Creating a Llama-3.1 LoRA adapter with the NeMo Framework and Deploy via NVIDIA NIM

It's Llama 3.1 Day and we're excited to share our newest notebook in collaboration with the NVIDIA for finetuning using the NeMo framework and deploying it using an NVIDIA NIM. In this notebook, we'll be finetuning our own LoRA with a cleaned up version of the Law StackExchange dataset using NeMo Framework. Law StackExchange is a dataset of legal question/answers. Each record consists of a question, its title, as well as human-provided answers. Given a Law StackExchange forum question our goal is to auto-generate an appropriate title for it.

NVIDIA NeMo Framework and NVIDIA NIM

NVIDIA NeMo Framework is a scalable and cloud-native generative AI framework built for researchers and developers working on Large Language Models, Multimodal, and Speech AI (e.g. Automatic Speech Recognition and Text-to-Speech). It enables users to efficiently create, customize, and deploy new generative AI models by leveraging existing code and pre-trained model checkpoints. After we finetune a LoRa using NeMo, we then deploy it using an NVIDIA NIM. An NVIDIA NIM is an accelerated inference solution for Generative AI models.

Prerequistes

Before you start this notebook, ensure that you have an NGC key available that is able to access the Llama3.1 NIM on NGC. To generate one, please visit build.nvidia.com and click Get API Key!

First we install the NGC CLI and docker and pull the `.nemo` checkpoint that we will use for finetuning. This can take about 5-7 minutes

In [1]:

```
%%bash
test -f setup-ngc.sh || (wget https://raw.githubusercontent.com/brevdev/note
./setup-ngc.sh
```

```
NGC CLI v3.49.0 installed. Restart terminal or source profile to use.
Alternatively, you can use an explicit path to: /root/verb-workspace/ngc-cl
i/ngc
```

In [2]:

```
!COLUMNS=400 ./ngc-cli/ngc registry model download-version "nvidia/nemo/llam
```

```
Getting files to download...
  ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ • 15.0/15.0 GiB • Remaining:
0:00:00 • 73.5 MB/s • Elapsed: 0:03:37 • Total: 1 – Completed: 1 – Failed: 0
0 – Failed: 0ed: 0

-----------------------------------------------------------------------------
--------
   Download status: COMPLETED
   Downloaded local path model: /root/verb-workspace/llama-3_1-8b-instruct-n
emo_v1.0
   Total files downloaded: 1
   Total transferred: 14.96 GB
   Started at: 2024-09-21 19:21:55
   Completed at: 2024-09-21 19:25:32
   Duration taken: 3m 37s
-----------------------------------------------------------------------------
--------
```

In [3]:

```
# this should the .nemo checkpoint that is saved
!ls ./llama-3_1-8b-instruct-nemo_v1.0
```

```
llama3_1_8b_instruct.nemo
```

In [4]:

```
import os
import json
import numpy as np
from rouge_score import rouge_scorer, scoring
```

Phase 1: Finetuning the LoRa adapter

Step-by-step PEFT finetuning instructions

1. Prepare the dataset
2. Run the PEFT finetuning script
3. Inference with NeMo Framework
4. Check the model accuracy

Step 1: Prepare the dataset

The dataset we used is a subset of the Law-StackExchange dataset. We've already filtered and processed this dataset and it can be used to train the model for various different tasks - question title generation (summarization), law domain question answering, and question tag generation (multi-label classification). To run your own data cleaning and prepreocessing, please refer to the data generation notebook. That tutorial also allows you to generate synthetic data and increase the size of the dataset.

This dataset is licensed under the CC BY-SA 4.0 license. You can use it for any purpose, including commercial use, without attribution. However, if you use the dataset in a publication, please cite the original authors and the Law-StackExchange dataset repository.

In [5]:

```
!wget https://huggingface.co/datasets/bigmlguy2234/hf-law-qa-dataset/resolve
```

```
--2024-09-21 19:25:34--  https://huggingface.co/datasets/bigmlguy2234/hf-law
-qa-dataset/resolve/main/law-qa-curated.zip
Resolving huggingface.co (huggingface.co)... 54.230.18.95, 54.230.18.110, 5
4.230.18.85, ...
Connecting to huggingface.co (huggingface.co)|54.230.18.95|:443... connecte
d.
HTTP request sent, awaiting response... 302 Found
Location: https://cdn-lfs-us-1.huggingface.co/repos/a6/d5/a6d5955c217c4e78e7
08cfea9bf52e46fb3c5cc93151c5447c804929b8db561a/b26fcd36ab38c6011cecb8f8d6f0e
9990441dfa9d1fa9f9a8d740612493c4a90?response-content-disposition=inline%3B+f
ilename*%3DUTF-8%27%27law-qa-curated.zip%3B+filename%3D%22law-qa-curated.zi
p%22%3B&response-content-type=application%2Fzip&Expires=1727205934&Policy=ey
JTdGF0ZW1lbnQiOlt7IkNvbmRpdGlvbiI6eyJEYXRlTGVzc1RoYW4iOnsiQVdTOkVwb2NoVGltZS
I6MTcyNzIwNTkzNH19LCJSZXNvdXJjZSI6Imh0dHBzOi8vY2RuLWxmcy11cy0xLmh1Z2dpbmdmYW
NlLmNvL3JlcG9zL2E2L2Q1L2E2ZDU5NTVjMjE3YzRlNzhlNzA4Y2ZlYTliZjUyZTQ2ZmIzYzVjYz
kzMTUxYzU0NDdjODA0OTI5YjhkYjU2MWEvYjI2ZmNkMzZhYjM4YzYwMTFjZWNiOGY4ZDZmMGU5OT
kwNDQxZGZhOWQxZmE5ZjlhOGQ3NDA2MTI0OTNjNGE5MD9yZXNwb25zZS1jb250ZW50LWRpc3Bvc2
l0aW9uPSomcmVzcG9uc2UtY29udGVudC10eXBlPSoifV19&Signature=XNy5Y-ytyPGN17237XH
t0yy3VG8XnqPtJvSkt5Q7r9wd6xOMa8QedOGbQYYsqLQz1WFys6IqszhZINe%7Ekxx1ZKQ8FCV1y
9l1Uk1nwf1g8rCCwkn0G9XsDFa7QmuBy4oVz9HoV7iX4fNMA8kzyYPgkyvYLLA851o2El0ZbTtTM
WiJKDgS%7E8o5iew%7ElDP69y12p2NmC3JmixwEIZhRnx0H%7EPeZGM5BFqye9V2sfQu4piBNLQD
RO8U8GFZZHAnffjKfJTrJJKUbKlY7wrPcosiby-mWU7zROhIcfnj1bAFpSdJDmf2NyTHCkUn68-o
lsMzlt-qmStnRA3DgsrZffYtItQ__&Key-Pair-Id=K24J24Z295AEI9 [following]
--2024-09-21 19:25:34--  https://cdn-lfs-us-1.huggingface.co/repos/a6/d5/a6d
5955c217c4e78e708cfea9bf52e46fb3c5cc93151c5447c804929b8db561a/b26fcd36ab38c6
011cecb8f8d6f0e9990441dfa9d1fa9f9a8d740612493c4a90?response-content-disposit
ion=inline%3B+filename*%3DUTF-8%27%27law-qa-curated.zip%3B+filename%3D%22law
-qa-curated.zip%22%3B&response-content-type=application%2Fzip&Expires=172720
5934&Policy=eyJTdGF0ZW1lbnQiOlt7IkNvbmRpdGlvbiI6eyJEYXRlTGVzc1RoYW4iOnsiQVdT
OkVwb2NoVGltZSI6MTcyNzIwNTkzNH19LCJSZXNvdXJjZSI6Imh0dHBzOi8vY2RuLWxmcy11cy0x
Lmh1Z2dpbmdmYWNlLmNvL3JlcG9zL2E2L2Q1L2E2ZDU5NTVjMjE3YzRlNzhlNzA4Y2ZlYTliZjUy
ZTQ2ZmIzYzVjYzkzMTUxYzU0NDdjODA0OTI5YjhkYjU2MWEvYjI2ZmNkMzZhYjM4YzYwMTFjZWNi
OGY4ZDZmMGU5OTkwNDQxZGZhOWQxZmE5ZjlhOGQ3NDA2MTI0OTNjNGE5MD9yZXNwb25zZS1jb250
ZW50LWRpc3Bvc2l0aW9uPSomcmVzcG9uc2UtY29udGVudC10eXBlPSoifV19&Signature=XNy5Y
-ytyPGN17237XHt0yy3VG8XnqPtJvSkt5Q7r9wd6xOMa8QedOGbQYYsqLQz1WFys6IqszhZINe%7
Ekxx1ZKQ8FCV1y9l1Uk1nwf1g8rCCwkn0G9XsDFa7QmuBy4oVz9HoV7iX4fNMA8kzyYPgkyvYLLA
851o2El0ZbTtTMWiJKDgS%7E8o5iew%7ElDP69y12p2NmC3JmixwEIZhRnx0H%7EPeZGM5BFqye9
V2sfQu4piBNLQDRO8U8GFZZHAnffjKfJTrJJKUbKlY7wrPcosiby-mWU7zROhIcfnj1bAFpSdJDm
f2NyTHCkUn68-olsMzlt-qmStnRA3DgsrZffYtItQ__&Key-Pair-Id=K24J24Z295AEI9
Resolving cdn-lfs-us-1.huggingface.co (cdn-lfs-us-1.huggingface.co)... 3.16
2.163.112, 3.162.163.26, 3.162.163.40, ...
Connecting to cdn-lfs-us-1.huggingface.co (cdn-lfs-us-1.huggingface.co)|3.16
2.163.112|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 14254627 (14M) [application/zip]
Saving to: ‘law-qa-curated.zip.1’

law-qa-curated.zip. 100%[===================>]  13.59M  --.-KB/s    in 0.1s

2024-09-21 19:25:34 (119 MB/s) - ‘law-qa-curated.zip.1’ saved [14254627/1425
4627]
```

In [6]:

```
!unzip -j law-qa-curated.zip -d curated-data
```

```
Archive:  law-qa-curated.zip
  inflating: curated-data/law-qa-test.jsonl
  inflating: curated-data/law-qa-val.jsonl
  inflating: curated-data/law-qa-train.jsonl
```

You should see the `law-qa-{train/val/test}.jsonl` splits in the curated folder

In [7]:

```
DATA_DIR = os.path.join("./curated-data")

TRAIN_DS = os.path.join(DATA_DIR, "law-qa-train.jsonl")
VAL_DS = os.path.join(DATA_DIR, "law-qa-val.jsonl")
TEST_DS = os.path.join(DATA_DIR, "law-qa-test.jsonl")
```

You will see several fields in the `.jsonl`, including `title`, `question`, `answer`, and other associated metadata.

For this tutorial, our input will be the `answer` field, and output will be it's `title`.

The following cell does two things -

- Adds a template - a prompt instruction (which is optional), and format `{PROMPT} \nQUESTION: {data["question"]} \nTITLE: `.
- Saves the data splits into the same location, also appending a `_preprocessed` marker to them.

In [8]:

```
 # Add a prompt instruction.
PROMPT='''Generate a concise, engaging title for the following legal questic

# Creates a preprocessed version of the data files
for input_file in [TRAIN_DS, VAL_DS, TEST_DS]:
    output_file = input_file.rsplit('.', 1)[0] + '_preprocessed.jsonl'
    with open(input_file, 'r') as infile, open(output_file, 'w') as outfile:
        for line in infile:
            # Parse each line as JSON
            data = json.loads(line)

            # Create a new dictionary with only the desired fields, renamed
            new_data = {
                "input": f'''{PROMPT} \nQUESTION: {data["question"]} \nTITLE
                "output": data['title']
            }

            # Write the new data as a JSON line to the output file
            json.dump(new_data, outfile)
            outfile.write('\n')  # Add a newline after each JSON object

    print(f"Processed {input_file} and created {output_file}")
```

```
Processed ./curated-data/law-qa-train.jsonl and created ./curated-data/law-q
a-train_preprocessed.jsonl
Processed ./curated-data/law-qa-val.jsonl and created ./curated-data/law-qa-
val_preprocessed.jsonl
Processed ./curated-data/law-qa-test.jsonl and created ./curated-data/law-qa
-test_preprocessed.jsonl
```

After running the above scripts, you will see `law-qa-{train/test/val}_preprocessed.jsonl` files appear in the data directory.

This is what an example will be formatted like -

```
{"input": "Generate a concise, engaging title for the following
legal question on an internet forum. The title should be legally
relevant, capture key aspects of the issue, and entice readers to
learn more. \nQUESTION: In order to be sued in a particular
jurisdiction, say New York, a company must have a minimal business
presence in the jurisdiction. What constitutes such a presence?
Suppose the company engaged a New York-based Plaintiff, and its
representatives signed the contract with the Plaintiff in New York
City. Does this satisfy the minimum presence rule? Suppose,
instead, the plaintiff and contract signing were in New Jersey, but
the company hired a law firm with offices in New York City. Does
this qualify? \nTITLE: ",
 "output": "What constitutes \"doing business in a jurisdiction?
\""}
```

Step 2: Run PEFT finetuning script for LoRA

NeMo framework includes a high level python script for fine-tuning megatron_gpt_finetuning.py that can abstract away some of the lower level API calls. Once you have your model downloaded and the dataset ready, LoRA fine-tuning with NeMo is essentially just running this script!

For this demonstration, this training run is capped by `max_steps`, and validation is carried out every `val_check_interval` steps. If the validation loss does not improve after a few checks, training is halted to avoid overfitting.

> `NOTE:` In the block of code below, pass the paths to your train, test and validation data files as well as path to the .nemo model.

