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/tensorflow2_p310/lib/python3.10/site-packa
ges (2.207.1)
Requirement already satisfied: datasets in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packag
es (2.16.1)
Requirement already satisfied: attrs<24,>=23.1.0 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/si
te-packages (from sagemaker) (23.1.0)
Requirement already satisfied: boto3<2.0,>=1.33.3 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/s
ite-packages (from sagemaker) (1.34.32)
Requirement already satisfied: cloudpickle==2.2.1 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/s
ite-packages (from sagemaker) (2.2.1)
Requirement already satisfied: google-pasta in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-pa
ckages (from sagemaker) (0.2.0)
Requirement already satisfied: numpy<2.0,>=1.9.0 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/si
te-packages (from sagemaker) (1.26.1)
Requirement already satisfied: protobuf<5.0,>=3.12 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/
site-packages (from sagemaker) (4.24.4)
Requirement already satisfied: smdebug-rulesconfig==1.0.1 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/pyth
on3.10/site-packages (from sagemaker) (1.0.1)
Requirement already satisfied: importlib-metadata<7.0,>=1.4.0 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/
python3.10/site-packages (from sagemaker) (6.8.0)
Requirement already satisfied: packaging>=20.0 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site
-packages (from sagemaker) (21.3)
Requirement already satisfied: pandas in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages
(from sagemaker) (1.5.3)
Requirement already satisfied: pathos in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages
(from sagemaker) (0.3.1)
Requirement already satisfied: schema in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages
(from sagemaker) (0.7.5)
Requirement already satisfied: PyYAML~=6.0 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-pac
kages (from sagemaker) (6.0.1)
Requirement already satisfied: jsonschema in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-pack
ages (from sagemaker) (4.19.1)
Requirement already satisfied: platformdirs in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-pa
ckages (from sagemaker) (3.11.0)
Requirement already satisfied: tblib<3,>=1.7.0 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site
-packages (from sagemaker) (1.7.0)
Requirement already satisfied: urllib3<3.0.0,>=1.26.8 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.
10/site-packages (from sagemaker) (1.26.18)
Requirement already satisfied: requests in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packag
es (from sagemaker) (2.31.0)
Requirement already satisfied: docker in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages
(from sagemaker) (6.1.3)
```

```
Requirement already satisfied: tqdm in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages
(from sagemaker) (4.66.1)
Requirement already satisfied: psutil in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages
(from sagemaker) (5.9.5)
Requirement already satisfied: filelock in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packag
es (from datasets) (3.12.4)
Requirement already satisfied: pyarrow>=8.0.0 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-
packages (from datasets) (13.0.0)
Requirement already satisfied: pyarrow-hotfix in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-
packages (from datasets) (0.6)
Requirement already satisfied: dill<0.3.8,>=0.3.0 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/s
ite-packages (from datasets) (0.3.7)
Requirement already satisfied: xxhash in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages
(from datasets) (3.4.1)
Requirement already satisfied: multiprocess in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-pa
ckages (from datasets) (0.70.15)
Requirement already satisfied: fsspec<=2023.10.0,>=2023.1.0 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/py
thon3.10/site-packages (from fsspec[http]<=2023.10.0,>=2023.1.0->datasets) (2023.10.0)
Requirement already satisfied: aiohttp in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-package
s (from datasets) (3.8.6)
Requirement already satisfied: huggingface-hub>=0.19.4 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python
3.10/site-packages (from datasets) (0.20.3)
Requirement already satisfied: botocore<1.35.0,>=1.34.32 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/pytho
n3.10/site-packages (from boto3<2.0,>=1.33.3->sagemaker) (1.34.32)
Requirement already satisfied: jmespath<2.0.0,>=0.7.1 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.
