In [1]:
```
!pip install transformers "datasets[s3]==2.18.0" "sagemaker>=2.190.0" "huggingface_hub
```

```
^C
```

In [ ]:
```
!huggingface-cli login --token YOUR_TOKEN
```

In [3]:
```
import sagemaker
import boto3
sess = sagemaker.Session()
# sagemaker session bucket -> used for uploading data, models and logs
# sagemaker will automatically create this bucket if it not exists
sagemaker_session_bucket=None
if sagemaker_session_bucket is None and sess is not None:
    # set to default bucket if a bucket name is not given
    sagemaker_session_bucket = sess.default_bucket()

try:
    role = sagemaker.get_execution_role()
except ValueError:
    iam = boto3.client('iam')
    role = iam.get_role(RoleName='sagemaker_execution_role')['Role']['Arn']

sess = sagemaker.Session(default_bucket=sagemaker_session_bucket)

print(f"sagemaker role arn: {role}")
print(f"sagemaker bucket: {sess.default_bucket()}")
print(f"sagemaker session region: {sess.boto_region_name}")
```

```
Couldn't call 'get_role' to get Role ARN from role name philippschmid to get Role pat
h.
sagemaker role arn: arn:aws:iam::558105141721:role/sagemaker_execution_role
sagemaker bucket: sagemaker-us-east-1-558105141721
sagemaker session region: us-east-1
```

In [4]:
```
from datasets import load_dataset

# Convert dataset to OAI messages
system_message = """You are an text to SQL query translator. Users will ask you questi
SCHEMA:
{schema}"""

def create_conversation(sample):
  return {
    "messages": [
      {"role": "system", "content": system_message.format(schema=sample["context"])},
      {"role": "user", "content": sample["question"]},
      {"role": "assistant", "content": sample["answer"]}
    ]
  }

# Load dataset from the hub
dataset = load_dataset("b-mc2/sql-create-context", split="train")
dataset = dataset.shuffle().select(range(12500))

# Convert dataset to OAI messages
dataset = dataset.map(create_conversation, remove_columns=dataset.features,batched=Fal
# split dataset into 10,000 training samples and 2,500 test samples
dataset = dataset.train_test_split(test_size=2500/12500)
```

```
print(dataset["train"][345]["messages"])
```

```
Map:   0%|          | 0/12500 [00:00<?, ? examples/s]
[{'content': 'You are an text to SQL query translator. Users will ask you questions i
n English and you will generate a SQL query based on the provided SCHEMA.\nSCHEMA:\nC
REATE TABLE table_name_33 (make VARCHAR, pos INTEGER)', 'role': 'system'}, {'conten
t': 'Which Make has a Pos larger than 9?', 'role': 'user'}, {'content': 'SELECT make
FROM table_name_33 WHERE pos > 9', 'role': 'assistant'}]
```

In [5]:
```
# save train_dataset to s3 using our SageMaker session
training_input_path = f's3://{sess.default_bucket()}/datasets/text-to-sql'

# save datasets to s3
dataset["train"].to_json(f"{training_input_path}/train_dataset.json", orient="records"
dataset["test"].to_json(f"{training_input_path}/test_dataset.json", orient="records")

print(f"Training data uploaded to:")
print(f"{training_input_path}/train_dataset.json")
print(f"https://s3.console.aws.amazon.com/s3/buckets/{sess.default_bucket()}/?region={
```

```
/home/ubuntu/miniconda3/envs/dev/lib/python3.9/site-packages/fsspec/registry.py:272:
UserWarning: Your installed version of s3fs is very old and known to cause
severe performance issues, see also https://github.com/dask/dask/issues/10276

To fix, you should specify a lower version bound on s3fs, or
update the current installation.

  warnings.warn(s3_msg)
Creating json from Arrow format:   0%|          | 0/10 [00:00<?, ?ba/s]
Creating json from Arrow format:   0%|          | 0/3 [00:00<?, ?ba/s]
Training data uploaded to:
s3://sagemaker-us-east-1-558105141721/datasets/text-to-sql/train_dataset.json
https://s3.console.aws.amazon.com/s3/buckets/sagemaker-us-east-1-558105141721/?region
=us-east-1&prefix=datasets/text-to-sql/
```

In [12]:
```
# hyperparameters, which are passed into the training job
hyperparameters = {
  ### SCRIPT PARAMETERS ###
  'dataset_path': '/opt/ml/input/data/training/train_dataset.json', # path where sagem
  'model_id': "codellama/CodeLlama-7b-hf",           # or `mistralai/Mistral-7B-v0.1`
  'max_seq_len': 3072,                               # max sequence length for model d
  'use_qlora': True,                                 # use QLoRA model
  ### TRAINING PARAMETERS ###
  'num_train_epochs': 3,                             # number of training epochs
  'per_device_train_batch_size': 1,                  # batch size per device during tr
  'gradient_accumulation_steps': 4,                  # number of steps before performi
  'gradient_checkpointing': True,                    # use gradient checkpointing to s
  'optim': "adamw_torch_fused",                      # use fused adamw optimizer
  'logging_steps': 10,                               # log every 10 steps
  'save_strategy': "epoch",                          # save checkpoint every epoch
  'learning_rate': 2e-4,                             # learning rate, based on QLoRA p
  'bf16': True,                                      # use bfloat16 precision
  'tf32': True,                                      # use tf32 precision
  'max_grad_norm': 0.3,                              # max gradient norm based on QLoR
  'warmup_ratio': 0.03,                              # warmup ratio based on QLoRA pap
  'lr_scheduler_type': "constant",                   # use constant learning rate sche
  'report_to': "tensorboard",                        # report metrics to tensorboard
  'output_dir': '/tmp/tun',                          # Temporary output directory for
```

```
  'merge_adapters': True,                          # merge LoRA adapters into model
}
```

In [13]:
```
from sagemaker.huggingface import HuggingFace

# define Training Job Name
job_name = f'codellama-7b-hf-text-to-sql-exp1'

# create the Estimator
huggingface_estimator = HuggingFace(
    entry_point          = 'run_sft.py',    # train script
    source_dir           = '../scripts/trl',      # directory which includes all the f
    instance_type        = 'ml.g5.2xlarge',   # instances type used for the training j
    instance_count       = 1,                 # the number of instances used for train
    max_run              = 2*24*60*60,        # maximum runtime in seconds (days * hou
    base_job_name        = job_name,          # the name of the training job
    role                 = role,              # Iam role used in training job to acces
    volume_size          = 300,               # the size of the EBS volume in GB
    transformers_version = '4.36',            # the transformers version used in the t
    pytorch_version      = '2.1',             # the pytorch_version version used in th
    py_version           = 'py310',           # the python version used in the trainin
    hyperparameters      =  hyperparameters,  # the hyperparameters passed to the trai
    disable_output_compression = True,        # not compress output to save training t
    environment          = {
                           "HUGGINGFACE_HUB_CACHE": "/tmp/.cache", # set env variable
                           # "HF_TOKEN": "REPALCE_WITH_YOUR_TOKEN" # huggingface toke
                           },
)
```