In [9]:

```
%%bash

# Set paths to the model, train, validation and test sets.
MODEL="./llama-3_1-8b-instruct-nemo_v1.0/llama3_1_8b_instruct.nemo"

TRAIN_DS="[./curated-data/law-qa-train_preprocessed.jsonl]"
VALID_DS="[./curated-data/law-qa-val_preprocessed.jsonl]"
TEST_DS="[./curated-data/law-qa-test_preprocessed.jsonl]"
TEST_NAMES="[law]"
```

```
SCHEME="lora"
TP_SIZE=1
PP_SIZE=1

rm -rf results
OUTPUT_DIR="./results/Meta-llama3.1-8B-Instruct-titlegen"

torchrun --nproc_per_node=1 \
/opt/NeMo/examples/nlp/language_modeling/tuning/megatron_gpt_finetuning.py \
    exp_manager.exp_dir=${OUTPUT_DIR} \
    exp_manager.explicit_log_dir=${OUTPUT_DIR} \
    trainer.devices=1 \
    trainer.num_nodes=1 \
    trainer.precision=bf16-mixed \
    trainer.val_check_interval=0.2 \
    trainer.max_steps=50 \
    model.megatron_amp_O2=True \
    ++model.mcore_gpt=True \
    model.tensor_model_parallel_size=${TP_SIZE} \
    model.pipeline_model_parallel_size=${PP_SIZE} \
    model.micro_batch_size=1 \
    model.global_batch_size=32 \
    model.restore_from_path=${MODEL} \
    model.data.train_ds.file_names=${TRAIN_DS} \
    model.data.train_ds.concat_sampling_probabilities=[1.0] \
    model.data.validation_ds.file_names=${VALID_DS} \
    model.peft.peft_scheme=${SCHEME}
```

```
[NeMo W 2024-09-21 19:26:04 nemo_logging:349] /usr/local/lib/python3.10/dist
-packages/hydra/_internal/hydra.py:119: UserWarning: Future Hydra versions w
ill no longer change working directory at job runtime by default.
    See https://hydra.cc/docs/1.2/upgrades/1.1_to_1.2/changes_to_job_working
_dir/ for more information.
      ret = run_job(
[NeMo I 2024-09-21 19:26:04 megatron_gpt_finetuning:56] 

    ************** Experiment configuration ***********
[NeMo I 2024-09-21 19:26:04 megatron_gpt_finetuning:57] 
    name: megatron_gpt_peft_${model.peft.peft_scheme}_tuning
    trainer:
      devices: 1
      accelerator: gpu
      num_nodes: 1
      precision: bf16-mixed
      logger: false
      enable_checkpointing: false
      use_distributed_sampler: false
      max_epochs: 9999
      max_steps: 50
      log_every_n_steps: 10
      val_check_interval: 0.2
      gradient_clip_val: 1.0
    exp_manager:
      explicit_log_dir: ./results/Meta-llama3.1-8B-Instruct-titlegen
      exp_dir: ./results/Meta-llama3.1-8B-Instruct-titlegen
      name: ${name}
      create_wandb_logger: false
      wandb_logger_kwargs:
        project: null
        name: null
      resume_if_exists: true
      resume_ignore_no_checkpoint: true
      create_checkpoint_callback: true
      checkpoint_callback_params:
        monitor: validation_${model.data.validation_ds.metric.name}
        save_top_k: 1
        mode: min
        save_nemo_on_train_end: true
        filename: ${name}--{${exp_manager.checkpoint_callback_params.monito
r}:.3f}-{step}-{consumed_samples}
        model_parallel_size: ${model.tensor_model_parallel_size}
        always_save_nemo: false
        save_best_model: true
      create_early_stopping_callback: true
      early_stopping_callback_params:
        monitor: val_loss
        mode: min
        min_delta: 0.001
        patience: 10
        verbose: true
        strict: false
    model:
      seed: 1234
      tensor_model_parallel_size: 1
      pipeline_model_parallel_size: 1
      global_batch_size: 32
      micro_batch_size: 1
      restore_from_path: ./llama-3_1-8b-instruct-nemo_v1.0/llama3_1_8b_instr
uct.nemo
      resume_from_checkpoint: null
```

```
save_nemo_on_validation_end: false
sync_batch_comm: false
megatron_amp_O2: true
sequence_parallel: false
activations_checkpoint_granularity: null
activations_checkpoint_method: null
activations_checkpoint_num_layers: null
activations_checkpoint_layers_per_pipeline: null
answer_only_loss: true
gradient_as_bucket_view: false
hidden_dropout: 0.0
attention_dropout: 0.0
ffn_dropout: 0.0
fsdp: false
fsdp_sharding_strategy: full
fsdp_grad_reduce_dtype: fp32
fsdp_sharded_checkpoint: false
fsdp_use_orig_params: false
peft:
  peft_scheme: lora
  restore_from_path: null
  adapter_tuning:
    type: parallel_adapter
    adapter_dim: 32
    adapter_dropout: 0.0
    norm_position: pre
    column_init_method: xavier
    row_init_method: zero
    norm_type: mixedfusedlayernorm
    layer_selection: null
    weight_tying: false
    position_embedding_strategy: null
  lora_tuning:
    variant: nemo
    target_modules:
    - attention_qkv
    adapter_dim: 32
    alpha: ${model.peft.lora_tuning.adapter_dim}
    adapter_dropout: 0.0
    column_init_method: xavier
    row_init_method: zero
    layer_selection: null
    weight_tying: false
    position_embedding_strategy: null
  p_tuning:
    virtual_tokens: 10
    bottleneck_dim: 1024
    embedding_dim: 1024
    init_std: 0.023
  ia3_tuning:
    layer_selection: null
  selective_tuning:
    tunable_base_param_names:
    - self_attention
    - word_embeddings
data:
```

```
train_ds:
  file_names:
  - ./curated-data/law-qa-train_preprocessed.jsonl
  global_batch_size: ${model.global_batch_size}
  micro_batch_size: ${model.micro_batch_size}
  shuffle: true
  num_workers: 0
  memmap_workers: 2
  pin_memory: true
  max_seq_length: 2048
  min_seq_length: 1
  drop_last: true
  concat_sampling_probabilities:
  - 1.0
  label_key: output
  add_eos: true
  add_sep: false
  add_bos: false
  truncation_field: input
  index_mapping_dir: null
  prompt_template: '{input} {output}'
  truncation_method: right
validation_ds:
  file_names:
  - ./curated-data/law-qa-val_preprocessed.jsonl
  names: null
  global_batch_size: ${model.global_batch_size}
  micro_batch_size: ${model.micro_batch_size}
  shuffle: false
  num_workers: 0
  memmap_workers: ${model.data.train_ds.memmap_workers}
  pin_memory: true
  max_seq_length: 2048
  min_seq_length: 1
  drop_last: false
  label_key: ${model.data.train_ds.label_key}
  add_eos: ${model.data.train_ds.add_eos}
  add_sep: ${model.data.train_ds.add_sep}
  add_bos: ${model.data.train_ds.add_bos}
  write_predictions_to_file: false
  output_file_path_prefix: null
  truncation_field: ${model.data.train_ds.truncation_field}
  index_mapping_dir: null
  prompt_template: ${model.data.train_ds.prompt_template}
  tokens_to_generate: 32
  truncation_method: right
  metric:
    name: loss
    average: null
    num_classes: null
test_ds:
  file_names: null
  names: null
  global_batch_size: ${model.global_batch_size}
  micro_batch_size: ${model.micro_batch_size}
  shuffle: false
```

```
          num_workers: 0
          memmap_workers: ${model.data.train_ds.memmap_workers}
          pin_memory: true
          max_seq_length: 2048
          min_seq_length: 1
          drop_last: false
          label_key: ${model.data.train_ds.label_key}
          add_eos: ${model.data.train_ds.add_eos}
          add_sep: ${model.data.train_ds.add_sep}
          add_bos: ${model.data.train_ds.add_bos}
          write_predictions_to_file: false
          output_file_path_prefix: null
          truncation_field: ${model.data.train_ds.truncation_field}
          index_mapping_dir: null
          prompt_template: ${model.data.train_ds.prompt_template}
          tokens_to_generate: 32
          truncation_method: right
          metric:
            name: loss
            average: null
            num_classes: null
      optim:
        name: fused_adam
        lr: 0.0001
        weight_decay: 0.01
        betas:
        - 0.9
        - 0.98
        sched:
          name: CosineAnnealing
          warmup_steps: 50
          min_lr: 0.0
          constant_steps: 0
          monitor: val_loss
          reduce_on_plateau: false
      mcore_gpt: true
```

```
[NeMo W 2024-09-21 19:26:04 nemo_logging:349] /usr/local/lib/python3.10/dist
-packages/pytorch_lightning/_graveyard/precision.py:49: The `MixedPrecisionP
lugin` is deprecated. Use `pytorch_lightning.plugins.precision.MixedPrecisio
n` instead.

GPU available: True (cuda), used: True
```

```
[NeMo I 2024-09-21 19:26:04 dist_ckpt_io:95] Using ('zarr', 1) dist-ckpt sav
e strategy.
```

```
TPU available: False, using: 0 TPU cores
HPU available: False, using: 0 HPUs
[NeMo E 2024-09-21 19:26:04 exp_manager:703] exp_manager received explicit_l
og_dir: ./results/Meta-llama3.1-8B-Instruct-titlegen and at least one of exp
_dir: ./results/Meta-llama3.1-8B-Instruct-titlegen, or version: None. Please
note that exp_dir, name, and version will be ignored.
[NeMo W 2024-09-21 19:26:04 exp_manager:630] There were no checkpoints found
in checkpoint_dir or no checkpoint folder at checkpoint_dir :results/Meta-ll
ama3.1-8B-Instruct-titlegen/checkpoints. Training from scratch.
```

```
[NeMo I 2024-09-21 19:26:04 exp_manager:396] Experiments will be logged at r
esults/Meta-llama3.1-8B-Instruct-titlegen
[NeMo I 2024-09-21 19:26:04 exp_manager:856] TensorboardLogger has been set
up
```

```
[NeMo W 2024-09-21 19:26:04 exp_manager:966] The checkpoint callback was tol
d to monitor a validation value and trainer's max_steps was set to 50. Pleas
e ensure that max_steps will run for at least 1 epochs to ensure that checkp
ointing will not error out.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: context_parallel_size in its cfg. Add t
his key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: expert_model_parallel_size in its cfg.
Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: moe_extended_tp in its cfg. Add this ke
y to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: finalize_model_grads_func in its cfg. A
dd this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: use_te_rng_tracker in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_bulk_wgrad in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_bulk_dgrad in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_overlap_ag in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_overlap_rs in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_overlap_rs_dgrad in its cfg. Ad
d this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_split_ag in its cfg. Add this k
ey to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_atomic_ag in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_split_rs in its cfg. Add this k
ey to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_atomic_rs in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: defer_embedding_wgrad_compute in its cf
g. Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: pipeline_model_parallel_split_rank in i
ts cfg. Add this key to cfg or config_mapping to make to make it configurabl
e.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: cpu_offloading in its cfg. Add this key
to cfg or config_mapping to make to make it configurable.
```

```
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPTSFTModel() does not have field.name: cpu_offloading_num_layers in its cfg. Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPTSFTModel() does not have field.name: _cpu_offloading_context in its cfg. Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPTSFTModel() does not have field.name: cpu_offloading_activations in its cfg. Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPTSFTModel() does not have field.name: cpu_offloading_weights in its cfg. Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPTSFTModel() does not have field.name: barrier_with_L1_time in its cfg. Add this key to cfg or config_mapping to make to make it configurable.
[W init.cpp:767] Warning: nvfuser is no longer supported in torch script, use _jit_set_nvfuser_enabled is deprecated and a no-op (function operator())
```

```
[NeMo I 2024-09-21 19:26:21 megatron_init:263] Rank 0 has data parallel group : [0]
[NeMo I 2024-09-21 19:26:21 megatron_init:269] Rank 0 has combined group of data parallel and context parallel : [0]
[NeMo I 2024-09-21 19:26:21 megatron_init:274] All data parallel group ranks with context parallel combined: [[0]]
[NeMo I 2024-09-21 19:26:21 megatron_init:277] Ranks 0 has data parallel rank: 0
[NeMo I 2024-09-21 19:26:21 megatron_init:285] Rank 0 has context parallel group: [0]
[NeMo I 2024-09-21 19:26:21 megatron_init:288] All context parallel group ranks: [[0]]
[NeMo I 2024-09-21 19:26:21 megatron_init:289] Ranks 0 has context parallel rank: 0
[NeMo I 2024-09-21 19:26:21 megatron_init:296] Rank 0 has model parallel group: [0]
[NeMo I 2024-09-21 19:26:21 megatron_init:297] All model parallel group ranks: [[0]]
[NeMo I 2024-09-21 19:26:21 megatron_init:306] Rank 0 has tensor model parallel group: [0]
[NeMo I 2024-09-21 19:26:21 megatron_init:310] All tensor model parallel group ranks: [[0]]
[NeMo I 2024-09-21 19:26:21 megatron_init:311] Rank 0 has tensor model parallel rank: 0
[NeMo I 2024-09-21 19:26:21 megatron_init:331] Rank 0 has pipeline model parallel group: [0]
[NeMo I 2024-09-21 19:26:21 megatron_init:343] Rank 0 has embedding group: [0]
[NeMo I 2024-09-21 19:26:21 megatron_init:349] All pipeline model parallel group ranks: [[0]]
[NeMo I 2024-09-21 19:26:21 megatron_init:350] Rank 0 has pipeline model parallel rank 0
[NeMo I 2024-09-21 19:26:21 megatron_init:351] All embedding group ranks: [[0]]
[NeMo I 2024-09-21 19:26:21 megatron_init:352] Rank 0 has embedding rank: 0
```

```
24-09-21 19:26:21 - PID:42349 - rank:(0, 0, 0, 0) - microbatches.py:39 - INF
O - setting number of micro-batches to constant 32
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: context_parallel_size in its cfg. Add t
his key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: expert_model_parallel_size in its cfg.
Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: moe_extended_tp in its cfg. Add this ke
y to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: finalize_model_grads_func in its cfg. A
dd this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: use_te_rng_tracker in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_bulk_wgrad in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_bulk_dgrad in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_overlap_ag in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_overlap_rs in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_overlap_rs_dgrad in its cfg. Ad
d this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_split_ag in its cfg. Add this k
ey to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_atomic_ag in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_split_rs in its cfg. Add this k
ey to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_atomic_rs in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: defer_embedding_wgrad_compute in its cf
g. Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: pipeline_model_parallel_split_rank in i
ts cfg. Add this key to cfg or config_mapping to make to make it configurabl
e.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: cpu_offloading in its cfg. Add this key
to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: cpu_offloading_num_layers in its cfg. A
```

```
dd this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: _cpu_offloading_context in its cfg. Add
this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: cpu_offloading_activations in its cfg.
Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: cpu_offloading_weights in its cfg. Add
this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: barrier_with_L1_time in its cfg. Add th
is key to cfg or config_mapping to make to make it configurable.
```

```
[NeMo I 2024-09-21 19:26:21 tokenizer_utils:178] Getting HuggingFace AutoTok
enizer with pretrained_model_name: meta-llama/Meta-Llama-3-8B
```

```
[NeMo W 2024-09-21 19:26:21 nemo_logging:349] /usr/local/lib/python3.10/dist
-packages/huggingface_hub/file_download.py:1132: FutureWarning: `resume_down
load` is deprecated and will be removed in version 1.0.0. Downloads always r
esume when possible. If you want to force a new download, use `force_downloa
d=True`.
      warnings.warn(

Special tokens have been added in the vocabulary, make sure the associated w
ord embeddings are fine-tuned or trained.
```

```
[NeMo I 2024-09-21 19:26:21 megatron_base_model:584] Padded vocab_size: 1282
56, original vocab_size: 128256, dummy tokens: 0.
```

```
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: context_parallel_size in its cfg. Add t
his key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: expert_model_parallel_size in its cfg.
Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: moe_extended_tp in its cfg. Add this ke
y to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: finalize_model_grads_func in its cfg. A
dd this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: use_te_rng_tracker in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_bulk_wgrad in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_bulk_dgrad in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_overlap_ag in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_overlap_rs in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_overlap_rs_dgrad in its cfg. Ad
d this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_split_ag in its cfg. Add this k
ey to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_atomic_ag in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_split_rs in its cfg. Add this k