10/site-packages (from boto3<2.0,>=1.33.3->sagemaker) (1.0.1)
Requirement already satisfied: s3transfer<0.11.0,>=0.10.0 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/pyth
on3.10/site-packages (from boto3<2.0,>=1.33.3->sagemaker) (0.10.0)
Requirement already satisfied: charset-normalizer<4.0,>=2.0 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/py
thon3.10/site-packages (from aiohttp->datasets) (3.3.1)
Requirement already satisfied: multidict<7.0,>=4.5 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/
site-packages (from aiohttp->datasets) (6.0.4)
Requirement already satisfied: async-timeout<5.0,>=4.0.0a3 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/pyt
hon3.10/site-packages (from aiohttp->datasets) (4.0.3)
Requirement already satisfied: yarl<2.0,>=1.0 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-
packages (from aiohttp->datasets) (1.9.2)
Requirement already satisfied: frozenlist>=1.1.1 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/si
te-packages (from aiohttp->datasets) (1.4.0)
Requirement already satisfied: aiosignal>=1.1.2 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/sit
e-packages (from aiohttp->datasets) (1.3.1)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/pyth
on3.10/site-packages (from huggingface-hub>=0.19.4->datasets) (4.5.0)
```

```
Requirement already satisfied: zipp>=0.5 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packa
ges (from importlib-metadata<7.0,>=1.4.0->sagemaker) (3.17.0)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python
3.10/site-packages (from packaging>=20.0->sagemaker) (3.1.1)
Requirement already satisfied: idna<4,>=2.5 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-pa
ckages (from requests->sagemaker) (3.4)
Requirement already satisfied: certifi>=2017.4.17 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/s
ite-packages (from requests->sagemaker) (2023.7.22)
Requirement already satisfied: websocket-client>=0.32.0 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python
3.10/site-packages (from docker->sagemaker) (1.6.4)
Requirement already satisfied: six in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (f
rom google-pasta->sagemaker) (1.16.0)
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in /home/ec2-user/anaconda3/envs/tensorflow2_p31
0/lib/python3.10/site-packages (from jsonschema->sagemaker) (2023.7.1)
Requirement already satisfied: referencing>=0.28.4 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/
site-packages (from jsonschema->sagemaker) (0.30.2)
Requirement already satisfied: rpds-py>=0.7.1 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-
packages (from jsonschema->sagemaker) (0.10.6)
Requirement already satisfied: python-dateutil>=2.8.1 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.
10/site-packages (from pandas->sagemaker) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-pa
ckages (from pandas->sagemaker) (2023.3.post1)
Requirement already satisfied: ppft>=1.7.6.7 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-p
ackages (from pathos->sagemaker) (1.7.6.7)
Requirement already satisfied: pox>=0.3.3 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-pack
ages (from pathos->sagemaker) (0.3.3)
Requirement already satisfied: contextlib2>=0.5.5 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/s
ite-packages (from schema->sagemaker) (21.6.0)
```

Select the model to fine-tune

In [2]:
```
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"`

In [3]:
```
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
#example: estimator.fit({"training": f"s3://genaiwithawsproject2024/training-datasets/medical"})
estimator.fit({"training": f"s3://genaiwithawsproject2024/training-datasets/medical"})
```

```
sagemaker.config INFO - Not applying SDK defaults from location: /etc/xdg/sagemaker/config.yaml
sagemaker.config INFO - Not applying SDK defaults from location: /home/ec2-user/.config/sagemaker/config.yaml
Using model 'meta-textgeneration-llama-2-7b' with wildcard version identifier '*'. You can pin to version '3.0.2' fo
r more stable results. Note that models may have different input/output signatures after a major version upgrade.
INFO:sagemaker:Creating training-job with name: meta-textgeneration-llama-2-7b-2024-02-07-07-41-17-830
```

```
2024-02-07 07:41:18 Starting - Starting the training job...
2024-02-07 07:41:35 Pending - Training job waiting for capacity...
2024-02-07 07:41:50 Pending - Preparing the instances for training......
2024-02-07 07:42:56 Downloading - Downloading input data.....................
2024-02-07 07:47:32 Training - Training image download completed. Training in progress.bash: cannot set terminal pro
cess group (-1): Inappropriate ioctl for device
bash: no job control in this shell
2024-02-07 07:47:33,526 sagemaker-training-toolkit INFO     Imported framework sagemaker_pytorch_container.training
2024-02-07 07:47:33,552 sagemaker-training-toolkit INFO     No Neurons detected (normal if no neurons installed)
2024-02-07 07:47:33,561 sagemaker_pytorch_container.training INFO     Block until all host DNS lookups succeed.
2024-02-07 07:47:33,565 sagemaker_pytorch_container.training INFO     Invoking user training script.
2024-02-07 07:47:41,445 sagemaker-training-toolkit INFO     Installing dependencies from requirements.txt:
/opt/conda/bin/python3.10 -m pip install -r requirements.txt
Processing ./lib/accelerate/accelerate-0.21.0-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 requirement
s.txt (line 3))
Processing ./lib/brotli/Brotli-1.0.9-cp310-cp310-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_12_x86_64.manyli
nux2010_x86_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/fire/fire-0.5.0.tar.gz
Preparing metadata (setup.py): started
Preparing metadata (setup.py): finished with status 'done'
Processing ./lib/inflate64/inflate64-0.3.1-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (from -r requi
rements.txt (line 7))
Processing ./lib/loralib/loralib-0.1.1-py3-none-any.whl (from -r requirements.txt (line 8))
Processing ./lib/multivolumefile/multivolumefile-0.2.3-py3-none-any.whl (from -r requirements.txt (line 9))
Processing ./lib/mypy-extensions/mypy_extensions-1.0.0-py3-none-any.whl (from -r requirements.txt (line 10))
Processing ./lib/nvidia-cublas-cu12/nvidia_cublas_cu12-12.1.3.1-py3-none-manylinux1_x86_64.whl (from -r requirement
s.txt (line 11))
Processing ./lib/nvidia-cuda-cupti-cu12/nvidia_cuda_cupti_cu12-12.1.105-py3-none-manylinux1_x86_64.whl (from -r requ
irements.txt (line 12))
Processing ./lib/nvidia-cuda-nvrtc-cu12/nvidia_cuda_nvrtc_cu12-12.1.105-py3-none-manylinux1_x86_64.whl (from -r requ
irements.txt (line 13))
Processing ./lib/nvidia-cuda-runtime-cu12/nvidia_cuda_runtime_cu12-12.1.105-py3-none-manylinux1_x86_64.whl (from -r
requirements.txt (line 14))
Processing ./lib/nvidia-cudnn-cu12/nvidia_cudnn_cu12-8.9.2.26-py3-none-manylinux1_x86_64.whl (from -r requirements.t
xt (line 15))
Processing ./lib/nvidia-cufft-cu12/nvidia_cufft_cu12-11.0.2.54-py3-none-manylinux1_x86_64.whl (from -r requirements.
txt (line 16))
Processing ./lib/nvidia-curand-cu12/nvidia_curand_cu12-10.3.2.106-py3-none-manylinux1_x86_64.whl (from -r requiremen
ts.txt (line 17))
```

```
Processing ./lib/nvidia-cusolver-cu12/nvidia_cusolver_cu12-11.4.5.107-py3-none-manylinux1_x86_64.whl (from -r requir
ements.txt (line 18))
Processing ./lib/nvidia-cusparse-cu12/nvidia_cusparse_cu12-12.1.0.106-py3-none-manylinux1_x86_64.whl (from -r requir
ements.txt (line 19))
Processing ./lib/nvidia-nccl-cu12/nvidia_nccl_cu12-2.18.1-py3-none-manylinux1_x86_64.whl (from -r requirements.txt
(line 20))
Processing ./lib/nvidia-nvjitlink-cu12/nvidia_nvjitlink_cu12-12.3.101-py3-none-manylinux1_x86_64.whl (from -r requir
ements.txt (line 21))
Processing ./lib/nvidia-nvtx-cu12/nvidia_nvtx_cu12-12.1.105-py3-none-manylinux1_x86_64.whl (from -r requirements.txt
(line 22))
Processing ./lib/pathspec/pathspec-0.11.1-py3-none-any.whl (from -r requirements.txt (line 23))
Processing ./lib/peft/peft-0.4.0-py3-none-any.whl (from -r requirements.txt (line 24))
Processing ./lib/py7zr/py7zr-0.20.5-py3-none-any.whl (from -r requirements.txt (line 25))
Processing ./lib/pybcj/pybcj-1.0.1-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (from -r requirements.
txt (line 26))
Processing ./lib/pycryptodomex/pycryptodomex-3.18.0-cp35-abi3-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (from -
r requirements.txt (line 27))
Processing ./lib/pyppmd/pyppmd-1.0.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (from -r requirement
s.txt (line 28))
Processing ./lib/pyzstd/pyzstd-0.15.9-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (from -r requiremen
ts.txt (line 29))
Processing ./lib/safetensors/safetensors-0.3.1-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (from -r r
equirements.txt (line 30))
Processing ./lib/scipy/scipy-1.11.1-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (from -r requirement
s.txt (line 31))
Processing ./lib/termcolor/termcolor-2.3.0-py3-none-any.whl (from -r requirements.txt (line 32))
Processing ./lib/texttable/texttable-1.6.7-py2.py3-none-any.whl (from -r requirements.txt (line 33))
Processing ./lib/tokenize-rt/tokenize_rt-5.1.0-py2.py3-none-any.whl (from -r requirements.txt (line 34))
Processing ./lib/tokenizers/tokenizers-0.13.3-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (from -r re
quirements.txt (line 35))
Processing ./lib/torch/torch-2.1.0-cp310-cp310-manylinux1_x86_64.whl (from -r requirements.txt (line 36))
Processing ./lib/transformers/transformers-4.31.0-py3-none-any.whl (from -r requirements.txt (line 37))
Processing ./lib/triton/triton-2.1.0-0-cp310-cp310-manylinux2014_x86_64.manylinux_2_17_x86_64.whl (from -r requireme
nts.txt (line 38))
Processing ./lib/typing-extensions/typing_extensions-4.8.0-py3-none-any.whl (from -r requirements.txt (line 39))
Processing ./lib/sagemaker_jumpstart_script_utilities/sagemaker_jumpstart_script_utilities-1.1.9-py2.py3-none-any.wh
l (from -r requirements.txt (line 40))
Processing ./lib/sagemaker_jumpstart_huggingface_script_utilities/sagemaker_jumpstart_huggingface_script_utilities-
1.1.4-py2.py3-none-any.whl (from -r requirements.txt (line 41))
Requirement already satisfied: numpy>=1.17 in /opt/conda/lib/python3.10/site-packages (from accelerate==0.21.0->-r r
equirements.txt (line 1)) (1.24.4)
Requirement already satisfied: packaging>=20.0 in /opt/conda/lib/python3.10/site-packages (from accelerate==0.21.0->
```

```
-r requirements.txt (line 1)) (23.1)
Requirement already satisfied: psutil in /opt/conda/lib/python3.10/site-packages (from accelerate==0.21.0->-r requir
ements.txt (line 1)) (5.9.5)
Requirement already satisfied: pyyaml in /opt/conda/lib/python3.10/site-packages (from accelerate==0.21.0->-r requir
ements.txt (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 requi
rements.txt (line 3)) (8.1.4)
Requirement already satisfied: platformdirs>=2 in /opt/conda/lib/python3.10/site-packages (from black==23.7.0->-r re
quirements.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 requi
rements.txt (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
requirements.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 requirements.txt (line 5)) (0.3.6)
Requirement already satisfied: pandas in /opt/conda/lib/python3.10/site-packages (from datasets==2.14.1->-r requirem
ents.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 requirements.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 re
quirements.txt (line 5)) (4.65.0)
Requirement already satisfied: xxhash in /opt/conda/lib/python3.10/site-packages (from datasets==2.14.1->-r requirem
ents.txt (line 5)) (3.4.1)
Requirement already satisfied: multiprocess in /opt/conda/lib/python3.10/site-packages (from datasets==2.14.1->-r re
quirements.txt (line 5)) (0.70.14)
Requirement already satisfied: fsspec>=2021.11.1 in /opt/conda/lib/python3.10/site-packages (from fsspec[http]>=202
1.11.1->datasets==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 require
ments.txt (line 5)) (3.9.3)
Requirement already satisfied: huggingface-hub<1.0.0,>=0.14.0 in /opt/conda/lib/python3.10/site-packages (from datas
ets==2.14.1->-r requirements.txt (line 5)) (0.20.3)
Requirement already satisfied: six in /opt/conda/lib/python3.10/site-packages (from fire==0.5.0->-r requirements.txt
(line 6)) (1.16.0)
Requirement already satisfied: filelock in /opt/conda/lib/python3.10/site-packages (from torch==2.1.0->-r requiremen
ts.txt (line 36)) (3.12.2)
Requirement already satisfied: sympy in /opt/conda/lib/python3.10/site-packages (from torch==2.1.0->-r requirements.
txt (line 36)) (1.12)
Requirement already satisfied: networkx in /opt/conda/lib/python3.10/site-packages (from torch==2.1.0->-r requiremen
ts.txt (line 36)) (3.1)
Requirement already satisfied: jinja2 in /opt/conda/lib/python3.10/site-packages (from torch==2.1.0->-r requirement
s.txt (line 36)) (3.1.2)
Requirement already satisfied: regex!=2019.12.17 in /opt/conda/lib/python3.10/site-packages (from transformers==4.3
```

```
1.0->-r requirements.txt (line 37)) (2023.12.25)
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==2.1
4.1->-r requirements.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->dataset
s==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->dat
asets==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->datasets==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->datas
ets==2.14.1->-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: MarkupSafe>=2.0 in /opt/conda/lib/python3.10/site-packages (from jinja2->torch==2.1.0
->-r requirements.txt (line 36)) (2.1.3)
Requirement already satisfied: python-dateutil>=2.8.2 in /opt/conda/lib/python3.10/site-packages (from pandas->datas
ets==2.14.1->-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 requirements.txt (line 5)) (2023.3)
Requirement already satisfied: tzdata>=2022.1 in /opt/conda/lib/python3.10/site-packages (from pandas->datasets==2.1
4.1->-r requirements.txt (line 5)) (2023.3)
Requirement already satisfied: mpmath>=0.19 in /opt/conda/lib/python3.10/site-packages (from sympy->torch==2.1.0->-r
requirements.txt (line 36)) (1.3.0)
scipy is already installed with the same version as the provided wheel. Use --force-reinstall to force an installati
on of the wheel.
tokenizers is already installed with the same version as the provided wheel. Use --force-reinstall to force an insta
llation 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=cc56f4cc6740660b9e7c64dbf80becd6
7aedf87e76dbe55b9558948356231377
Stored in directory: /root/.cache/pip/wheels/db/3d/41/7e69dca5f61e37d109a4457082ffc5c6edb55ab633bafded38
Successfully built fire
```

```
Installing collected packages: texttable, safetensors, Brotli, bitsandbytes, typing-extensions, triton, tokenize-rt,
termcolor, sagemaker-jumpstart-script-utilities, sagemaker-jumpstart-huggingface-script-utilities, pyzstd, pyppmd, p
ycryptodomex, 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-exte
nsions, multivolumefile, loralib, inflate64, py7zr, nvidia-cusparse-cu12, nvidia-cudnn-cu12, fire, black, transforme
rs, nvidia-cusolver-cu12, torch, datasets, accelerate, 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: 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 beha
viour is the source of the following dependency conflicts.
fastai 2.7.12 requires torch<2.1,>=1.7, but you have torch 2.1.0 which is incompatible.
Successfully installed Brotli-1.0.9 accelerate-0.21.0 bitsandbytes-0.39.1 black-23.7.0 datasets-2.14.1 fire-0.5.0 in
flate64-0.3.1 loralib-0.1.1 multivolumefile-0.2.3 mypy-extensions-1.0.0 nvidia-cublas-cu12-12.1.3.1 nvidia-cuda-cupt
i-cu12-12.1.105 nvidia-cuda-nvrtc-cu12-12.1.105 nvidia-cuda-runtime-cu12-12.1.105 nvidia-cudnn-cu12-8.9.2.26 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 n
vidia-nccl-cu12-2.18.1 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-0.3.1 sagemaker-jumpstart-huggingface-sc
ript-utilities-1.1.4 sagemaker-jumpstart-script-utilities-1.1.9 termcolor-2.3.0 texttable-1.6.7 tokenize-rt-5.1.0 to
rch-2.1.0 transformers-4.31.0 triton-2.1.0 typing-extensions-4.8.0
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system p
```

```
ackage manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv
2024-02-07 07:48:48,739 sagemaker-training-toolkit INFO     Waiting for the process to finish and give a return cod
e.
2024-02-07 07:48:48,739 sagemaker-training-toolkit INFO     Done waiting for a return code. Received 0 from exiting
process.
2024-02-07 07:48:48,786 sagemaker-training-toolkit INFO     No Neurons detected (normal if no neurons installed)
2024-02-07 07:48:48,822 sagemaker-training-toolkit INFO     No Neurons detected (normal if no neurons installed)
2024-02-07 07:48:48,858 sagemaker-training-toolkit INFO     No Neurons detected (normal if no neurons installed)
2024-02-07 07:48:48,868 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",
    "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-2024-02-07-07-41-17-830",
"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","instructi
on_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","pe
r_device_train_batch_size":"4","preprocessing_num_workers":"None","seed":"10","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.g
5.2xlarge","hosts":["algo-1"],"instance_groups":[{"hosts":["algo-1"],"instance_group_name":"homogeneousCluster","ins
tance_type":"ml.g5.2xlarge"}],"network_interface_name":"eth0"}
SM_INPUT_DATA_CONFIG={"code":{"RecordWrapperType":"None","S3DistributionType":"FullyReplicated","TrainingInputMod
```

e":"File"},"training":{"RecordWrapperType":"None","S3DistributionType":"FullyReplicated","TrainingInputMode":"Fil
e"}}
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","instan
ce_type":"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","traini
ng":"/opt/ml/input/data/training"},"current_host":"algo-1","current_instance_group":"homogeneousCluster","current_in
stance_group_hosts":["algo-1"],"current_instance_type":"ml.g5.2xlarge","distribution_hosts":[],"distribution_instanc
e_groups":[],"framework_module":"sagemaker_pytorch_container.training:main","hosts":["algo-1"],"hyperparameters":{"a
dd_input_output_demarcation_key":"True","chat_dataset":"False","enable_fsdp":"True","epoch":"5","instruction_tune
d":"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","train_data_split_seed":"0","validation
_split_ratio":"0.2"},"input_config_dir":"/opt/ml/input/config","input_data_config":{"code":{"RecordWrapperType":"Non
e","S3DistributionType":"FullyReplicated","TrainingInputMode":"File"},"training":{"RecordWrapperType":"None","S3Dist
ributionType":"FullyReplicated","TrainingInputMode":"File"}},"input_dir":"/opt/ml/input","instance_groups":["homogen
eousCluster"],"instance_groups_dict":{"homogeneousCluster":{"hosts":["algo-1"],"instance_group_name":"homogeneousClu
ster","instance_type":"ml.g5.2xlarge"}},"is_hetero":false,"is_master":true,"is_modelparallel_enabled":null,"is_smddp
mprun_installed":true,"job_name":"meta-textgeneration-llama-2-7b-2024-02-07-07-41-17-830","log_level":20,"master_hos
tname":"algo-1","model_dir":"/opt/ml/model","module_dir":"/opt/ml/input/data/code/sourcedir.tar.gz","module_name":"t
ransfer_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_conf
ig":{"current_group_name":"homogeneousCluster","current_host":"algo-1","current_instance_type":"ml.g5.2xlarge","host

s":["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","--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_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-dynload:/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 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 --train_data_split_seed 0 --validation_split_ratio 0.2
2024-02-07 07:48:48,907 sagemaker-training-toolkit INFO   Exceptions not imported for SageMaker TF as Tensorflow is not installed.
===================================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 dir
ectories listed in your path were found to be non-existent: {PosixPath('/usr/local/nvidia/lib'), PosixPath('/usr/loc
al/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: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.p
y', '--model_name', '/opt/ml/additonals3data', '--num_gpus', '1', '--pure_bf16', '--dist_checkpoint_root_folder', 'm
odel_checkpoints', '--dist_checkpoint_folder', 'fine-tuned', '--batch_size_training', '4', '--micro_batch_size',
'4', '--train_file', '/opt/ml/input/data/training', '--lr', '0.0001', '--do_train', '--output_dir', 'saved_peft_mode
l', '--num_epochs', '5', '--use_peft', '--peft_method', 'lora', '--max_train_samples', '-1', '--max_val_samples', '-
1', '--seed', '10', '--per_device_eval_batch_size', '1', '--max_input_length', '-1', '--preprocessing_num_workers',
'--None', '--validation_split_ratio', '0.2', '--train_data_split_seed', '0', '--num_workers_dataloader', '0', '--wei
ght_decay', '0.1', '--lora_r', '8', '--lora_alpha', '32', '--lora_dropout', '0.05', '--enable_fsdp', '--add_input_ou
tput_demarcation_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 dir
ectories listed in your path were found to be non-existent: {PosixPath('/usr/local/nvidia/lib'), PosixPath('/usr/loc
al/nvidia/lib64')}
  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: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, 1330.68it/s]
Extracting data files:   0%|          | 0/1 [00:00<?, ?it/s]
Extracting data files: 100%|██████████| 1/1 [00:00<00:00, 739.87it/s]
Generating train split: 0 examples [00:00, ? examples/s]
Generating train split: 113 examples [00:00, 42934.72 examples/s]
Training data is identified. The corresponded column names are ['text'].
INFO:sagemaker_jumpstart_huggingface_script_utilities.fine_tuning.data_preprocessor:Training data is identified. The
corresponded column names are ['text'].
The tokenizer picked has a `model_max_length` (1000000000000000019884624838656) larger than maximum input length cap
1024. Picking 1024 instead.
The max sequence length is set as 1024.
WARNING:sagemaker_jumpstart_huggingface_script_utilities.fine_tuning.data_preprocessor:The tokenizer picked has a `m
odel_max_length` (1000000000000000019884624838656) larger than maximum input length cap 1024. Picking 1024 instead.
INFO:sagemaker_jumpstart_huggingface_script_utilities.fine_tuning.data_preprocessor:The max sequence length is set a
s 1024.
Running tokenizer on dataset:   0%|          | 0/113 [00:00<?, ? examples/s]
Running tokenizer on dataset: 100%|██████████| 113/113 [00:00<00:00, 8986.66 examples/s]
Grouping texts in chunks of 1024:   0%|          | 0/113 [00:00<?, ? examples/s]
Grouping texts in chunks of 1024: 100%|██████████| 113/113 [00:00<00:00, 5408.36 examples/s]
Test data is not identified. Split the data into train and test data respectively.
INFO:sagemaker_jumpstart_huggingface_script_utilities.fine_tuning.data_preprocessor:Test data is not identified. Spl
it 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 [00:29<00:29, 29.90s/it]
Loading checkpoint shards: 100%|██████████| 2/2 [00:40<00:00, 18.66s/it]
Loading checkpoint shards: 100%|██████████| 2/2 [00:40<00:00, 20.34s/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 = 8
INFO:root:--> Validation Set Length = 2
/opt/conda/lib/python3.10/site-packages/torch/cuda/memory.py:329: FutureWarning: torch.cuda.reset_max_memory_allocat
ed now calls 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]
NCCL version 2.18.1+cuda12.1
```

```
algo-1:53:76 [0] nccl_net_ofi_init:1444 NCCL WARN NET/OFI Only EFA provider is supported
algo-1:53:76 [0] nccl_net_ofi_init:1483 NCCL WARN NET/OFI aws-ofi-nccl initialization failed
step 0 is completed and loss is 2.1504554748535156
Training Epoch0:  50%|#033[34m[redacted]     #033[0m| 1/2 [00:06<00:06,  6.99s/it]
step 1 is completed and loss is 2.0601589679718018
Training Epoch0: 100%|#033[34m[redacted]#033[0m| 2/2 [00:11<00:00,  5.52s/it]
Training Epoch0: 100%|#033[34m[redacted]#033[0m| 2/2 [00:11<00:00,  5.74s/it]
Max CUDA memory allocated was 17 GB
Max CUDA memory reserved was 18 GB
Peak active CUDA memory was 17 GB
Cuda Malloc retires : 0
CPU Total Peak Memory consumed during the train (max): 1 GB
evaluating Epoch:   0%|#033[32m          #033[0m| 0/2 [00:00<?, ?it/s]
evaluating Epoch:  50%|#033[32m[redacted]      #033[0m| 1/2 [00:00<00:00,  2.49it/s]
evaluating Epoch: 100%|#033[32m[redacted]#033[0m| 2/2 [00:00<00:00,  2.50it/s]
evaluating Epoch: 100%|#033[32m[redacted]#033[0m| 2/2 [00:00<00:00,  2.50it/s]
eval_ppl=tensor(12.3674, device='cuda:0') eval_epoch_loss=tensor(2.5151, 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 2.5150630474090576
Epoch 1: train_perplexity=8.2096, train_epoch_loss=2.1053, epcoh time 11.835746732999995s
Training Epoch1:   0%|#033[34m          #033[0m| 0/2 [00:00<?, ?it/s]
step 0 is completed and loss is 2.135524034500122
Training Epoch1:  50%|#033[34m[redacted]     #033[0m| 1/2 [00:04<00:04,  4.50s/it]
step 1 is completed and loss is 2.045548915863037
Training Epoch1: 100%|#033[34m[redacted]#033[0m| 2/2 [00:08<00:00,  4.49s/it]
Training Epoch1: 100%|#033[34m[redacted]#033[0m| 2/2 [00:08<00:00,  4.49s/it]
Max CUDA memory allocated was 17 GB
Max CUDA memory reserved was 18 GB
Peak active CUDA memory was 17 GB
Cuda Malloc retires : 62
CPU Total Peak Memory consumed during the train (max): 2 GB
evaluating Epoch:   0%|#033[32m          #033[0m| 0/2 [00:00<?, ?it/s]
evaluating Epoch:  50%|#033[32m[redacted]      #033[0m| 1/2 [00:00<00:00,  2.49it/s]
evaluating Epoch: 100%|#033[32m[redacted]#033[0m| 2/2 [00:00<00:00,  2.51it/s]
evaluating Epoch: 100%|#033[32m[redacted]#033[0m| 2/2 [00:00<00:00,  2.51it/s]
eval_ppl=tensor(12.1604, device='cuda:0') eval_epoch_loss=tensor(2.4982, 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 2.498188018798828
Epoch 2: train_perplexity=8.0893, train_epoch_loss=2.0905, epcoh time 9.493111386999999s
Training Epoch2:   0%|#033[34m          #033[0m| 0/2 [00:00<?, ?it/s]
```

```
step 0 is completed and loss is 2.1169276237487793
Training Epoch2:  50%|#033[34m█████     #033[0m| 1/2 [00:04<00:04,  4.50s/it]
step 1 is completed and loss is 2.027191638946533
Training Epoch2: 100%|#033[34m██████████#033[0m| 2/2 [00:08<00:00,  4.49s/it]
Training Epoch2: 100%|#033[34m██████████#033[0m| 2/2 [00:08<00:00,  4.50s/it]
Max CUDA memory allocated was 17 GB
Max CUDA memory reserved was 18 GB
Peak active CUDA memory was 17 GB
Cuda Malloc retires : 124
CPU Total Peak Memory consumed during the train (max): 2 GB
evaluating Epoch:   0%|#033[32m          #033[0m| 0/2 [00:00<?, ?it/s]
evaluating Epoch:  50%|#033[32m█████     #033[0m| 1/2 [00:00<00:00,  2.49it/s]
evaluating Epoch: 100%|#033[32m██████████#033[0m| 2/2 [00:00<00:00,  2.50it/s]
evaluating Epoch: 100%|#033[32m██████████#033[0m| 2/2 [00:00<00:00,  2.50it/s]
eval_ppl=tensor(11.9348, device='cuda:0') eval_epoch_loss=tensor(2.4795, 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 2.4794626235961914
Epoch 3: train_perplexity=7.9412, train_epoch_loss=2.0721, epcoh time 9.509339330999978s
Training Epoch3:   0%|#033[34m          #033[0m| 0/2 [00:00<?, ?it/s]
step 0 is completed and loss is 2.098294734954834
Training Epoch3:  50%|#033[34m█████     #033[0m| 1/2 [00:04<00:04,  4.50s/it]
step 1 is completed and loss is 2.0087413787841797
Training Epoch3: 100%|#033[34m██████████#033[0m| 2/2 [00:08<00:00,  4.50s/it]
Training Epoch3: 100%|#033[34m██████████#033[0m| 2/2 [00:08<00:00,  4.50s/it]
Max CUDA memory allocated was 17 GB
Max CUDA memory reserved was 18 GB
Peak active CUDA memory was 17 GB
Cuda Malloc retires : 186
CPU Total Peak Memory consumed during the train (max): 2 GB
evaluating Epoch:   0%|#033[32m          #033[0m| 0/2 [00:00<?, ?it/s]
evaluating Epoch:  50%|#033[32m█████     #033[0m| 1/2 [00:00<00:00,  2.49it/s]
evaluating Epoch: 100%|#033[32m██████████#033[0m| 2/2 [00:00<00:00,  2.50it/s]
evaluating Epoch: 100%|#033[32m██████████#033[0m| 2/2 [00:00<00:00,  2.50it/s]
eval_ppl=tensor(11.7107, device='cuda:0') eval_epoch_loss=tensor(2.4605, 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 2.460505723953247
Epoch 4: train_perplexity=7.7953, train_epoch_loss=2.0535, epcoh time 9.52560391499992s
Training Epoch4:   0%|#033[34m          #033[0m| 0/2 [00:00<?, ?it/s]
step 0 is completed and loss is 2.0797555446624756
Training Epoch4:  50%|#033[34m█████     #033[0m| 1/2 [00:04<00:04,  4.50s/it]
```

```
step 1 is completed and loss is 1.989830732345581
Training Epoch4: 100%|#033[34m###########033[0m| 2/2 [00:08<00:00,  4.50s/it]
Training Epoch4: 100%|#033[34m###########033[0m| 2/2 [00:08<00:00,  4.50s/it]
Max CUDA memory allocated was 17 GB
Max CUDA memory reserved was 18 GB
Peak active CUDA memory was 17 GB
Cuda Malloc retires : 248
CPU Total Peak Memory consumed during the train (max): 2 GB
evaluating Epoch:   0%|#033[32m          #033[0m| 0/2 [00:00<?, ?it/s]
evaluating Epoch:  50%|#033[32m#####     #033[0m| 1/2 [00:00<00:00,  2.49it/s]
evaluating Epoch: 100%|#033[32m###########033[0m| 2/2 [00:00<00:00,  2.51it/s]
evaluating Epoch: 100%|#033[32m###########033[0m| 2/2 [00:00<00:00,  2.50it/s]
eval_ppl=tensor(11.5074, device='cuda:0') eval_epoch_loss=tensor(2.4430, 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 2.4429869651794434
Epoch 5: train_perplexity=7.6507, train_epoch_loss=2.0348, epcoh time 9.520108025000013s
INFO:root:Key: avg_train_prep, Value: 7.937197208404541
INFO:root:Key: avg_train_loss, Value: 2.0712430477142334
INFO:root:Key: avg_eval_prep, Value: 11.93615436553955
INFO:root:Key: avg_eval_loss, Value: 2.479241132736206
INFO:root:Key: avg_epoch_time, Value: 9.976781878199972
INFO:root:Key: avg_checkpoint_time, Value: 0.9318586372000255
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, 29.81s/it]
Loading checkpoint shards: 100%|##########| 2/2 [00:35<00:00, 15.43s/it]
Loading checkpoint shards: 100%|##########| 2/2 [00:35<00:00, 17.58s/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.
2024-02-07 07:54:59,688 sagemaker-training-toolkit INFO     Waiting for the process to finish and give a return cod
e.
2024-02-07 07:54:59,688 sagemaker-training-toolkit INFO     Done waiting for a return code. Received 0 from exiting
process.
2024-02-07 07:54:59,688 sagemaker-training-toolkit INFO     Reporting training SUCCESS

2024-02-07 07:55:03 Uploading - Uploading generated training model
2024-02-07 07:55:54 Completed - Training job completed
Training seconds: 779
Billable seconds: 779
```

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 [4]:
```
finetuned_predictor = estimator.deploy()
```

```
No instance type selected for inference hosting endpoint. Defaulting to ml.g5.2xlarge.
INFO:sagemaker.jumpstart:No instance type selected for inference hosting endpoint. Defaulting to ml.g5.2xlarge.
INFO:sagemaker:Creating model with name: meta-textgeneration-llama-2-7b-2024-02-07-07-56-12-404
INFO:sagemaker:Creating endpoint-config with name meta-textgeneration-llama-2-7b-2024-02-07-07-56-12-389
INFO:sagemaker:Creating endpoint with name meta-textgeneration-llama-2-7b-2024-02-07-07-56-12-389
-------!
```

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 [5]:
```
def print_response(payload, response):
    print(payload["inputs"])
    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 [6]:
```
payload = {
    "inputs": "Myeloid neoplasms and acute leukemias derive from",
    "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)
```

```
Myeloid neoplasms and acute leukemias derive from
> [{'generated_text': ' a common progenitor cell, the myeloid stem cell, which resides in the bone marrow. These cel
ls undergo continuous proliferation and differentiation, as well as apoptosis, to maintain the stem cell pool.\nThe
molecular basis of myeloid neoplasms'}]

==================================
```

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 [7]:
```
finetuned_predictor.delete_model()
finetuned_predictor.delete_endpoint()
```

```
INFO:sagemaker:Deleting model with name: meta-textgeneration-llama-2-7b-2024-02-07-07-56-12-404
INFO:sagemaker:Deleting endpoint configuration with name: meta-textgeneration-llama-2-7b-2024-02-07-07-56-12-389
INFO:sagemaker:Deleting endpoint with name: meta-textgeneration-llama-2-7b-2024-02-07-07-56-12-389
```