In [14]:
```
# define a data input dictonary with our uploaded s3 uris
data = {'training': training_input_path}

# starting the train job with our uploaded datasets as input
huggingface_estimator.fit(data, wait=True)
```

```
INFO:sagemaker.image_uris:image_uri is not presented, retrieving image_uri based on i
nstance_type, framework etc.
INFO:sagemaker:Creating training-job with name: codellama-7b-hf-text-to-sql-exp1-2024
-03-08-08-05-53-957
```

```
2024-03-08 08:05:55 Starting - Starting the training job
2024-03-08 08:05:55 Pending - Training job waiting for capacity......
2024-03-08 08:06:33 Pending - Preparing the instances for training...
2024-03-08 08:07:15 Downloading - Downloading input data...
2024-03-08 08:07:34 Downloading - Downloading the training image..............
2024-03-08 08:10:25 Training - Training image download completed. Training in progres
s.....bash: cannot set terminal process group (-1): Inappropriate ioctl for device
bash: no job control in this shell
2024-03-08 08:11:01,570 sagemaker-training-toolkit INFO     Imported framework sagema
ker_pytorch_container.training
2024-03-08 08:11:01,588 sagemaker-training-toolkit INFO     No Neurons detected (norm
al if no neurons installed)
2024-03-08 08:11:01,598 sagemaker_pytorch_container.training INFO     Block until all
host DNS lookups succeed.
2024-03-08 08:11:01,600 sagemaker_pytorch_container.training INFO     Invoking user t
raining script.
2024-03-08 08:11:03,036 sagemaker-training-toolkit INFO     Installing dependencies f
rom requirements.txt:
/opt/conda/bin/python3.10 -m pip install -r requirements.txt
Collecting transformers==4.38.2 (from -r requirements.txt (line 1))
Downloading transformers-4.38.2-py3-none-any.whl.metadata (130 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 130.7/130.7 kB 9.5 MB/s eta 0:00:00
Collecting datasets==2.18.0 (from -r requirements.txt (line 2))
Downloading datasets-2.18.0-py3-none-any.whl.metadata (20 kB)
Collecting accelerate==0.27.2 (from -r requirements.txt (line 3))
Downloading accelerate-0.27.2-py3-none-any.whl.metadata (18 kB)
Requirement already satisfied: evaluate==0.4.1 in /opt/conda/lib/python3.10/site-pack
ages (from -r requirements.txt (line 4)) (0.4.1)
Requirement already satisfied: bitsandbytes==0.42.0 in /opt/conda/lib/python3.10/site
-packages (from -r requirements.txt (line 5)) (0.42.0)
Collecting trl==0.7.11 (from -r requirements.txt (line 6))
Downloading trl-0.7.11-py3-none-any.whl.metadata (10 kB)
Collecting peft==0.8.2 (from -r requirements.txt (line 7))
Downloading peft-0.8.2-py3-none-any.whl.metadata (25 kB)
Collecting flash-attn==2.5.6 (from -r requirements.txt (line 8))
Downloading flash_attn-2.5.6.tar.gz (2.5 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2.5/2.5 MB 62.3 MB/s eta 0:00:00
Preparing metadata (setup.py): started
Preparing metadata (setup.py): finished with status 'done'
Requirement already satisfied: filelock in /opt/conda/lib/python3.10/site-packages (f
rom transformers==4.38.2->-r requirements.txt (line 1)) (3.13.1)
Requirement already satisfied: huggingface-hub<1.0,>=0.19.3 in /opt/conda/lib/python
3.10/site-packages (from transformers==4.38.2->-r requirements.txt (line 1)) (0.20.3)
Requirement already satisfied: numpy>=1.17 in /opt/conda/lib/python3.10/site-packages
(from transformers==4.38.2->-r requirements.txt (line 1)) (1.24.4)
Requirement already satisfied: packaging>=20.0 in /opt/conda/lib/python3.10/site-pack
ages (from transformers==4.38.2->-r requirements.txt (line 1)) (23.1)
Requirement already satisfied: pyyaml>=5.1 in /opt/conda/lib/python3.10/site-packages
(from transformers==4.38.2->-r requirements.txt (line 1)) (6.0.1)
Requirement already satisfied: regex!=2019.12.17 in /opt/conda/lib/python3.10/site-pa
ckages (from transformers==4.38.2->-r requirements.txt (line 1)) (2023.12.25)
Requirement already satisfied: requests in /opt/conda/lib/python3.10/site-packages (f
rom transformers==4.38.2->-r requirements.txt (line 1)) (2.31.0)
Requirement already satisfied: tokenizers<0.19,>=0.14 in /opt/conda/lib/python3.10/si
te-packages (from transformers==4.38.2->-r requirements.txt (line 1)) (0.15.1)
Requirement already satisfied: safetensors>=0.4.1 in /opt/conda/lib/python3.10/site-p
ackages (from transformers==4.38.2->-r requirements.txt (line 1)) (0.4.2)
Requirement already satisfied: tqdm>=4.27 in /opt/conda/lib/python3.10/site-packages
(from transformers==4.38.2->-r requirements.txt (line 1)) (4.66.1)
Requirement already satisfied: pyarrow>=12.0.0 in /opt/conda/lib/python3.10/site-pack
```

```
ages (from datasets==2.18.0->-r requirements.txt (line 2)) (15.0.0)
Requirement already satisfied: pyarrow-hotfix in /opt/conda/lib/python3.10/site-packa
ges (from datasets==2.18.0->-r requirements.txt (line 2)) (0.6)
Requirement already satisfied: dill<0.3.9,>=0.3.0 in /opt/conda/lib/python3.10/site-p
ackages (from datasets==2.18.0->-r requirements.txt (line 2)) (0.3.7)
Requirement already satisfied: pandas in /opt/conda/lib/python3.10/site-packages (fro
m datasets==2.18.0->-r requirements.txt (line 2)) (2.1.2)
Requirement already satisfied: xxhash in /opt/conda/lib/python3.10/site-packages (fro
m datasets==2.18.0->-r requirements.txt (line 2)) (3.4.1)
Requirement already satisfied: multiprocess in /opt/conda/lib/python3.10/site-package
s (from datasets==2.18.0->-r requirements.txt (line 2)) (0.70.15)
Requirement already satisfied: fsspec<=2024.2.0,>=2023.1.0 in /opt/conda/lib/python3.
10/site-packages (from fsspec[http]<=2024.2.0,>=2023.1.0->datasets==2.18.0->-r requir
ements.txt (line 2)) (2023.10.0)
Requirement already satisfied: aiohttp in /opt/conda/lib/python3.10/site-packages (fr
om datasets==2.18.0->-r requirements.txt (line 2)) (3.9.3)
Requirement already satisfied: psutil in /opt/conda/lib/python3.10/site-packages (fro
m accelerate==0.27.2->-r requirements.txt (line 3)) (5.9.5)
Requirement already satisfied: torch>=1.10.0 in /opt/conda/lib/python3.10/site-packag
es (from accelerate==0.27.2->-r requirements.txt (line 3)) (2.1.0)
Requirement already satisfied: responses<0.19 in /opt/conda/lib/python3.10/site-packa
ges (from evaluate==0.4.1->-r requirements.txt (line 4)) (0.18.0)
Requirement already satisfied: scipy in /opt/conda/lib/python3.10/site-packages (from
bitsandbytes==0.42.0->-r requirements.txt (line 5)) (1.11.3)
Requirement already satisfied: tyro>=0.5.11 in /opt/conda/lib/python3.10/site-package
s (from trl==0.7.11->-r requirements.txt (line 6)) (0.7.0)
Requirement already satisfied: einops in /opt/conda/lib/python3.10/site-packages (fro
m flash-attn==2.5.6->-r requirements.txt (line 8)) (0.7.0)
Requirement already satisfied: ninja in /opt/conda/lib/python3.10/site-packages (from
flash-attn==2.5.6->-r requirements.txt (line 8)) (1.11.1.1)
Requirement already satisfied: aiosignal>=1.1.2 in /opt/conda/lib/python3.10/site-pac
kages (from aiohttp->datasets==2.18.0->-r requirements.txt (line 2)) (1.3.1)
Requirement already satisfied: attrs>=17.3.0 in /opt/conda/lib/python3.10/site-packag
es (from aiohttp->datasets==2.18.0->-r requirements.txt (line 2)) (23.1.0)
Requirement already satisfied: frozenlist>=1.1.1 in /opt/conda/lib/python3.10/site-pa
ckages (from aiohttp->datasets==2.18.0->-r requirements.txt (line 2)) (1.4.1)
Requirement already satisfied: multidict<7.0,>=4.5 in /opt/conda/lib/python3.10/site-
packages (from aiohttp->datasets==2.18.0->-r requirements.txt (line 2)) (6.0.4)
Requirement already satisfied: yarl<2.0,>=1.0 in /opt/conda/lib/python3.10/site-packa
ges (from aiohttp->datasets==2.18.0->-r requirements.txt (line 2)) (1.9.4)
Requirement already satisfied: async-timeout<5.0,>=4.0 in /opt/conda/lib/python3.10/s
ite-packages (from aiohttp->datasets==2.18.0->-r requirements.txt (line 2)) (4.0.3)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /opt/conda/lib/python3.1
0/site-packages (from huggingface-hub<1.0,>=0.19.3->transformers==4.38.2->-r requirem
ents.txt (line 1)) (4.8.0)
Requirement already satisfied: charset-normalizer<4,>=2 in /opt/conda/lib/python3.10/
site-packages (from requests->transformers==4.38.2->-r requirements.txt (line 1)) (3.
2.0)
Requirement already satisfied: idna<4,>=2.5 in /opt/conda/lib/python3.10/site-package
s (from requests->transformers==4.38.2->-r requirements.txt (line 1)) (3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in /opt/conda/lib/python3.10/site-p
ackages (from requests->transformers==4.38.2->-r requirements.txt (line 1)) (1.26.18)
Requirement already satisfied: certifi>=2017.4.17 in /opt/conda/lib/python3.10/site-p
ackages (from requests->transformers==4.38.2->-r requirements.txt (line 1)) (2023.7.2
2)
Requirement already satisfied: sympy in /opt/conda/lib/python3.10/site-packages (from
torch>=1.10.0->accelerate==0.27.2->-r requirements.txt (line 3)) (1.12)
Requirement already satisfied: networkx in /opt/conda/lib/python3.10/site-packages (f
rom torch>=1.10.0->accelerate==0.27.2->-r requirements.txt (line 3)) (3.2.1)
Requirement already satisfied: jinja2 in /opt/conda/lib/python3.10/site-packages (fro
```

```
m torch>=1.10.0->accelerate==0.27.2->-r requirements.txt (line 3)) (3.1.2)
Requirement already satisfied: docstring-parser>=0.14.1 in /opt/conda/lib/python3.10/
site-packages (from tyro>=0.5.11->trl==0.7.11->-r requirements.txt (line 6)) (0.15)
Requirement already satisfied: rich>=11.1.0 in /opt/conda/lib/python3.10/site-package
s (from tyro>=0.5.11->trl==0.7.11->-r requirements.txt (line 6)) (13.6.0)
Requirement already satisfied: shtab>=1.5.6 in /opt/conda/lib/python3.10/site-package
s (from tyro>=0.5.11->trl==0.7.11->-r requirements.txt (line 6)) (1.6.5)
Requirement already satisfied: python-dateutil>=2.8.2 in /opt/conda/lib/python3.10/si
te-packages (from pandas->datasets==2.18.0->-r requirements.txt (line 2)) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /opt/conda/lib/python3.10/site-package
s (from pandas->datasets==2.18.0->-r requirements.txt (line 2)) (2023.3.post1)
Requirement already satisfied: tzdata>=2022.1 in /opt/conda/lib/python3.10/site-packa
ges (from pandas->datasets==2.18.0->-r requirements.txt (line 2)) (2023.3)
Requirement already satisfied: six>=1.5 in /opt/conda/lib/python3.10/site-packages (f
rom python-dateutil>=2.8.2->pandas->datasets==2.18.0->-r requirements.txt (line 2))
(1.16.0)
Requirement already satisfied: markdown-it-py>=2.2.0 in /opt/conda/lib/python3.10/sit
e-packages (from rich>=11.1.0->tyro>=0.5.11->trl==0.7.11->-r requirements.txt (line
6)) (3.0.0)
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in /opt/conda/lib/python3.10/s
ite-packages (from rich>=11.1.0->tyro>=0.5.11->trl==0.7.11->-r requirements.txt (line
6)) (2.16.1)
Requirement already satisfied: MarkupSafe>=2.0 in /opt/conda/lib/python3.10/site-pack
ages (from jinja2->torch>=1.10.0->accelerate==0.27.2->-r requirements.txt (line 3))
(2.1.3)
Requirement already satisfied: mpmath>=0.19 in /opt/conda/lib/python3.10/site-package
s (from sympy->torch>=1.10.0->accelerate==0.27.2->-r requirements.txt (line 3)) (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>=0.5.11->trl==0.7.11->-r requirement
s.txt (line 6)) (0.1.0)
Downloading transformers-4.38.2-py3-none-any.whl (8.5 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 8.5/8.5 MB 112.2 MB/s eta 0:00:00
Downloading datasets-2.18.0-py3-none-any.whl (510 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 510.5/510.5 kB 58.0 MB/s eta 0:00:00
Downloading accelerate-0.27.2-py3-none-any.whl (279 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 280.0/280.0 kB 43.2 MB/s eta 0:00:00
Downloading trl-0.7.11-py3-none-any.whl (155 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 155.3/155.3 kB 27.7 MB/s eta 0:00:00
Downloading peft-0.8.2-py3-none-any.whl (183 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 183.4/183.4 kB 31.4 MB/s eta 0:00:00
Building wheels for collected packages: flash-attn
Building wheel for flash-attn (setup.py): started
Building wheel for flash-attn (setup.py): finished with status 'done'
Created wheel for flash-attn: filename=flash_attn-2.5.6-cp310-cp310-linux_x86_64.whl
size=120352136 sha256=63ab5a3883b67719671e154beb91522e2901cbe4af0c5e031306a06a85df0be
5
Stored in directory: /root/.cache/pip/wheels/a8/1c/88/b959d6818b98a46d61ba231683abb75
23b89ac1a7ed1e0c206
Successfully built flash-attn
Installing collected packages: flash-attn, accelerate, transformers, datasets, trl, p
eft
Attempting uninstall: flash-attn
Found existing installation: flash-attn 2.3.6
Uninstalling flash-attn-2.3.6:
Successfully uninstalled flash-attn-2.3.6
Attempting uninstall: accelerate
Found existing installation: accelerate 0.25.0
Uninstalling accelerate-0.25.0:
Successfully uninstalled accelerate-0.25.0
```

```
Attempting uninstall: transformers
Found existing installation: transformers 4.36.0
Uninstalling transformers-4.36.0:
Successfully uninstalled transformers-4.36.0
Attempting uninstall: datasets
Found existing installation: datasets 2.15.0
Uninstalling datasets-2.15.0:
Successfully uninstalled datasets-2.15.0
Attempting uninstall: trl
Found existing installation: trl 0.7.4
Uninstalling trl-0.7.4:
Successfully uninstalled trl-0.7.4
Attempting uninstall: peft
Found existing installation: peft 0.7.1
Uninstalling peft-0.7.1:
Successfully uninstalled peft-0.7.1
Successfully installed accelerate-0.27.2 datasets-2.18.0 flash-attn-2.5.6 peft-0.8.2
transformers-4.38.2 trl-0.7.11
WARNING: Running pip as the 'root' user can result in broken permissions and conflict
ing behaviour with the system package manager. It is recommended to use a virtual env
ironment instead: https://pip.pypa.io/warnings/venv
[notice] A new release of pip is available: 23.3.2 -> 24.0
[notice] To update, run: pip install --upgrade pip
2024-03-08 08:11:21,714 sagemaker-training-toolkit INFO     Waiting for the process t
o finish and give a return code.
2024-03-08 08:11:21,714 sagemaker-training-toolkit INFO     Done waiting for a return
code. Received 0 from exiting process.
2024-03-08 08:11:21,751 sagemaker-training-toolkit INFO     No Neurons detected (norm
al if no neurons installed)
2024-03-08 08:11:21,779 sagemaker-training-toolkit INFO     No Neurons detected (norm
al if no neurons installed)
2024-03-08 08:11:21,807 sagemaker-training-toolkit INFO     No Neurons detected (norm
al if no neurons installed)
2024-03-08 08:11:21,819 sagemaker-training-toolkit INFO     Invoking user script
Training Env:
{
    "additional_framework_parameters": {},
    "channel_input_dirs": {
        "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": {
        "bf16": true,
        "dataset_path": "/opt/ml/input/data/training/train_dataset.json",
        "gradient_accumulation_steps": 4,
        "gradient_checkpointing": true,
        "learning_rate": 0.0002,
        "logging_steps": 10,
        "lr_scheduler_type": "constant",
```

```
            "max_grad_norm": 0.3,
            "max_seq_len": 3072,
            "merge_adapters": true,
            "model_id": "codellama/CodeLlama-7b-hf",
            "num_train_epochs": 3,
            "optim": "adamw_torch_fused",
            "output_dir": "/tmp/tun",
            "per_device_train_batch_size": 1,
            "report_to": "tensorboard",
            "save_strategy": "epoch",
            "tf32": true,
            "use_qlora": true,
            "warmup_ratio": 0.03
        },
        "input_config_dir": "/opt/ml/input/config",
        "input_data_config": {
            "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": false,
        "is_smddprun_installed": false,
        "job_name": "codellama-7b-hf-text-to-sql-exp1-2024-03-08-08-05-53-957",
        "log_level": 20,
        "master_hostname": "algo-1",
        "model_dir": "/opt/ml/model",
        "module_dir": "s3://sagemaker-us-east-1-558105141721/codellama-7b-hf-text-to-sql-
    exp1-2024-03-08-08-05-53-957/source/sourcedir.tar.gz",
        "module_name": "run_sft",
        "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": "run_sft.py"
}
Environment variables:
SM_HOSTS=["algo-1"]
SM_NETWORK_INTERFACE_NAME=eth0
SM_HPS={"bf16":true,"dataset_path":"/opt/ml/input/data/training/train_dataset.jso
n","gradient_accumulation_steps":4,"gradient_checkpointing":true,"learning_rate":0.00
02,"logging_steps":10,"lr_scheduler_type":"constant","max_grad_norm":0.3,"max_seq_le
n":3072,"merge_adapters":true,"model_id":"codellama/CodeLlama-7b-hf","num_train_epoch
s":3,"optim":"adamw_torch_fused","output_dir":"/tmp/tun","per_device_train_batch_siz
e":1,"report_to":"tensorboard","save_strategy":"epoch","tf32":true,"use_qlora":tru
e,"warmup_ratio":0.03}
SM_USER_ENTRY_POINT=run_sft.py
SM_FRAMEWORK_PARAMS={}
SM_RESOURCE_CONFIG={"current_group_name":"homogeneousCluster","current_host":"algo-
1","current_instance_type":"ml.g5.2xlarge","hosts":["algo-1"],"instance_groups":[{"ho
sts":["algo-1"],"instance_group_name":"homogeneousCluster","instance_type":"ml.g5.2xl
arge"}],"network_interface_name":"eth0"}
SM_INPUT_DATA_CONFIG={"training":{"RecordWrapperType":"None","S3DistributionType":"Fu
llyReplicated","TrainingInputMode":"File"}}
SM_OUTPUT_DATA_DIR=/opt/ml/output/data
SM_CHANNELS=["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_nam
e":"homogeneousCluster","instance_type":"ml.g5.2xlarge"}}
SM_DISTRIBUTION_INSTANCE_GROUPS=[]
SM_IS_HETERO=false
SM_MODULE_NAME=run_sft
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=s3://sagemaker-us-east-1-558105141721/codellama-7b-hf-text-to-sql-exp1-
2024-03-08-08-05-53-957/source/sourcedir.tar.gz
SM_TRAINING_ENV={"additional_framework_parameters":{},"channel_input_dirs":{"trainin
g":"/opt/ml/input/data/training"},"current_host":"algo-1","current_instance_group":"h
omogeneousCluster","current_instance_group_hosts":["algo-1"],"current_instance_typ
e":"ml.g5.2xlarge","distribution_hosts":[],"distribution_instance_groups":[],"framewo
rk_module":"sagemaker_pytorch_container.training:main","hosts":["algo-1"],"hyperparam
```