ey to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_atomic_rs in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: defer_embedding_wgrad_compute in its cf
g. Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: pipeline_model_parallel_split_rank in i
ts cfg. Add this key to cfg or config_mapping to make to make it configurabl
e.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: cpu_offloading in its cfg. Add this key
to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: cpu_offloading_num_layers in its cfg. A
dd this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
```

```
SFTModel() does not have field.name: _cpu_offloading_context in its cfg. Add
this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: cpu_offloading_activations in its cfg.
Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: cpu_offloading_weights in its cfg. Add
this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: barrier_with_L1_time in its cfg. Add th
is key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:498] apply_query_key_layer_s
caling is only enabled when using FP16, setting it to False and setting NVTE
_APPLY_QK_LAYER_SCALING=0
[NeMo W 2024-09-21 19:26:21 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: activation_func_fp8_input_store in its c
fg. Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: num_moe_experts in its cfg. Add this key
to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: window_size in its cfg. Add this key to
cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: qk_layernorm in its cfg. Add this key to
cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: test_mode in its cfg. Add this key to cf
g or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: calculate_per_token_loss in its cfg. Add
this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: memory_efficient_layer_norm in its cfg.
Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: fp8_wgrad in its cfg. Add this key to cf
g or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: fp8_dot_product_attention in its cfg. Ad
d this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: fp8_multi_head_attention in its cfg. Add
this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: moe_router_load_balancing_type in its cf
g. Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: moe_router_topk in its cfg. Add this key
to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: moe_grouped_gemm in its cfg. Add this ke
y to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: moe_aux_loss_coeff in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
```

```
[NeMo W 2024-09-21 19:26:21 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: moe_z_loss_coeff in its cfg. Add this ke
y to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: moe_input_jitter_eps in its cfg. Add thi
s key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: moe_token_dropping in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: moe_token_dispatcher_type in its cfg. Ad
d this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: moe_per_layer_logging in its cfg. Add th
is key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: moe_expert_capacity_factor in its cfg. A
dd this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: moe_pad_expert_input_to_capacity in its
cfg. Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: moe_token_drop_policy in its cfg. Add th
is key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: moe_layer_recompute in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: clone_scatter_output_in_embedding in its
cfg. Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: disable_parameter_transpose_cache in its
cfg. Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: enable_cuda_graph in its cfg. Add this k
ey to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:26:21 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: rotary_percent in its cfg. Add this key
to cfg or config_mapping to make to make it configurable.
Initializing distributed: GLOBAL_RANK: 0, MEMBER: 1/1
----------------------------------------------------------------------------
------------------------
distributed_backend=nccl
All distributed processes registered. Starting with 1 processes
----------------------------------------------------------------------------
------------------------
```

```
[NeMo I 2024-09-21 19:26:41 dist_ckpt_io:95] Using ('zarr', 1) dist-ckpt sav
e strategy.
Loading distributed checkpoint with TensorStoreLoadShardedStrategy
Loading distributed checkpoint directly on the GPU
[NeMo I 2024-09-21 19:27:29 nlp_overrides:1180] Model MegatronGPTSFTModel wa
s successfully restored from /root/verb-workspace/llama-3_1-8b-instruct-nemo
_v1.0/llama3_1_8b_instruct.nemo.
[NeMo I 2024-09-21 19:27:29 megatron_gpt_finetuning:72] Adding adapter weigh
ts to the model for PEFT
[NeMo I 2024-09-21 19:27:29 nlp_adapter_mixins:203] Before adding PEFT param
s:
      | Name  | Type          | Params | Mode 
    ------------------------------------------------
    0 | model | Float16Module | 8.0 B  | train
    ------------------------------------------------
    0         Trainable params
    8.0 B     Non-trainable params
    8.0 B     Total params
    32,121.045Total estimated model params size (MB)
[NeMo I 2024-09-21 19:27:33 nlp_adapter_mixins:208] After adding PEFT param
s:
      | Name  | Type          | Params | Mode 
    ------------------------------------------------
    0 | model | Float16Module | 8.0 B  | train
    ------------------------------------------------
    10.5 M    Trainable params
    8.0 B     Non-trainable params
    8.0 B     Total params
    32,162.988Total estimated model params size (MB)
```

```
[NeMo W 2024-09-21 19:27:33 nemo_logging:349] /usr/local/lib/python3.10/dist
-packages/pytorch_lightning/trainer/configuration_validator.py:161: You have
overridden `MegatronGPTSFTModel.configure_sharded_model` which is deprecate
d. Please override the `configure_model` hook instead. Instantiation with th
e newer hook will be created on the device right away and have the right dat
a type depending on the precision setting in the Trainer.

[NeMo W 2024-09-21 19:27:33 nemo_logging:349] /usr/local/lib/python3.10/dist
-packages/pytorch_lightning/trainer/configuration_validator.py:143: You are
using the `dataloader_iter` step flavor. If you consume the iterator more th
an once per step, the `batch_idx` argument in any hook that takes it will no
t match with the batch index of the last batch consumed. This might have unf
oreseen effects on callbacks or code that expects to get the correct index.
This will also not work well with gradient accumulation. This feature is ver
y experimental and subject to change. Here be dragons.
```

```
[NeMo I 2024-09-21 19:27:33 megatron_gpt_sft_model:811] Building GPT SFT val
idation datasets.
[NeMo I 2024-09-21 19:27:33 text_memmap_dataset:116] Building data files
[NeMo I 2024-09-21 19:27:33 text_memmap_dataset:525] Processing 1 data files
using 2 workers
```

```
huggingface/tokenizers: The current process just got forked, after paralleli
sm has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(tr
ue | false)
huggingface/tokenizers: The current process just got forked, after paralleli
sm has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(tr
ue | false)
[NeMo I 2024-09-21 19:27:33 text_memmap_dataset:495] Building indexing for f
n = ./curated-data/law-qa-val_preprocessed.jsonl
[NeMo I 2024-09-21 19:27:33 text_memmap_dataset:507] Saving idx file = ./cur
ated-data/law-qa-val_preprocessed.jsonl.idx.npy
[NeMo I 2024-09-21 19:27:33 text_memmap_dataset:509] Saving metadata file = 
./curated-data/law-qa-val_preprocessed.jsonl.idx.info
[NeMo I 2024-09-21 19:27:33 text_memmap_dataset:535] Time building 1 / 1 mem
-mapped files: 0:00:00.064523
[NeMo I 2024-09-21 19:27:33 text_memmap_dataset:525] Processing 1 data files
using 2 workers
huggingface/tokenizers: The current process just got forked, after paralleli
sm has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(tr
ue | false)
huggingface/tokenizers: The current process just got forked, after paralleli
sm has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(tr
ue | false)
[NeMo I 2024-09-21 19:27:33 text_memmap_dataset:535] Time building 0 / 1 mem
-mapped files: 0:00:00.063430
[NeMo I 2024-09-21 19:27:33 text_memmap_dataset:158] Loading data files
[NeMo I 2024-09-21 19:27:33 text_memmap_dataset:249] Loading ./curated-data/
law-qa-val_preprocessed.jsonl
[NeMo I 2024-09-21 19:27:33 text_memmap_dataset:161] Time loading 1 mem-mapp
ed files: 0:00:00.001905
[NeMo I 2024-09-21 19:27:33 text_memmap_dataset:165] Computing global indice
s
[NeMo I 2024-09-21 19:27:33 megatron_gpt_sft_model:815] Length of val datase
t: 2434
[NeMo I 2024-09-21 19:27:33 megatron_gpt_sft_model:822] Building GPT SFT tra
ing datasets.
[NeMo I 2024-09-21 19:27:33 text_memmap_dataset:116] Building data files
[NeMo I 2024-09-21 19:27:33 text_memmap_dataset:525] Processing 1 data files
using 2 workers
```

```
huggingface/tokenizers: The current process just got forked, after paralleli
sm has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(tr
ue | false)
huggingface/tokenizers: The current process just got forked, after paralleli
sm has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(tr
ue | false)
```

```
[NeMo I 2024-09-21 19:27:33 text_memmap_dataset:495] Building indexing for f
n = ./curated-data/law-qa-train_preprocessed.jsonl
[NeMo I 2024-09-21 19:27:33 text_memmap_dataset:507] Saving idx file = ./cur
ated-data/law-qa-train_preprocessed.jsonl.idx.npy
[NeMo I 2024-09-21 19:27:33 text_memmap_dataset:509] Saving metadata file =
./curated-data/law-qa-train_preprocessed.jsonl.idx.info
[NeMo I 2024-09-21 19:27:33 text_memmap_dataset:535] Time building 1 / 1 mem
-mapped files: 0:00:00.070394
[NeMo I 2024-09-21 19:27:33 text_memmap_dataset:525] Processing 1 data files
using 2 workers
```

```
huggingface/tokenizers: The current process just got forked, after paralleli
sm has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(tr
ue | false)
huggingface/tokenizers: The current process just got forked, after paralleli
sm has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(tr
ue | false)
```

```
[NeMo I 2024-09-21 19:27:33 text_memmap_dataset:535] Time building 0 / 1 mem
-mapped files: 0:00:00.048431
[NeMo I 2024-09-21 19:27:33 text_memmap_dataset:158] Loading data files
[NeMo I 2024-09-21 19:27:33 text_memmap_dataset:249] Loading ./curated-data/
law-qa-train_preprocessed.jsonl
[NeMo I 2024-09-21 19:27:33 text_memmap_dataset:161] Time loading 1 mem-mapp
ed files: 0:00:00.001379
[NeMo I 2024-09-21 19:27:33 text_memmap_dataset:165] Computing global indice
s
```

```
[NeMo W 2024-09-21 19:27:33 nemo_logging:349] /opt/NeMo/nemo/collections/nl
p/data/language_modeling/megatron/dataset_utils.py:1332: UserWarning: The to
rch.cuda.*DtypeTensor constructors are no longer recommended. It's best to u
se methods such as torch.tensor(data, dtype=*, device='cuda') to create tens
ors. (Triggered internally at /opt/pytorch/pytorch/torch/csrc/tensor/python_
tensor.cpp:83.)
      counts = torch.cuda.LongTensor([1])
```

```
make: Entering directory '/opt/NeMo/nemo/collections/nlp/data/language_model
ing/megatron'
make: Nothing to be done for 'default'.
make: Leaving directory '/opt/NeMo/nemo/collections/nlp/data/language_modeli
ng/megatron'
> building indices for blendable datasets ...
 > sample ratios:
   dataset 0, input: 1, achieved: 1
[NeMo I 2024-09-21 19:27:33 blendable_dataset:67] > elapsed time for buildin
g blendable dataset indices: 0.05 (sec)
[NeMo I 2024-09-21 19:27:33 megatron_gpt_sft_model:824] Length of train data
set: 1608
[NeMo I 2024-09-21 19:27:33 megatron_gpt_sft_model:829] Building dataloader
with consumed samples: 0
[NeMo I 2024-09-21 19:27:33 megatron_gpt_sft_model:829] Building dataloader
with consumed samples: 0
```

```
LOCAL_RANK: 0 - CUDA_VISIBLE_DEVICES: [0]
[NeMo W 2024-09-21 19:27:33 megatron_base_model:1199] Ignoring `trainer.max_
epochs` when computing `max_steps` because `trainer.max_steps` is already se
t to 50.
```

```
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
```

```
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 adapter_mixins:435] Unfrozen adapter : lora_kqv_
adapter
[NeMo I 2024-09-21 19:27:33 nlp_adapter_mixins:269] Optimizer groups set:
      | Name  | Type          | Params | Mode
    ------------------------------------------------
    0 | model | Float16Module | 8.0 B  | train
    ------------------------------------------------
    10.5 M    Trainable params
    8.0 B     Non-trainable params
    8.0 B     Total params
    32,162.988Total estimated model params size (MB)
[NeMo I 2024-09-21 19:27:33 modelPT:770] Optimizer config = FusedAdam (
    Parameter Group 0
        betas: [0.9, 0.98]
        bias_correction: True
        eps: 1e-08
        lr: 0.0001
        weight_decay: 0.01
    )
[NeMo I 2024-09-21 19:27:33 lr_scheduler:923] Scheduler "<nemo.core.optim.lr
_scheduler.CosineAnnealing object at 0x7efc6d6fac80>"
    will be used during training (effective maximum steps = 50) -
    Parameters :
    (warmup_steps: 50
    min_lr: 0.0
    constant_steps: 0
    max_steps: 50
    )
[NeMo I 2024-09-21 19:27:33 lr_scheduler:923] Scheduler "<nemo.core.optim.lr
_scheduler.CosineAnnealing object at 0x7efc6d703e50>"
    will be used during training (effective maximum steps = 50) -
    Parameters :
    (warmup_steps: 50
    min_lr: 0.0
    constant_steps: 0
    max_steps: 50
    )
```

```
  | Name  | Type          | Params | Mode 
------------------------------------------------
0 | model | Float16Module | 8.0 B  | train
------------------------------------------------
10.5 M    Trainable params
8.0 B     Non-trainable params
8.0 B     Total params
32,162.988Total estimated model params size (MB)
[NeMo W 2024-09-21 19:27:33 nemo_logging:349] /usr/local/lib/python3.10/dist
-packages/pytorch_lightning/trainer/connectors/data_connector.py:424: The 'v
al_dataloader' does not have many workers which may be a bottleneck. Conside
r increasing the value of the `num_workers` argument` to `num_workers=11` in
the `DataLoader` to improve performance.

[NeMo W 2024-09-21 19:27:33 nemo_logging:349] /usr/local/lib/python3.10/dist
-packages/pytorch_lightning/loops/utilities.py:149: Found `dataloader_iter`
argument in the `validation_step`. Note that the support for this signature
is experimental and the behavior is subject to change.

[NeMo W 2024-09-21 19:27:33 nemo_logging:349] /opt/apex/apex/transformer/pip
eline_parallel/utils.py:81: UserWarning: This function is only for unittest
      warnings.warn("This function is only for unittest")

[NeMo W 2024-09-21 19:27:40 nemo_logging:349] /usr/local/lib/python3.10/dist
-packages/pytorch_lightning/trainer/connectors/logger_connector/result.py:43
9: It is recommended to use `self.log('val_loss', ..., sync_dist=True)` when
logging on epoch level in distributed setting to accumulate the metric acros
s devices.

[NeMo W 2024-09-21 19:27:40 nemo_logging:349] /usr/local/lib/python3.10/dist
-packages/pytorch_lightning/trainer/connectors/logger_connector/result.py:43
9: It is recommended to use `self.log('validation_loss_dataloader0', ..., sy
nc_dist=True)` when logging on epoch level in distributed setting to accumul
ate the metric across devices.

[NeMo W 2024-09-21 19:27:40 nemo_logging:349] /usr/local/lib/python3.10/dist
-packages/pytorch_lightning/trainer/connectors/logger_connector/result.py:43
9: It is recommended to use `self.log('validation_loss', ..., sync_dist=Tru
e)` when logging on epoch level in distributed setting to accumulate the met
ric across devices.

[NeMo W 2024-09-21 19:27:40 nemo_logging:349] /usr/local/lib/python3.10/dist
-packages/pytorch_lightning/trainer/connectors/data_connector.py:424: The 't
rain_dataloader' does not have many workers which may be a bottleneck. Consi
der increasing the value of the `num_workers` argument` to `num_workers=11`
in the `DataLoader` to improve performance.

[NeMo W 2024-09-21 19:27:40 nemo_logging:349] /usr/local/lib/python3.10/dist
-packages/pytorch_lightning/loops/utilities.py:149: Found `dataloader_iter`
argument in the `training_step`. Note that the support for this signature is
experimental and the behavior is subject to change.
```

```
Epoch 0: :  20%|██        | 10/50 [00:57<03:48, reduced_train_loss=3.340, gl
obal_step=9.000, consumed_samples=320.0, train_step_timing in s=5.650]
Validation: |          | 0/? [00:00<?, ?it/s]
Validation:   0%|          | 0/77 [00:00<?, ?it/s]
Validation DataLoader 0:   0%|          | 0/77 [00:00<?, ?it/s]
Validation DataLoader 0:   1%|▏         | 1/77 [00:03<04:02,  0.31it/s]
Validation DataLoader 0:   3%|▎         | 2/77 [00:06<03:55,  0.32it/s]
Validation DataLoader 0:   4%|▍         | 3/77 [00:09<03:51,  0.32it/s]
Validation DataLoader 0:   5%|▌         | 4/77 [00:13<04:05,  0.30it/s]
Validation DataLoader 0:   6%|▋         | 5/77 [00:16<03:58,  0.30it/s]
Validation DataLoader 0:   8%|▊         | 6/77 [00:21<04:15,  0.28it/s]
Validation DataLoader 0:   9%|▉         | 7/77 [00:24<04:08,  0.28it/s]
Validation DataLoader 0:  10%|█         | 8/77 [00:27<04:01,  0.29it/s]
Validation DataLoader 0:  12%|█▏        | 9/77 [00:31<03:54,  0.29it/s]
Validation DataLoader 0:  13%|█▎        | 10/77 [00:34<03:48,  0.29it/s]
Validation DataLoader 0:  14%|█▍        | 11/77 [00:37<03:44,  0.29it/s]
Validation DataLoader 0:  16%|█▌        | 12/77 [00:40<03:39,  0.30it/s]
Validation DataLoader 0:  17%|█▋        | 13/77 [00:43<03:34,  0.30it/s]
Validation DataLoader 0:  18%|█▊        | 14/77 [00:46<03:30,  0.30it/s]
Validation DataLoader 0:  19%|█▉        | 15/77 [00:49<03:26,  0.30it/s]
Validation DataLoader 0:  21%|██        | 16/77 [00:53<03:22,  0.30it/s]
Validation DataLoader 0:  22%|██▏       | 17/77 [00:56<03:18,  0.30it/s]
Validation DataLoader 0:  23%|██▎       | 18/77 [00:59<03:14,  0.30it/s]
Validation DataLoader 0:  25%|██▍       | 19/77 [01:04<03:17,  0.29it/s]
Validation DataLoader 0:  26%|██▌       | 20/77 [01:07<03:13,  0.29it/s]
Validation DataLoader 0:  27%|██▋       | 21/77 [01:11<03:09,  0.30it/s]
Validation DataLoader 0:  29%|██▊       | 22/77 [01:14<03:05,  0.30it/s]
Validation DataLoader 0:  30%|██▉       | 23/77 [01:17<03:01,  0.30it/s]
Validation DataLoader 0:  31%|███       | 24/77 [01:20<02:57,  0.30it/s]
Validation DataLoader 0:  32%|███▏      | 25/77 [01:23<02:53,  0.30it/s]
Validation DataLoader 0:  34%|███▍      | 26/77 [01:26<02:49,  0.30it/s]
Validation DataLoader 0:  35%|███▌      | 27/77 [01:29<02:46,  0.30it/s]
Validation DataLoader 0:  36%|███▋      | 28/77 [01:35<02:46,  0.29it/s]
Validation DataLoader 0:  38%|███▊      | 29/77 [01:38<02:42,  0.30it/s]
Validation DataLoader 0:  39%|███▉      | 30/77 [01:41<02:38,  0.30it/s]
Validation DataLoader 0:  40%|████      | 31/77 [01:44<02:34,  0.30it/s]
Validation DataLoader 0:  42%|████▏     | 32/77 [01:47<02:31,  0.30it/s]
Validation DataLoader 0:  43%|████▎     | 33/77 [01:50<02:27,  0.30it/s]
Validation DataLoader 0:  44%|████▍     | 34/77 [01:53<02:23,  0.30it/s]
Validation DataLoader 0:  45%|████▌     | 35/77 [01:56<02:20,  0.30it/s]
Validation DataLoader 0:  47%|████▋     | 36/77 [02:00<02:17,  0.30it/s]
Validation DataLoader 0:  48%|████▊     | 37/77 [02:03<02:13,  0.30it/s]
Validation DataLoader 0:  49%|████▉     | 38/77 [02:06<02:10,  0.30it/s]
Validation DataLoader 0:  51%|█████     | 39/77 [02:09<02:06,  0.30it/s]
Validation DataLoader 0:  52%|█████▏    | 40/77 [02:12<02:02,  0.30it/s]
Validation DataLoader 0:  53%|█████▎    | 41/77 [02:15<01:59,  0.30it/s]
Validation DataLoader 0:  55%|█████▍    | 42/77 [02:19<01:55,  0.30it/s]
Validation DataLoader 0:  56%|█████▌    | 43/77 [02:22<01:52,  0.30it/s]
Validation DataLoader 0:  57%|█████▋    | 44/77 [02:25<01:49,  0.30it/s]
Validation DataLoader 0:  58%|█████▊    | 45/77 [02:28<01:45,  0.30it/s]
Validation DataLoader 0:  60%|█████▉    | 46/77 [02:31<01:42,  0.30it/s]
Validation DataLoader 0:  61%|██████    | 47/77 [02:35<01:39,  0.30it/s]
Validation DataLoader 0:  62%|██████▏   | 48/77 [02:39<01:36,  0.30it/s]
Validation DataLoader 0:  64%|██████▎   | 49/77 [02:42<01:32,  0.30it/s]
Validation DataLoader 0:  65%|██████▍   | 50/77 [02:45<01:29,  0.30it/s]
Validation DataLoader 0:  66%|██████▌   | 51/77 [02:48<01:25,  0.30it/s]
Validation DataLoader 0:  68%|██████▊   | 52/77 [02:51<01:22,  0.30it/s]
Validation DataLoader 0:  69%|██████▉   | 53/77 [02:54<01:19,  0.30it/s]
Validation DataLoader 0:  70%|███████   | 54/77 [02:58<01:15,  0.30it/s]
Validation DataLoader 0:  71%|███████▏  | 55/77 [03:01<01:12,  0.30it/s]
Validation DataLoader 0:  73%|███████▎  | 56/77 [03:04<01:09,  0.30it/s]
Validation DataLoader 0:  74%|███████▍  | 57/77 [03:07<01:05,  0.30it/s]
Validation DataLoader 0:  75%|███████▌  | 58/77 [03:10<01:02,  0.30it/s]
Validation DataLoader 0:  77%|███████▋  | 59/77 [03:13<00:59,  0.30it/s]
Validation DataLoader 0:  78%|███████▊  | 60/77 [03:16<00:55,  0.30it/s]
Validation DataLoader 0:  79%|███████▉  | 61/77 [03:20<00:52,  0.30it/s]
Validation DataLoader 0:  81%|████████  | 62/77 [03:23<00:49,  0.30it/s]
Validation DataLoader 0:  82%|████████▏ | 63/77 [03:26<00:45,  0.30it/s]
Validation DataLoader 0:  83%|████████▎ | 64/77 [03:29<00:42,  0.31it/s]
Validation DataLoader 0:  84%|████████▍ | 65/77 [03:32<00:39,  0.31it/s]
Validation DataLoader 0:  86%|████████▌ | 66/77 [03:36<00:36,  0.31it/s]
Validation DataLoader 0:  87%|████████▋ | 67/77 [03:39<00:32,  0.31it/s]
Validation DataLoader 0:  88%|████████▊ | 68/77 [03:42<00:29,  0.31it/s]
Validation DataLoader 0:  90%|████████▉ | 69/77 [03:45<00:26,  0.31it/s]
Validation DataLoader 0:  91%|█████████ | 70/77 [03:49<00:22,  0.31it/s]
Validation DataLoader 0:  92%|█████████▏| 71/77 [03:52<00:19,  0.31it/s]
Validation DataLoader 0:  94%|█████████▎| 72/77 [03:55<00:16,  0.31it/s]
Validation DataLoader 0:  95%|█████████▍| 73/77 [03:58<00:13,  0.31it/s]
Validation DataLoader 0:  96%|█████████▌| 74/77 [04:01<00:09,  0.31it/s]
Validation DataLoader 0:  97%|█████████▋| 75/77 [04:05<00:06,  0.31it/s]
Validation DataLoader 0:  99%|█████████▊| 76/77 [04:08<00:03,  0.31it/s]
Validation DataLoader 0: 100%|██████████| 77/77 [04:11<00:00,  0.31it/s]
```

```
Metric val_loss improved. New best score: 3.313
Epoch 0, global step 10: 'validation_loss' reached 3.31265 (best 3.31265), s
aving model to '/root/verb-workspace/results/Meta-llama3.1-8B-Instruct-title
gen/checkpoints/megatron_gpt_peft_lora_tuning--validation_loss=3.313-step=10
-consumed_samples=320.0.ckpt' as top 1
[NeMo W 2024-09-21 19:32:49 nlp_overrides:480] DistributedCheckpointIO confi
gured but should not be used. Reverting back to TorchCheckpointIO
```

```
Epoch 0: :  40%|████      | 20/50 [06:06<09:09, reduced_train_loss=2.870, gl
obal_step=19.00, consumed_samples=640.0, train_step_timing in s=5.660, val_l
oss=3.310]
Validation: |          | 0/? [00:00<?, ?it/s]
Validation:   0%|          | 0/77 [00:00<?, ?it/s]
Validation DataLoader 0:   0%|          | 0/77 [00:00<?, ?it/s]
Validation DataLoader 0:   1%|▏         | 1/77 [00:03<03:58,  0.32it/s]
Validation DataLoader 0:   3%|▎         | 2/77 [00:06<03:53,  0.32it/s]
Validation DataLoader 0:   4%|▍         | 3/77 [00:09<03:48,  0.32it/s]
Validation DataLoader 0:   5%|▌         | 4/77 [00:13<04:02,  0.30it/s]
Validation DataLoader 0:   6%|▋         | 5/77 [00:16<03:56,  0.30it/s]
Validation DataLoader 0:   8%|▊         | 6/77 [00:21<04:15,  0.28it/s]
Validation DataLoader 0:   9%|▉         | 7/77 [00:24<04:07,  0.28it/s]
Validation DataLoader 0:  10%|█         | 8/77 [00:27<04:00,  0.29it/s]
Validation DataLoader 0:  12%|█▏        | 9/77 [00:31<03:54,  0.29it/s]
Validation DataLoader 0:  13%|█▎        | 10/77 [00:34<03:48,  0.29it/s]
Validation DataLoader 0:  14%|█▍        | 11/77 [00:37<03:44,  0.29it/s]
Validation DataLoader 0:  16%|█▌        | 12/77 [00:40<03:38,  0.30it/s]
Validation DataLoader 0:  17%|█▋        | 13/77 [00:43<03:35,  0.30it/s]
Validation DataLoader 0:  18%|█▊        | 14/77 [00:47<03:31,  0.30it/s]
Validation DataLoader 0:  19%|█▉        | 15/77 [00:50<03:27,  0.30it/s]
Validation DataLoader 0:  21%|██        | 16/77 [00:53<03:23,  0.30it/s]
Validation DataLoader 0:  22%|██▏       | 17/77 [00:56<03:19,  0.30it/s]
Validation DataLoader 0:  23%|██▎       | 18/77 [00:59<03:15,  0.30it/s]
Validation DataLoader 0:  25%|██▍       | 19/77 [01:05<03:19,  0.29it/s]
Validation DataLoader 0:  26%|██▌       | 20/77 [01:08<03:15,  0.29it/s]
Validation DataLoader 0:  27%|██▋       | 21/77 [01:11<03:11,  0.29it/s]
Validation DataLoader 0:  29%|██▊       | 22/77 [01:14<03:07,  0.29it/s]
Validation DataLoader 0:  30%|██▉       | 23/77 [01:18<03:03,  0.29it/s]
Validation DataLoader 0:  31%|███       | 24/77 [01:21<02:59,  0.30it/s]
Validation DataLoader 0:  32%|███▏      | 25/77 [01:24<02:55,  0.30it/s]
Validation DataLoader 0:  34%|███▍      | 26/77 [01:27<02:52,  0.30it/s]
Validation DataLoader 0:  35%|███▌      | 27/77 [01:30<02:48,  0.30it/s]
Validation DataLoader 0:  36%|███▋      | 28/77 [01:36<02:48,  0.29it/s]
Validation DataLoader 0:  38%|███▊      | 29/77 [01:39<02:45,  0.29it/s]
Validation DataLoader 0:  39%|███▉      | 30/77 [01:42<02:41,  0.29it/s]
Validation DataLoader 0:  40%|████      | 31/77 [01:45<02:37,  0.29it/s]
Validation DataLoader 0:  42%|████▏     | 32/77 [01:49<02:33,  0.29it/s]
Validation DataLoader 0:  43%|████▎     | 33/77 [01:52<02:29,  0.29it/s]
Validation DataLoader 0:  44%|████▍     | 34/77 [01:55<02:25,  0.29it/s]
Validation DataLoader 0:  45%|████▌     | 35/77 [01:58<02:22,  0.29it/s]
Validation DataLoader 0:  47%|████▋     | 36/77 [02:02<02:19,  0.29it/s]
Validation DataLoader 0:  48%|████▊     | 37/77 [02:05<02:15,  0.30it/s]
Validation DataLoader 0:  49%|████▉     | 38/77 [02:08<02:11,  0.30it/s]
Validation DataLoader 0:  51%|█████     | 39/77 [02:11<02:08,  0.30it/s]
Validation DataLoader 0:  52%|█████▏    | 40/77 [02:15<02:04,  0.30it/s]
Validation DataLoader 0:  53%|█████▎    | 41/77 [02:18<02:01,  0.30it/s]
Validation DataLoader 0:  55%|█████▍    | 42/77 [02:21<01:57,  0.30it/s]
Validation DataLoader 0:  56%|█████▌    | 43/77 [02:24<01:54,  0.30it/s]
Validation DataLoader 0:  57%|█████▋    | 44/77 [02:27<01:50,  0.30it/s]
Validation DataLoader 0:  58%|█████▊    | 45/77 [02:30<01:47,  0.30it/s]
Validation DataLoader 0:  60%|█████▉    | 46/77 [02:33<01:43,  0.30it/s]
Validation DataLoader 0:  61%|██████    | 47/77 [02:38<01:40,  0.30it/s]
Validation DataLoader 0:  62%|██████▏   | 48/77 [02:41<01:37,  0.30it/s]
Validation DataLoader 0:  64%|██████▎   | 49/77 [02:44<01:33,  0.30it/s]
Validation DataLoader 0:  65%|██████▍   | 50/77 [02:47<01:30,  0.30it/s]
Validation DataLoader 0:  66%|██████▌   | 51/77 [02:50<01:27,  0.30it/s]
Validation DataLoader 0:  68%|██████▊   | 52/77 [02:54<01:23,  0.30it/s]
Validation DataLoader 0:  69%|██████▉   | 53/77 [02:57<01:20,  0.30it/s]
Validation DataLoader 0:  70%|███████   | 54/77 [03:00<01:16,  0.30it/s]
Validation DataLoader 0:  71%|███████▏  | 55/77 [03:03<01:13,  0.30it/s]
Validation DataLoader 0:  73%|███████▎  | 56/77 [03:06<01:10,  0.30it/s]
Validation DataLoader 0:  74%|███████▍  | 57/77 [03:10<01:06,  0.30it/s]
Validation DataLoader 0:  75%|███████▌  | 58/77 [03:13<01:03,  0.30it/s]
Validation DataLoader 0:  77%|███████▋  | 59/77 [03:16<00:59,  0.30it/s]
Validation DataLoader 0:  78%|███████▊  | 60/77 [03:19<00:56,  0.30it/s]
Validation DataLoader 0:  79%|███████▉  | 61/77 [03:22<00:53,  0.30it/s]
Validation DataLoader 0:  81%|████████  | 62/77 [03:25<00:49,  0.30it/s]
Validation DataLoader 0:  82%|████████▏ | 63/77 [03:29<00:46,  0.30it/s]
Validation DataLoader 0:  83%|████████▎ | 64/77 [03:32<00:43,  0.30it/s]
Validation DataLoader 0:  84%|████████▍ | 65/77 [03:35<00:39,  0.30it/s]
Validation DataLoader 0:  86%|████████▌ | 66/77 [03:38<00:36,  0.30it/s]
Validation DataLoader 0:  87%|████████▋ | 67/77 [03:41<00:33,  0.30it/s]
Validation DataLoader 0:  88%|████████▊ | 68/77 [03:45<00:29,  0.30it/s]
Validation DataLoader 0:  90%|████████▉ | 69/77 [03:48<00:26,  0.30it/s]
Validation DataLoader 0:  91%|█████████ | 70/77 [03:51<00:23,  0.30it/s]
Validation DataLoader 0:  92%|█████████▏| 71/77 [03:54<00:19,  0.30it/s]
Validation DataLoader 0:  94%|█████████▎| 72/77 [03:57<00:16,  0.30it/s]
Validation DataLoader 0:  95%|█████████▍| 73/77 [04:00<00:13,  0.30it/s]
Validation DataLoader 0:  96%|█████████▌| 74/77 [04:04<00:09,  0.30it/s]
Validation DataLoader 0:  97%|█████████▋| 75/77 [04:07<00:06,  0.30it/s]
Validation DataLoader 0:  99%|█████████▊| 76/77 [04:11<00:03,  0.30it/s]
Validation DataLoader 0: 100%|██████████| 77/77 [04:14<00:00,  0.30it/s]
```

```
Metric val_loss improved by 0.749 >= min_delta = 0.001. New best score: 2.56
4
Epoch 0, global step 20: 'validation_loss' reached 2.56401 (best 2.56401), s
aving model to '/root/verb-workspace/results/Meta-llama3.1-8B-Instruct-title
gen/checkpoints/megatron_gpt_peft_lora_tuning--validation_loss=2.564-step=20
-consumed_samples=640.0.ckpt' as top 1
```

```
Epoch 0: :  40%|████      | 20/50 [10:20<15:30, reduced_train_loss=2.870, gl
obal_step=19.00, consumed_samples=640.0, train_step_timing in s=5.660, val_l
oss=2.560][NeMo I 2024-09-21 19:38:01 nlp_overrides:464] Removing checkpoin
t: /root/verb-workspace/results/Meta-llama3.1-8B-Instruct-titlegen/checkpoin
ts/megatron_gpt_peft_lora_tuning--validation_loss=3.313-step=10-consumed_sam
ples=320.0.ckpt
[NeMo I 2024-09-21 19:38:02 nlp_overrides:464] Removing checkpoint: /root/ve
rb-workspace/results/Meta-llama3.1-8B-Instruct-titlegen/checkpoints/megatron
_gpt_peft_lora_tuning--validation_loss=3.313-step=10-consumed_samples=320.0-
last.ckpt
Epoch 0: :  60%|██████    | 30/50 [11:18<07:32, reduced_train_loss=2.080, gl
obal_step=29.00, consumed_samples=960.0, train_step_timing in s=5.670, val_l
oss=2.560]
Validation: |          | 0/? [00:00<?, ?it/s]
Validation:   0%|          | 0/77 [00:00<?, ?it/s]
Validation DataLoader 0:   0%|          | 0/77 [00:00<?, ?it/s]
Validation DataLoader 0:   1%|▏         | 1/77 [00:03<03:53,  0.33it/s]
Validation DataLoader 0:   3%|▎         | 2/77 [00:06<03:51,  0.32it/s]
Validation DataLoader 0:   4%|▍         | 3/77 [00:09<03:48,  0.32it/s]
Validation DataLoader 0:   5%|▌         | 4/77 [00:13<04:02,  0.30it/s]
Validation DataLoader 0:   6%|▋         | 5/77 [00:16<03:55,  0.31it/s]
Validation DataLoader 0:   8%|▊         | 6/77 [00:21<04:17,  0.28it/s]
Validation DataLoader 0:   9%|▉         | 7/77 [00:24<04:09,  0.28it/s]
Validation DataLoader 0:  10%|█         | 8/77 [00:28<04:01,  0.29it/s]
Validation DataLoader 0:  12%|█▏        | 9/77 [00:31<03:55,  0.29it/s]
Validation DataLoader 0:  13%|█▎        | 10/77 [00:34<03:49,  0.29it/s]
Validation DataLoader 0:  14%|█▍        | 11/77 [00:37<03:45,  0.29it/s]
Validation DataLoader 0:  16%|█▌        | 12/77 [00:40<03:40,  0.29it/s]
Validation DataLoader 0:  17%|█▋        | 13/77 [00:43<03:35,  0.30it/s]
Validation DataLoader 0:  18%|█▊        | 14/77 [00:46<03:31,  0.30it/s]
Validation DataLoader 0:  19%|█▉        | 15/77 [00:50<03:27,  0.30it/s]
Validation DataLoader 0:  21%|██        | 16/77 [00:53<03:22,  0.30it/s]
Validation DataLoader 0:  22%|██▏       | 17/77 [00:56<03:18,  0.30it/s]
Validation DataLoader 0:  23%|██▎       | 18/77 [00:59<03:15,  0.30it/s]
Validation DataLoader 0:  25%|██▍       | 19/77 [01:05<03:19,  0.29it/s]
Validation DataLoader 0:  26%|██▌       | 20/77 [01:08<03:14,  0.29it/s]
Validation DataLoader 0:  27%|██▋       | 21/77 [01:11<03:10,  0.29it/s]
Validation DataLoader 0:  29%|██▊       | 22/77 [01:14<03:07,  0.29it/s]
Validation DataLoader 0:  30%|██▉       | 23/77 [01:18<03:03,  0.29it/s]
Validation DataLoader 0:  31%|███       | 24/77 [01:21<02:59,  0.30it/s]
Validation DataLoader 0:  32%|███▏      | 25/77 [01:24<02:55,  0.30it/s]
Validation DataLoader 0:  34%|███▍      | 26/77 [01:27<02:51,  0.30it/s]
Validation DataLoader 0:  35%|███▌      | 27/77 [01:30<02:48,  0.30it/s]
Validation DataLoader 0:  36%|███▋      | 28/77 [01:36<02:48,  0.29it/s]
Validation DataLoader 0:  38%|███▊      | 29/77 [01:39<02:44,  0.29it/s]
Validation DataLoader 0:  39%|███▉      | 30/77 [01:42<02:40,  0.29it/s]
Validation DataLoader 0:  40%|████      | 31/77 [01:45<02:37,  0.29it/s]
Validation DataLoader 0:  42%|████▏     | 32/77 [01:48<02:33,  0.29it/s]
Validation DataLoader 0:  43%|████▎     | 33/77 [01:52<02:29,  0.29it/s]
Validation DataLoader 0:  44%|████▍     | 34/77 [01:55<02:25,  0.30it/s]
Validation DataLoader 0:  45%|████▌     | 35/77 [01:58<02:22,  0.30it/s]
Validation DataLoader 0:  47%|████▋     | 36/77 [02:01<02:18,  0.30it/s]
Validation DataLoader 0:  48%|████▊     | 37/77 [02:05<02:15,  0.30it/s]
Validation DataLoader 0:  49%|████▉     | 38/77 [02:08<02:11,  0.30it/s]
Validation DataLoader 0:  51%|█████     | 39/77 [02:11<02:07,  0.30it/s]
Validation DataLoader 0:  52%|█████▏    | 40/77 [02:14<02:04,  0.30it/s]
Validation DataLoader 0:  53%|█████▎    | 41/77 [02:17<02:00,  0.30it/s]
Validation DataLoader 0:  55%|█████▍    | 42/77 [02:20<01:57,  0.30it/s]
Validation DataLoader 0:  56%|█████▌    | 43/77 [02:23<01:53,  0.30it/s]
Validation DataLoader 0:  57%|█████▋    | 44/77 [02:26<01:50,  0.30it/s]
Validation DataLoader 0:  58%|█████▊    | 45/77 [02:29<01:46,  0.30it/s]
Validation DataLoader 0:  60%|█████▉    | 46/77 [02:33<01:43,  0.30it/s]
Validation DataLoader 0:  61%|██████    | 47/77 [02:37<01:40,  0.30it/s]
Validation DataLoader 0:  62%|██████▏   | 48/77 [02:40<01:36,  0.30it/s]
Validation DataLoader 0:  64%|██████▎   | 49/77 [02:43<01:33,  0.30it/s]
Validation DataLoader 0:  65%|██████▍   | 50/77 [02:46<01:30,  0.30it/s]
Validation DataLoader 0:  66%|██████▌   | 51/77 [02:49<01:26,  0.30it/s]
Validation DataLoader 0:  68%|██████▊   | 52/77 [02:53<01:23,  0.30it/s]
Validation DataLoader 0:  69%|██████▉   | 53/77 [02:56<01:19,  0.30it/s]
Validation DataLoader 0:  70%|███████   | 54/77 [02:59<01:16,  0.30it/s]
Validation DataLoader 0:  71%|███████▏  | 55/77 [03:02<01:13,  0.30it/s]
Validation DataLoader 0:  73%|███████▎  | 56/77 [03:06<01:09,  0.30it/s]
Validation DataLoader 0:  74%|███████▍  | 57/77 [03:09<01:06,  0.30it/s]
Validation DataLoader 0:  75%|███████▌  | 58/77 [03:12<01:02,  0.30it/s]
Validation DataLoader 0:  77%|███████▋  | 59/77 [03:15<00:59,  0.30it/s]
Validation DataLoader 0:  78%|███████▊  | 60/77 [03:18<00:56,  0.30it/s]
Validation DataLoader 0:  79%|███████▉  | 61/77 [03:21<00:52,  0.30it/s]
Validation DataLoader 0:  81%|████████  | 62/77 [03:24<00:49,  0.30it/s]
Validation DataLoader 0:  82%|████████▏ | 63/77 [03:27<00:46,  0.30it/s]
Validation DataLoader 0:  83%|████████▎ | 64/77 [03:31<00:42,  0.30it/s]
Validation DataLoader 0:  84%|████████▍ | 65/77 [03:34<00:39,  0.30it/s]
Validation DataLoader 0:  86%|████████▌ | 66/77 [03:37<00:36,  0.30it/s]
Validation DataLoader 0:  87%|████████▋ | 67/77 [03:40<00:32,  0.30it/s]
Validation DataLoader 0:  88%|████████▊ | 68/77 [03:43<00:29,  0.30it/s]
Validation DataLoader 0:  90%|████████▉ | 69/77 [03:46<00:26,  0.30it/s]
Validation DataLoader 0:  91%|█████████ | 70/77 [03:50<00:23,  0.30it/s]
Validation DataLoader 0:  92%|█████████▏| 71/77 [03:53<00:19,  0.30it/s]
Validation DataLoader 0:  94%|█████████▎| 72/77 [03:56<00:16,  0.30it/s]
Validation DataLoader 0:  95%|█████████▍| 73/77 [03:59<00:13,  0.30it/s]
Validation DataLoader 0:  96%|█████████▌| 74/77 [04:02<00:09,  0.30it/s]
Validation DataLoader 0:  97%|█████████▋| 75/77 [04:06<00:06,  0.30it/s]
Validation DataLoader 0:  99%|█████████▊| 76/77 [04:09<00:03,  0.30it/s]
Validation DataLoader 0: 100%|██████████| 77/77 [04:12<00:00,  0.31it/s]
```

```
Metric val_loss improved by 0.586 >= min_delta = 0.001. New best score: 1.97
8
Epoch 0, global step 30: 'validation_loss' reached 1.97761 (best 1.97761), s
aving model to '/root/verb-workspace/results/Meta-llama3.1-8B-Instruct-title
gen/checkpoints/megatron_gpt_peft_lora_tuning--validation_loss=1.978-step=30
-consumed_samples=960.0.ckpt' as top 1
```

```
Epoch 0: :  60%|██████    | 30/50 [15:30<10:20, reduced_train_loss=2.080, gl
obal_step=29.00, consumed_samples=960.0, train_step_timing in s=5.670, val_l
oss=1.980][NeMo I 2024-09-21 19:43:11 nlp_overrides:464] Removing checkpoin
t: /root/verb-workspace/results/Meta-llama3.1-8B-Instruct-titlegen/checkpoin
ts/megatron_gpt_peft_lora_tuning--validation_loss=2.564-step=20-consumed_sam
ples=640.0.ckpt
[NeMo I 2024-09-21 19:43:12 nlp_overrides:464] Removing checkpoint: /root/ve
rb-workspace/results/Meta-llama3.1-8B-Instruct-titlegen/checkpoints/megatron
_gpt_peft_lora_tuning--validation_loss=2.564-step=20-consumed_samples=640.0-
last.ckpt
Epoch 0: :  80%|████████  | 40/50 [16:27<04:06, reduced_train_loss=1.790, gl
obal_step=39.00, consumed_samples=1280.0, train_step_timing in s=5.600, val_
loss=1.980]
Validation: |          | 0/? [00:00<?, ?it/s]
Validation:   0%|          | 0/77 [00:00<?, ?it/s]
Validation DataLoader 0:   0%|          | 0/77 [00:00<?, ?it/s]
Validation DataLoader 0:   1%|▏         | 1/77 [00:03<03:55,  0.32it/s]
Validation DataLoader 0:   3%|▎         | 2/77 [00:06<03:53,  0.32it/s]
Validation DataLoader 0:   4%|▍         | 3/77 [00:09<03:49,  0.32it/s]
Validation DataLoader 0:   5%|▌         | 4/77 [00:13<04:03,  0.30it/s]
Validation DataLoader 0:   6%|▋         | 5/77 [00:16<03:56,  0.30it/s]
Validation DataLoader 0:   8%|▊         | 6/77 [00:21<04:14,  0.28it/s]
Validation DataLoader 0:   9%|▉         | 7/77 [00:24<04:06,  0.28it/s]
Validation DataLoader 0:  10%|█         | 8/77 [00:27<03:59,  0.29it/s]
Validation DataLoader 0:  12%|█▏        | 9/77 [00:30<03:53,  0.29it/s]
Validation DataLoader 0:  13%|█▎        | 10/77 [00:34<03:48,  0.29it/s]
Validation DataLoader 0:  14%|█▍        | 11/77 [00:37<03:43,  0.30it/s]
Validation DataLoader 0:  16%|█▌        | 12/77 [00:40<03:38,  0.30it/s]
Validation DataLoader 0:  17%|█▋        | 13/77 [00:43<03:33,  0.30it/s]
Validation DataLoader 0:  18%|█▊        | 14/77 [00:46<03:29,  0.30it/s]
Validation DataLoader 0:  19%|█▉        | 15/77 [00:49<03:24,  0.30it/s]
Validation DataLoader 0:  21%|██        | 16/77 [00:52<03:20,  0.30it/s]
Validation DataLoader 0:  22%|██▏       | 17/77 [00:55<03:17,  0.30it/s]
Validation DataLoader 0:  23%|██▎       | 18/77 [00:59<03:13,  0.30it/s]
Validation DataLoader 0:  25%|██▍       | 19/77 [01:04<03:17,  0.29it/s]
Validation DataLoader 0:  26%|██▌       | 20/77 [01:07<03:13,  0.30it/s]
Validation DataLoader 0:  27%|██▋       | 21/77 [01:10<03:08,  0.30it/s]
Validation DataLoader 0:  29%|██▊       | 22/77 [01:14<03:05,  0.30it/s]
Validation DataLoader 0:  30%|██▉       | 23/77 [01:17<03:01,  0.30it/s]
Validation DataLoader 0:  31%|███       | 24/77 [01:20<02:57,  0.30it/s]
Validation DataLoader 0:  32%|███▏      | 25/77 [01:23<02:53,  0.30it/s]
Validation DataLoader 0:  34%|███▍      | 26/77 [01:26<02:49,  0.30it/s]
Validation DataLoader 0:  35%|███▌      | 27/77 [01:29<02:46,  0.30it/s]
Validation DataLoader 0:  36%|███▋      | 28/77 [01:35<02:47,  0.29it/s]
Validation DataLoader 0:  38%|███▊      | 29/77 [01:38<02:43,  0.29it/s]
Validation DataLoader 0:  39%|███▉      | 30/77 [01:41<02:39,  0.30it/s]
Validation DataLoader 0:  40%|████      | 31/77 [01:44<02:35,  0.30it/s]
Validation DataLoader 0:  42%|████▏     | 32/77 [01:47<02:31,  0.30it/s]
Validation DataLoader 0:  43%|████▎     | 33/77 [01:50<02:27,  0.30it/s]
Validation DataLoader 0:  44%|████▍     | 34/77 [01:54<02:24,  0.30it/s]
Validation DataLoader 0:  45%|████▌     | 35/77 [01:57<02:21,  0.30it/s]
Validation DataLoader 0:  47%|████▋     | 36/77 [02:00<02:17,  0.30it/s]
Validation DataLoader 0:  48%|████▊     | 37/77 [02:03<02:13,  0.30it/s]
Validation DataLoader 0:  49%|████▉     | 38/77 [02:06<02:10,  0.30it/s]
Validation DataLoader 0:  51%|█████     | 39/77 [02:10<02:06,  0.30it/s]
Validation DataLoader 0:  52%|█████▏    | 40/77 [02:13<02:03,  0.30it/s]
Validation DataLoader 0:  53%|█████▎    | 41/77 [02:16<01:59,  0.30it/s]
Validation DataLoader 0:  55%|█████▍    | 42/77 [02:19<01:56,  0.30it/s]
Validation DataLoader 0:  56%|█████▌    | 43/77 [02:22<01:52,  0.30it/s]
Validation DataLoader 0:  57%|█████▋    | 44/77 [02:26<01:49,  0.30it/s]
Validation DataLoader 0:  58%|█████▊    | 45/77 [02:29<01:46,  0.30it/s]
Validation DataLoader 0:  60%|█████▉    | 46/77 [02:32<01:42,  0.30it/s]
Validation DataLoader 0:  61%|██████    | 47/77 [02:36<01:39,  0.30it/s]
Validation DataLoader 0:  62%|██████▏   | 48/77 [02:39<01:36,  0.30it/s]
Validation DataLoader 0:  64%|██████▎   | 49/77 [02:42<01:32,  0.30it/s]
Validation DataLoader 0:  65%|██████▍   | 50/77 [02:45<01:29,  0.30it/s]
Validation DataLoader 0:  66%|██████▌   | 51/77 [02:49<01:26,  0.30it/s]
Validation DataLoader 0:  68%|██████▊   | 52/77 [02:52<01:22,  0.30it/s]
Validation DataLoader 0:  69%|██████▉   | 53/77 [02:55<01:19,  0.30it/s]
Validation DataLoader 0:  70%|███████   | 54/77 [02:58<01:16,  0.30it/s]
Validation DataLoader 0:  71%|███████▏  | 55/77 [03:02<01:12,  0.30it/s]
Validation DataLoader 0:  73%|███████▎  | 56/77 [03:05<01:09,  0.30it/s]
Validation DataLoader 0:  74%|███████▍  | 57/77 [03:08<01:06,  0.30it/s]
Validation DataLoader 0:  75%|███████▌  | 58/77 [03:11<01:02,  0.30it/s]
Validation DataLoader 0:  77%|███████▋  | 59/77 [03:14<00:59,  0.30it/s]
Validation DataLoader 0:  78%|███████▊  | 60/77 [03:17<00:56,  0.30it/s]
Validation DataLoader 0:  79%|███████▉  | 61/77 [03:20<00:52,  0.30it/s]
Validation DataLoader 0:  81%|████████  | 62/77 [03:24<00:49,  0.30it/s]
Validation DataLoader 0:  82%|████████▏ | 63/77 [03:27<00:46,  0.30it/s]
Validation DataLoader 0:  83%|████████▎ | 64/77 [03:30<00:42,  0.30it/s]
Validation DataLoader 0:  84%|████████▍ | 65/77 [03:33<00:39,  0.30it/s]
Validation DataLoader 0:  86%|████████▌ | 66/77 [03:36<00:36,  0.30it/s]
Validation DataLoader 0:  87%|████████▋ | 67/77 [03:39<00:32,  0.30it/s]
Validation DataLoader 0:  88%|████████▊ | 68/77 [03:42<00:29,  0.31it/s]
Validation DataLoader 0:  90%|████████▉ | 69/77 [03:45<00:26,  0.31it/s]
Validation DataLoader 0:  91%|█████████ | 70/77 [03:48<00:22,  0.31it/s]
Validation DataLoader 0:  92%|█████████▏| 71/77 [03:52<00:19,  0.31it/s]
Validation DataLoader 0:  94%|█████████▎| 72/77 [03:55<00:16,  0.31it/s]
Validation DataLoader 0:  95%|█████████▍| 73/77 [03:58<00:13,  0.31it/s]
Validation DataLoader 0:  96%|█████████▌| 74/77 [04:01<00:09,  0.31it/s]
Validation DataLoader 0:  97%|█████████▋| 75/77 [04:04<00:06,  0.31it/s]
Validation DataLoader 0:  99%|█████████▊| 76/77 [04:07<00:03,  0.31it/s]
Validation DataLoader 0: 100%|██████████| 77/77 [04:10<00:00,  0.31it/s]
```

```
Metric val_loss improved by 0.216 >= min_delta = 0.001. New best score: 1.76
1
Epoch 0, global step 40: 'validation_loss' reached 1.76125 (best 1.76125), s
aving model to '/root/verb-workspace/results/Meta-llama3.1-8B-Instruct-title
gen/checkpoints/megatron_gpt_peft_lora_tuning--validation_loss=1.761-step=40
-consumed_samples=1280.0.ckpt' as top 1
```

```
Epoch 0: :  80%|████████  | 40/50 [20:38<05:09, reduced_train_loss=1.790, gl
obal_step=39.00, consumed_samples=1280.0, train_step_timing in s=5.600, val_
loss=1.760][NeMo I 2024-09-21 19:48:20 nlp_overrides:464] Removing checkpoin
t: /root/verb-workspace/results/Meta-llama3.1-8B-Instruct-titlegen/checkpoin
ts/megatron_gpt_peft_lora_tuning--validation_loss=1.978-step=30-consumed_sam
ples=960.0.ckpt
[NeMo I 2024-09-21 19:48:20 nlp_overrides:464] Removing checkpoint: /root/ve
rb-workspace/results/Meta-llama3.1-8B-Instruct-titlegen/checkpoints/megatron
_gpt_peft_lora_tuning--validation_loss=1.978-step=30-consumed_samples=960.0-
last.ckpt
Epoch 0: : 100%|██████████| 50/50 [21:36<00:00, reduced_train_loss=1.710, gl
obal_step=49.00, consumed_samples=1600.0, train_step_timing in s=5.580, val_
loss=1.760]
Validation: |          | 0/? [00:00<?, ?it/s]
Validation:   0%|          | 0/77 [00:00<?, ?it/s]
Validation DataLoader 0:   0%|          | 0/77 [00:00<?, ?it/s]
Validation DataLoader 0:   1%|▏         | 1/77 [00:03<03:52,  0.33it/s]
Validation DataLoader 0:   3%|▎         | 2/77 [00:06<03:50,  0.33it/s]
Validation DataLoader 0:   4%|▍         | 3/77 [00:09<03:49,  0.32it/s]
Validation DataLoader 0:   5%|▌         | 4/77 [00:13<04:03,  0.30it/s]
Validation DataLoader 0:   6%|▋         | 5/77 [00:16<03:56,  0.30it/s]
Validation DataLoader 0:   8%|▊         | 6/77 [00:21<04:13,  0.28it/s]
Validation DataLoader 0:   9%|▉         | 7/77 [00:24<04:06,  0.28it/s]
Validation DataLoader 0:  10%|█         | 8/77 [00:27<03:58,  0.29it/s]
Validation DataLoader 0:  12%|█▏        | 9/77 [00:30<03:52,  0.29it/s]
Validation DataLoader 0:  13%|█▎        | 10/77 [00:33<03:47,  0.29it/s]
Validation DataLoader 0:  14%|█▍        | 11/77 [00:37<03:42,  0.30it/s]
Validation DataLoader 0:  16%|█▌        | 12/77 [00:40<03:37,  0.30it/s]
Validation DataLoader 0:  17%|█▋        | 13/77 [00:43<03:32,  0.30it/s]
Validation DataLoader 0:  18%|█▊        | 14/77 [00:46<03:29,  0.30it/s]
Validation DataLoader 0:  19%|█▉        | 15/77 [00:49<03:25,  0.30it/s]
Validation DataLoader 0:  21%|██        | 16/77 [00:52<03:21,  0.30it/s]
Validation DataLoader 0:  22%|██▏       | 17/77 [00:56<03:18,  0.30it/s]
Validation DataLoader 0:  23%|██▎       | 18/77 [00:59<03:14,  0.30it/s]
Validation DataLoader 0:  25%|██▍       | 19/77 [01:04<03:17,  0.29it/s]
Validation DataLoader 0:  26%|██▌       | 20/77 [01:07<03:13,  0.29it/s]
Validation DataLoader 0:  27%|██▋       | 21/77 [01:11<03:09,  0.30it/s]
Validation DataLoader 0:  29%|██▊       | 22/77 [01:14<03:05,  0.30it/s]
Validation DataLoader 0:  30%|██▉       | 23/77 [01:17<03:01,  0.30it/s]
Validation DataLoader 0:  31%|███       | 24/77 [01:20<02:57,  0.30it/s]
Validation DataLoader 0:  32%|███▏      | 25/77 [01:23<02:53,  0.30it/s]
Validation DataLoader 0:  34%|███▍      | 26/77 [01:26<02:49,  0.30it/s]
Validation DataLoader 0:  35%|███▌      | 27/77 [01:30<02:47,  0.30it/s]
Validation DataLoader 0:  36%|███▋      | 28/77 [01:35<02:47,  0.29it/s]
Validation DataLoader 0:  38%|███▊      | 29/77 [01:38<02:43,  0.29it/s]
Validation DataLoader 0:  39%|███▉      | 30/77 [01:41<02:39,  0.29it/s]
Validation DataLoader 0:  40%|████      | 31/77 [01:45<02:35,  0.30it/s]
Validation DataLoader 0:  42%|████▏     | 32/77 [01:48<02:32,  0.30it/s]
Validation DataLoader 0:  43%|████▎     | 33/77 [01:51<02:28,  0.30it/s]
Validation DataLoader 0:  44%|████▍     | 34/77 [01:54<02:24,  0.30it/s]
Validation DataLoader 0:  45%|████▌     | 35/77 [01:57<02:21,  0.30it/s]
Validation DataLoader 0:  47%|████▋     | 36/77 [02:01<02:17,  0.30it/s]
Validation DataLoader 0:  48%|████▊     | 37/77 [02:04<02:14,  0.30it/s]
Validation DataLoader 0:  49%|████▉     | 38/77 [02:07<02:10,  0.30it/s]
Validation DataLoader 0:  51%|█████     | 39/77 [02:10<02:07,  0.30it/s]
Validation DataLoader 0:  52%|█████▏    | 40/77 [02:13<02:03,  0.30it/s]
Validation DataLoader 0:  53%|█████▎    | 41/77 [02:16<01:59,  0.30it/s]
Validation DataLoader 0:  55%|█████▍    | 42/77 [02:19<01:56,  0.30it/s]
Validation DataLoader 0:  56%|█████▌    | 43/77 [02:22<01:52,  0.30it/s]
Validation DataLoader 0:  57%|█████▋    | 44/77 [02:25<01:49,  0.30it/s]
Validation DataLoader 0:  58%|█████▊    | 45/77 [02:29<01:45,  0.30it/s]
Validation DataLoader 0:  60%|█████▉    | 46/77 [02:32<01:42,  0.30it/s]
Validation DataLoader 0:  61%|██████    | 47/77 [02:36<01:39,  0.30it/s]
Validation DataLoader 0:  62%|██████▏   | 48/77 [02:39<01:36,  0.30it/s]
Validation DataLoader 0:  64%|██████▎   | 49/77 [02:42<01:33,  0.30it/s]
Validation DataLoader 0:  65%|██████▍   | 50/77 [02:46<01:29,  0.30it/s]
Validation DataLoader 0:  66%|██████▌   | 51/77 [02:49<01:26,  0.30it/s]
Validation DataLoader 0:  68%|██████▊   | 52/77 [02:52<01:22,  0.30it/s]
Validation DataLoader 0:  69%|██████▉   | 53/77 [02:55<01:19,  0.30it/s]
Validation DataLoader 0:  70%|███████   | 54/77 [02:58<01:16,  0.30it/s]
Validation DataLoader 0:  71%|███████▏  | 55/77 [03:01<01:12,  0.30it/s]
Validation DataLoader 0:  73%|███████▎  | 56/77 [03:04<01:09,  0.30it/s]
Validation DataLoader 0:  74%|███████▍  | 57/77 [03:08<01:06,  0.30it/s]
Validation DataLoader 0:  75%|███████▌  | 58/77 [03:11<01:02,  0.30it/s]
Validation DataLoader 0:  77%|███████▋  | 59/77 [03:14<00:59,  0.30it/s]
Validation DataLoader 0:  78%|███████▊  | 60/77 [03:17<00:55,  0.30it/s]
Validation DataLoader 0:  79%|███████▉  | 61/77 [03:20<00:52,  0.30it/s]
Validation DataLoader 0:  81%|████████  | 62/77 [03:23<00:49,  0.30it/s]
Validation DataLoader 0:  82%|████████▏ | 63/77 [03:26<00:45,  0.30it/s]
Validation DataLoader 0:  83%|████████▎ | 64/77 [03:30<00:42,  0.30it/s]
Validation DataLoader 0:  84%|████████▍ | 65/77 [03:34<00:39,  0.30it/s]
Validation DataLoader 0:  86%|████████▌ | 66/77 [03:37<00:36,  0.30it/s]
Validation DataLoader 0:  87%|████████▋ | 67/77 [03:40<00:32,  0.30it/s]
Validation DataLoader 0:  88%|████████▊ | 68/77 [03:43<00:29,  0.30it/s]
Validation DataLoader 0:  90%|████████▉ | 69/77 [03:46<00:26,  0.30it/s]
Validation DataLoader 0:  91%|█████████ | 70/77 [03:49<00:22,  0.31it/s]
Validation DataLoader 0:  92%|█████████▏| 71/77 [03:52<00:19,  0.31it/s]
Validation DataLoader 0:  94%|█████████▎| 72/77 [03:55<00:16,  0.31it/s]
Validation DataLoader 0:  95%|█████████▍| 73/77 [03:58<00:13,  0.31it/s]
Validation DataLoader 0:  96%|█████████▌| 74/77 [04:02<00:09,  0.31it/s]
Validation DataLoader 0:  97%|█████████▋| 75/77 [04:05<00:06,  0.31it/s]
Validation DataLoader 0:  99%|█████████▊| 76/77 [04:08<00:03,  0.31it/s]
Validation DataLoader 0: 100%|██████████| 77/77 [04:11<00:00,  0.31it/s]
```

```
Metric val_loss improved by 0.045 >= min_delta = 0.001. New best score: 1.717
Epoch 0, global step 50: 'validation_loss' reached 1.71671 (best 1.71671), saving model to '/root/verb-workspace/results/Meta-llama3.1-8B-Instruct-titlegen/checkpoints/megatron_gpt_peft_lora_tuning--validation_loss=1.717-step=50-consumed_samples=1600.0.ckpt' as top 1
```

```
Epoch 0: : 100%|██████████| 50/50 [25:47<00:00, reduced_train_loss=1.710, global_step=49.00, consumed_samples=1600.0, train_step_timing in s=5.580, val_loss=1.720][NeMo I 2024-09-21 19:53:28 nlp_overrides:464] Removing checkpoint: /root/verb-workspace/results/Meta-llama3.1-8B-Instruct-titlegen/checkpoints/megatron_gpt_peft_lora_tuning--validation_loss=1.761-step=40-consumed_samples=1280.0.ckpt
[NeMo I 2024-09-21 19:53:29 nlp_overrides:464] Removing checkpoint: /root/verb-workspace/results/Meta-llama3.1-8B-Instruct-titlegen/checkpoints/megatron_gpt_peft_lora_tuning--validation_loss=1.761-step=40-consumed_samples=1280.0-last.ckpt
```

```
`Trainer.fit` stopped: `max_steps=50` reached.
```

```
Epoch 0: : 100%|██████████| 50/50 [25:48<00:00, reduced_train_loss=1.710, gl
obal_step=49.00, consumed_samples=1600.0, train_step_timing in s=5.580, val_
loss=1.720]
```

```
Restoring states from the checkpoint path at /root/verb-workspace/results/Me
ta-llama3.1-8B-Instruct-titlegen/checkpoints/megatron_gpt_peft_lora_tuning--
validation_loss=1.717-step=50-consumed_samples=1600.0.ckpt
Restored all states from the checkpoint at /root/verb-workspace/results/Meta
-llama3.1-8B-Instruct-titlegen/checkpoints/megatron_gpt_peft_lora_tuning--va
lidation_loss=1.717-step=50-consumed_samples=1600.0.ckpt
```

This will create a LoRA adapter - a file named `megatron_gpt_peft_lora_tuning.nemo` in `./results/Meta-Llama-3-8B-Instruct/checkpoints/` . We'll use this later.

To further configure the run above -

- **A different PEFT technique**: The `peft.peft_scheme` parameter determines the technique being used. In this case, we did LoRA, but NeMo Framework supports other techniques as well - such as P-tuning, Adapters, and IA3. For more information, refer to the PEFT support matrix. For example, for P-tuning, simply set

```
model.peft.peft_scheme="ptuning" # instead of "lora"
```

- **Tuning Llama-3.1 70B**: You will need 8xA100 or 8xH100 GPUs. Provide the path to it's .nemo checkpoint (similar to the download and conversion steps earlier), and change the model parallelization settings for Llama-3.1 70B PEFT to distribute across the GPUs. It is also recommended to run the fine-tuning script from a terminal directly instead of Jupyter when using more than 1 GPU.

```
# Change the following settings, and run from a terminal directly
trainer.devices=8
model.tensor_model_parallel_size=8
model.pipeline_model_parallel_size=1
```

You can override many such configurations while running the script. A full set of possible configurations is located in NeMo Framework Github.

Step 3: Inference with NeMo Framework

Running text generation within the framework is also possible with running a Python script. Note that is more for testing and validation, not a full-fledged deployment solution like NVIDIA NIM.

In [10]:

```
# Check that the LORA model file exists
!ls -l ./results/Meta-llama3.1-8B-Instruct-titlegen/checkpoints
```

```
total 307504
-rw-r--r-- 1 root root 146928238 Sep 21 19:53 'megatron_gpt_peft_lora_tuning
--validation_loss=1.717-step=50-consumed_samples=1600.0-last.ckpt'
-rw-r--r-- 1 root root 146928238 Sep 21 19:53 'megatron_gpt_peft_lora_tuning
--validation_loss=1.717-step=50-consumed_samples=1600.0.ckpt'
-rw-r--r-- 1 root root  21012480 Sep 21 19:53  megatron_gpt_peft_lora_tunin
g.nemo
```

In the code snippet below, the following configurations are worth noting -

1. `model.restore_from_path` to the path for the Meta-Llama-3.1-8B-Instruct.nemo file.
2. `model.peft.restore_from_path` to the path for the PEFT checkpoint that was created in the fine-tuning run in the last step.
3. `model.test_ds.file_names` to the path of the preprocessed test file.

If you have made any changes in model or experiment paths, please ensure they are configured correctly below.

In [11]:

```
# Create a smaller test subset for a quick eval demonstration.
!head -n 128 ./curated-data/law-qa-test_preprocessed.jsonl > ./curated-data/
```

In [12]:

```
%%bash
MODEL="./llama-3_1-8b-instruct-nemo_v1.0/llama3_1_8b_instruct.nemo"

TEST_DS="[./curated-data/law-qa-test_preprocessed-n128.jsonl]" # Smaller tes
# TEST_DS="[./curated-data/law-qa-test_preprocessed.jsonl]" # Full test set
TEST_NAMES="[law]"

TP_SIZE=1
PP_SIZE=1

# This is where your LoRA checkpoint was saved
PATH_TO_TRAINED_MODEL="./results/Meta-llama3.1-8B-Instruct-titlegen/checkpoi

# The generation run will save the generated outputs over the test dataset i
OUTPUT_PREFIX="law_titlegen_lora"

python /opt/NeMo/examples/nlp/language_modeling/tuning/megatron_gpt_generate
    model.restore_from_path=${MODEL} \
    model.peft.restore_from_path=${PATH_TO_TRAINED_MODEL} \
    trainer.devices=1 \
    trainer.num_nodes=1 \
    model.data.test_ds.file_names=${TEST_DS} \
    model.data.test_ds.names=${TEST_NAMES} \
    model.data.test_ds.global_batch_size=32 \
    model.data.test_ds.micro_batch_size=1 \
    model.data.test_ds.tokens_to_generate=50 \
    model.tensor_model_parallel_size=${TP_SIZE} \
    model.pipeline_model_parallel_size=${PP_SIZE} \
    inference.greedy=True  \
    model.data.test_ds.output_file_path_prefix=${OUTPUT_PREFIX} \
    model.data.test_ds.write_predictions_to_file=True \
    model.data.test_ds.add_bos=False \
    model.data.test_ds.add_eos=True \
    model.data.test_ds.add_sep=False \
    model.data.test_ds.label_key="output" \
    model.data.test_ds.prompt_template="\{input\}\ \{output\}"
```

```
[NeMo W 2024-09-21 19:54:15 nemo_logging:349] /usr/local/lib/python3.10/dist
-packages/hydra/_internal/hydra.py:119: UserWarning: Future Hydra versions w
ill no longer change working directory at job runtime by default.
    See https://hydra.cc/docs/1.2/upgrades/1.1_to_1.2/changes_to_job_working
_dir/ for more information.
      ret = run_job(
[NeMo I 2024-09-21 19:54:15 megatron_gpt_generate:125]

    ************** Experiment configuration ***********
[NeMo I 2024-09-21 19:54:15 megatron_gpt_generate:126]
    name: megatron_gpt_peft_${model.peft.peft_scheme}_tuning
    trainer:
      devices: 1
      accelerator: gpu
      num_nodes: 1
      precision: 16
      logger: false
      enable_checkpointing: false
      use_distributed_sampler: false
      max_epochs: 9999
      max_steps: 20000
      log_every_n_steps: 10
      val_check_interval: 200
      gradient_clip_val: 1.0
    exp_manager:
      explicit_log_dir: null
      exp_dir: null
      name: ${name}
      create_wandb_logger: false
      wandb_logger_kwargs:
        project: null
        name: null
      resume_if_exists: true
      resume_ignore_no_checkpoint: true
      create_checkpoint_callback: true
      checkpoint_callback_params:
        monitor: validation_${model.data.test_ds.metric.name}
        save_top_k: 1
        mode: max
        save_nemo_on_train_end: true
        filename: ${name}--{${exp_manager.checkpoint_callback_params.monito
r}:.3f}-{step}-{consumed_samples}
        model_parallel_size: ${model.tensor_model_parallel_size}
        always_save_nemo: true
        save_best_model: false
    model:
      seed: 1234
      tensor_model_parallel_size: 1
      pipeline_model_parallel_size: 1
      global_batch_size: 1
      micro_batch_size: 1
      restore_from_path: ./llama-3_1-8b-instruct-nemo_v1.0/llama3_1_8b_instr
uct.nemo
      resume_from_checkpoint: null
      save_nemo_on_validation_end: true
      sync_batch_comm: false
      megatron_amp_O2: false
      sequence_parallel: false
      activations_checkpoint_granularity: null
      activations_checkpoint_method: null
      activations_checkpoint_num_layers: null
      activations_checkpoint_layers_per_pipeline: null
      answer_only_loss: true
      gradient_as_bucket_view: false
      hidden_dropout: 0.0
      attention_dropout: 0.0
      ffn_dropout: 0.0
      peft:
        peft_scheme: adapter
        restore_from_path: ./results/Meta-llama3.1-8B-Instruct-titlegen/chec
kpoints/megatron_gpt_peft_lora_tuning.nemo
        restore_from_ckpt:
          checkpoint_dir: null
          checkpoint_name: null
        adapter_tuning:
          type: parallel_adapter
          adapter_dim: 32
          adapter_dropout: 0.0
          norm_position: pre
          column_init_method: xavier
          row_init_method: zero
          norm_type: mixedfusedlayernorm
          layer_selection: null
          weight_tying: false
          position_embedding_strategy: null
        lora_tuning:
          variant: nemo
          target_modules:
          - attention_qkv
          adapter_dim: 32
          adapter_dropout: 0.0
          column_init_method: xavier
          row_init_method: zero
          layer_selection: null
          weight_tying: false
          position_embedding_strategy: null
        p_tuning:
          virtual_tokens: 10
          bottleneck_dim: 1024
          embedding_dim: 1024
          init_std: 0.023
        ia3_tuning:
          layer_selection: null
      data:
        test_ds:
          file_names:
          - ./curated-data/law-qa-test_preprocessed-n128.jsonl
          names:
          - law
          global_batch_size: 32
          micro_batch_size: 1
          shuffle: false
          num_workers: 0
          pin_memory: true
          max_seq_length: 2048
          min_seq_length: 1
          drop_last: false
          context_key: input
          label_key: output
          add_eos: true
          add_sep: false
          add_bos: false
          write_predictions_to_file: true
          output_file_path_prefix: law_titlegen_lora
          truncation_field: ${data.train_ds.truncation_field}
          index_mapping_dir: null
          prompt_template: '{input} {output}'
          tokens_to_generate: 50
          truncation_method: right
          metric:
            name: loss
            average: null
            num_classes: null
    inference:
      greedy: true
      top_k: 0
      top_p: 0.9
      temperature: 1.0
      all_probs: false
      repetition_penalty: 1.0
      min_tokens_to_generate: 0
      compute_logprob: false
      outfile_path: output.txt
      compute_attention_mask: true
    server: false
    port: 5555
    web_server: false
    share: true
    username: test
    password: test2
    web_port: 9889
    chat: false
    chatbot_config:
      value: false
      attributes:
      - name: Quality
        min: 0
        max: 4
        key: quality
        type: int
        default: 4
      - name: Toxicity
        min: 0
        max: 4
        key: toxcity
        type: int
        default: 0
      - name: Humor
        min: 0
        max: 4
        key: humor
        type: int
        default: 0
      - name: Creativity
```

```
      min: 0
      max: 4
      key: creativity
      type: int
      default: 0
    - name: Violence
      min: 0
      max: 4
      key: violence
      type: int
      default: 0
    - name: Helpfulness
      min: 0
      max: 4
      key: helpfulness
      type: int
      default: 4
    - name: Not_Appropriate
      min: 0
      max: 4
      key: not_appropriate
      type: int
      default: 0
    - name: Language
      choices:
      - ar
      - bg
      - bn
      - ca
      - cs
      - da
      - de
      - el
      - en
      - eo
      - es
      - eu
      - fa
      - fi
      - fr
      - gl
      - he
      - hu
      - id
      - it
      - ja
      - ko
      - nb
      - nl
      - pl
      - pt
      - ro
      - ru
      - sk
      - sv
      - th
```

```
        - tr
        - uk
        - vi
        - zh
        key: lang
        type: list
        default: en
      user: User
      assistant: Assistant
      system: 'A chat between a curious human and an artificial intelligence
assistant.
        The assistant gives helpful, detailed, and polite answers to the hum
an''s questions.


        '
```

```
[NeMo W 2024-09-21 19:54:15 nemo_logging:349] /usr/local/lib/python3.10/dist
-packages/pytorch_lightning/_graveyard/precision.py:49: The `MixedPrecisionP
lugin` is deprecated. Use `pytorch_lightning.plugins.precision.MixedPrecisio
n` instead.

GPU available: True (cuda), used: True
TPU available: False, using: 0 TPU cores
HPU available: False, using: 0 HPUs
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: context_parallel_size in its cfg. Add t
his key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: expert_model_parallel_size in its cfg.
Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: moe_extended_tp in its cfg. Add this ke
y to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: finalize_model_grads_func in its cfg. A
dd this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: use_te_rng_tracker in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_bulk_wgrad in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_bulk_dgrad in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_overlap_ag in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_overlap_rs in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_overlap_rs_dgrad in its cfg. Ad
d this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_split_ag in its cfg. Add this k
ey to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_atomic_ag in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_split_rs in its cfg. Add this k
ey to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_atomic_rs in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: defer_embedding_wgrad_compute in its cf
g. Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: pipeline_model_parallel_split_rank in i
ts cfg. Add this key to cfg or config_mapping to make to make it configurabl
```

```
e.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: cpu_offloading in its cfg. Add this key
to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: cpu_offloading_num_layers in its cfg. A
dd this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: _cpu_offloading_context in its cfg. Add
this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: cpu_offloading_activations in its cfg.
Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: cpu_offloading_weights in its cfg. Add
this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: barrier_with_L1_time in its cfg. Add th
is key to cfg or config_mapping to make to make it configurable.
[W init.cpp:767] Warning: nvfuser is no longer supported in torch script, us
e _jit_set_nvfuser_enabled is deprecated and a no-op (function operator())
```

```
[NeMo I 2024-09-21 19:54:31 megatron_init:263] Rank 0 has data parallel grou
p : [0]
[NeMo I 2024-09-21 19:54:31 megatron_init:269] Rank 0 has combined group of
data parallel and context parallel : [0]
[NeMo I 2024-09-21 19:54:31 megatron_init:274] All data parallel group ranks
with context parallel combined: [[0]]
[NeMo I 2024-09-21 19:54:31 megatron_init:277] Ranks 0 has data parallel ran
k: 0
[NeMo I 2024-09-21 19:54:31 megatron_init:285] Rank 0 has context parallel g
roup: [0]
[NeMo I 2024-09-21 19:54:31 megatron_init:288] All context parallel group ra
nks: [[0]]
[NeMo I 2024-09-21 19:54:31 megatron_init:289] Ranks 0 has context parallel
rank: 0
[NeMo I 2024-09-21 19:54:31 megatron_init:296] Rank 0 has model parallel gro
up: [0]
[NeMo I 2024-09-21 19:54:31 megatron_init:297] All model parallel group rank
s: [[0]]
[NeMo I 2024-09-21 19:54:31 megatron_init:306] Rank 0 has tensor model paral
lel group: [0]
[NeMo I 2024-09-21 19:54:31 megatron_init:310] All tensor model parallel gro
up ranks: [[0]]
[NeMo I 2024-09-21 19:54:31 megatron_init:311] Rank 0 has tensor model paral
lel rank: 0
[NeMo I 2024-09-21 19:54:31 megatron_init:331] Rank 0 has pipeline model par
allel group: [0]
[NeMo I 2024-09-21 19:54:31 megatron_init:343] Rank 0 has embedding group:
[0]
[NeMo I 2024-09-21 19:54:31 megatron_init:349] All pipeline model parallel g
roup ranks: [[0]]
[NeMo I 2024-09-21 19:54:31 megatron_init:350] Rank 0 has pipeline model par
allel rank 0
[NeMo I 2024-09-21 19:54:31 megatron_init:351] All embedding group ranks:
[[0]]
[NeMo I 2024-09-21 19:54:31 megatron_init:352] Rank 0 has embedding rank: 0
[NeMo I 2024-09-21 19:54:31 tokenizer_utils:178] Getting HuggingFace AutoTok
enizer with pretrained_model_name: meta-llama/Meta-Llama-3-8B
```

```
24-09-21 19:54:31 - PID:51172 - rank:(0, 0, 0, 0) - microbatches.py:39 - INF
O - setting number of micro-batches to constant 1
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: context_parallel_size in its cfg. Add t
his key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: expert_model_parallel_size in its cfg.
Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: moe_extended_tp in its cfg. Add this ke
y to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: finalize_model_grads_func in its cfg. A
dd this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: use_te_rng_tracker in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_bulk_wgrad in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_bulk_dgrad in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_overlap_ag in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_overlap_rs in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_overlap_rs_dgrad in its cfg. Ad
d this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_split_ag in its cfg. Add this k
ey to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_atomic_ag in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_split_rs in its cfg. Add this k
ey to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_atomic_rs in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: defer_embedding_wgrad_compute in its cf
g. Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: pipeline_model_parallel_split_rank in i
ts cfg. Add this key to cfg or config_mapping to make to make it configurabl
e.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: cpu_offloading in its cfg. Add this key
to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: cpu_offloading_num_layers in its cfg. A
```

```
dd this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: _cpu_offloading_context in its cfg. Add
this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: cpu_offloading_activations in its cfg.
Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: cpu_offloading_weights in its cfg. Add
this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: barrier_with_L1_time in its cfg. Add th
is key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:31 nemo_logging:349] /usr/local/lib/python3.10/dist
-packages/huggingface_hub/file_download.py:1132: FutureWarning: `resume_down
load` is deprecated and will be removed in version 1.0.0. Downloads always r
esume when possible. If you want to force a new download, use `force_downloa
d=True`.
      warnings.warn(

Special tokens have been added in the vocabulary, make sure the associated w
ord embeddings are fine-tuned or trained.
[NeMo I 2024-09-21 19:54:32 megatron_base_model:584] Padded vocab_size: 1282
56, original vocab_size: 128256, dummy tokens: 0.
```

```
[NeMo W 2024-09-21 19:54:32 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: context_parallel_size in its cfg. Add t
his key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: expert_model_parallel_size in its cfg.
Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: moe_extended_tp in its cfg. Add this ke
y to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: finalize_model_grads_func in its cfg. A
dd this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: use_te_rng_tracker in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_bulk_wgrad in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_bulk_dgrad in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_overlap_ag in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_overlap_rs in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_overlap_rs_dgrad in its cfg. Ad
d this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_split_ag in its cfg. Add this k
ey to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_atomic_ag in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_split_rs in its cfg. Add this k
ey to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: tp_comm_atomic_rs in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: defer_embedding_wgrad_compute in its cf
g. Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: pipeline_model_parallel_split_rank in i
ts cfg. Add this key to cfg or config_mapping to make to make it configurabl
e.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: cpu_offloading in its cfg. Add this key
to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: cpu_offloading_num_layers in its cfg. A
dd this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:1158] The model: MegatronGPT
```

```
SFTModel() does not have field.name: _cpu_offloading_context in its cfg. Add
this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: cpu_offloading_activations in its cfg.
Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: cpu_offloading_weights in its cfg. Add
this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:1158] The model: MegatronGPT
SFTModel() does not have field.name: barrier_with_L1_time in its cfg. Add th
is key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: activation_func_fp8_input_store in its c
fg. Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: num_moe_experts in its cfg. Add this key
to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: window_size in its cfg. Add this key to
cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: qk_layernorm in its cfg. Add this key to
cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: test_mode in its cfg. Add this key to cf
g or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: calculate_per_token_loss in its cfg. Add
this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: memory_efficient_layer_norm in its cfg.
Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: fp8_wgrad in its cfg. Add this key to cf
g or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: fp8_dot_product_attention in its cfg. Ad
d this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: fp8_multi_head_attention in its cfg. Add
this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: moe_router_load_balancing_type in its cf
g. Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: moe_router_topk in its cfg. Add this key
to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: moe_grouped_gemm in its cfg. Add this ke
y to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: moe_aux_loss_coeff in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: moe_z_loss_coeff in its cfg. Add this ke
y to cfg or config_mapping to make to make it configurable.
```

```
[NeMo W 2024-09-21 19:54:32 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: moe_input_jitter_eps in its cfg. Add thi
s key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: moe_token_dropping in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: moe_token_dispatcher_type in its cfg. Ad
d this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: moe_per_layer_logging in its cfg. Add th
is key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: moe_expert_capacity_factor in its cfg. A
dd this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: moe_pad_expert_input_to_capacity in its
cfg. Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: moe_token_drop_policy in its cfg. Add th
is key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: moe_layer_recompute in its cfg. Add this
key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: clone_scatter_output_in_embedding in its
cfg. Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: disable_parameter_transpose_cache in its
cfg. Add this key to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: enable_cuda_graph in its cfg. Add this k
ey to cfg or config_mapping to make to make it configurable.
[NeMo W 2024-09-21 19:54:32 megatron_base_model:556] The model: MegatronGPTS
FTModel() does not have field.name: rotary_percent in its cfg. Add this key
to cfg or config_mapping to make to make it configurable.
Initializing distributed: GLOBAL_RANK: 0, MEMBER: 1/1
----------------------------------------------------------------------------
------------------------
distributed_backend=nccl
All distributed processes registered. Starting with 1 processes
----------------------------------------------------------------------------
------------------------
```

```
[NeMo I 2024-09-21 19:54:51 dist_ckpt_io:95] Using ('zarr', 1) dist-ckpt sav
e strategy.
Loading distributed checkpoint with TensorStoreLoadShardedStrategy
Loading distributed checkpoint directly on the GPU
[NeMo I 2024-09-21 19:55:46 nlp_overrides:1180] Model MegatronGPTSFTModel wa
s successfully restored from /root/verb-workspace/llama-3_1-8b-instruct-nemo
_v1.0/llama3_1_8b_instruct.nemo.
[NeMo I 2024-09-21 19:55:46 nlp_adapter_mixins:203] Before adding PEFT param
s:
      | Name  | Type     | Params | Mode
    -------------------------------------------
    0 | model | GPTModel | 8.0 B  | train
    -------------------------------------------
    0         Trainable params
    8.0 B     Non-trainable params
    8.0 B     Total params
    32,121.045Total estimated model params size (MB)
[NeMo I 2024-09-21 19:55:50 nlp_adapter_mixins:208] After adding PEFT param
s:
      | Name  | Type     | Params | Mode
    -------------------------------------------
    0 | model | GPTModel | 8.0 B  | train
    -------------------------------------------
    10.5 M    Trainable params
    8.0 B     Non-trainable params
    8.0 B     Total params
    32,162.988Total estimated model params size (MB)
[NeMo I 2024-09-21 19:55:50 megatron_gpt_generate:156] Freezing parameters f
or PEFT eval:
      | Name  | Type     | Params | Mode
    ------------------------------------------
    0 | model | GPTModel | 8.0 B  | eval
    ------------------------------------------
    0         Trainable params
    8.0 B     Non-trainable params
    8.0 B     Total params
    32,162.988Total estimated model params size (MB)
```

```
[NeMo W 2024-09-21 19:55:50 nemo_logging:349] /usr/local/lib/python3.10/dist
-packages/pytorch_lightning/trainer/configuration_validator.py:161: You have
overridden `MegatronGPTSFTModel.configure_sharded_model` which is deprecate
d. Please override the `configure_model` hook instead. Instantiation with th
e newer hook will be created on the device right away and have the right dat
a type depending on the precision setting in the Trainer.

[NeMo W 2024-09-21 19:55:50 nemo_logging:349] /usr/local/lib/python3.10/dist
-packages/pytorch_lightning/trainer/configuration_validator.py:143: You are
using the `dataloader_iter` step flavor. If you consume the iterator more th
an once per step, the `batch_idx` argument in any hook that takes it will no
t match with the batch index of the last batch consumed. This might have unf
oreseen effects on callbacks or code that expects to get the correct index.
This will also not work well with gradient accumulation. This feature is ver
y experimental and subject to change. Here be dragons.
```

```
[NeMo I 2024-09-21 19:55:50 megatron_gpt_sft_model:803] Building GPT SFT tes
t datasets.
[NeMo I 2024-09-21 19:55:50 text_memmap_dataset:116] Building data files
[NeMo I 2024-09-21 19:55:50 text_memmap_dataset:525] Processing 1 data files
using 6 workers
```

```
huggingface/tokenizers: The current process just got forked, after paralleli
sm has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(tr
ue | false)
huggingface/tokenizers: The current process just got forked, after paralleli
sm has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(tr
ue | false)
huggingface/tokenizers: The current process just got forked, after paralleli
sm has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(tr
ue | false)
huggingface/tokenizers: The current process just got forked, after paralleli
sm has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(tr
ue | false)
huggingface/tokenizers: The current process just got forked, after paralleli
sm has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(tr
ue | false)
huggingface/tokenizers: The current process just got forked, after paralleli
sm has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(tr
ue | false)
```

```
[NeMo I 2024-09-21 19:55:50 text_memmap_dataset:495] Building indexing for f
n = ./curated-data/law-qa-test_preprocessed-n128.jsonl
[NeMo I 2024-09-21 19:55:50 text_memmap_dataset:507] Saving idx file = ./cur
ated-data/law-qa-test_preprocessed-n128.jsonl.idx.npy
[NeMo I 2024-09-21 19:55:50 text_memmap_dataset:509] Saving metadata file =
./curated-data/law-qa-test_preprocessed-n128.jsonl.idx.info
[NeMo I 2024-09-21 19:55:50 text_memmap_dataset:535] Time building 1 / 1 mem
-mapped files: 0:00:00.181687
[NeMo I 2024-09-21 19:55:50 text_memmap_dataset:525] Processing 1 data files
using 6 workers
```

```
huggingface/tokenizers: The current process just got forked, after paralleli
sm has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(tr
ue | false)
huggingface/tokenizers: The current process just got forked, after paralleli
sm has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(tr
ue | false)
huggingface/tokenizers: The current process just got forked, after paralleli
sm has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(tr
ue | false)
huggingface/tokenizers: The current process just got forked, after paralleli
sm has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(tr
ue | false)
huggingface/tokenizers: The current process just got forked, after paralleli
sm has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(tr
ue | false)
huggingface/tokenizers: The current process just got forked, after paralleli
sm has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(tr
ue | false)
```

```
[NeMo I 2024-09-21 19:55:50 text_memmap_dataset:535] Time building 0 / 1 mem
-mapped files: 0:00:00.162028
[NeMo I 2024-09-21 19:55:50 text_memmap_dataset:158] Loading data files
[NeMo I 2024-09-21 19:55:50 text_memmap_dataset:249] Loading ./curated-data/
law-qa-test_preprocessed-n128.jsonl
[NeMo I 2024-09-21 19:55:50 text_memmap_dataset:161] Time loading 1 mem-mapp
ed files: 0:00:00.001437
[NeMo I 2024-09-21 19:55:50 text_memmap_dataset:165] Computing global indice
s
[NeMo I 2024-09-21 19:55:50 megatron_gpt_sft_model:806] Length of test datas
et: 128
[NeMo I 2024-09-21 19:55:50 megatron_gpt_sft_model:829] Building dataloader
with consumed samples: 0
```

```
LOCAL_RANK: 0 - CUDA_VISIBLE_DEVICES: [0]
[NeMo W 2024-09-21 19:55:50 nemo_logging:349] /usr/local/lib/python3.10/dist
-packages/pytorch_lightning/trainer/connectors/data_connector.py:424: The 't
est_dataloader' does not have many workers which may be a bottleneck. Consid
er increasing the value of the `num_workers` argument` to `num_workers=11` i
n the `DataLoader` to improve performance.

[NeMo W 2024-09-21 19:55:50 nemo_logging:349] /usr/local/lib/python3.10/dist
-packages/pytorch_lightning/loops/utilities.py:149: Found `dataloader_iter`
argument in the `test_step`. Note that the support for this signature is exp
erimental and the behavior is subject to change.

[NeMo W 2024-09-21 19:55:50 nemo_logging:349] /opt/apex/apex/transformer/pip
eline_parallel/utils.py:81: UserWarning: This function is only for unittest
      warnings.warn("This function is only for unittest")

[NeMo W 2024-09-21 19:55:55 nemo_logging:349] /opt/NeMo/nemo/collections/nl
p/modules/common/text_generation_utils.py:395: UserWarning: The torch.cuda.*
DtypeTensor constructors are no longer recommended. It's best to use methods
such as torch.tensor(data, dtype=*, device='cuda') to create tensors. (Trigg
ered internally at /opt/pytorch/pytorch/torch/csrc/tensor/python_tensor.cpp:
83.)
      input_info_tensor = torch.cuda.FloatTensor(input_info)

[NeMo W 2024-09-21 19:55:55 nemo_logging:349] /opt/NeMo/nemo/collections/nl
p/modules/common/text_generation_utils.py:403: UserWarning: The given NumPy
array is not writable, and PyTorch does not support non-writable tensors. Th
is means writing to this tensor will result in undefined behavior. You may w
ant to copy the array to protect its data or make it writable before convert
ing it to a tensor. This type of warning will be suppressed for the rest of
this program. (Triggered internally at /opt/pytorch/pytorch/torch/csrc/util
s/tensor_numpy.cpp:206.)
      string_tensor = torch.as_tensor(
```

```
Testing DataLoader 0: 100%|██████████| 4/4 [05:52<00:00,  0.01it/s][NeMo I 2
024-09-21 20:01:42 megatron_gpt_sft_model:561] Total deduplicated inference
data size: 128 to 128
[NeMo I 2024-09-21 20:01:42 megatron_gpt_sft_model:712] Predictions saved to
law_titlegen_lora_test_law_inputs_preds_labels.jsonl
```

```
[NeMo W 2024-09-21 20:01:42 megatron_gpt_sft_model:652] No training data fou
nd, reconfiguring microbatches based on validation batch sizes.
[NeMo W 2024-09-21 20:01:42 nemo_logging:349] /usr/local/lib/python3.10/dist
-packages/pytorch_lightning/trainer/connectors/logger_connector/result.py:43
9: It is recommended to use `self.log('val_loss', ..., sync_dist=True)` when
logging on epoch level in distributed setting to accumulate the metric acros
s devices.

[NeMo W 2024-09-21 20:01:42 nemo_logging:349] /usr/local/lib/python3.10/dist
-packages/pytorch_lightning/trainer/connectors/logger_connector/result.py:43
9: It is recommended to use `self.log('test_loss_law', ..., sync_dist=True)`
when logging on epoch level in distributed setting to accumulate the metric
across devices.

[NeMo W 2024-09-21 20:01:42 nemo_logging:349] /usr/local/lib/python3.10/dist
-packages/pytorch_lightning/trainer/connectors/logger_connector/result.py:43
9: It is recommended to use `self.log('test_loss', ..., sync_dist=True)` whe
n logging on epoch level in distributed setting to accumulate the metric acr
oss devices.
```

```
Testing DataLoader 0: 100%|██████████| 4/4 [05:52<00:00,  0.01it/s]
┏━━━━━━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━━━━┓
┃        Test metric        ┃       DataLoader 0        ┃
┡━━━━━━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━━━━┩
│         test_loss         │    1.6104145050048828     │
│       test_loss_law       │    1.6104145050048828     │
│         val_loss          │    1.6104145050048828     │
└───────────────────────────┴───────────────────────────┘
```

Step 4: Check the model accuracy

Now that the results are in, let's read the results and calculate the accuracy on the question title generation task. Let's take a look at one of the predictions in the generated output file. The pred key indicates what was generated.

In [13]:
```
# Take a look at predictions
!head -n1  law_titlegen_lora_test_law_inputs_preds_labels.jsonl
```

```
{"input": "Generate a concise, engaging title for the following legal questi
on on an internet forum. The title should be legally relevant, capture key a
spects of the issue, and entice readers to learn more. \nQUESTION: In order
to be sued in a particular jurisdiction, say New York, a company must have a
minimal business presence in the jurisdiction. What constitutes such a prese
nce? Suppose the company engaged a New York-based Plaintiff, and its represe
ntatives signed the contract with the Plaintiff in New York City. Does this
satisfy the minimum presence rule? Suppose, instead, the plaintiff and contr
act signing were in New Jersey, but the company hired a law firm with office
s in New York City. Does this qualify? \nTITLE:", "pred": " What constitutes
a minimal business presence in a jurisdiction?", "label": " What constitutes
\"doing business in a jurisdiction?\""}
```

For evaluating this task, we will use ROUGE. It measures overlap of ngrams, and a higher score is better. While it's not perfect and it misses capturing the semantics of the prediction, it is a popular metric in academia and industry for evaluating such systems.

The following method uses the rouge_score library to implement scoring. It will report `ROUGE_{1/2/L/Lsum}` metrics.

In [14]:
```
def compute_rouge(input_file: str) -> dict:
    ROUGE_KEYS = ["rouge1", "rouge2", "rougeL", "rougeLsum"]
    scorer = rouge_scorer.RougeScorer(ROUGE_KEYS, use_stemmer=True)
    aggregator = scoring.BootstrapAggregator()
    lines = [json.loads(line) for line in open(input_file)]
    num_response_words = []
    num_ref_words = []
    for idx, line in enumerate(lines):
        prompt = line['input']
        response = line['pred']
        answer = line['label']
        scores = scorer.score(response, answer)
        aggregator.add_scores(scores)
        num_response_words.append(len(response.split()))
        num_ref_words.append(len(answer.split()))

    result = aggregator.aggregate()
    rouge_scores = {k: round(v.mid.fmeasure * 100, 4) for k, v in result.ite
    print(rouge_scores)
    print(f"Average and stddev of response length: {np.mean(num_response_wor
    print(f"Average and stddev of ref length: {np.mean(num_ref_words):.2f},

    return rouge_scores
```

In [15]:
```
compute_rouge("./law_titlegen_lora_test_law_inputs_preds_labels.jsonl")
```

```
{'rouge1': 40.0619, 'rouge2': 20.3573, 'rougeL': 36.1957, 'rougeLsum': 36.19
38}
Average and stddev of response length: 11.70, 4.55
Average and stddev of ref length: 11.26, 4.97
```

Out[15]:
```
{'rouge1': 40.0619, 'rouge2': 20.3573, 'rougeL': 36.1957, 'rougeLsum': 36.1
938}
```

For the Llama-3.1-8B-Instruct model, you should see accuracy comparable to the below:

`{'rouge1': 39.2082, 'rouge2': 18.8573, 'rougeL': 35.4098, 'rougeLsum': 35.3906}`

LoRA inference with NVIDIA NIM

Now that we've trained our LoRA, lets go ahead and deploy them with NVIDIA NIM. NIM's let you deploy multiple LoRA adapters and supports the .nemo and Hugging Face model formats. We will deploy the Law LoRA adapter.

Before you begin

Lets download the NIM from NGC and get it up and running with the LoRa's that we've trained.

Note this cell might take a few minutes as it pulls the NIM

In [16]:

```
%%bash

wget https://raw.githubusercontent.com/brevdev/notebooks/main/assets/setup-r
chmod +x setup-nim
export NGC_API_KEY=
./setup-nim
```

```
--2024-09-21 20:02:12--  https://raw.githubusercontent.com/brevdev/notebook
s/main/assets/setup-nim.sh
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.1
11.133, 185.199.109.133, 185.199.110.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.
111.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 1713 (1.7K) [text/plain]
Saving to: ‘setup-nim’

     0K .                                                     100% 23.2M=0s

2024-09-21 20:02:12 (23.2 MB/s) - ‘setup-nim’ saved [1713/1713]

./setup-nim: line 5: docker: command not found
```

```
---------------------------------------------------------------------------
CalledProcessError                        Traceback (most recent call last)
Cell In[16], line 1
----> 1 get_ipython().run_cell_magic('bash', '', '\nwget https://raw.githubusercontent.com/brevdev/notebooks/main/assets/setup-nim.sh -O setup-nim\nchmod +x setup-nim\nexport NGC_API_KEY=\n./setup-nim\n')

File /usr/local/lib/python3.10/dist-packages/IPython/core/interactiveshell.py:2517, in InteractiveShell.run_cell_magic(self, magic_name, line, cell)
   2515 with self.builtin_trap:
   2516     args = (magic_arg_s, cell)
-> 2517     result = fn(*args, **kwargs)
   2519 # The code below prevents the output from being displayed
   2520 # when using magics with decorator @output_can_be_silenced
   2521 # when the last Python token in the expression is a ';'.
   2522 if getattr(fn, magic.MAGIC_OUTPUT_CAN_BE_SILENCED, False):

File /usr/local/lib/python3.10/dist-packages/IPython/core/magics/script.py:154, in ScriptMagics._make_script_magic.<locals>.named_script_magic(line, cell)
    152 else:
    153     line = script
--> 154 return self.shebang(line, cell)

File /usr/local/lib/python3.10/dist-packages/IPython/core/magics/script.py:314, in ScriptMagics.shebang(self, line, cell)
    309 if args.raise_error and p.returncode != 0:
    310     # If we get here and p.returncode is still None, we must have
    311     # killed it but not yet seen its return code. We don't wait for
it,
    312     # in case it's stuck in uninterruptible sleep. -9 = SIGKILL
    313     rc = p.returncode or -9
--> 314     raise CalledProcessError(rc, cell)

CalledProcessError: Command 'b'\nwget https://raw.githubusercontent.com/brevdev/notebooks/main/assets/setup-nim.sh -O setup-nim\nchmod +x setup-nim\nexport NGC_API_KEY=\n./setup-nim\n'' returned non-zero exit status 127.
```

This notebook includes instructions to send an inference call to NVIDIA NIM using the Python `requests` library.

In [ ]:

```
import requests
import json
```

Check available LoRA models

Once the NIM server is up and running, check the available models as follows:

In [ ]:

```
url = 'http://0.0.0.0:8000/v1/models'

response = requests.get(url)
data = response.json()
```

```
print(json.dumps(data, indent=4))
```

This will return all the models available for inference by NIM. In this case, it will return the base model, as well as the LoRA adapters that were provided during NIM deployment - `llama3.1-8b-law-titlegen`.

Inference

Inference can be performed by sending POST requests to the `/completions` endpoint.

A few things to note:

- The `model` parameter in the payload specifies the model that the request will be directed to. This can be the base model `meta/llama3.1-8b-instruct`, or any of the LoRA models, such as `llama3.1-8b-law-titlegen`.
- `max_tokens` parameter specifies the maximum number of tokens to generate. At any point, the cumulative number of input prompt tokens and specified number of output tokens to generate should not exceed the model's maximum context limit. For llama3-8b-instruct, the context length supported is 8192 tokens.

Following code snippets show how it's possible to send requests belonging to different LoRAs (or tasks). NIM dynamically loads the LoRA adapters and serves the requests. It also internally handles the batching of requests belonging to different LoRAs to allow better performance and more efficient of compute.

Title Generation

Try sending an example from the test set.

In [ ]:
```
url = 'http://0.0.0.0:8000/v1/completions'
headers = {
    'accept': 'application/json',
    'Content-Type': 'application/json'
}

# Example from the test set
prompt="Generate a concise, engaging title for the following legal question
data = {
    "model": "llama3.1-8b-law-titlegen",
    "prompt": prompt,
    "max_tokens": 50
}

response = requests.post(url, headers=headers, json=data)
```

```
response_data = response.json()

print(json.dumps(response_data, indent=4))
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: