INSTANCE_GROUPS=["homogeneousCluster"]
SM_INSTANCE_GROUPS_DICT={"homogeneousCluster":{"hosts":["algo-1"],"instance_group_name":"homogeneousCluster","instance_typ
e":"ml.g5.2xlarge"}}
SM_DISTRIBUTION_INSTANCE_GROUPS=[]
SM_IS_HETERO=false
SM_MODULE_NAME=transfer_learning
SM_LOG_LEVEL=20
SM_FRAMEWORK_MODULE=sagemaker_pytorch_container.training:main
SM_INPUT_DIR=/opt/ml/input
SM_INPUT_CONFIG_DIR=/opt/ml/input/config
SM_OUTPUT_DIR=/opt/ml/output
SM_NUM_CPUS=8
```

```
SM_NUM_GPUS=1
SM_NUM_NEURONS=0
SM_MODEL_DIR=/opt/ml/model
SM_MODULE_DIR=/opt/ml/input/data/code/sourcedir.tar.gz
SM_TRAINING_ENV={"additional_framework_parameters":{},"channel_input_dirs":{"code":"/opt/ml/input/data/code","training":"/opt/ml/input/data/training"},"current_host":"algo-1","current_instance_group":"homogeneousCluster","current_instance_group_hosts":["algo-1"],"current_instance_type":"ml.g5.2xlarge","distribution_hosts":[],"distribution_instance_groups":[],"framework_module":"sagemaker_pytorch_container.training:main","hosts":["algo-1"],"hyperparameters":{"add_input_output_demarcation_key":"True","chat_dataset":"False","enable_fsdp":"True","epoch":5,"instruction_tuned":false,"int8_quantization":"False","learning_rate":"0.0001","lora_alpha":"32","lora_dropout":"0.05","lora_r":"8","max_input_length":"-1","max_train_samples":"-1","max_val_samples":"-1","per_device_eval_batch_size":"1","per_device_train_batch_size":"4","preprocessing_num_workers":"None","seed":"10","target_modules":"q_proj,v_proj","train_data_split_seed":"0","validation_split_ratio":"0.2"},"input_config_dir":"/opt/ml/input/config","input_data_config":{"code":{"RecordWrapperType":"None","S3DistributionType":"FullyReplicated","TrainingInputMode":"File"},"training":{"RecordWrapperType":"None","S3DistributionType":"FullyReplicated","TrainingInputMode":"File"}},"input_dir":"/opt/ml/input","instance_groups":["homogeneousCluster"],"instance_groups_dict":{"homogeneousCluster":{"hosts":["algo-1"],"instance_group_name":"homogeneousCluster","instance_type":"ml.g5.2xlarge"}},"is_hetero":false,"is_master":true,"is_model_parallel_enabled":null,"is_smddpmprun_installed":true,"job_name":"meta-textgeneration-llama-2-7b-2024-12-24-18-54-28-800","log_level":20,"master_hostname":"algo-1","model_dir":"/opt/ml/model","module_dir":"/opt/ml/input/data/code/sourcedir.tar.gz","module_name":"transfer_learning","network_interface_name":"eth0","num_cpus":8,"num_gpus":1,"num_neurons":0,"output_data_dir":"/opt/ml/output/data","output_dir":"/opt/ml/output","output_intermediate_dir":"/opt/ml/output/intermediate","resource_config":{"current_group_name":"homogeneousCluster","current_host":"algo-1","current_instance_type":"ml.g5.2xlarge","hosts":["algo-1"],"instance_groups":[{"hosts":["algo-1"],"instance_group_name":"homogeneousCluster","instance_type":"ml.g5.2xlarge"}],"network_interface_name":"eth0"},"user_entry_point":"transfer_learning.py"}
SM_USER_ARGS=["--add_input_output_demarcation_key","True","--chat_dataset","False","--enable_fsdp","True","--epoch","5","--instruction_tuned","False","--int8_quantization","False","--learning_rate","0.0001","--lora_alpha","32","--lora_dropout","0.05","--lora_r","8","--max_input_length","-1","--max_train_samples","-1","--max_val_samples","-1","--per_device_eval_batch_size","1","--per_device_train_batch_size","4","--preprocessing_num_workers","None","--seed","10","--target_modules","q_proj,v_proj","--train_data_split_seed","0","--validation_split_ratio","0.2"]
SM_OUTPUT_INTERMEDIATE_DIR=/opt/ml/output/intermediate
SM_CHANNEL_CODE=/opt/ml/input/data/code
SM_CHANNEL_TRAINING=/opt/ml/input/data/training
SM_HP_ADD_INPUT_OUTPUT_DEMARCATION_KEY=True
SM_HP_CHAT_DATASET=False
SM_HP_ENABLE_FSDP=True
SM_HP_EPOCH=5
SM_HP_INSTRUCTION_TUNED=false
SM_HP_INT8_QUANTIZATION=False
SM_HP_LEARNING_RATE=0.0001
SM_HP_LORA_ALPHA=32
SM_HP_LORA_DROPOUT=0.05
SM_HP_LORA_R=8
```

```
SM_HP_MAX_INPUT_LENGTH=-1
SM_HP_MAX_TRAIN_SAMPLES=-1
SM_HP_MAX_VAL_SAMPLES=-1
SM_HP_PER_DEVICE_EVAL_BATCH_SIZE=1
SM_HP_PER_DEVICE_TRAIN_BATCH_SIZE=4
SM_HP_PREPROCESSING_NUM_WORKERS=None
SM_HP_SEED=10
SM_HP_TARGET_MODULES=q_proj,v_proj
SM_HP_TRAIN_DATA_SPLIT_SEED=0
SM_HP_VALIDATION_SPLIT_RATIO=0.2
PYTHONPATH=/opt/ml/code:/opt/conda/bin:/opt/conda/lib/python310.zip:/opt/conda/lib/python3.10:/opt/conda/lib/python3.10/lib-d
ynload:/opt/conda/lib/python3.10/site-packages
Invoking script with the following command:
/opt/conda/bin/python3.10 transfer_learning.py --add_input_output_demarcation_key True --chat_dataset False --enable_fsdp Tru
e --epoch 5 --instruction_tuned False --int8_quantization False --learning_rate 0.0001 --lora_alpha 32 --lora_dropout 0.05 --
lora_r 8 --max_input_length -1 --max_train_samples -1 --max_val_samples -1 --per_device_eval_batch_size 1 --per_device_train_
batch_size 4 --preprocessing_num_workers None --seed 10 --target_modules q_proj,v_proj --train_data_split_seed 0 --validation
_split_ratio 0.2
2024-12-24 19:01:51,559 sagemaker-training-toolkit INFO     Exceptions not imported for SageMaker TF as Tensorflow is not ins
talled.
================================BUG REPORT================================
Welcome to bitsandbytes. For bug reports, please run
python -m bitsandbytes
 and submit this information together with your error trace to: https://github.com/TimDettmers/bitsandbytes/issues
==========================================================================
bin /opt/conda/lib/python3.10/site-packages/bitsandbytes/libbitsandbytes_cuda118.so
/opt/conda/lib/python3.10/site-packages/bitsandbytes/cuda_setup/main.py:149: UserWarning: WARNING: The following directories
listed in your path were found to be non-existent: {PosixPath('/usr/local/nvidia/lib64'), PosixPath('/usr/local/nvidia/lib')}
  warn(msg)
CUDA SETUP: CUDA runtime path found: /opt/conda/lib/libcudart.so.11.0
CUDA SETUP: Highest compute capability among GPUs detected: 8.6
CUDA SETUP: Detected CUDA version 118
CUDA SETUP: Loading binary /opt/conda/lib/python3.10/site-packages/bitsandbytes/libbitsandbytes_cuda118.so...
INFO:root:Using pre-trained artifacts in SAGEMAKER_ADDITIONAL_S3_DATA_PATH=/opt/ml/additonals3data
INFO:root:Identify file serving.properties in the un-tar directory /opt/ml/additonals3data. Copying it over to /opt/ml/model
for model deployment after training is finished.
INFO:root:Invoking the training command ['torchrun', '--nnodes', '1', '--nproc_per_node', '1', 'llama_finetuning.py', '--mode
l_name', '/opt/ml/additonals3data', '--num_gpus', '1', '--pure_bf16', '--dist_checkpoint_root_folder', 'model_checkpoints',
'--dist_checkpoint_folder', 'fine-tuned', '--batch_size_training', '4', '--micro_batch_size', '4', '--train_file', '/opt/ml/i
nput/data/training', '--lr', '0.0001', '--do_train', '--output_dir', 'saved_peft_model', '--num_epochs', '5', '--use_peft',
'--peft_method', 'lora', '--max_train_samples', '-1', '--max_val_samples', '-1', '--seed', '10', '--per_device_eval_batch_siz
```

```
e', '1', '--max_input_length', '-1', '--preprocessing_num_workers', '--None', '--validation_split_ratio', '0.2', '--train_dat
a_split_seed', '0', '--num_workers_dataloader', '0', '--weight_decay', '0.1', '--lora_r', '8', '--lora_alpha', '32', '--lora_
dropout', '0.05', '--target_modules', 'q_proj,v_proj', '--chat_template', 'None', '--enable_fsdp', '--add_input_output_demarc
ation_key'].
================================BUG REPORT================================
Welcome to bitsandbytes. For bug reports, please run
python -m bitsandbytes
 and submit this information together with your error trace to: https://github.com/TimDettmers/bitsandbytes/issues
==========================================================================
bin /opt/conda/lib/python3.10/site-packages/bitsandbytes/libbitsandbytes_cuda118.so
/opt/conda/lib/python3.10/site-packages/bitsandbytes/cuda_setup/main.py:149: UserWarning: WARNING: The following directories
listed in your path were found to be non-existent: {PosixPath('/usr/local/nvidia/lib'), PosixPath('/usr/local/nvidia/lib64')}
  warn(msg)
CUDA SETUP: CUDA runtime path found: /opt/conda/lib/libcudart.so
CUDA SETUP: Highest compute capability among GPUs detected: 8.6
CUDA SETUP: Detected CUDA version 118
CUDA SETUP: Loading binary /opt/conda/lib/python3.10/site-packages/bitsandbytes/libbitsandbytes_cuda118.so...
INFO:root:Local rank is 0. Rank is 0
INFO:root:Setting torch device = 0
INFO:root:Loading the tokenizer.
--> Running with torch dist debug set to detail
INFO:root:Loading the data.
INFO:root:Both instruction_tuned and chat_dataset are set to False.Assuming domain adaptation dataset format.
Downloading data files:   0%|          | 0/1 [00:00<?, ?it/s]
Downloading data files: 100%|██████████| 1/1 [00:00<00:00, 13443.28it/s]
Extracting data files:   0%|          | 0/1 [00:00<?, ?it/s]
Extracting data files: 100%|██████████| 1/1 [00:00<00:00, 1257.29it/s]
Generating train split: 0 examples [00:00, ? examples/s]
Generating train split: 797 examples [00:00, 280723.91 examples/s]
INFO:jumpstart:Training data is identified. The corresponded column names are ['text'].
WARNING:jumpstart:The tokenizer picked has a `model_max_length` (100000000000000019884624838656) larger than maximum input l
ength cap 1024. Picking 1024 instead.
INFO:jumpstart:The max sequence length is set as 1024.
Running tokenizer on dataset:   0%|          | 0/797 [00:00<?, ? examples/s]
Running tokenizer on dataset: 100%|██████████| 797/797 [00:00<00:00, 44784.65 examples/s]
Grouping texts in chunks of 1024:   0%|          | 0/797 [00:00<?, ? examples/s]
Grouping texts in chunks of 1024: 100%|██████████| 797/797 [00:00<00:00, 12091.00 examples/s]
INFO:jumpstart:Test data is not identified. Split the data into train and test data respectively.
INFO:root:Loading the pre-trained model.
Loading checkpoint shards:   0%|          | 0/2 [00:00<?, ?it/s]
Loading checkpoint shards:  50%|█████     | 1/2 [00:27<00:27, 27.81s/it]
```

```
Loading checkpoint shards: 100%|██████████| 2/2 [00:38<00:00, 17.64s/it]
Loading checkpoint shards: 100%|██████████| 2/2 [00:38<00:00, 19.17s/it]
--> Model /opt/ml/additonals3data
--> /opt/ml/additonals3data has 6738.415616 Million params
trainable params: 4,194,304 || all params: 6,742,609,920 || trainable%: 0.06220594176090199
bFloat16 enabled for mixed precision - using bfSixteen policy
--> applying fsdp activation checkpointing...
INFO:root:--> Training Set Length = 10
INFO:root:--> Validation Set Length = 3
/opt/conda/lib/python3.10/site-packages/torch/cuda/memory.py:330: FutureWarning: torch.cuda.reset_max_memory_allocated now ca
lls torch.cuda.reset_peak_memory_stats, which resets /all/ peak memory stats.
  warnings.warn(
Training Epoch0:   0%|#033[34m          #033[0m| 0/2 [00:00<?, ?it/s]
`use_cache=True` is incompatible with gradient checkpointing. Setting `use_cache=False`.
NCCL version 2.19.3+cuda12.3
algo-1:57:79 [0] nccl_net_ofi_init:1444 NCCL WARN NET/OFI Only EFA provider is supported
algo-1:57:79 [0] nccl_net_ofi_init:1483 NCCL WARN NET/OFI aws-ofi-nccl initialization failed
step 0 is completed and loss is 3.415360927581787
Training Epoch0:  50%|#033[34m█████     #033[0m| 1/2 [00:04<00:04,  4.69s/it]
step 1 is completed and loss is 2.7928905487060547
Training Epoch0: 100%|#033[34m██████████#033[0m| 2/2 [00:08<00:00,  3.96s/it]
Training Epoch0: 100%|#033[34m██████████#033[0m| 2/2 [00:08<00:00,  4.07s/it]
Max CUDA memory allocated was 15 GB
Max CUDA memory reserved was 15 GB
Peak active CUDA memory was 15 GB
Cuda Malloc retires : 0
CPU Total Peak Memory consumed during the train (max): 1 GB
evaluating Epoch:   0%|#033[32m          #033[0m| 0/3 [00:00<?, ?it/s]
We detected that you are passing `past_key_values` as a tuple and this is deprecated and will be removed in v4.43. Please use
an appropriate `Cache` class (https://huggingface.co/docs/transformers/v4.41.3/en/internal/generation_utils#transformers.Cach
e)
evaluating Epoch:  33%|#033[32m███▎       #033[0m| 1/3 [00:00<00:00,  3.05it/s]
evaluating Epoch:  67%|#033[32m██████▋    #033[0m| 2/3 [00:00<00:00,  3.07it/s]
evaluating Epoch: 100%|#033[32m██████████#033[0m| 3/3 [00:00<00:00,  3.08it/s]
evaluating Epoch: 100%|#033[32m██████████#033[0m| 3/3 [00:00<00:00,  3.08it/s]
eval_ppl=tensor(14.1326, device='cuda:0') eval_epoch_loss=tensor(2.6485, device='cuda:0')
we are about to save the PEFT modules
PEFT modules are saved in saved_peft_model directory
best eval loss on epoch 0 is 2.6484813690185547
Epoch 1: train_perplexity=22.2897, train_epoch_loss=3.1041, epcoh time 8.457441443999983s
Training Epoch1:   0%|#033[34m          #033[0m| 0/2 [00:00<?, ?it/s]
```

```
step 0 is completed and loss is 3.3271257877349854
Training Epoch1:  50%|#033[34m█████     #033[0m| 1/2 [00:03<00:03,  3.45s/it]
step 1 is completed and loss is 2.7320451736450195
Training Epoch1: 100%|#033[34m██████████#033[0m| 2/2 [00:06<00:00,  3.45s/it]
Training Epoch1: 100%|#033[34m██████████#033[0m| 2/2 [00:06<00:00,  3.45s/it]
Max CUDA memory allocated was 15 GB
Max CUDA memory reserved was 15 GB
Peak active CUDA memory was 15 GB
Cuda Malloc retires : 59
CPU Total Peak Memory consumed during the train (max): 2 GB
evaluating Epoch:   0%|#033[32m          #033[0m| 0/3 [00:00<?, ?it/s]
evaluating Epoch:  33%|#033[32m███        #033[0m| 1/3 [00:00<00:00,  3.10it/s]
evaluating Epoch:  67%|#033[32m██████     #033[0m| 2/3 [00:00<00:00,  3.11it/s]
evaluating Epoch: 100%|#033[32m██████████#033[0m| 3/3 [00:00<00:00,  3.11it/s]
evaluating Epoch: 100%|#033[32m██████████#033[0m| 3/3 [00:00<00:00,  3.11it/s]
eval_ppl=tensor(13.3125, device='cuda:0') eval_epoch_loss=tensor(2.5887, device='cuda:0')
we are about to save the PEFT modules
PEFT modules are saved in saved_peft_model directory
best eval loss on epoch 1 is 2.5887045860290527
Epoch 2: train_perplexity=20.6887, train_epoch_loss=3.0296, epcoh time 7.373499697999989s
Training Epoch2:   0%|#033[34m          #033[0m| 0/2 [00:00<?, ?it/s]
step 0 is completed and loss is 3.219477891921997
Training Epoch2:  50%|#033[34m█████     #033[0m| 1/2 [00:03<00:03,  3.45s/it]
step 1 is completed and loss is 2.663853406906128
Training Epoch2: 100%|#033[34m██████████#033[0m| 2/2 [00:06<00:00,  3.45s/it]
Training Epoch2: 100%|#033[34m██████████#033[0m| 2/2 [00:06<00:00,  3.45s/it]
Max CUDA memory allocated was 15 GB
Max CUDA memory reserved was 15 GB
Peak active CUDA memory was 15 GB
Cuda Malloc retires : 118
CPU Total Peak Memory consumed during the train (max): 2 GB
evaluating Epoch:   0%|#033[32m          #033[0m| 0/3 [00:00<?, ?it/s]
evaluating Epoch:  33%|#033[32m███        #033[0m| 1/3 [00:00<00:00,  3.10it/s]
evaluating Epoch:  67%|#033[32m██████     #033[0m| 2/3 [00:00<00:00,  3.11it/s]
evaluating Epoch: 100%|#033[32m██████████#033[0m| 3/3 [00:00<00:00,  3.12it/s]
evaluating Epoch: 100%|#033[32m██████████#033[0m| 3/3 [00:00<00:00,  3.12it/s]
eval_ppl=tensor(12.5904, device='cuda:0') eval_epoch_loss=tensor(2.5329, device='cuda:0')
we are about to save the PEFT modules
PEFT modules are saved in saved_peft_model directory
best eval loss on epoch 2 is 2.532933235168457
Epoch 3: train_perplexity=18.9474, train_epoch_loss=2.9417, epcoh time 7.369341618000021s
```

```
Training Epoch3:   0%|#033[34m          #033[0m| 0/2 [00:00<?, ?it/s]
step 0 is completed and loss is 3.1138834953308105
Training Epoch3:  50%|#033[34m█████     #033[0m| 1/2 [00:03<00:03,  3.45s/it]
step 1 is completed and loss is 2.598935604095459
Training Epoch3: 100%|#033[34m██████████#033[0m| 2/2 [00:06<00:00,  3.45s/it]
Training Epoch3: 100%|#033[34m██████████#033[0m| 2/2 [00:06<00:00,  3.45s/it]
Max CUDA memory allocated was 15 GB
Max CUDA memory reserved was 15 GB
Peak active CUDA memory was 15 GB
Cuda Malloc retires : 177
CPU Total Peak Memory consumed during the train (max): 2 GB
evaluating Epoch:   0%|#033[32m          #033[0m| 0/3 [00:00<?, ?it/s]
evaluating Epoch:  33%|#033[32m███▎      #033[0m| 1/3 [00:00<00:00,  3.10it/s]
evaluating Epoch:  67%|#033[32m██████▋   #033[0m| 2/3 [00:00<00:00,  3.10it/s]
evaluating Epoch: 100%|#033[32m██████████#033[0m| 3/3 [00:00<00:00,  3.11it/s]
evaluating Epoch: 100%|#033[32m██████████#033[0m| 3/3 [00:00<00:00,  3.11it/s]
eval_ppl=tensor(11.9411, device='cuda:0') eval_epoch_loss=tensor(2.4800, device='cuda:0')
we are about to save the PEFT modules
PEFT modules are saved in saved_peft_model directory
best eval loss on epoch 3 is 2.4799888134002686
Epoch 4: train_perplexity=17.3989, train_epoch_loss=2.8564, epcoh time 7.372977058000004s
Training Epoch4:   0%|#033[34m          #033[0m| 0/2 [00:00<?, ?it/s]
step 0 is completed and loss is 3.0026698112487793
Training Epoch4:  50%|#033[34m█████     #033[0m| 1/2 [00:03<00:03,  3.45s/it]
step 1 is completed and loss is 2.5367534160614014
Training Epoch4: 100%|#033[34m██████████#033[0m| 2/2 [00:06<00:00,  3.45s/it]
Training Epoch4: 100%|#033[34m██████████#033[0m| 2/2 [00:06<00:00,  3.45s/it]
Max CUDA memory allocated was 15 GB
Max CUDA memory reserved was 15 GB
Peak active CUDA memory was 15 GB
Cuda Malloc retires : 236
CPU Total Peak Memory consumed during the train (max): 2 GB
evaluating Epoch:   0%|#033[32m          #033[0m| 0/3 [00:00<?, ?it/s]
evaluating Epoch:  33%|#033[32m███▎      #033[0m| 1/3 [00:00<00:00,  3.10it/s]
evaluating Epoch:  67%|#033[32m██████▋   #033[0m| 2/3 [00:00<00:00,  3.11it/s]
evaluating Epoch: 100%|#033[32m██████████#033[0m| 3/3 [00:00<00:00,  3.12it/s]
evaluating Epoch: 100%|#033[32m██████████#033[0m| 3/3 [00:00<00:00,  3.11it/s]
eval_ppl=tensor(11.3263, device='cuda:0') eval_epoch_loss=tensor(2.4271, device='cuda:0')
we are about to save the PEFT modules
PEFT modules are saved in saved_peft_model directory
best eval loss on epoch 4 is 2.4271256923675537
```

```
Epoch 5: train_perplexity=15.9540, train_epoch_loss=2.7697, epcoh time 7.372671649999916s
INFO:root:Key: avg_train_prep, Value: 19.05574607849121
INFO:root:Key: avg_train_loss, Value: 2.9402995109558105
INFO:root:Key: avg_eval_prep, Value: 12.660573959350586
INFO:root:Key: avg_eval_loss, Value: 2.535446882247925
INFO:root:Key: avg_epoch_time, Value: 7.589186293599982
INFO:root:Key: avg_checkpoint_time, Value: 0.739669894400015
INFO:root:Combining pre-trained base model with the PEFT adapter module.
Loading checkpoint shards:   0%|          | 0/2 [00:00<?, ?it/s]
Loading checkpoint shards:  50%|█████     | 1/2 [00:29<00:29, 29.73s/it]
Loading checkpoint shards: 100%|██████████| 2/2 [00:35<00:00, 15.62s/it]
Loading checkpoint shards: 100%|██████████| 2/2 [00:35<00:00, 17.74s/it]
INFO:root:Saving the combined model in safetensors format.
INFO:root:Saving complete.
INFO:root:Copying tokenizer to the output directory.
INFO:root:Putting inference code with the fine-tuned model directory.
2024-12-24 19:06:15,161 sagemaker-training-toolkit INFO     Waiting for the process to finish and give a return code.
2024-12-24 19:06:15,161 sagemaker-training-toolkit INFO     Done waiting for a return code. Received 0 from exiting process.
2024-12-24 19:06:15,161 sagemaker-training-toolkit INFO     Reporting training SUCCESS

2024-12-24 19:06:24 Uploading - Uploading generated training model
2024-12-24 19:07:08 Completed - Training job completed
Training seconds: 697
Billable seconds: 697
```

In [9]:

```
'''
# Do not use estimator.deploy() without mentioning the instance_type.
# It's because when you call estimator.deploy() without explicitly setting the instance_type for the endpoint,
# SageMaker selects a default instance type for hosting, which, in this case, is ml.g5.12xlarge.
# However, Udacity doesn't allow instance type more than "ml.*.2xlarge".
'''
finetuned_predictor = estimator.deploy(instance_type="ml.g5.2xlarge", initial_instance_count=1)
```

```
[12/24/24 19:07:26] INFO     Creating model with name:                                    session.py:4094
                             meta-textgeneration-llama-2-7b-2024-12-24-19-07-26-509
[12/24/24 19:07:27] INFO     Creating endpoint-config with name                           session.py:5889
                             meta-textgeneration-llama-2-7b-2024-12-24-19-07-26-504
                    INFO     Creating endpoint with name                                  session.py:4711
                             meta-textgeneration-llama-2-7b-2024-12-24-19-07-26-504
```

```
----------!
```

Evaluate the pre-trained and fine-tuned model

Next, we use the same input from the model evaluation step to evaluate the performance of the fine-tuned model and compare it with the base pre-trained model.

Create a function to print the response from the model

In [10]:
```
def print_response(payload, response):
    print(payload["The investment tests performed indicate"])
    print(f"> {response}")
    print("\n==================================\n")
```

Now we can run the same prompts on the fine-tuned model to evaluate it's domain knowledge.

Replace "inputs" in the next cell with the input to send the model based on the domain you've chosen.

For financial domain:

"inputs": "Replace with sentence below from text"

- "The investment tests performed indicate"
- "the relative volume for the long out of the money options, indicates"
- "The results for the short in the money options"
- "The results are encouraging for aggressive investors"

For medical domain:

"inputs": "Replace with sentence below from text"

- "Myeloid neoplasms and acute leukemias derive from"
- "Genomic characterization is essential for"
- "Certain germline disorders may be associated with"

- "In contrast to targeted approaches, genome-wide sequencing"

For IT domain:

"inputs": "Replace with sentence below from text"

- "Traditional approaches to data management such as"
- "A second important aspect of ubiquitous computing environments is"
- "because ubiquitous computing is intended to"
- "outline the key aspects of ubiquitous computing from a data management perspective."

In [11]:
```
payload = {
    "inputs": "Domain specific input chosen from above",
    "parameters": {
        "max_new_tokens": 64,
        "top_p": 0.9,
        "temperature": 0.6,
        "return_full_text": False,
    },
}
try:
    response = finetuned_predictor.predict(payload, custom_attributes="accept_eula=true")
    print_response(payload, response)
except Exception as e:
    print(e)
```

```
'The investment tests performed indicate'
```

Do the outputs from the fine-tuned model provide domain-specific insightful and relevant content? You can continue experimenting with the inputs of the model to test it's domain knowledge.

Use the output from this notebook to fill out the "model fine-tuning" section of the project documentation report

After you've filled out the report, run the cells below to delete the model deployment

`IF YOU FAIL TO RUN THE CELLS BELOW YOU WILL RUN OUT OF BUDGET TO COMPLETE THE PROJECT`

In [ ]:
```
finetuned_predictor.delete_model()
finetuned_predictor.delete_endpoint()
```