eters":{"bf16":true,"dataset_path":"/opt/ml/input/data/training/train_dataset.jso
n","gradient_accumulation_steps":4,"gradient_checkpointing":true,"learning_rate":0.00
02,"logging_steps":10,"lr_scheduler_type":"constant","max_grad_norm":0.3,"max_seq_le
n":3072,"merge_adapters":true,"model_id":"codellama/CodeLlama-7b-hf","num_train_epoch
s":3,"optim":"adamw_torch_fused","output_dir":"/tmp/tun","per_device_train_batch_siz
e":1,"report_to":"tensorboard","save_strategy":"epoch","tf32":true,"use_qlora":tru
e,"warmup_ratio":0.03},"input_config_dir":"/opt/ml/input/config","input_data_config":
{"training":{"RecordWrapperType":"None","S3DistributionType":"FullyReplicated","Train
ingInputMode":"File"}},"input_dir":"/opt/ml/input","instance_groups":["homogeneousClu
ster"],"instance_groups_dict":{"homogeneousCluster":{"hosts":["algo-1"],"instance_gro
up_name":"homogeneousCluster","instance_type":"ml.g5.2xlarge"}},"is_hetero":false,"is
_master":true,"is_modelparallel_enabled":null,"is_smddpmprun_installed":false,"is_smd
dprun_installed":false,"job_name":"codellama-7b-hf-text-to-sql-exp1-2024-03-08-08-05-
53-957","log_level":20,"master_hostname":"algo-1","model_dir":"/opt/ml/model","module
_dir":"s3://sagemaker-us-east-1-558105141721/codellama-7b-hf-text-to-sql-exp1-2024-03
-08-08-05-53-957/source/sourcedir.tar.gz","module_name":"run_sft","network_interface_
name":"eth0","num_cpus":8,"num_gpus":1,"num_neurons":0,"output_data_dir":"/opt/ml/out
put/data","output_dir":"/opt/ml/output","output_intermediate_dir":"/opt/ml/output/int
ermediate","resource_config":{"current_group_name":"homogeneousCluster","current_hos
t":"algo-1","current_instance_type":"ml.g5.2xlarge","hosts":["algo-1"],"instance_grou
ps":[{"hosts":["algo-1"],"instance_group_name":"homogeneousCluster","instance_typ
e":"ml.g5.2xlarge"}],"network_interface_name":"eth0"},"user_entry_point":"run_sft.p
y"}
SM_USER_ARGS=["--bf16","True","--dataset_path","/opt/ml/input/data/training/train_dat
aset.json","--gradient_accumulation_steps","4","--gradient_checkpointing","True","--l
earning_rate","0.0002","--logging_steps","10","--lr_scheduler_type","constant","--max
_grad_norm","0.3","--max_seq_len","3072","--merge_adapters","True","--model_id","code
llama/CodeLlama-7b-hf","--num_train_epochs","3","--optim","adamw_torch_fused","--outp
ut_dir","/tmp/tun","--per_device_train_batch_size","1","--report_to","tensorboard","-
-save_strategy","epoch","--tf32","True","--use_qlora","True","--warmup_ratio","0.03"]
SM_OUTPUT_INTERMEDIATE_DIR=/opt/ml/output/intermediate
SM_CHANNEL_TRAINING=/opt/ml/input/data/training
SM_HP_BF16=true
SM_HP_DATASET_PATH=/opt/ml/input/data/training/train_dataset.json
SM_HP_GRADIENT_ACCUMULATION_STEPS=4
SM_HP_GRADIENT_CHECKPOINTING=true
SM_HP_LEARNING_RATE=0.0002
SM_HP_LOGGING_STEPS=10
SM_HP_LR_SCHEDULER_TYPE=constant
SM_HP_MAX_GRAD_NORM=0.3
SM_HP_MAX_SEQ_LEN=3072
SM_HP_MERGE_ADAPTERS=true
SM_HP_MODEL_ID=codellama/CodeLlama-7b-hf
SM_HP_NUM_TRAIN_EPOCHS=3
SM_HP_OPTIM=adamw_torch_fused
SM_HP_OUTPUT_DIR=/tmp/tun
SM_HP_PER_DEVICE_TRAIN_BATCH_SIZE=1
SM_HP_REPORT_TO=tensorboard
SM_HP_SAVE_STRATEGY=epoch
SM_HP_TF32=true
SM_HP_USE_QLORA=true
SM_HP_WARMUP_RATIO=0.03
PYTHONPATH=/opt/ml/code:/opt/conda/bin:/opt/conda/lib/python310.zip:/opt/conda/lib/py
thon3.10:/opt/conda/lib/python3.10/lib-dynload:/opt/conda/lib/python3.10/site-package
s
Invoking script with the following command:
/opt/conda/bin/python3.10 run_sft.py --bf16 True --dataset_path /opt/ml/input/data/tr
aining/train_dataset.json --gradient_accumulation_steps 4 --gradient_checkpointing Tr
ue --learning_rate 0.0002 --logging_steps 10 --lr_scheduler_type constant --max_grad_
norm 0.3 --max_seq_len 3072 --merge_adapters True --model_id codellama/CodeLlama-7b-h

```
f --num_train_epochs 3 --optim adamw_torch_fused --output_dir /tmp/tun --per_device_t
rain_batch_size 1 --report_to tensorboard --save_strategy epoch --tf32 True --use_qlo
ra True --warmup_ratio 0.03
2024-03-08 08:11:21,820 sagemaker-training-toolkit INFO     Exceptions not imported f
or SageMaker Debugger as it is not installed.
2024-03-08 08:11:21,820 sagemaker-training-toolkit INFO     Exceptions not imported f
or SageMaker TF as Tensorflow is not installed.
Generating train split: 0 examples [00:00, ? examples/s]
Generating train split: 10000 examples [00:00, 688538.97 examples/s]
Using QLoRA
config.json:   0%|          | 0.00/637 [00:00<?, ?B/s]
config.json: 100%|██████████| 637/637 [00:00<00:00, 7.30MB/s]
model.safetensors.index.json:   0%|          | 0.00/25.1k [00:00<?, ?B/s]
model.safetensors.index.json: 100%|██████████| 25.1k/25.1k [00:00<00:00, 159MB/s]
Downloading shards:   0%|          | 0/2 [00:00<?, ?it/s]
model-00001-of-00002.safetensors:   0%|          | 0.00/9.98G [00:00<?, ?B/s]#033[A
model-00001-of-00002.safetensors:   1%|          | 62.9M/9.98G [00:00<00:18, 551MB/s]
#033[A
model-00001-of-00002.safetensors:   1%|▏         | 126M/9.98G [00:00<00:20, 487MB/s]
#033[A
model-00001-of-00002.safetensors:   2%|▏         | 178M/9.98G [00:00<00:20, 474MB/s]#
033[A
model-00001-of-00002.safetensors:   2%|▏         | 231M/9.98G [00:00<00:20, 483MB/s]#
033[A
model-00001-of-00002.safetensors:   3%|▎         | 283M/9.98G [00:00<00:19, 488MB/s]#
033[A
model-00001-of-00002.safetensors:   3%|▎         | 336M/9.98G [00:00<00:19, 488MB/s]#
033[A
model-00001-of-00002.safetensors:   4%|▍         | 388M/9.98G [00:00<00:19, 491MB/s]#
033[A
model-00001-of-00002.safetensors:   5%|▍         | 451M/9.98G [00:00<00:18, 517MB/s]#
033[A
model-00001-of-00002.safetensors:   5%|▌         | 503M/9.98G [00:01<00:18, 508MB/s]#
033[A
model-00001-of-00002.safetensors:   6%|▌         | 556M/9.98G [00:01<00:19, 484MB/s]#
033[A
model-00001-of-00002.safetensors:   6%|▋         | 608M/9.98G [00:01<00:19, 471MB/s]#
033[A
model-00001-of-00002.safetensors:   7%|▋         | 661M/9.98G [00:01<00:20, 458MB/s]#
033[A
model-00001-of-00002.safetensors:   7%|▋         | 713M/9.98G [00:01<00:20, 463MB/s]#
033[A
model-00001-of-00002.safetensors:   8%|▊         | 765M/9.98G [00:01<00:20, 454MB/s]#
033[A
model-00001-of-00002.safetensors:   8%|▊         | 818M/9.98G [00:01<00:20, 442MB/s]#
033[A
model-00001-of-00002.safetensors:   9%|▉         | 870M/9.98G [00:01<00:20, 454MB/s]#
033[A
model-00001-of-00002.safetensors:   9%|▉         | 923M/9.98G [00:01<00:20, 449MB/s]#
033[A
model-00001-of-00002.safetensors:  10%|▉         | 975M/9.98G [00:02<00:19, 460MB/s]#
033[A
model-00001-of-00002.safetensors:  10%|▉         | 1.03G/9.98G [00:02<00:19, 451MB/s]
#033[A
model-00001-of-00002.safetensors:  11%|▉         | 1.08G/9.98G [00:02<00:20, 430MB/s]
#033[A
model-00001-of-00002.safetensors:  11%|█         | 1.13G/9.98G [00:02<00:20, 438MB/s]
#033[A
model-00001-of-00002.safetensors:  12%|█         | 1.18G/9.98G [00:02<00:19, 441MB/s]
#033[A
```

```
model-00001-of-00002.safetensors:  12%|▉          | 1.24G/9.98G [00:02<00:19, 443MB/s]
#033[A
model-00001-of-00002.safetensors:  13%|▉          | 1.29G/9.98G [00:02<00:19, 444MB/s]
#033[A
model-00001-of-00002.safetensors:  13%|▉          | 1.34G/9.98G [00:02<00:18, 458MB/s]
#033[A
model-00001-of-00002.safetensors:  14%|▉          | 1.39G/9.98G [00:03<00:19, 442MB/s]
#033[A
model-00001-of-00002.safetensors:  15%|▉          | 1.45G/9.98G [00:03<00:18, 454MB/s]
#033[A
model-00001-of-00002.safetensors:  15%|▉          | 1.50G/9.98G [00:03<00:18, 466MB/s]
#033[A
model-00001-of-00002.safetensors:  16%|▉          | 1.55G/9.98G [00:03<00:19, 440MB/s]
#033[A
model-00001-of-00002.safetensors:  16%|▉          | 1.60G/9.98G [00:03<00:18, 447MB/s]
#033[A
model-00001-of-00002.safetensors:  17%|▉▏         | 1.66G/9.98G [00:03<00:19, 425MB/s]
#033[A
model-00001-of-00002.safetensors:  17%|▉▏         | 1.72G/9.98G [00:03<00:17, 463MB/s]
#033[A
model-00001-of-00002.safetensors:  18%|▉▏         | 1.77G/9.98G [00:03<00:18, 456MB/s]
#033[A
model-00001-of-00002.safetensors:  18%|▉▏         | 1.82G/9.98G [00:03<00:18, 432MB/s]
#033[A
model-00001-of-00002.safetensors:  19%|▉▏         | 1.88G/9.98G [00:04<00:18, 444MB/s]
#033[A
model-00001-of-00002.safetensors:  19%|▉▏         | 1.93G/9.98G [00:04<00:17, 456MB/s]
#033[A
model-00001-of-00002.safetensors:  20%|▉▏         | 1.98G/9.98G [00:04<00:17, 450MB/s]
#033[A
model-00001-of-00002.safetensors:  20%|▉▏         | 2.03G/9.98G [00:04<00:17, 463MB/s]
#033[A
model-00001-of-00002.safetensors:  21%|▉▏         | 2.09G/9.98G [00:04<00:17, 457MB/s]
#033[A
model-00001-of-00002.safetensors:  21%|▉▏         | 2.14G/9.98G [00:04<00:17, 447MB/s]
#033[A
model-00001-of-00002.safetensors:  22%|▉▏         | 2.19G/9.98G [00:04<00:17, 440MB/s]
#033[A
model-00001-of-00002.safetensors:  22%|▉▏         | 2.24G/9.98G [00:04<00:16, 455MB/s]
#033[A
model-00001-of-00002.safetensors:  23%|▉▏         | 2.30G/9.98G [00:05<00:16, 467MB/s]
#033[A
model-00001-of-00002.safetensors:  24%|▉▏         | 2.35G/9.98G [00:05<00:16, 465MB/s]
#033[A
model-00001-of-00002.safetensors:  24%|▉▏         | 2.40G/9.98G [00:05<00:16, 454MB/s]
#033[A
model-00001-of-00002.safetensors:  25%|▉▏         | 2.45G/9.98G [00:05<00:16, 449MB/s]
#033[A
model-00001-of-00002.safetensors:  25%|▉▏         | 2.51G/9.98G [00:05<00:15, 469MB/s]
#033[A
model-00001-of-00002.safetensors:  26%|▉▏         | 2.56G/9.98G [00:05<00:15, 471MB/s]
#033[A
model-00001-of-00002.safetensors:  26%|▉▏         | 2.61G/9.98G [00:05<00:15, 461MB/s]
#033[A
model-00001-of-00002.safetensors:  27%|▉▏         | 2.66G/9.98G [00:05<00:16, 452MB/s]
#033[A
model-00001-of-00002.safetensors:  27%|▉▏         | 2.73G/9.98G [00:05<00:15, 470MB/s]
#033[A
model-00001-of-00002.safetensors:  28%|▉▏         | 2.78G/9.98G [00:06<00:15, 459MB/s]
#033[A
```

```
model-00001-of-00002.safetensors:  28%|██▊       | 2.83G/9.98G [00:06<00:15, 469MB/s]
#033[A
model-00001-of-00002.safetensors:  29%|██▉       | 2.88G/9.98G [00:06<00:14, 473MB/s]
#033[A
model-00001-of-00002.safetensors:  29%|██▉       | 2.94G/9.98G [00:06<00:15, 465MB/s]
#033[A
model-00001-of-00002.safetensors:  30%|██▉       | 2.99G/9.98G [00:06<00:14, 476MB/s]
#033[A
model-00001-of-00002.safetensors:  30%|███       | 3.04G/9.98G [00:06<00:14, 479MB/s]
#033[A
model-00001-of-00002.safetensors:  31%|███       | 3.09G/9.98G [00:06<00:14, 465MB/s]
#033[A
model-00001-of-00002.safetensors:  32%|███       | 3.15G/9.98G [00:06<00:14, 471MB/s]
#033[A
model-00001-of-00002.safetensors:  32%|███       | 3.20G/9.98G [00:06<00:14, 460MB/s]
#033[A
model-00001-of-00002.safetensors:  33%|███       | 3.25G/9.98G [00:07<00:14, 469MB/s]
#033[A
model-00001-of-00002.safetensors:  33%|███       | 3.30G/9.98G [00:07<00:14, 470MB/s]
#033[A
model-00001-of-00002.safetensors:  34%|███       | 3.36G/9.98G [00:07<00:14, 459MB/s]
#033[A
model-00001-of-00002.safetensors:  34%|███       | 3.41G/9.98G [00:07<00:13, 470MB/s]
#033[A
model-00001-of-00002.safetensors:  35%|███       | 3.46G/9.98G [00:07<00:13, 472MB/s]
#033[A
model-00001-of-00002.safetensors:  35%|███       | 3.52G/9.98G [00:07<00:13, 495MB/s]
#033[A
model-00001-of-00002.safetensors:  36%|███       | 3.58G/9.98G [00:07<00:13, 476MB/s]
#033[A
model-00001-of-00002.safetensors:  36%|███▌      | 3.63G/9.98G [00:07<00:13, 487MB/s]
#033[A
model-00001-of-00002.safetensors:  37%|███▌      | 3.69G/9.98G [00:07<00:12, 507MB/s]
#033[A
model-00001-of-00002.safetensors:  38%|███▊      | 3.74G/9.98G [00:08<00:12, 483MB/s]
#033[A
model-00001-of-00002.safetensors:  38%|███▊      | 3.80G/9.98G [00:08<00:12, 487MB/s]
#033[A
model-00001-of-00002.safetensors:  39%|███▊      | 3.86G/9.98G [00:08<00:12, 499MB/s]
#033[A
model-00001-of-00002.safetensors:  39%|███▉      | 3.91G/9.98G [00:08<00:12, 495MB/s]
#033[A
model-00001-of-00002.safetensors:  40%|███▉      | 3.96G/9.98G [00:08<00:12, 477MB/s]
#033[A
model-00001-of-00002.safetensors:  40%|████      | 4.02G/9.98G [00:08<00:12, 489MB/s]
#033[A
model-00001-of-00002.safetensors:  41%|████      | 4.07G/9.98G [00:08<00:12, 489MB/s]
#033[A
model-00001-of-00002.safetensors:  41%|████      | 4.12G/9.98G [00:08<00:11, 495MB/s]
#033[A
model-00001-of-00002.safetensors:  42%|████      | 4.17G/9.98G [00:08<00:12, 475MB/s]
#033[A
model-00001-of-00002.safetensors:  42%|████      | 4.23G/9.98G [00:09<00:12, 466MB/s]
#033[A
model-00001-of-00002.safetensors:  43%|████      | 4.28G/9.98G [00:09<00:12, 467MB/s]
#033[A
model-00001-of-00002.safetensors:  44%|████      | 4.34G/9.98G [00:09<00:11, 482MB/s]
#033[A
model-00001-of-00002.safetensors:  44%|████▍     | 4.39G/9.98G [00:09<00:11, 487MB/s]
#033[A
```

```
model-00001-of-00002.safetensors:  45%|████▌      | 4.46G/9.98G [00:09<00:10, 507MB/s]
#033[A
model-00001-of-00002.safetensors:  45%|████▌      | 4.52G/9.98G [00:09<00:10, 514MB/s]
#033[A
model-00001-of-00002.safetensors:  46%|████▌      | 4.58G/9.98G [00:09<00:10, 499MB/s]
#033[A
model-00001-of-00002.safetensors:  46%|████▋      | 4.63G/9.98G [00:09<00:10, 501MB/s]
#033[A
model-00001-of-00002.safetensors:  47%|████▋      | 4.69G/9.98G [00:09<00:10, 501MB/s]
#033[A
model-00001-of-00002.safetensors:  48%|████▊      | 4.74G/9.98G [00:10<00:11, 474MB/s]
#033[A
model-00001-of-00002.safetensors:  48%|████▊      | 4.79G/9.98G [00:10<00:10, 485MB/s]
#033[A
model-00001-of-00002.safetensors:  49%|████▊      | 4.84G/9.98G [00:10<00:10, 481MB/s]
#033[A
model-00001-of-00002.safetensors:  49%|████▉      | 4.90G/9.98G [00:10<00:11, 438MB/s]
#033[A
model-00001-of-00002.safetensors:  50%|████▉      | 4.95G/9.98G [00:10<00:11, 448MB/s]
#033[A
model-00001-of-00002.safetensors:  50%|█████      | 5.00G/9.98G [00:10<00:10, 459MB/s]
#033[A
model-00001-of-00002.safetensors:  51%|█████      | 5.05G/9.98G [00:10<00:12, 387MB/s]
#033[A
model-00001-of-00002.safetensors:  51%|█████      | 5.12G/9.98G [00:10<00:11, 434MB/s]
#033[A
model-00001-of-00002.safetensors:  52%|█████▏     | 5.17G/9.98G [00:11<00:10, 450MB/s]
#033[A
model-00001-of-00002.safetensors:  52%|█████▏     | 5.22G/9.98G [00:11<00:10, 465MB/s]
#033[A
model-00001-of-00002.safetensors:  53%|█████▎     | 5.27G/9.98G [00:11<00:10, 470MB/s]
#033[A
model-00001-of-00002.safetensors:  53%|█████▎     | 5.33G/9.98G [00:11<00:09, 471MB/s]
#033[A
model-00001-of-00002.safetensors:  54%|█████▎     | 5.38G/9.98G [00:11<00:09, 475MB/s]
#033[A
model-00001-of-00002.safetensors:  54%|█████▍     | 5.43G/9.98G [00:11<00:09, 471MB/s]
#033[A
model-00001-of-00002.safetensors:  55%|█████▍     | 5.48G/9.98G [00:11<00:09, 463MB/s]
#033[A
model-00001-of-00002.safetensors:  55%|█████▌     | 5.54G/9.98G [00:11<00:09, 469MB/s]
#033[A
model-00001-of-00002.safetensors:  56%|█████▌     | 5.59G/9.98G [00:12<00:11, 397MB/s]
#033[A
model-00001-of-00002.safetensors:  56%|█████▌     | 5.63G/9.98G [00:12<00:11, 378MB/s]
#033[A
model-00001-of-00002.safetensors:  57%|█████▋     | 5.68G/9.98G [00:12<00:10, 392MB/s]
#033[A
model-00001-of-00002.safetensors:  57%|█████▋     | 5.73G/9.98G [00:12<00:13, 311MB/s]
#033[A
model-00001-of-00002.safetensors:  58%|█████▊     | 5.77G/9.98G [00:12<00:13, 323MB/s]
#033[A
model-00001-of-00002.safetensors:  58%|█████▊     | 5.82G/9.98G [00:12<00:12, 339MB/s]
#033[A
model-00001-of-00002.safetensors:  59%|█████▊     | 5.86G/9.98G [00:12<00:14, 282MB/s]
#033[A
model-00001-of-00002.safetensors:  59%|█████▉     | 5.91G/9.98G [00:13<00:12, 329MB/s]
#033[A
model-00001-of-00002.safetensors:  60%|█████▉     | 5.96G/9.98G [00:13<00:18, 214MB/s]
#033[A
```

```
model-00001-of-00002.safetensors:  60%|[redacted]    | 6.02G/9.98G [00:13<00:14, 271MB/s]
#033[A
model-00001-of-00002.safetensors:  61%|[redacted]    | 6.08G/9.98G [00:13<00:11, 328MB/s]
#033[A
model-00001-of-00002.safetensors:  61%|[redacted]    | 6.12G/9.98G [00:13<00:12, 307MB/s]
#033[A
model-00001-of-00002.safetensors:  62%|[redacted]    | 6.17G/9.98G [00:13<00:11, 328MB/s]
#033[A
model-00001-of-00002.safetensors:  62%|[redacted]    | 6.21G/9.98G [00:14<00:11, 326MB/s]
#033[A
model-00001-of-00002.safetensors:  63%|[redacted]    | 6.25G/9.98G [00:14<00:14, 255MB/s]
#033[A
model-00001-of-00002.safetensors:  63%|[redacted]    | 6.28G/9.98G [00:14<00:16, 225MB/s]
#033[A
model-00001-of-00002.safetensors:  63%|[redacted]    | 6.31G/9.98G [00:14<00:18, 203MB/s]
#033[A
model-00001-of-00002.safetensors:  64%|[redacted]    | 6.34G/9.98G [00:14<00:19, 191MB/s]
#033[A
model-00001-of-00002.safetensors:  64%|[redacted]    | 6.36G/9.98G [00:15<00:20, 180MB/s]
#033[A
model-00001-of-00002.safetensors:  64%|[redacted]    | 6.39G/9.98G [00:15<00:19, 183MB/s]
#033[A
model-00001-of-00002.safetensors:  64%|[redacted]    | 6.41G/9.98G [00:15<00:19, 179MB/s]
#033[A
model-00001-of-00002.safetensors:  64%|[redacted]    | 6.43G/9.98G [00:15<00:20, 169MB/s]
#033[A
model-00001-of-00002.safetensors:  65%|[redacted]    | 6.46G/9.98G [00:15<00:19, 182MB/s]
#033[A
model-00001-of-00002.safetensors:  65%|[redacted]    | 6.48G/9.98G [00:15<00:18, 187MB/s]
#033[A
model-00001-of-00002.safetensors:  65%|[redacted]    | 6.51G/9.98G [00:15<00:18, 183MB/s]
#033[A
model-00001-of-00002.safetensors:  66%|[redacted]    | 6.54G/9.98G [00:16<00:17, 199MB/s]
#033[A
model-00001-of-00002.safetensors:  66%|[redacted]    | 6.57G/9.98G [00:16<00:16, 202MB/s]
#033[A
model-00001-of-00002.safetensors:  66%|[redacted]    | 6.60G/9.98G [00:16<00:16, 202MB/s]
#033[A
model-00001-of-00002.safetensors:  66%|[redacted]    | 6.63G/9.98G [00:16<00:16, 201MB/s]
#033[A
model-00001-of-00002.safetensors:  67%|[redacted]    | 6.66G/9.98G [00:16<00:15, 209MB/s]
#033[A
model-00001-of-00002.safetensors:  67%|[redacted]    | 6.69G/9.98G [00:16<00:16, 204MB/s]
#033[A
model-00001-of-00002.safetensors:  67%|[redacted]    | 6.72G/9.98G [00:16<00:15, 208MB/s]
#033[A
model-00001-of-00002.safetensors:  68%|[redacted]    | 6.74G/9.98G [00:16<00:15, 204MB/s]
#033[A
model-00001-of-00002.safetensors:  68%|[redacted]    | 6.77G/9.98G [00:17<00:15, 207MB/s]
#033[A
model-00001-of-00002.safetensors:  68%|[redacted]    | 6.79G/9.98G [00:17<00:15, 204MB/s]
#033[A
model-00001-of-00002.safetensors:  68%|[redacted]    | 6.83G/9.98G [00:17<00:15, 201MB/s]
#033[A
model-00001-of-00002.safetensors:  69%|[redacted]    | 6.86G/9.98G [00:17<00:15, 206MB/s]
#033[A
model-00001-of-00002.safetensors:  69%|[redacted]    | 6.89G/9.98G [00:17<00:15, 204MB/s]
#033[A
model-00001-of-00002.safetensors:  69%|[redacted]    | 6.92G/9.98G [00:17<00:14, 207MB/s]
#033[A
```

```
model-00001-of-00002.safetensors:  70%|██████▉   | 6.94G/9.98G [00:17<00:14, 206MB/s]
#033[A
model-00001-of-00002.safetensors:  70%|██████▉   | 6.97G/9.98G [00:18<00:14, 204MB/s]
#033[A
model-00001-of-00002.safetensors:  70%|██████▉   | 7.00G/9.98G [00:18<00:14, 204MB/s]
#033[A
model-00001-of-00002.safetensors:  71%|███████   | 7.04G/9.98G [00:18<00:14, 208MB/s]
#033[A
model-00001-of-00002.safetensors:  71%|███████   | 7.06G/9.98G [00:18<00:14, 208MB/s]
#033[A
model-00001-of-00002.safetensors:  71%|███████   | 7.08G/9.98G [00:18<00:14, 205MB/s]
#033[A
model-00001-of-00002.safetensors:  71%|███████   | 7.11G/9.98G [00:18<00:13, 208MB/s]
#033[A
model-00001-of-00002.safetensors:  71%|███████   | 7.13G/9.98G [00:18<00:13, 204MB/s]
#033[A
model-00001-of-00002.safetensors:  72%|███████   | 7.16G/9.98G [00:19<00:13, 206MB/s]
#033[A
model-00001-of-00002.safetensors:  72%|███████   | 7.18G/9.98G [00:19<00:13, 204MB/s]
#033[A
model-00001-of-00002.safetensors:  72%|███████   | 7.21G/9.98G [00:19<00:13, 208MB/s]
#033[A
model-00001-of-00002.safetensors:  73%|███████   | 7.24G/9.98G [00:19<00:13, 203MB/s]
#033[A
model-00001-of-00002.safetensors:  73%|███████   | 7.27G/9.98G [00:19<00:13, 207MB/s]
#033[A
model-00001-of-00002.safetensors:  73%|███████   | 7.29G/9.98G [00:19<00:13, 203MB/s]
#033[A
model-00001-of-00002.safetensors:  73%|███████   | 7.32G/9.98G [00:19<00:12, 207MB/s]
#033[A
model-00001-of-00002.safetensors:  74%|███████   | 7.34G/9.98G [00:19<00:12, 203MB/s]
#033[A
model-00001-of-00002.safetensors:  74%|███████   | 7.36G/9.98G [00:20<00:12, 204MB/s]
#033[A
model-00001-of-00002.safetensors:  74%|███████   | 7.39G/9.98G [00:20<00:12, 207MB/s]
#033[A
model-00001-of-00002.safetensors:  74%|███████   | 7.41G/9.98G [00:20<00:12, 207MB/s]
#033[A
model-00001-of-00002.safetensors:  75%|███████   | 7.44G/9.98G [00:20<00:12, 209MB/s]
#033[A
model-00001-of-00002.safetensors:  75%|███████   | 7.47G/9.98G [00:20<00:12, 209MB/s]
#033[A
model-00001-of-00002.safetensors:  75%|███████   | 7.49G/9.98G [00:20<00:12, 200MB/s]
#033[A
model-00001-of-00002.safetensors:  75%|███████   | 7.52G/9.98G [00:20<00:11, 206MB/s]
#033[A
model-00001-of-00002.safetensors:  76%|███████   | 7.54G/9.98G [00:20<00:12, 201MB/s]
#033[A
model-00001-of-00002.safetensors:  76%|███████   | 7.57G/9.98G [00:21<00:11, 205MB/s]
#033[A
model-00001-of-00002.safetensors:  76%|███████   | 7.60G/9.98G [00:21<00:11, 202MB/s]
#033[A
model-00001-of-00002.safetensors:  77%|███████▋  | 7.63G/9.98G [00:21<00:11, 208MB/s]
#033[A
model-00001-of-00002.safetensors:  77%|███████▋  | 7.67G/9.98G [00:21<00:11, 207MB/s]
#033[A
model-00001-of-00002.safetensors:  77%|███████▋  | 7.70G/9.98G [00:21<00:10, 209MB/s]
#033[A
model-00001-of-00002.safetensors:  77%|███████▋  | 7.72G/9.98G [00:21<00:11, 202MB/s]
#033[A
```

```
model-00001-of-00002.safetensors:  78%|███████▊  | 7.75G/9.98G [00:21<00:10, 206MB/s]
#033[A
model-00001-of-00002.safetensors:  78%|███████▊  | 7.77G/9.98G [00:22<00:11, 200MB/s]
#033[A
model-00001-of-00002.safetensors:  78%|███████▊  | 7.79G/9.98G [00:22<00:10, 199MB/s]
#033[A
model-00001-of-00002.safetensors:  78%|███████▊  | 7.81G/9.98G [00:22<00:10, 200MB/s]
#033[A
model-00001-of-00002.safetensors:  79%|███████▊  | 7.83G/9.98G [00:22<00:11, 195MB/s]
#033[A
model-00001-of-00002.safetensors:  79%|███████▊  | 7.85G/9.98G [00:22<00:11, 189MB/s]
#033[A
model-00001-of-00002.safetensors:  79%|███████▊  | 7.87G/9.98G [00:22<00:10, 194MB/s]
#033[A
model-00001-of-00002.safetensors:  79%|███████▊  | 7.90G/9.98G [00:22<00:10, 192MB/s]
#033[A
model-00001-of-00002.safetensors:  79%|███████▊  | 7.92G/9.98G [00:22<00:11, 187MB/s]
#033[A
model-00001-of-00002.safetensors:  80%|███████▊  | 7.94G/9.98G [00:22<00:10, 187MB/s]
#033[A
model-00001-of-00002.safetensors:  80%|███████▊  | 7.96G/9.98G [00:23<00:11, 183MB/s]
#033[A
model-00001-of-00002.safetensors:  80%|███████▊  | 7.98G/9.98G [00:23<00:10, 188MB/s]
#033[A
model-00001-of-00002.safetensors:  80%|████████  | 8.00G/9.98G [00:23<00:10, 182MB/s]
#033[A
model-00001-of-00002.safetensors:  80%|████████  | 8.02G/9.98G [00:23<00:10, 188MB/s]
#033[A
model-00001-of-00002.safetensors:  81%|████████  | 8.04G/9.98G [00:23<00:16, 120MB/s]
#033[A
model-00001-of-00002.safetensors:  81%|████████  | 8.11G/9.98G [00:23<00:08, 210MB/s]
#033[A
model-00001-of-00002.safetensors:  82%|████████  | 8.14G/9.98G [00:23<00:09, 199MB/s]
#033[A
model-00001-of-00002.safetensors:  82%|████████  | 8.17G/9.98G [00:24<00:09, 194MB/s]
#033[A
model-00001-of-00002.safetensors:  82%|████████  | 8.20G/9.98G [00:24<00:09, 191MB/s]
#033[A
model-00001-of-00002.safetensors:  82%|████████  | 8.22G/9.98G [00:24<00:09, 190MB/s]
#033[A
model-00001-of-00002.safetensors:  83%|████████  | 8.24G/9.98G [00:24<00:09, 188MB/s]
#033[A
model-00001-of-00002.safetensors:  83%|████████  | 8.26G/9.98G [00:24<00:08, 192MB/s]
#033[A
model-00001-of-00002.safetensors:  83%|████████  | 8.28G/9.98G [00:24<00:09, 187MB/s]
#033[A
model-00001-of-00002.safetensors:  83%|████████  | 8.30G/9.98G [00:24<00:08, 187MB/s]
#033[A
model-00001-of-00002.safetensors:  83%|████████  | 8.33G/9.98G [00:24<00:08, 185MB/s]
#033[A
model-00001-of-00002.safetensors:  84%|████████  | 8.35G/9.98G [00:25<00:08, 185MB/s]
#033[A
model-00001-of-00002.safetensors:  84%|████████  | 8.37G/9.98G [00:25<00:08, 183MB/s]
#033[A
model-00001-of-00002.safetensors:  84%|████████  | 8.39G/9.98G [00:25<00:08, 184MB/s]
#033[A
model-00001-of-00002.safetensors:  84%|████████  | 8.41G/9.98G [00:25<00:08, 178MB/s]
#033[A
model-00001-of-00002.safetensors:  85%|████████  | 8.43G/9.98G [00:25<00:08, 178MB/s]
#033[A
```

```
model-00001-of-00002.safetensors:  85%|████████▌  | 8.45G/9.98G [00:25<00:08, 182MB/s]
#033[A
model-00001-of-00002.safetensors:  85%|████████▌  | 8.47G/9.98G [00:25<00:08, 172MB/s]
#033[A
model-00001-of-00002.safetensors:  85%|████████▌  | 8.49G/9.98G [00:25<00:08, 181MB/s]
#033[A
model-00001-of-00002.safetensors:  85%|████████▌  | 8.51G/9.98G [00:26<00:08, 179MB/s]
#033[A
model-00001-of-00002.safetensors:  86%|████████▌  | 8.54G/9.98G [00:26<00:07, 185MB/s]
#033[A
model-00001-of-00002.safetensors:  86%|████████▌  | 8.56G/9.98G [00:26<00:08, 176MB/s]
#033[A
model-00001-of-00002.safetensors:  86%|████████▌  | 8.58G/9.98G [00:26<00:07, 176MB/s]
#033[A
model-00001-of-00002.safetensors:  86%|████████▌  | 8.60G/9.98G [00:26<00:08, 171MB/s]
#033[A
model-00001-of-00002.safetensors:  86%|████████▌  | 8.62G/9.98G [00:26<00:07, 173MB/s]
#033[A
model-00001-of-00002.safetensors:  87%|████████▋  | 8.65G/9.98G [00:26<00:07, 179MB/s]
#033[A
model-00001-of-00002.safetensors:  87%|████████▋  | 8.67G/9.98G [00:26<00:07, 179MB/s]
#033[A
model-00001-of-00002.safetensors:  87%|████████▋  | 8.69G/9.98G [00:27<00:07, 183MB/s]
#033[A
model-00001-of-00002.safetensors:  87%|████████▋  | 8.71G/9.98G [00:27<00:06, 183MB/s]
#033[A
model-00001-of-00002.safetensors:  88%|████████▊  | 8.73G/9.98G [00:27<00:06, 184MB/s]
#033[A
model-00001-of-00002.safetensors:  88%|████████▊  | 8.76G/9.98G [00:27<00:06, 179MB/s]
#033[A
model-00001-of-00002.safetensors:  88%|████████▊  | 8.78G/9.98G [00:27<00:06, 180MB/s]
#033[A
model-00001-of-00002.safetensors:  88%|████████▊  | 8.80G/9.98G [00:27<00:06, 172MB/s]
#033[A
model-00001-of-00002.safetensors:  88%|████████▊  | 8.82G/9.98G [00:27<00:06, 178MB/s]
#033[A
model-00001-of-00002.safetensors:  89%|████████▊  | 8.84G/9.98G [00:27<00:06, 180MB/s]
#033[A
model-00001-of-00002.safetensors:  89%|████████▉  | 8.86G/9.98G [00:27<00:06, 179MB/s]
#033[A
model-00001-of-00002.safetensors:  89%|████████▉  | 8.88G/9.98G [00:28<00:06, 176MB/s]
#033[A
model-00001-of-00002.safetensors:  89%|████████▉  | 8.90G/9.98G [00:28<00:06, 173MB/s]
#033[A
model-00001-of-00002.safetensors:  89%|████████▉  | 8.92G/9.98G [00:28<00:06, 173MB/s]
#033[A
model-00001-of-00002.safetensors:  90%|████████▉  | 8.94G/9.98G [00:28<00:06, 168MB/s]
#033[A
model-00001-of-00002.safetensors:  90%|████████▉  | 8.97G/9.98G [00:28<00:05, 175MB/s]
#033[A
model-00001-of-00002.safetensors:  90%|█████████  | 8.99G/9.98G [00:28<00:05, 171MB/s]
#033[A
model-00001-of-00002.safetensors:  90%|█████████  | 9.01G/9.98G [00:28<00:05, 168MB/s]
#033[A
model-00001-of-00002.safetensors:  90%|█████████  | 9.03G/9.98G [00:28<00:05, 168MB/s]
#033[A
model-00001-of-00002.safetensors:  91%|█████████  | 9.05G/9.98G [00:29<00:05, 164MB/s]
#033[A
model-00001-of-00002.safetensors:  91%|█████████  | 9.08G/9.98G [00:29<00:05, 175MB/s]
#033[A
```

```
model-00001-of-00002.safetensors:  91%|[redacted] | 9.10G/9.98G [00:29<00:05, 172MB/s]
#033[A
model-00001-of-00002.safetensors:  91%|[redacted] | 9.12G/9.98G [00:29<00:04, 174MB/s]
#033[A
model-00001-of-00002.safetensors:  92%|[redacted] | 9.14G/9.98G [00:29<00:04, 171MB/s]
#033[A
model-00001-of-00002.safetensors:  92%|[redacted] | 9.16G/9.98G [00:29<00:04, 168MB/s]
#033[A
model-00001-of-00002.safetensors:  92%|[redacted] | 9.19G/9.98G [00:29<00:04, 173MB/s]
#033[A
model-00001-of-00002.safetensors:  92%|[redacted] | 9.21G/9.98G [00:30<00:04, 168MB/s]
#033[A
model-00001-of-00002.safetensors:  92%|[redacted] | 9.23G/9.98G [00:30<00:04, 171MB/s]
#033[A
model-00001-of-00002.safetensors:  93%|[redacted] | 9.25G/9.98G [00:30<00:04, 178MB/s]
#033[A
model-00001-of-00002.safetensors:  93%|[redacted] | 9.27G/9.98G [00:30<00:04, 172MB/s]
#033[A
model-00001-of-00002.safetensors:  93%|[redacted] | 9.29G/9.98G [00:30<00:03, 180MB/s]
#033[A
model-00001-of-00002.safetensors:  93%|[redacted] | 9.31G/9.98G [00:30<00:03, 172MB/s]
#033[A
model-00001-of-00002.safetensors:  94%|[redacted] | 9.33G/9.98G [00:30<00:03, 165MB/s]
#033[A
model-00001-of-00002.safetensors:  94%|[redacted] | 9.35G/9.98G [00:30<00:03, 168MB/s]
#033[A
model-00001-of-00002.safetensors:  94%|[redacted] | 9.38G/9.98G [00:31<00:03, 175MB/s]
#033[A
model-00001-of-00002.safetensors:  94%|[redacted] | 9.41G/9.98G [00:31<00:03, 175MB/s]
#033[A
model-00001-of-00002.safetensors:  94%|[redacted] | 9.43G/9.98G [00:31<00:03, 175MB/s]
#033[A
model-00001-of-00002.safetensors:  95%|[redacted] | 9.45G/9.98G [00:31<00:03, 169MB/s]
#033[A
model-00001-of-00002.safetensors:  95%|[redacted] | 9.47G/9.98G [00:31<00:02, 170MB/s]
#033[A
model-00001-of-00002.safetensors:  95%|[redacted] | 9.49G/9.98G [00:31<00:02, 167MB/s]
#033[A
model-00001-of-00002.safetensors:  95%|[redacted] | 9.51G/9.98G [00:31<00:02, 171MB/s]
#033[A
model-00001-of-00002.safetensors:  96%|[redacted] | 9.53G/9.98G [00:31<00:02, 164MB/s]
#033[A
model-00001-of-00002.safetensors:  96%|[redacted] | 9.55G/9.98G [00:32<00:02, 165MB/s]
#033[A
model-00001-of-00002.safetensors:  96%|[redacted] | 9.57G/9.98G [00:32<00:02, 169MB/s]
#033[A
model-00001-of-00002.safetensors:  96%|[redacted] | 9.59G/9.98G [00:32<00:02, 166MB/s]
#033[A
model-00001-of-00002.safetensors:  96%|[redacted] | 9.63G/9.98G [00:32<00:02, 171MB/s]
#033[A
model-00001-of-00002.safetensors:  97%|[redacted] | 9.65G/9.98G [00:32<00:01, 172MB/s]
#033[A
model-00001-of-00002.safetensors:  97%|[redacted] | 9.67G/9.98G [00:32<00:01, 173MB/s]
#033[A
model-00001-of-00002.safetensors:  97%|[redacted] | 9.69G/9.98G [00:32<00:01, 173MB/s]
#033[A
model-00001-of-00002.safetensors:  97%|[redacted] | 9.71G/9.98G [00:32<00:01, 174MB/s]
#033[A
model-00001-of-00002.safetensors:  98%|[redacted] | 9.73G/9.98G [00:33<00:01, 174MB/s]
#033[A
```

```
model-00001-of-00002.safetensors:  98%|▉| 9.75G/9.98G [00:33<00:01, 163MB/s]
#033[A
model-00001-of-00002.safetensors:  98%|▉| 9.77G/9.98G [00:33<00:01, 170MB/s]
#033[A
model-00001-of-00002.safetensors:  98%|▉| 9.79G/9.98G [00:33<00:01, 171MB/s]
#033[A
model-00001-of-00002.safetensors:  98%|▉| 9.81G/9.98G [00:33<00:00, 172MB/s]
#033[A
model-00001-of-00002.safetensors:  99%|▉| 9.84G/9.98G [00:33<00:01, 117MB/s]
#033[A
model-00001-of-00002.safetensors:  99%|▉| 9.89G/9.98G [00:34<00:00, 193MB/s]
#033[A
model-00001-of-00002.safetensors:  99%|▉| 9.92G/9.98G [00:34<00:00, 212MB/s]
#033[A
model-00001-of-00002.safetensors: 100%|▉| 9.95G/9.98G [00:34<00:00, 200MB/s]
#033[A
model-00001-of-00002.safetensors: 100%|▉| 9.98G/9.98G [00:34<00:00, 176MB/s]
#033[A
model-00001-of-00002.safetensors: 100%|▉| 9.98G/9.98G [00:34<00:00, 289MB/s]
Downloading shards:  50%|▌    | 1/2 [00:34<00:34, 34.58s/it]
model-00002-of-00002.safetensors:   0%|          | 0.00/3.50G [00:00<?, ?B/s]#033[A
model-00002-of-00002.safetensors:   1%|          | 41.9M/3.50G [00:00<00:11, 306MB/s]
#033[A
model-00002-of-00002.safetensors:   2%|▏         | 73.4M/3.50G [00:00<00:15, 223MB/s]
#033[A
model-00002-of-00002.safetensors:   3%|▏         | 105M/3.50G [00:00<00:17, 198MB/s]
#033[A
model-00002-of-00002.safetensors:   4%|▏         | 126M/3.50G [00:00<00:17, 192MB/s]#
033[A
model-00002-of-00002.safetensors:   4%|▏         | 147M/3.50G [00:00<00:17, 187MB/s]#
033[A
model-00002-of-00002.safetensors:   5%|▏         | 168M/3.50G [00:00<00:18, 182MB/s]#
033[A
model-00002-of-00002.safetensors:   5%|▏         | 189M/3.50G [00:00<00:18, 180MB/s]#
033[A
model-00002-of-00002.safetensors:   6%|▏         | 210M/3.50G [00:01<00:18, 178MB/s]
#033[A
model-00002-of-00002.safetensors:   7%|▏         | 241M/3.50G [00:01<00:16, 194MB/s]
#033[A
model-00002-of-00002.safetensors:   7%|▏         | 262M/3.50G [00:01<00:16, 197MB/s]#
033[A
model-00002-of-00002.safetensors:   9%|▎         | 304M/3.50G [00:01<00:13, 233MB/s]#
033[A
model-00002-of-00002.safetensors:  10%|▎         | 336M/3.50G [00:01<00:13, 241MB/s]
#033[A
model-00002-of-00002.safetensors:  10%|▎         | 367M/3.50G [00:01<00:12, 247MB/s]#
033[A
model-00002-of-00002.safetensors:  11%|▎         | 398M/3.50G [00:01<00:12, 252MB/s]#
033[A
model-00002-of-00002.safetensors:  12%|▎         | 430M/3.50G [00:01<00:12, 253MB/s]#
033[A
model-00002-of-00002.safetensors:  13%|▍         | 461M/3.50G [00:02<00:11, 253MB/s]#
033[A
model-00002-of-00002.safetensors:  14%|▍         | 493M/3.50G [00:02<00:11, 252MB/s]#
033[A
model-00002-of-00002.safetensors:  15%|▍         | 524M/3.50G [00:02<00:11, 266MB/s]#
033[A
model-00002-of-00002.safetensors:  16%|▍         | 556M/3.50G [00:02<00:11, 263MB/s]#
033[A
model-00002-of-00002.safetensors:  17%|▍         | 587M/3.50G [00:02<00:11, 264MB/s]#
```

```
033[A
model-00002-of-00002.safetensors:  18%|██         | 619M/3.50G [00:02<00:10, 264MB/s]#
033[A
model-00002-of-00002.safetensors:  19%|██         | 650M/3.50G [00:02<00:10, 262MB/s]#
033[A
model-00002-of-00002.safetensors:  19%|██         | 682M/3.50G [00:02<00:11, 254MB/s]#
033[A
model-00002-of-00002.safetensors:  20%|██        | 713M/3.50G [00:03<00:10, 261MB/s]#
033[A
model-00002-of-00002.safetensors:  21%|██         | 744M/3.50G [00:03<00:10, 263MB/s]
#033[A
model-00002-of-00002.safetensors:  22%|██         | 776M/3.50G [00:03<00:11, 242MB/s]#
033[A
model-00002-of-00002.safetensors:  23%|██         | 807M/3.50G [00:03<00:11, 225MB/s]#
033[A
model-00002-of-00002.safetensors:  24%|██▏        | 839M/3.50G [00:03<00:13, 195MB/s]#
033[A
model-00002-of-00002.safetensors:  25%|██▏        | 860M/3.50G [00:03<00:13, 192MB/s]#
033[A
model-00002-of-00002.safetensors:  25%|██        | 881M/3.50G [00:03<00:13, 188MB/s]#
033[A
model-00002-of-00002.safetensors:  26%|██        | 902M/3.50G [00:04<00:13, 193MB/s]#
033[A
model-00002-of-00002.safetensors:  26%|██▏        | 923M/3.50G [00:04<00:13, 192MB/s]#
033[A
model-00002-of-00002.safetensors:  27%|██▏        | 944M/3.50G [00:04<00:14, 182MB/s]#
033[A
model-00002-of-00002.safetensors:  28%|██▏        | 965M/3.50G [00:04<00:14, 176MB/s]#
033[A
model-00002-of-00002.safetensors:  28%|██▏        | 986M/3.50G [00:04<00:13, 182MB/s]#
033[A
model-00002-of-00002.safetensors:  29%|██▎        | 1.01G/3.50G [00:04<00:13, 185MB/s]
#033[A
model-00002-of-00002.safetensors:  29%|██▎        | 1.03G/3.50G [00:04<00:13, 181MB/s]
#033[A
model-00002-of-00002.safetensors:  30%|██▎        | 1.05G/3.50G [00:04<00:13, 185MB/s]
#033[A
model-00002-of-00002.safetensors:  31%|██▎       | 1.07G/3.50G [00:04<00:13, 185MB/s]
#033[A
model-00002-of-00002.safetensors:  31%|██▎       | 1.09G/3.50G [00:05<00:13, 180MB/s]
#033[A
model-00002-of-00002.safetensors:  32%|██▍        | 1.11G/3.50G [00:05<00:13, 181MB/s]
#033[A
model-00002-of-00002.safetensors:  32%|██▍        | 1.13G/3.50G [00:05<00:13, 174MB/s]
#033[A
model-00002-of-00002.safetensors:  33%|██▍        | 1.15G/3.50G [00:05<00:13, 177MB/s]
#033[A
model-00002-of-00002.safetensors:  34%|██▍        | 1.17G/3.50G [00:05<00:13, 176MB/s]
#033[A
model-00002-of-00002.safetensors:  34%|██▌        | 1.20G/3.50G [00:05<00:13, 173MB/s]
#033[A
model-00002-of-00002.safetensors:  35%|██▌        | 1.22G/3.50G [00:05<00:13, 175MB/s]
#033[A
model-00002-of-00002.safetensors:  35%|██▌       | 1.24G/3.50G [00:05<00:12, 176MB/s]
#033[A
model-00002-of-00002.safetensors:  36%|██▌       | 1.26G/3.50G [00:06<00:12, 173MB/s]
#033[A
model-00002-of-00002.safetensors:  37%|██▋        | 1.28G/3.50G [00:06<00:13, 171MB/s]
#033[A
model-00002-of-00002.safetensors:  37%|██▋        | 1.30G/3.50G [00:06<00:12, 178MB/s]
```

```
#033[A
model-00002-of-00002.safetensors:  38%|███▊      | 1.32G/3.50G [00:06<00:12, 180MB/s]
#033[A
model-00002-of-00002.safetensors:  38%|███▊      | 1.34G/3.50G [00:06<00:12, 170MB/s]
#033[A
model-00002-of-00002.safetensors:  39%|███▊      | 1.36G/3.50G [00:06<00:13, 164MB/s]
#033[A
model-00002-of-00002.safetensors:  40%|███▉      | 1.38G/3.50G [00:06<00:12, 167MB/s]
#033[A
model-00002-of-00002.safetensors:  40%|███▉      | 1.41G/3.50G [00:06<00:12, 171MB/s]
#033[A
model-00002-of-00002.safetensors:  41%|████      | 1.43G/3.50G [00:07<00:12, 169MB/s]
#033[A
model-00002-of-00002.safetensors:  42%|████      | 1.46G/3.50G [00:07<00:12, 170MB/s]
#033[A
model-00002-of-00002.safetensors:  42%|████      | 1.48G/3.50G [00:07<00:11, 174MB/s]
#033[A
model-00002-of-00002.safetensors:  43%|████      | 1.50G/3.50G [00:07<00:11, 170MB/s]
#033[A
model-00002-of-00002.safetensors:  43%|████      | 1.52G/3.50G [00:07<00:11, 168MB/s]
#033[A
model-00002-of-00002.safetensors:  44%|████      | 1.54G/3.50G [00:07<00:11, 176MB/s]
#033[A
model-00002-of-00002.safetensors:  45%|████      | 1.56G/3.50G [00:07<00:11, 171MB/s]
#033[A
model-00002-of-00002.safetensors:  45%|████      | 1.58G/3.50G [00:07<00:11, 170MB/s]
#033[A
model-00002-of-00002.safetensors:  46%|████      | 1.60G/3.50G [00:08<00:11, 168MB/s]
#033[A
model-00002-of-00002.safetensors:  46%|████▏     | 1.63G/3.50G [00:08<00:11, 169MB/s]
#033[A
model-00002-of-00002.safetensors:  47%|████▏     | 1.65G/3.50G [00:08<00:10, 173MB/s]
#033[A
model-00002-of-00002.safetensors:  48%|████▏     | 1.67G/3.50G [00:08<00:10, 168MB/s]
#033[A
model-00002-of-00002.safetensors:  48%|████▏     | 1.69G/3.50G [00:08<00:10, 167MB/s]
#033[A
model-00002-of-00002.safetensors:  49%|████▏     | 1.71G/3.50G [00:08<00:10, 165MB/s]
#033[A
model-00002-of-00002.safetensors:  49%|████▏     | 1.73G/3.50G [00:08<00:10, 173MB/s]
#033[A
model-00002-of-00002.safetensors:  50%|████▏     | 1.75G/3.50G [00:08<00:10, 167MB/s]
#033[A
model-00002-of-00002.safetensors:  51%|████▏     | 1.77G/3.50G [00:09<00:10, 167MB/s]
#033[A
model-00002-of-00002.safetensors:  51%|████▏     | 1.79G/3.50G [00:09<00:10, 162MB/s]
#033[A
model-00002-of-00002.safetensors:  52%|████▎     | 1.81G/3.50G [00:09<00:10, 166MB/s]
#033[A
model-00002-of-00002.safetensors:  52%|████▎     | 1.84G/3.50G [00:09<00:13, 121MB/s]
#033[A
model-00002-of-00002.safetensors:  54%|████▍     | 1.90G/3.50G [00:09<00:07, 202MB/s]
#033[A
model-00002-of-00002.safetensors:  55%|████▍     | 1.93G/3.50G [00:09<00:08, 177MB/s]
#033[A
model-00002-of-00002.safetensors:  56%|████▍     | 1.95G/3.50G [00:10<00:09, 168MB/s]
#033[A
model-00002-of-00002.safetensors:  56%|████▍     | 1.97G/3.50G [00:10<00:09, 166MB/s]
#033[A
model-00002-of-00002.safetensors:  57%|████▌     | 1.99G/3.50G [00:10<00:08, 170MB/s]
```

```
#033[A
model-00002-of-00002.safetensors:  58%|[redacted]  | 2.01G/3.50G [00:10<00:08, 169MB/s]
#033[A
model-00002-of-00002.safetensors:  58%|[redacted]  | 2.03G/3.50G [00:10<00:08, 168MB/s]
#033[A
model-00002-of-00002.safetensors:  59%|[redacted]  | 2.06G/3.50G [00:10<00:08, 169MB/s]
#033[A
model-00002-of-00002.safetensors:  59%|[redacted]  | 2.08G/3.50G [00:10<00:08, 162MB/s]
#033[A
model-00002-of-00002.safetensors:  60%|[redacted]  | 2.10G/3.50G [00:11<00:08, 166MB/s]
#033[A
model-00002-of-00002.safetensors:  61%|[redacted]  | 2.12G/3.50G [00:11<00:08, 166MB/s]
#033[A
model-00002-of-00002.safetensors:  61%|[redacted]  | 2.14G/3.50G [00:11<00:08, 169MB/s]
#033[A
model-00002-of-00002.safetensors:  62%|[redacted]  | 2.16G/3.50G [00:11<00:07, 172MB/s]
#033[A
model-00002-of-00002.safetensors:  62%|[redacted]  | 2.18G/3.50G [00:11<00:08, 162MB/s]
#033[A
model-00002-of-00002.safetensors:  63%|[redacted]  | 2.20G/3.50G [00:11<00:07, 164MB/s]
#033[A
model-00002-of-00002.safetensors:  64%|[redacted]  | 2.22G/3.50G [00:11<00:07, 163MB/s]
#033[A
model-00002-of-00002.safetensors:  64%|[redacted]  | 2.24G/3.50G [00:11<00:07, 167MB/s]
#033[A
model-00002-of-00002.safetensors:  65%|[redacted]  | 2.26G/3.50G [00:12<00:07, 164MB/s]
#033[A
model-00002-of-00002.safetensors:  65%|[redacted]  | 2.29G/3.50G [00:12<00:07, 164MB/s]
#033[A
model-00002-of-00002.safetensors:  66%|[redacted]  | 2.31G/3.50G [00:12<00:07, 165MB/s]
#033[A
model-00002-of-00002.safetensors:  67%|[redacted]  | 2.33G/3.50G [00:12<00:07, 159MB/s]
#033[A
model-00002-of-00002.safetensors:  67%|[redacted]  | 2.35G/3.50G [00:12<00:06, 166MB/s]
#033[A
model-00002-of-00002.safetensors:  68%|[redacted]  | 2.37G/3.50G [00:12<00:06, 164MB/s]
#033[A
model-00002-of-00002.safetensors:  68%|[redacted]  | 2.39G/3.50G [00:12<00:06, 161MB/s]
#033[A
model-00002-of-00002.safetensors:  69%|[redacted]  | 2.41G/3.50G [00:12<00:06, 163MB/s]
#033[A
model-00002-of-00002.safetensors:  69%|[redacted]  | 2.43G/3.50G [00:13<00:06, 167MB/s]
#033[A
model-00002-of-00002.safetensors:  70%|[redacted]  | 2.45G/3.50G [00:13<00:06, 163MB/s]
#033[A
model-00002-of-00002.safetensors:  71%|[redacted]  | 2.47G/3.50G [00:13<00:06, 166MB/s]
#033[A
model-00002-of-00002.safetensors:  71%|[redacted]  | 2.50G/3.50G [00:13<00:06, 161MB/s]
#033[A
model-00002-of-00002.safetensors:  72%|[redacted]  | 2.52G/3.50G [00:13<00:05, 168MB/s]
#033[A
model-00002-of-00002.safetensors:  72%|[redacted]  | 2.54G/3.50G [00:13<00:05, 161MB/s]
#033[A
model-00002-of-00002.safetensors:  73%|[redacted]  | 2.56G/3.50G [00:13<00:05, 164MB/s]
#033[A
model-00002-of-00002.safetensors:  74%|[redacted]  | 2.58G/3.50G [00:13<00:05, 162MB/s]
#033[A
model-00002-of-00002.safetensors:  74%|[redacted]  | 2.60G/3.50G [00:14<00:05, 163MB/s]
#033[A
model-00002-of-00002.safetensors:  75%|[redacted]  | 2.62G/3.50G [00:14<00:05, 162MB/s]
```

```
#033[A
model-00002-of-00002.safetensors:  75%|███████▌  | 2.64G/3.50G [00:14<00:05, 162MB/s]
#033[A
model-00002-of-00002.safetensors:  76%|███████▌  | 2.66G/3.50G [00:14<00:05, 164MB/s]
#033[A
model-00002-of-00002.safetensors:  77%|███████▋  | 2.68G/3.50G [00:14<00:05, 162MB/s]
#033[A
model-00002-of-00002.safetensors:  77%|███████▋  | 2.71G/3.50G [00:14<00:04, 164MB/s]
#033[A
model-00002-of-00002.safetensors:  78%|███████▊  | 2.73G/3.50G [00:14<00:04, 164MB/s]
#033[A
model-00002-of-00002.safetensors:  78%|███████▊  | 2.75G/3.50G [00:14<00:04, 163MB/s]
#033[A
model-00002-of-00002.safetensors:  79%|███████▊  | 2.77G/3.50G [00:15<00:04, 158MB/s]
#033[A
model-00002-of-00002.safetensors:  80%|███████▉  | 2.79G/3.50G [00:15<00:04, 168MB/s]
#033[A
model-00002-of-00002.safetensors:  80%|███████▉  | 2.81G/3.50G [00:15<00:04, 165MB/s]
#033[A
model-00002-of-00002.safetensors:  81%|████████  | 2.83G/3.50G [00:15<00:04, 160MB/s]
#033[A
model-00002-of-00002.safetensors:  81%|████████  | 2.85G/3.50G [00:15<00:03, 165MB/s]
#033[A
model-00002-of-00002.safetensors:  82%|████████▏ | 2.87G/3.50G [00:15<00:03, 159MB/s]
#033[A
model-00002-of-00002.safetensors:  83%|████████▎ | 2.89G/3.50G [00:15<00:03, 170MB/s]
#033[A
model-00002-of-00002.safetensors:  83%|████████▎ | 2.92G/3.50G [00:16<00:03, 162MB/s]
#033[A
model-00002-of-00002.safetensors:  84%|████████▍ | 2.94G/3.50G [00:16<00:03, 166MB/s]
#033[A
model-00002-of-00002.safetensors:  84%|████████▍ | 2.96G/3.50G [00:16<00:03, 160MB/s]
#033[A
model-00002-of-00002.safetensors:  85%|████████▍ | 2.99G/3.50G [00:16<00:03, 166MB/s]
#033[A
model-00002-of-00002.safetensors:  86%|████████▌ | 3.01G/3.50G [00:16<00:02, 164MB/s]
#033[A
model-00002-of-00002.safetensors:  87%|████████▋ | 3.03G/3.50G [00:16<00:02, 164MB/s]
#033[A
model-00002-of-00002.safetensors:  87%|████████▋ | 3.05G/3.50G [00:16<00:02, 161MB/s]
#033[A
model-00002-of-00002.safetensors:  88%|████████▊ | 3.07G/3.50G [00:16<00:02, 164MB/s]
#033[A
model-00002-of-00002.safetensors:  88%|████████▊ | 3.09G/3.50G [00:17<00:02, 166MB/s]
#033[A
model-00002-of-00002.safetensors:  89%|████████▉ | 3.11G/3.50G [00:17<00:02, 160MB/s]
#033[A
model-00002-of-00002.safetensors:  90%|████████▉ | 3.14G/3.50G [00:17<00:02, 164MB/s]
#033[A
model-00002-of-00002.safetensors:  90%|█████████ | 3.16G/3.50G [00:17<00:02, 160MB/s]
#033[A
model-00002-of-00002.safetensors:  91%|█████████ | 3.18G/3.50G [00:17<00:01, 163MB/s]
#033[A
model-00002-of-00002.safetensors:  91%|█████████ | 3.20G/3.50G [00:17<00:01, 160MB/s]
#033[A
model-00002-of-00002.safetensors:  92%|█████████▏| 3.22G/3.50G [00:17<00:01, 159MB/s]
#033[A
model-00002-of-00002.safetensors:  93%|█████████▎| 3.24G/3.50G [00:18<00:01, 166MB/s]
#033[A
model-00002-of-00002.safetensors:  93%|█████████▎| 3.26G/3.50G [00:18<00:01, 161MB/s]
```

```
#033[A
model-00002-of-00002.safetensors:  94%|█████████▍| 3.28G/3.50G [00:18<00:01, 161MB/s]
#033[A
model-00002-of-00002.safetensors:  94%|█████████▍| 3.30G/3.50G [00:18<00:01, 163MB/s]
#033[A
model-00002-of-00002.safetensors:  95%|█████████▍| 3.32G/3.50G [00:18<00:01, 161MB/s]
#033[A
model-00002-of-00002.safetensors:  96%|█████████▌| 3.34G/3.50G [00:18<00:00, 162MB/s]
#033[A
model-00002-of-00002.safetensors:  96%|█████████▌| 3.37G/3.50G [00:18<00:00, 165MB/s]
#033[A
model-00002-of-00002.safetensors:  97%|█████████▋| 3.39G/3.50G [00:18<00:00, 160MB/s]
#033[A
model-00002-of-00002.safetensors:  97%|█████████▋| 3.41G/3.50G [00:19<00:00, 161MB/s]
#033[A
model-00002-of-00002.safetensors:  98%|█████████▊| 3.43G/3.50G [00:19<00:00, 173MB/s]
#033[A
model-00002-of-00002.safetensors:  99%|█████████▊| 3.45G/3.50G [00:19<00:00, 164MB/s]
#033[A
model-00002-of-00002.safetensors:  99%|█████████▉| 3.47G/3.50G [00:19<00:00, 164MB/s]
#033[A
model-00002-of-00002.safetensors: 100%|█████████▉| 3.49G/3.50G [00:19<00:00, 162MB/s]
#033[A
model-00002-of-00002.safetensors: 100%|██████████| 3.50G/3.50G [00:19<00:00, 177MB/s]
Downloading shards: 100%|██████████| 2/2 [00:54<00:00, 25.91s/it]
Downloading shards: 100%|██████████| 2/2 [00:54<00:00, 27.21s/it]
Loading checkpoint shards:   0%|          | 0/2 [00:00<?, ?it/s]
Loading checkpoint shards:  50%|█████     | 1/2 [00:09<00:09,  9.13s/it]
Loading checkpoint shards: 100%|██████████| 2/2 [00:10<00:00,  4.43s/it]
Loading checkpoint shards: 100%|██████████| 2/2 [00:10<00:00,  5.13s/it]
generation_config.json:   0%|          | 0.00/116 [00:00<?, ?B/s]
generation_config.json: 100%|██████████| 116/116 [00:00<00:00, 1.33MB/s]
tokenizer_config.json:   0%|          | 0.00/749 [00:00<?, ?B/s]
tokenizer_config.json: 100%|██████████| 749/749 [00:00<00:00, 7.87MB/s]
tokenizer.model:   0%|          | 0.00/500k [00:00<?, ?B/s]
tokenizer.model: 100%|██████████| 500k/500k [00:00<00:00, 481MB/s]
tokenizer.json:   0%|          | 0.00/1.84M [00:00<?, ?B/s]
tokenizer.json: 100%|██████████| 1.84M/1.84M [00:00<00:00, 32.3MB/s]
special_tokens_map.json:   0%|          | 0.00/411 [00:00<?, ?B/s]
special_tokens_map.json: 100%|██████████| 411/411 [00:00<00:00, 4.84MB/s]
Generating train split: 0 examples [00:00, ? examples/s]
No chat template is defined for this tokenizer - using the default template for the C
odeLlamaTokenizerFast class. If the default is not appropriate for your model, please
set `tokenizer.chat_template` to an appropriate template. See https://huggingface.co/
docs/transformers/main/chat_templating for more information.
No chat template is defined for this tokenizer - using the default template for the C
odeLlamaTokenizerFast class. If the default is not appropriate for your model, please
set `tokenizer.chat_template` to an appropriate template. See https://huggingface.co/
docs/transformers/main/chat_templating for more information.
Generating train split: 1 examples [00:01,  1.86s/ examples]
Generating train split: 245 examples [00:01, 172.86 examples/s]
Generating train split: 416 examples [00:02, 197.78 examples/s]
/opt/conda/lib/python3.10/site-packages/trl/trainer/sft_trainer.py:294: UserWarning:
You passed a tokenizer with `padding_side` not equal to `right` to the SFTTrainer. Th
is might lead to some unexpected behaviour due to overflow issues when training a mod
el in half-precision. You might consider adding `tokenizer.padding_side = 'right'` to
your code.
  warnings.warn(
0%|          | 0/312 [00:00<?, ?it/s]
`use_cache=True` is incompatible with gradient checkpointing. Setting `use_cache=Fals
```

```
e`.
`use_cache=True` is incompatible with gradient checkpointing. Setting `use_cache=Fals
e`.
The input hidden states seems to be silently casted in float32, this might be related
to the fact you have upcasted embedding or layer norm layers in float32. We will cast
back the input in torch.bfloat16.
The input hidden states seems to be silently casted in float32, this might be related
to the fact you have upcasted embedding or layer norm layers in float32. We will cast
back the input in torch.bfloat16.
0%|          | 1/312 [00:18<1:37:18, 18.77s/it]
1%|          | 2/312 [00:36<1:35:14, 18.43s/it]
1%|          | 3/312 [00:55<1:34:22, 18.33s/it]
1%|▏         | 4/312 [01:13<1:33:48, 18.28s/it]
2%|▏         | 5/312 [01:31<1:33:22, 18.25s/it]
2%|▏         | 6/312 [01:49<1:32:58, 18.23s/it]
2%|▏         | 7/312 [02:07<1:32:37, 18.22s/it]
3%|▎         | 8/312 [02:26<1:32:17, 18.21s/it]
3%|▎         | 9/312 [02:44<1:31:57, 18.21s/it]
3%|▎         | 10/312 [03:02<1:31:38, 18.21s/it]
{'loss': 0.9418, 'grad_norm': 0.07568359375, 'learning_rate': 0.0002, 'epoch': 0.1}
3%|▎         | 10/312 [03:02<1:31:38, 18.21s/it]
4%|▎         | 11/312 [03:20<1:31:19, 18.20s/it]
4%|▍         | 12/312 [03:38<1:31:00, 18.20s/it]
4%|▍         | 13/312 [03:57<1:30:41, 18.20s/it]
4%|▍         | 14/312 [04:15<1:30:23, 18.20s/it]
5%|▍         | 15/312 [04:33<1:30:05, 18.20s/it]
5%|▌         | 16/312 [04:51<1:29:46, 18.20s/it]
5%|▌         | 17/312 [05:09<1:29:28, 18.20s/it]
6%|▌         | 18/312 [05:28<1:29:10, 18.20s/it]
6%|▌         | 19/312 [05:46<1:28:51, 18.20s/it]
6%|▋         | 20/312 [06:04<1:28:33, 18.20s/it]
{'loss': 0.7638, 'grad_norm': 0.0634765625, 'learning_rate': 0.0002, 'epoch': 0.19}
6%|▋         | 20/312 [06:04<1:28:33, 18.20s/it]
7%|▋         | 21/312 [06:22<1:28:15, 18.20s/it]
7%|▋         | 22/312 [06:40<1:27:57, 18.20s/it]
7%|▋         | 23/312 [06:59<1:27:38, 18.20s/it]
8%|▊         | 24/312 [07:17<1:27:20, 18.20s/it]
8%|▊         | 25/312 [07:35<1:27:02, 18.20s/it]
8%|▊         | 26/312 [07:53<1:26:44, 18.20s/it]
9%|▊         | 27/312 [08:11<1:26:26, 18.20s/it]
9%|▉         | 28/312 [08:30<1:26:07, 18.20s/it]
9%|▉         | 29/312 [08:48<1:25:49, 18.20s/it]
10%|▉         | 30/312 [09:06<1:25:31, 18.20s/it]
{'loss': 0.6835, 'grad_norm': 0.0810546875, 'learning_rate': 0.0002, 'epoch': 0.29}
10%|▉         | 30/312 [09:06<1:25:31, 18.20s/it]
10%|▉         | 31/312 [09:24<1:25:13, 18.20s/it]
10%|█         | 32/312 [09:42<1:24:55, 18.20s/it]
11%|█         | 33/312 [10:01<1:24:37, 18.20s/it]
11%|█         | 34/312 [10:19<1:24:18, 18.20s/it]
11%|█         | 35/312 [10:37<1:24:00, 18.20s/it]
12%|█▏        | 36/312 [10:55<1:23:42, 18.20s/it]
12%|█▏        | 37/312 [11:13<1:23:23, 18.20s/it]
12%|█▏        | 38/312 [11:32<1:23:11, 18.22s/it]
12%|█▎        | 39/312 [11:50<1:22:51, 18.21s/it]
13%|█▎        | 40/312 [12:08<1:22:31, 18.21s/it]
{'loss': 0.6212, 'grad_norm': 0.08984375, 'learning_rate': 0.0002, 'epoch': 0.38}
13%|█▎        | 40/312 [12:08<1:22:31, 18.21s/it]
13%|█▎        | 41/312 [12:26<1:22:13, 18.20s/it]
13%|█▎        | 42/312 [12:44<1:21:54, 18.20s/it]
14%|█▍        | 43/312 [13:03<1:21:35, 18.20s/it]
```

```
14%|▏         | 44/312 [13:21<1:21:17, 18.20s/it]
14%|▏         | 45/312 [13:39<1:20:58, 18.20s/it]
15%|▏         | 46/312 [13:57<1:20:40, 18.20s/it]
15%|▏         | 47/312 [14:15<1:20:22, 18.20s/it]
15%|▏         | 48/312 [14:34<1:20:03, 18.20s/it]
16%|▏         | 49/312 [14:52<1:19:45, 18.20s/it]
16%|▏         | 50/312 [15:10<1:19:27, 18.20s/it]
{'loss': 0.5625, 'grad_norm': 0.1279296875, 'learning_rate': 0.0002, 'epoch': 0.48}
16%|▏         | 50/312 [15:10<1:19:27, 18.20s/it]
16%|▏         | 51/312 [15:28<1:19:09, 18.20s/it]
17%|▏         | 52/312 [15:46<1:18:51, 18.20s/it]
17%|▏         | 53/312 [16:05<1:18:32, 18.20s/it]
17%|▏         | 54/312 [16:23<1:18:14, 18.20s/it]
18%|▏         | 55/312 [16:41<1:17:56, 18.20s/it]
18%|▏         | 56/312 [16:59<1:17:38, 18.20s/it]
18%|▏         | 57/312 [17:17<1:17:20, 18.20s/it]
19%|▏         | 58/312 [17:36<1:17:01, 18.20s/it]
19%|▏         | 59/312 [17:54<1:16:43, 18.20s/it]
19%|▏         | 60/312 [18:12<1:16:25, 18.20s/it]
{'loss': 0.5458, 'grad_norm': 0.05859375, 'learning_rate': 0.0002, 'epoch': 0.58}
19%|▏         | 60/312 [18:12<1:16:25, 18.20s/it]
20%|▏         | 61/312 [18:30<1:16:07, 18.20s/it]
20%|▏         | 62/312 [18:48<1:15:49, 18.20s/it]
20%|▏         | 63/312 [19:07<1:15:30, 18.20s/it]
21%|▏         | 64/312 [19:25<1:15:12, 18.20s/it]
21%|▏         | 65/312 [19:43<1:14:54, 18.20s/it]
21%|▏         | 66/312 [20:01<1:14:36, 18.20s/it]
21%|▏         | 67/312 [20:19<1:14:18, 18.20s/it]
22%|▏         | 68/312 [20:38<1:14:00, 18.20s/it]
22%|▏         | 69/312 [20:56<1:13:41, 18.20s/it]
22%|▏         | 70/312 [21:14<1:13:23, 18.20s/it]
{'loss': 0.526, 'grad_norm': 0.061767578125, 'learning_rate': 0.0002, 'epoch': 0.67}
22%|▏         | 70/312 [21:14<1:13:23, 18.20s/it]
23%|▏         | 71/312 [21:32<1:13:05, 18.20s/it]
23%|▏         | 72/312 [21:50<1:12:47, 18.20s/it]
23%|▏         | 73/312 [22:09<1:12:28, 18.20s/it]
24%|▏         | 74/312 [22:27<1:12:10, 18.20s/it]
24%|▏         | 75/312 [22:45<1:11:52, 18.20s/it]
24%|▏         | 76/312 [23:03<1:11:34, 18.20s/it]
25%|▏         | 77/312 [23:21<1:11:15, 18.20s/it]
25%|▏         | 78/312 [23:39<1:10:57, 18.20s/it]
25%|▏         | 79/312 [23:58<1:10:39, 18.20s/it]
26%|▏         | 80/312 [24:16<1:10:21, 18.20s/it]
{'loss': 0.5189, 'grad_norm': 0.06787109375, 'learning_rate': 0.0002, 'epoch': 0.77}
26%|▏         | 80/312 [24:16<1:10:21, 18.20s/it]
26%|▏         | 81/312 [24:34<1:10:03, 18.20s/it]
26%|▏         | 82/312 [24:52<1:09:45, 18.20s/it]
27%|▏         | 83/312 [25:10<1:09:27, 18.20s/it]
27%|▏         | 84/312 [25:29<1:09:08, 18.20s/it]
27%|▏         | 85/312 [25:47<1:08:50, 18.20s/it]
28%|▏         | 86/312 [26:05<1:08:32, 18.20s/it]
28%|▏         | 87/312 [26:23<1:08:14, 18.20s/it]
28%|▏         | 88/312 [26:41<1:07:55, 18.20s/it]
29%|▏         | 89/312 [27:00<1:07:37, 18.20s/it]
29%|▏         | 90/312 [27:18<1:07:19, 18.20s/it]
{'loss': 0.5029, 'grad_norm': 0.06982421875, 'learning_rate': 0.0002, 'epoch': 0.87}
29%|▏         | 90/312 [27:18<1:07:19, 18.20s/it]
29%|▏         | 91/312 [27:36<1:07:01, 18.20s/it]
29%|▏         | 92/312 [27:54<1:06:43, 18.20s/it]
30%|▏         | 93/312 [28:12<1:06:25, 18.20s/it]
```

```
30%|███       | 94/312 [28:31<1:06:06, 18.20s/it]
30%|███       | 95/312 [28:49<1:05:48, 18.20s/it]
31%|███       | 96/312 [29:07<1:05:30, 18.20s/it]
31%|███       | 97/312 [29:25<1:05:12, 18.20s/it]
31%|███       | 98/312 [29:43<1:04:54, 18.20s/it]
32%|███       | 99/312 [30:02<1:04:35, 18.20s/it]
32%|███       | 100/312 [30:20<1:04:17, 18.20s/it]
{'loss': 0.5035, 'grad_norm': 0.0703125, 'learning_rate': 0.0002, 'epoch': 0.96}
32%|███       | 100/312 [30:20<1:04:17, 18.20s/it]
32%|███       | 101/312 [30:38<1:04:04, 18.22s/it]
33%|███       | 102/312 [30:56<1:03:44, 18.21s/it]
33%|███       | 103/312 [31:14<1:03:25, 18.21s/it]
33%|███       | 104/312 [31:33<1:03:06, 18.20s/it]
34%|███       | 105/312 [31:51<1:03:16, 18.34s/it]
34%|███       | 106/312 [32:10<1:02:49, 18.30s/it]
34%|███       | 107/312 [32:28<1:02:24, 18.27s/it]
35%|███       | 108/312 [32:46<1:02:02, 18.25s/it]
35%|███       | 109/312 [33:04<1:01:40, 18.23s/it]
35%|███       | 110/312 [33:22<1:01:20, 18.22s/it]
{'loss': 0.4836, 'grad_norm': 0.072265625, 'learning_rate': 0.0002, 'epoch': 1.06}
35%|███       | 110/312 [33:22<1:01:20, 18.22s/it]
36%|███       | 111/312 [33:41<1:01:01, 18.21s/it]
36%|███       | 112/312 [33:59<1:00:41, 18.21s/it]
36%|███       | 113/312 [34:17<1:00:22, 18.21s/it]
37%|███       | 114/312 [34:35<1:00:04, 18.20s/it]
37%|███       | 115/312 [34:53<59:45, 18.20s/it]
37%|███       | 116/312 [35:12<59:27, 18.20s/it]
38%|███       | 117/312 [35:30<59:08, 18.20s/it]
38%|███       | 118/312 [35:48<58:50, 18.20s/it]
38%|███       | 119/312 [36:06<58:32, 18.20s/it]
38%|███       | 120/312 [36:24<58:14, 18.20s/it]
{'loss': 0.4738, 'grad_norm': 0.07763671875, 'learning_rate': 0.0002, 'epoch': 1.15}
38%|███       | 120/312 [36:24<58:14, 18.20s/it]
39%|███       | 121/312 [36:43<57:55, 18.20s/it]
39%|███       | 122/312 [37:01<57:37, 18.20s/it]
39%|███       | 123/312 [37:19<57:19, 18.20s/it]
40%|███       | 124/312 [37:37<57:01, 18.20s/it]
40%|████      | 125/312 [37:55<56:42, 18.20s/it]
40%|████      | 126/312 [38:13<56:24, 18.20s/it]
41%|████      | 127/312 [38:32<56:06, 18.20s/it]
41%|████      | 128/312 [38:50<55:48, 18.20s/it]
41%|████      | 129/312 [39:08<55:30, 18.20s/it]
42%|████      | 130/312 [39:26<55:11, 18.20s/it]
{'loss': 0.4665, 'grad_norm': 0.0791015625, 'learning_rate': 0.0002, 'epoch': 1.25}
42%|████      | 130/312 [39:26<55:11, 18.20s/it]
42%|████      | 131/312 [39:44<54:53, 18.20s/it]
42%|████      | 132/312 [40:03<54:35, 18.20s/it]
43%|████      | 133/312 [40:21<54:17, 18.20s/it]
43%|████      | 134/312 [40:39<53:59, 18.20s/it]
43%|████      | 135/312 [40:57<53:40, 18.20s/it]
44%|████      | 136/312 [41:15<53:22, 18.20s/it]
44%|████      | 137/312 [41:34<53:04, 18.20s/it]
44%|████      | 138/312 [41:52<52:46, 18.20s/it]
45%|████      | 139/312 [42:10<52:28, 18.20s/it]
45%|████      | 140/312 [42:28<52:09, 18.20s/it]
{'loss': 0.4693, 'grad_norm': 0.07861328125, 'learning_rate': 0.0002, 'epoch': 1.35}
45%|████      | 140/312 [42:28<52:09, 18.20s/it]
45%|████      | 141/312 [42:46<51:51, 18.20s/it]
46%|████      | 142/312 [43:05<51:33, 18.20s/it]
46%|████      | 143/312 [43:23<51:15, 18.20s/it]
```

```
46%|████▋      | 144/312 [43:41<50:57, 18.20s/it]
46%|████▊      | 145/312 [43:59<50:39, 18.20s/it]
47%|████▊      | 146/312 [44:17<50:20, 18.20s/it]
47%|████▊      | 147/312 [44:36<50:02, 18.20s/it]
47%|████▊      | 148/312 [44:54<49:44, 18.20s/it]
48%|████▊      | 149/312 [45:12<49:26, 18.20s/it]
48%|████▊      | 150/312 [45:30<49:07, 18.20s/it]
{'loss': 0.4592, 'grad_norm': 0.08203125, 'learning_rate': 0.0002, 'epoch': 1.44}
48%|████▊      | 150/312 [45:30<49:07, 18.20s/it]
48%|████▊      | 151/312 [45:48<48:49, 18.20s/it]
49%|████▊      | 152/312 [46:07<48:31, 18.20s/it]
49%|████▉      | 153/312 [46:25<48:13, 18.20s/it]
49%|████▉      | 154/312 [46:43<47:58, 18.22s/it]
50%|████▉      | 155/312 [47:01<47:39, 18.21s/it]
50%|█████      | 156/312 [47:19<47:20, 18.21s/it]
50%|█████      | 157/312 [47:38<47:01, 18.20s/it]
51%|█████      | 158/312 [47:56<46:43, 18.20s/it]
51%|█████      | 159/312 [48:14<46:24, 18.20s/it]
51%|█████      | 160/312 [48:32<46:06, 18.20s/it]
{'loss': 0.4591, 'grad_norm': 0.07568359375, 'learning_rate': 0.0002, 'epoch': 1.54}
51%|█████      | 160/312 [48:32<46:06, 18.20s/it]
52%|█████▏     | 161/312 [48:50<45:48, 18.20s/it]
52%|█████▏     | 162/312 [49:09<45:29, 18.20s/it]
52%|█████▏     | 163/312 [49:27<45:11, 18.20s/it]
53%|█████▎     | 164/312 [49:45<44:53, 18.20s/it]
53%|█████▎     | 165/312 [50:03<44:34, 18.20s/it]
53%|█████▎     | 166/312 [50:21<44:16, 18.20s/it]
54%|█████▎     | 167/312 [50:40<43:58, 18.20s/it]
54%|█████▍     | 168/312 [50:58<43:40, 18.20s/it]
54%|█████▍     | 169/312 [51:16<43:22, 18.20s/it]
54%|█████▍     | 170/312 [51:34<43:03, 18.20s/it]
{'loss': 0.4585, 'grad_norm': 0.083984375, 'learning_rate': 0.0002, 'epoch': 1.63}
54%|█████▍     | 170/312 [51:34<43:03, 18.20s/it]
55%|█████▍     | 171/312 [51:52<42:45, 18.20s/it]
55%|█████▌     | 172/312 [52:11<42:27, 18.20s/it]
55%|█████▌     | 173/312 [52:29<42:09, 18.20s/it]
56%|█████▌     | 174/312 [52:47<41:51, 18.20s/it]
56%|█████▌     | 175/312 [53:05<41:32, 18.20s/it]
56%|█████▋     | 176/312 [53:23<41:14, 18.20s/it]
57%|█████▋     | 177/312 [53:42<40:56, 18.20s/it]
57%|█████▋     | 178/312 [54:00<40:38, 18.20s/it]
57%|█████▋     | 179/312 [54:18<40:20, 18.20s/it]
58%|█████▊     | 180/312 [54:36<40:01, 18.20s/it]
{'loss': 0.4561, 'grad_norm': 0.08203125, 'learning_rate': 0.0002, 'epoch': 1.73}
58%|█████▊     | 180/312 [54:36<40:01, 18.20s/it]
58%|█████▊     | 181/312 [54:54<39:43, 18.20s/it]
58%|█████▊     | 182/312 [55:13<39:25, 18.20s/it]
59%|█████▊     | 183/312 [55:31<39:07, 18.20s/it]
59%|█████▉     | 184/312 [55:49<38:49, 18.20s/it]
59%|█████▉     | 185/312 [56:07<38:31, 18.20s/it]
60%|█████▉     | 186/312 [56:25<38:12, 18.20s/it]
60%|█████▉     | 187/312 [56:44<37:54, 18.20s/it]
60%|██████     | 188/312 [57:02<37:36, 18.20s/it]
61%|██████     | 189/312 [57:20<37:18, 18.20s/it]
61%|██████     | 190/312 [57:38<36:59, 18.20s/it]
{'loss': 0.4543, 'grad_norm': 0.09033203125, 'learning_rate': 0.0002, 'epoch': 1.83}
61%|██████     | 190/312 [57:38<36:59, 18.20s/it]
61%|██████     | 191/312 [57:56<36:41, 18.20s/it]
62%|██████▏    | 192/312 [58:15<36:23, 18.20s/it]
62%|██████▏    | 193/312 [58:33<36:05, 18.20s/it]
```

```
62%|██████▏     | 194/312 [58:51<35:47, 18.20s/it]
62%|██████▏     | 195/312 [59:09<35:29, 18.20s/it]
63%|██████▎     | 196/312 [59:27<35:11, 18.20s/it]
63%|██████▎     | 197/312 [59:46<34:52, 18.20s/it]
63%|██████▎     | 198/312 [1:00:04<34:34, 18.20s/it]
64%|██████▍     | 199/312 [1:00:22<34:16, 18.20s/it]
64%|██████▍     | 200/312 [1:00:40<33:58, 18.20s/it]
{'loss': 0.45, 'grad_norm': 0.08984375, 'learning_rate': 0.0002, 'epoch': 1.92}
64%|██████▍     | 200/312 [1:00:40<33:58, 18.20s/it]
64%|██████▍     | 201/312 [1:00:58<33:39, 18.20s/it]
65%|██████▍     | 202/312 [1:01:17<33:21, 18.20s/it]
65%|██████▌     | 203/312 [1:01:35<33:03, 18.20s/it]
65%|██████▌     | 204/312 [1:01:53<32:45, 18.20s/it]
66%|██████▌     | 205/312 [1:02:11<32:27, 18.20s/it]
66%|██████▌     | 206/312 [1:02:29<32:08, 18.20s/it]
66%|██████▋     | 207/312 [1:02:48<31:50, 18.20s/it]
67%|██████▋     | 208/312 [1:03:06<31:32, 18.20s/it]
67%|██████▋     | 209/312 [1:03:24<31:28, 18.34s/it]
67%|██████▋     | 210/312 [1:03:43<31:06, 18.30s/it]
{'loss': 0.4493, 'grad_norm': 0.09130859375, 'learning_rate': 0.0002, 'epoch': 2.02}
67%|██████▋     | 210/312 [1:03:43<31:06, 18.30s/it]
68%|██████▊     | 211/312 [1:04:01<30:46, 18.29s/it]
68%|██████▊     | 212/312 [1:04:19<30:25, 18.26s/it]
68%|██████▊     | 213/312 [1:04:37<30:05, 18.24s/it]
69%|██████▊     | 214/312 [1:04:55<29:46, 18.23s/it]
69%|██████▉     | 215/312 [1:05:14<29:27, 18.22s/it]
69%|██████▉     | 216/312 [1:05:32<29:08, 18.21s/it]
70%|██████▉     | 217/312 [1:05:50<28:49, 18.21s/it]
70%|██████▉     | 218/312 [1:06:08<28:31, 18.20s/it]
70%|███████     | 219/312 [1:06:26<28:12, 18.20s/it]
71%|███████     | 220/312 [1:06:45<27:54, 18.20s/it]
{'loss': 0.4245, 'grad_norm': 0.09228515625, 'learning_rate': 0.0002, 'epoch': 2.12}
71%|███████     | 220/312 [1:06:45<27:54, 18.20s/it]
71%|███████     | 221/312 [1:07:03<27:36, 18.20s/it]
71%|███████     | 222/312 [1:07:21<27:17, 18.20s/it]
71%|███████▏    | 223/312 [1:07:39<26:59, 18.20s/it]
72%|███████▏    | 224/312 [1:07:57<26:41, 18.20s/it]
72%|███████▏    | 225/312 [1:08:16<26:23, 18.20s/it]
72%|███████▏    | 226/312 [1:08:34<26:04, 18.20s/it]
73%|███████▎    | 227/312 [1:08:52<25:46, 18.20s/it]
73%|███████▎    | 228/312 [1:09:10<25:28, 18.20s/it]
73%|███████▎    | 229/312 [1:09:28<25:10, 18.20s/it]
74%|███████▎    | 230/312 [1:09:47<24:52, 18.20s/it]
{'loss': 0.4221, 'grad_norm': 0.0986328125, 'learning_rate': 0.0002, 'epoch': 2.21}
74%|███████▎    | 230/312 [1:09:47<24:52, 18.20s/it]
74%|███████▍    | 231/312 [1:10:05<24:33, 18.20s/it]
74%|███████▍    | 232/312 [1:10:23<24:15, 18.20s/it]
75%|███████▍    | 233/312 [1:10:41<23:57, 18.20s/it]
75%|███████▌    | 234/312 [1:10:59<23:39, 18.20s/it]
75%|███████▌    | 235/312 [1:11:18<23:21, 18.20s/it]
76%|███████▌    | 236/312 [1:11:36<23:02, 18.20s/it]
76%|███████▌    | 237/312 [1:11:54<22:44, 18.20s/it]
76%|███████▋    | 238/312 [1:12:12<22:26, 18.20s/it]
77%|███████▋    | 239/312 [1:12:30<22:08, 18.20s/it]
77%|███████▋    | 240/312 [1:12:49<21:50, 18.20s/it]
{'loss': 0.4168, 'grad_norm': 0.0986328125, 'learning_rate': 0.0002, 'epoch': 2.31}
77%|███████▋    | 240/312 [1:12:49<21:50, 18.20s/it]
77%|███████▋    | 241/312 [1:13:07<21:32, 18.20s/it]
78%|███████▊    | 242/312 [1:13:25<21:13, 18.20s/it]
78%|███████▊    | 243/312 [1:13:43<20:55, 18.20s/it]
```

```
78%|███████▊  | 244/312 [1:14:01<20:37, 18.20s/it]
79%|███████▉  | 245/312 [1:14:20<20:19, 18.20s/it]
79%|███████▉  | 246/312 [1:14:38<20:01, 18.20s/it]
79%|███████▉  | 247/312 [1:14:56<19:42, 18.20s/it]
79%|███████▉  | 248/312 [1:15:14<19:24, 18.20s/it]
80%|████████  | 249/312 [1:15:32<19:06, 18.20s/it]
80%|████████  | 250/312 [1:15:51<18:48, 18.20s/it]
{'loss': 0.4241, 'grad_norm': 0.099609375, 'learning_rate': 0.0002, 'epoch': 2.4}
80%|████████  | 250/312 [1:15:51<18:48, 18.20s/it]
80%|████████  | 251/312 [1:16:09<18:30, 18.20s/it]
81%|████████  | 252/312 [1:16:27<18:11, 18.20s/it]
81%|████████  | 253/312 [1:16:45<17:53, 18.20s/it]
81%|████████▏ | 254/312 [1:17:03<17:35, 18.20s/it]
82%|████████▏ | 255/312 [1:17:22<17:17, 18.20s/it]
82%|████████▏ | 256/312 [1:17:40<16:58, 18.20s/it]
82%|████████▏ | 257/312 [1:17:58<16:40, 18.20s/it]
83%|████████▎ | 258/312 [1:18:16<16:22, 18.20s/it]
83%|████████▎ | 259/312 [1:18:34<16:04, 18.20s/it]
83%|████████▎ | 260/312 [1:18:53<15:46, 18.20s/it]
{'loss': 0.424, 'grad_norm': 0.1005859375, 'learning_rate': 0.0002, 'epoch': 2.5}
83%|████████▎ | 260/312 [1:18:53<15:46, 18.20s/it]
84%|████████▎ | 261/312 [1:19:11<15:27, 18.20s/it]
84%|████████▍ | 262/312 [1:19:29<15:09, 18.20s/it]
84%|████████▍ | 263/312 [1:19:47<14:51, 18.20s/it]
85%|████████▍ | 264/312 [1:20:05<14:33, 18.20s/it]
85%|████████▍ | 265/312 [1:20:24<14:15, 18.20s/it]
85%|████████▌ | 266/312 [1:20:42<13:57, 18.20s/it]
86%|████████▌ | 267/312 [1:21:00<13:38, 18.20s/it]
86%|████████▌ | 268/312 [1:21:18<13:20, 18.20s/it]
86%|████████▌ | 269/312 [1:21:36<13:02, 18.20s/it]
87%|████████▋ | 270/312 [1:21:54<12:44, 18.20s/it]
{'loss': 0.4198, 'grad_norm': 0.10009765625, 'learning_rate': 0.0002, 'epoch': 2.6}
87%|████████▋ | 270/312 [1:21:55<12:44, 18.20s/it]
87%|████████▋ | 271/312 [1:22:13<12:26, 18.20s/it]
87%|████████▋ | 272/312 [1:22:31<12:07, 18.20s/it]
88%|████████▊ | 273/312 [1:22:49<11:49, 18.20s/it]
88%|████████▊ | 274/312 [1:23:07<11:32, 18.22s/it]
88%|████████▊ | 275/312 [1:23:26<11:13, 18.21s/it]
88%|████████▊ | 276/312 [1:23:44<10:55, 18.21s/it]
89%|████████▉ | 277/312 [1:24:02<10:37, 18.20s/it]
89%|████████▉ | 278/312 [1:24:20<10:18, 18.20s/it]
89%|████████▉ | 279/312 [1:24:38<10:00, 18.20s/it]
90%|████████▉ | 280/312 [1:24:57<09:42, 18.20s/it]
{'loss': 0.4202, 'grad_norm': 0.10009765625, 'learning_rate': 0.0002, 'epoch': 2.69}
90%|████████▉ | 280/312 [1:24:57<09:42, 18.20s/it]
90%|█████████ | 281/312 [1:25:15<09:24, 18.20s/it]
90%|█████████ | 282/312 [1:25:33<09:05, 18.20s/it]
91%|█████████ | 283/312 [1:25:51<08:47, 18.20s/it]
91%|█████████ | 284/312 [1:26:09<08:29, 18.20s/it]
91%|█████████▏| 285/312 [1:26:28<08:11, 18.20s/it]
92%|█████████▏| 286/312 [1:26:46<07:53, 18.20s/it]
92%|█████████▏| 287/312 [1:27:04<07:34, 18.20s/it]
92%|█████████▏| 288/312 [1:27:22<07:16, 18.20s/it]
93%|█████████▎| 289/312 [1:27:40<06:58, 18.20s/it]
93%|█████████▎| 290/312 [1:27:58<06:40, 18.20s/it]
{'loss': 0.4223, 'grad_norm': 0.10302734375, 'learning_rate': 0.0002, 'epoch': 2.79}
93%|█████████▎| 290/312 [1:27:58<06:40, 18.20s/it]
93%|█████████▎| 291/312 [1:28:17<06:22, 18.20s/it]
94%|█████████▎| 292/312 [1:28:35<06:03, 18.20s/it]
94%|█████████▍| 293/312 [1:28:53<05:45, 18.20s/it]
```

```
94%|█████████▍| 294/312 [1:29:11<05:27, 18.20s/it]
95%|█████████▍| 295/312 [1:29:29<05:09, 18.20s/it]
95%|█████████▍| 296/312 [1:29:48<04:51, 18.20s/it]
95%|█████████▌| 297/312 [1:30:06<04:32, 18.20s/it]
96%|█████████▌| 298/312 [1:30:24<04:14, 18.20s/it]
96%|█████████▌| 299/312 [1:30:42<03:56, 18.20s/it]
96%|█████████▌| 300/312 [1:31:00<03:38, 18.20s/it]
{'loss': 0.4189, 'grad_norm': 0.10498046875, 'learning_rate': 0.0002, 'epoch': 2.88}
96%|█████████▌| 300/312 [1:31:00<03:38, 18.20s/it]
96%|█████████▋| 301/312 [1:31:19<03:20, 18.20s/it]
97%|█████████▋| 302/312 [1:31:37<03:01, 18.20s/it]
97%|█████████▋| 303/312 [1:31:55<02:43, 18.20s/it]
97%|█████████▋| 304/312 [1:32:13<02:25, 18.20s/it]
98%|█████████▊| 305/312 [1:32:31<02:07, 18.20s/it]
98%|█████████▊| 306/312 [1:32:50<01:49, 18.20s/it]
98%|█████████▊| 307/312 [1:33:08<01:30, 18.20s/it]
99%|█████████▊| 308/312 [1:33:26<01:12, 18.20s/it]
99%|█████████▉| 309/312 [1:33:44<00:54, 18.20s/it]
99%|█████████▉| 310/312 [1:34:02<00:36, 18.20s/it]
{'loss': 0.4132, 'grad_norm': 0.1044921875, 'learning_rate': 0.0002, 'epoch': 2.98}
99%|█████████▉| 310/312 [1:34:02<00:36, 18.20s/it]
100%|█████████▉| 311/312 [1:34:21<00:18, 18.20s/it]
100%|██████████| 312/312 [1:34:39<00:00, 18.20s/it]
{'train_runtime': 5679.7905, 'train_samples_per_second': 0.22, 'train_steps_per_secon
d': 0.055, 'train_loss': 0.4980104019244512, 'epoch': 3.0}
100%|██████████| 312/312 [1:34:39<00:00, 18.20s/it]
100%|██████████| 312/312 [1:34:39<00:00, 18.20s/it]
['adapter_config.json', 'runs', 'checkpoint-208', 'tokenizer_config.json', 'tokenize
r.json', 'README.md', 'adapter_model.safetensors', 'checkpoint-312', 'checkpoint-10
4', 'tokenizer.model', 'special_tokens_map.json']
Loading checkpoint shards:   0%|          | 0/2 [00:00<?, ?it/s]
Loading checkpoint shards:  50%|█████     | 1/2 [00:20<00:20, 20.81s/it]
Loading checkpoint shards: 100%|██████████| 2/2 [00:31<00:00, 14.79s/it]
Loading checkpoint shards: 100%|██████████| 2/2 [00:31<00:00, 15.70s/it]

2024-03-08 09:49:15 Uploading - Uploading generated training model2024-03-08 09:49:1
0,847 sagemaker-training-toolkit INFO     Waiting for the process to finish and give
a return code.
2024-03-08 09:49:10,848 sagemaker-training-toolkit INFO     Done waiting for a return
code. Received 0 from exiting process.
2024-03-08 09:49:10,848 sagemaker-training-toolkit INFO     Reporting training SUCCES
S

2024-03-08 09:49:56 Completed - Training job completed
Training seconds: 6162
Billable seconds: 6162
```

In [16]:
```
huggingface_estimator.model_data["S3DataSource"]["S3Uri"].replace("s3://", "https://s3
```

Out[16]:
```
'https://s3.console.aws.amazon.com/s3/buckets/sagemaker-us-east-1-558105141721/codell
ama-7b-hf-text-to-sql-exp1-2024-03-08-08-05-53-957/output/model/'
```

In [17]:
```
from sagemaker.huggingface import get_huggingface_llm_image_uri

# retrieve the llm image uri
llm_image = get_huggingface_llm_image_uri(
  "huggingface",
  version="1.4.0",
  session=sess,
)
```

```
# print ecr image uri
print(f"llm image uri: {llm_image}")
```

```
INFO:sagemaker.image_uris:Defaulting to only available Python version: py310
INFO:sagemaker.image_uris:Defaulting to only supported image scope: gpu.
llm image uri: 763104351884.dkr.ecr.us-east-1.amazonaws.com/huggingface-pytorch-tgi-i
nference:2.1.1-tgi1.4.0-gpu-py310-cu121-ubuntu20.04
```

In [18]:
```
import json
from sagemaker.huggingface import HuggingFaceModel

# s3 path where the model will be uploaded
# if you try to deploy the model to a different time add the s3 path here
model_s3_path = huggingface_estimator.model_data["S3DataSource"]["S3Uri"]

# sagemaker config
instance_type = "ml.g5.2xlarge"
number_of_gpu = 1
health_check_timeout = 300

# Define Model and Endpoint configuration parameter
config = {
  'HF_MODEL_ID': "/opt/ml/model", # path to where sagemaker stores the model
  'SM_NUM_GPUS': json.dumps(number_of_gpu), # Number of GPU used per replica
  'MAX_INPUT_LENGTH': json.dumps(1024), # Max length of input text
  'MAX_TOTAL_TOKENS': json.dumps(2048), # Max length of the generation (including inpu
}

# create HuggingFaceModel with the image uri
llm_model = HuggingFaceModel(
  role=role,
  image_uri=llm_image,
  model_data={'S3DataSource':{'S3Uri': model_s3_path,'S3DataType': 'S3Prefix','Compres
  env=config
)
```

In [19]:
```
# Deploy model to an endpoint
# https://sagemaker.readthedocs.io/en/stable/api/inference/model.html#sagemaker.model.
llm = llm_model.deploy(
  initial_instance_count=1,
  instance_type=instance_type,
  container_startup_health_check_timeout=health_check_timeout, # 10 minutes to give Sa
)
```

```
INFO:sagemaker:Creating model with name: huggingface-pytorch-tgi-inference-2024-03-08
-10-50-17-803
INFO:sagemaker:Creating endpoint-config with name huggingface-pytorch-tgi-inference-2
024-03-08-10-50-18-669
INFO:sagemaker:Creating endpoint with name huggingface-pytorch-tgi-inference-2024-03-
08-10-50-18-669
---------!
```

In [24]:
```
from transformers import AutoTokenizer
from sagemaker.s3 import S3Downloader

# Load the tokenizer
tokenizer = AutoTokenizer.from_pretrained("codellama/CodeLlama-7b-hf")

# Load the test dataset from s3
```

```
S3Downloader.download(f"{training_input_path}/test_dataset.json", ".")
test_dataset = load_dataset("json", data_files="test_dataset.json",split="train")
random_sample = test_dataset[345]

def request(sample):
    prompt = tokenizer.apply_chat_template(sample, tokenize=False, add_generation_prom
    outputs = llm.predict({
      "inputs": prompt,
      "parameters": {
        "max_new_tokens": 512,
        "do_sample": False,
        "return_full_text": False,
        "stop": ["<|im_end|>"],
      }
    })
    return {"role": "assistant", "content": outputs[0]["generated_text"].strip()}

print(random_sample["messages"][1])
request(random_sample["messages"][:2])
```

```
Generating train split: 0 examples [00:00, ? examples/s]
{'content': 'For the attendance of 2 january 1999 with a home team of plymouth argyle
what is the tie no. ?', 'role': 'user'}
```

Out[24]:

```
{'role': 'assistant',
 'content': 'SELECT tie_no FROM table_name_5 WHERE attendance = "2 january 1999" AND
home_team = "plymouth argyle"'}
```

Awesome! Our model is working as expected. Now we can evaluate our model on 1000 samples from test dataset.

In [28]:

```
from tqdm import tqdm

def evaluate(sample):
    predicted_answer = request(sample["messages"][:2])
    if predicted_answer["content"] == sample["messages"][2]["content"]:
        return 1
    else:
        return 0

success_rate = []
number_of_eval_samples = 1000
# iterate over eval dataset and predict
for s in tqdm(test_dataset.shuffle().select(range(number_of_eval_samples))):
    success_rate.append(evaluate(s))

# compute accuracy
accuracy = sum(success_rate)/len(success_rate)

print(f"Accuracy: {accuracy*100:.2f}%")
```

```
100%|██████████| 1000/1000 [16:10<00:00,  1.03it/s]
Accuracy: 78.40%
```

In [29]:

```
llm.delete_model()
llm.delete_endpoint()
```

```
INFO:sagemaker:Deleting model with name: huggingface-pytorch-tgi-inference-2024-03-08
-10-50-17-803
INFO:sagemaker:Deleting endpoint configuration with name: huggingface-pytorch-tgi-inf
erence-2024-03-08-10-50-18-669
INFO:sagemaker:Deleting endpoint with name: huggingface-pytorch-tgi-inference-2024-03
-08-10-50-18-669
```

In [ ]: