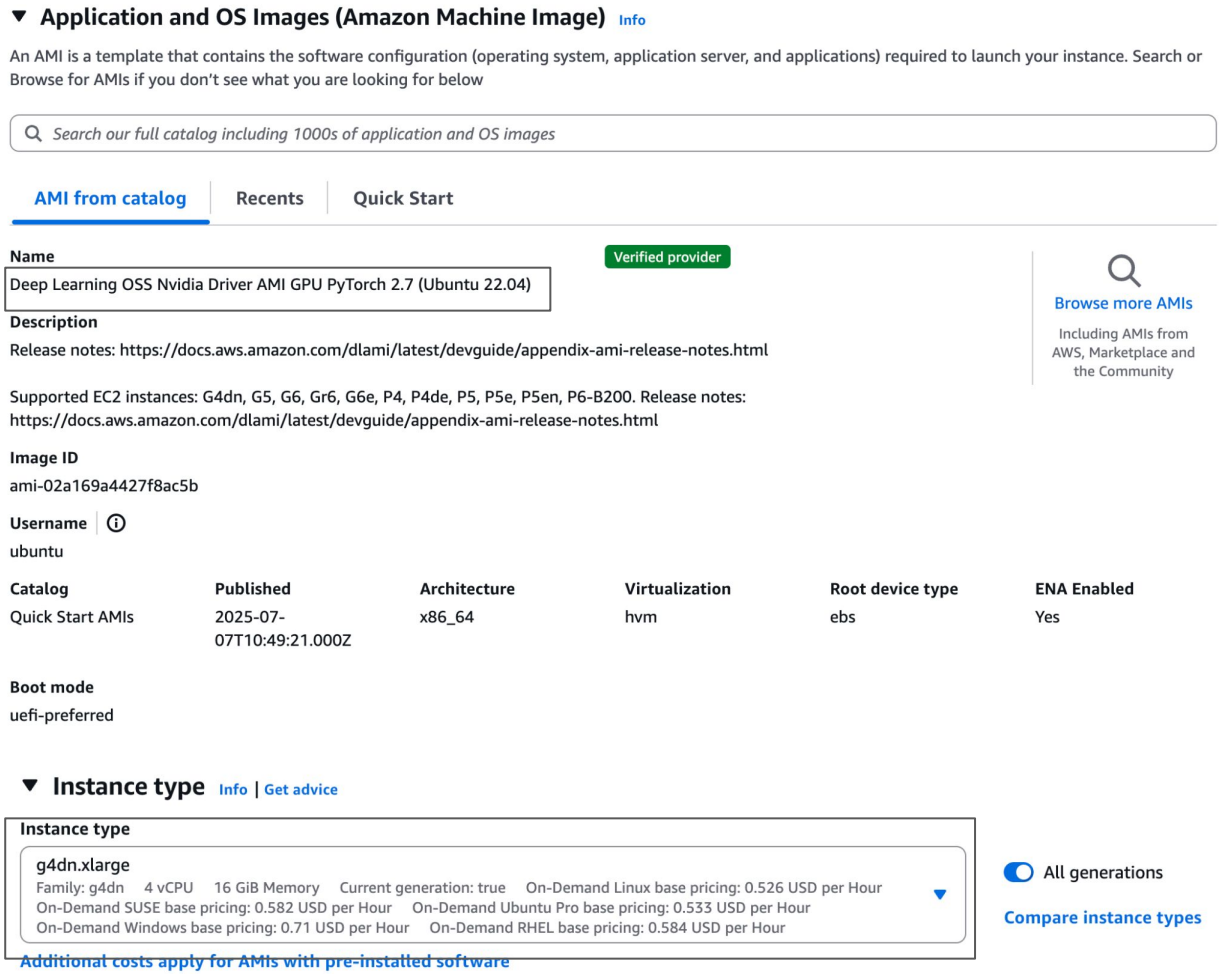
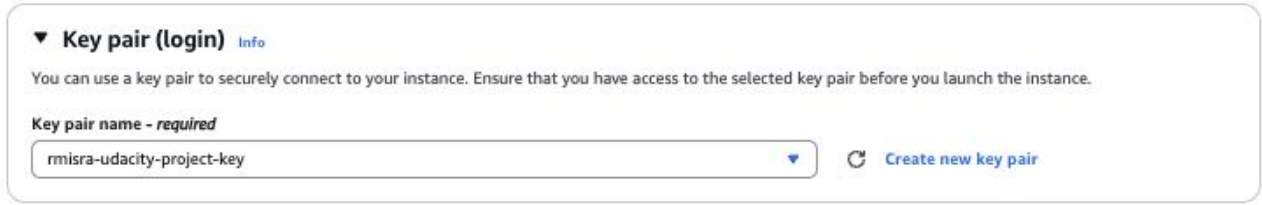
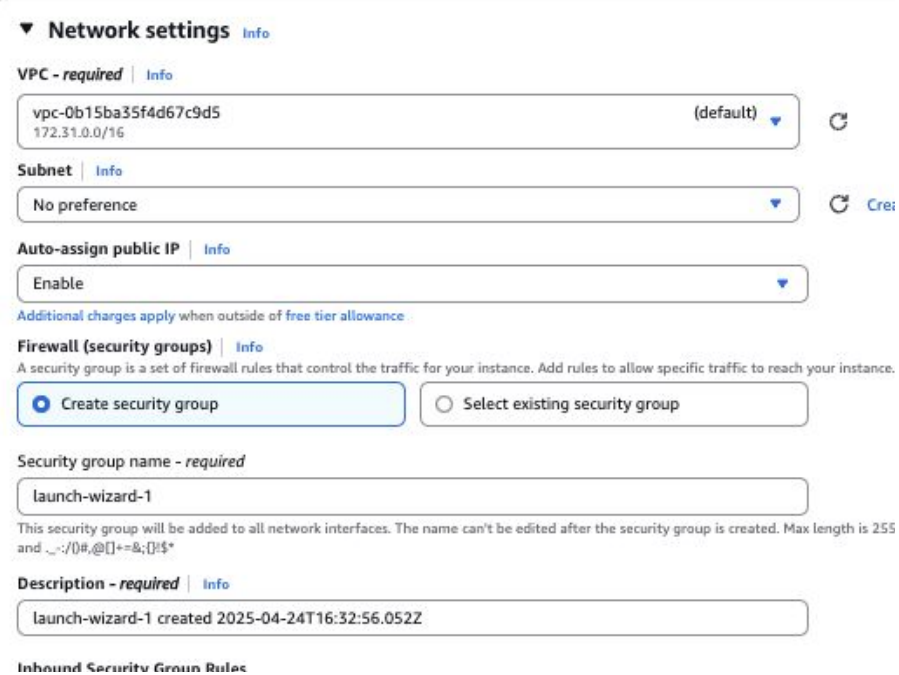
EC2 Instance Creation

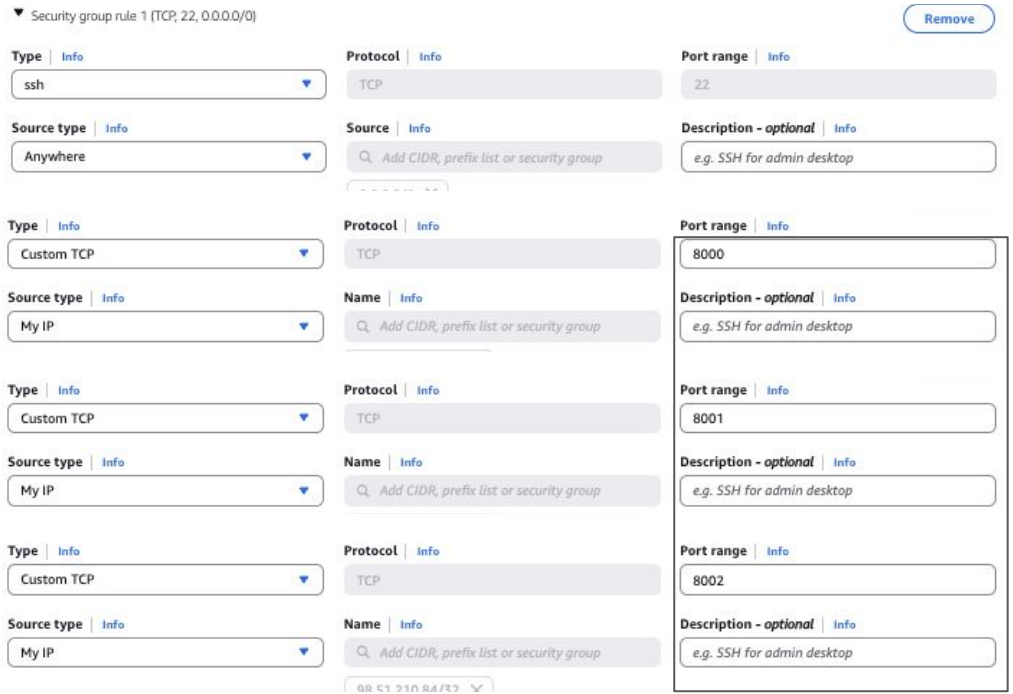
AMI and Instance Type



Access and Network Setup







Post Setup

Instance summary for i-0a534409c73486e94 (UdaciHeadlineEC2) Info

Updated less than a minute ago

Connect | Instance state | Actions

Instance ID
i-0a534409c73486e94

IPv6 address
–

Hostname type
IP name: ip-172-31-70-50.ec2.internal

~~Answer private resource DNS name~~
IPv4 (A)

Auto-assigned IP address
3.235.60.162 [Public IP]

IAM Role
–

IMDSv2
Optional
EC2 recommends setting IMDSv2 to required | Learn more

Operator
–

Public IPv4 address
3.235.60.162 | open address

Instance state
Running

Private IP DNS name (IPv4 only)
ip-172-31-70-50.ec2.internal

Instance type
g4dn.xlarge

VPC ID
vpc-05cf067a5c90c4b0b

Subnet ID
subnet-00299464f44a70cb1

Instance ARN
arn:aws:ec2:us-east-1:710271915376:instance/i-0a534409c73486e94

Private IPv4 addresses
172.31.70.50

Public DNS
ec2-3-235-60-162.compute-1.amazonaws.com | open address

Elastic IP addresses
–

AWS Compute Optimizer finding
Opt-in to AWS Compute Optimizer for recommendations. | Learn more

Auto Scaling Group name
–

Managed
false

Details | Status and alarms | Monitoring | Security | Networking | Storage | Tags

Instance details Info

AMI ID
ami-05ee60afff9d0a480

AMI name
Deep Learning OSS Nvidia Driver AMI GPU PyTorch 2.7 (Ubuntu 22.04) 20250602

Stop protection
Disabled

Instance reboot migration
Default (On)

Stop-hibernate behavior
Disabled

State transition reason

Monitoring
disabled

Allowed image
-

Launch time
Tue Jul 15 2025 07:37:15 GMT-0700 (Pacific Daylight Time) (1 minute)

Instance auto-recovery
Default

AMI Launch index
0

Credit specification

Platform details
Linux/UNIX

Termination protection
Disabled

AMI location
amazon/Deep Learning OSS Nvidia Driver AMI GPU PyTorch 2.7 (Ubuntu 22.04) 20250602

Lifecycle
normal

Key pair assigned at launch
vockey

Kernel ID

Access EC2 Instance

```
rmisra@MBP-RMISRA-7FJJVW3 ~ % ssh -i /Users/[redacted]4-key-pair.pem ubuntu@35.168.112.51
Welcome to Ubuntu 22.04.5 LTS (GNU/Linux 6.8.0-1030-aws x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/pro

 System information as of Sat Jul 12 19:42:45 UTC 2025

  System load:  0.28               Processes:             164
  Usage of /:   63.2% of 145.19GB  Users logged in:       1
  Memory usage: 4%                 IPv4 address for ens5: 172.31.68.64
  Swap usage:   0%

  => There are 2 zombie processes.

 * Ubuntu Pro delivers the most comprehensive open source security and
   compliance features.

   https://ubuntu.com/aws/pro

Expanded Security Maintenance for Applications is not enabled.

15 updates can be applied immediately.
To see these additional updates run: apt list --upgradable

3 additional security updates can be applied with ESM Apps.
Learn more about enabling ESM Apps service at https://ubuntu.com/esm


The list of available updates is more than a week old.
To check for new updates run: sudo apt update
New release '24.04.2 LTS' available.
Run 'do-release-upgrade' to upgrade to it.


3 updates could not be installed automatically. For more details,
see /var/log/unattended-upgrades/unattended-upgrades.log

================================================================================
AMI Name: Deep Learning OSS Nvidia Driver AMI GPU PyTorch 2.7 (Ubuntu 22.04)
Supported EC2 instances: G4dn, G5, G6, Gr6, G6e, P4, P4de, P5, P5e, P5en, P6-B200
* To activate pre-built pytorch environment, run: 'source /opt/pytorch/bin/activate '
NVIDIA driver version: 570.133.20
CUDA versions available: cuda-12.8
Default CUDA version is 12.8

Release notes: https://docs.aws.amazon.com/dlami/latest/devguide/appendix-ami-release-notes.html
AWS Deep Learning AMI Homepage: https://aws.amazon.com/machine-learning/amis/
Developer Guide and Release Notes: https://docs.aws.amazon.com/dlami/latest/devguide/what-is-dlami.html
Support: https://forums.aws.amazon.com/forum.jspa?forumID=263
For a fully managed experience, check out Amazon SageMaker at https://aws.amazon.com/sagemaker
================================================================================
Last login: Sat Jul 12 19:32:36 2025 from 98.51.210.84
ubuntu@ip-172-31-68-64:~$
```

```
ubuntu@ip-172-31-72-27:~$ curl https://pyenv.run | bash
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100   270  100   270    0     0   1953      0 --:--:-- --:--:-- --:--:--  1956

WARNING: Can not proceed with installation. Kindly remove the '/home/ubuntu/.pyenv' directory first.

ubuntu@ip-172-31-72-27:~$ echo 'export PYENV_ROOT="$HOME/.pyenv"' >> ~/.bashrc
ubuntu@ip-172-31-72-27:~$ echo 'export PATH="$PYENV_ROOT/bin:$PATH"' >> ~/.bashrc
ubuntu@ip-172-31-72-27:~$ echo 'eval "$(pyenv init --path)"' >> ~/.bashrc
ubuntu@ip-172-31-72-27:~$ echo 'eval "$(pyenv init -)"' >> ~/.bashrc
ubuntu@ip-172-31-72-27:~$ exec "$SHELL"
ubuntu@ip-172-31-72-27:~$
ubuntu@ip-172-31-72-27:~$ pyenv install 3.10.14
pyenv: /home/ubuntu/.pyenv/versions/3.10.14 already exists
continue with installation? (y/N) ubuntu@ip-172-31-72-27:~$
ubuntu@ip-172-31-72-27:~$
ubuntu@ip-172-31-72-27:~$
ubuntu@ip-172-31-72-27:~$ python --version
Python 3.10.14
ubuntu@ip-172-31-72-27:~$ Python 3.10.14
Command 'Python' not found, did you mean:
  command 'jython' from deb jython (2.7.2+repack1-4)
Try: sudo apt install <deb name>
ubuntu@ip-172-31-72-27:~$
ubuntu@ip-172-31-72-27:~$ python -m venv ~/trt-llm-env
ubuntu@ip-172-31-72-27:~$ source ~/trt-llm-env/bin/activate
(trt-llm-env) ubuntu@ip-172-31-72-27:~$
(trt-llm-env) ubuntu@ip-172-31-72-27:~$ sudo apt-get update
sudo apt-get install git-lfs -y

Hit:1 http://us-east-1.ec2.archive.ubuntu.com/ubuntu jammy InRelease
Get:2 http://security.ubuntu.com/ubuntu jammy-security InRelease [129 kB]
Get:3 http://us-east-1.ec2.archive.ubuntu.com/ubuntu jammy-updates InRelease [128 kB]
Get:4 http://us-east-1.ec2.archive.ubuntu.com/ubuntu jammy-backports InRelease [127 kB]
Hit:5 https://nvidia.github.io/libnvidia-container/stable/deb/amd64  InRelease
Hit:6 https://apt.corretto.aws stable InRelease
Get:7 https://download.docker.com/linux/ubuntu jammy InRelease [48.8 kB]
Get:8 https://fsx-lustre-client-repo.s3.amazonaws.com/ubuntu jammy InRelease [3823 B]
Get:9 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64  InRelease [1581 B]
Get:10 http://us-east-1.ec2.archive.ubuntu.com/ubuntu jammy-updates/main amd64 Packages [2723 kB]
Get:11 http://us-east-1.ec2.archive.ubuntu.com/ubuntu jammy-updates/main Translation-en [434 kB]
Get:12 http://us-east-1.ec2.archive.ubuntu.com/ubuntu jammy-updates/restricted amd64 Packages [3898 kB]
Get:13 http://us-east-1.ec2.archive.ubuntu.com/ubuntu jammy-updates/restricted Translation-en [698 kB]
Get:14 http://us-east-1.ec2.archive.ubuntu.com/ubuntu jammy-updates/universe amd64 Packages [1221 kB]
Get:15 http://us-east-1.ec2.archive.ubuntu.com/ubuntu jammy-updates/universe Translation-en [301 kB]
Get:16 http://us-east-1.ec2.archive.ubuntu.com/ubuntu jammy-updates/multiverse amd64 Packages [47.1 kB]
Get:17 http://us-east-1.ec2.archive.ubuntu.com/ubuntu jammy-updates/multiverse Translation-en [12.0 kB]
Get:18 http://us-east-1.ec2.archive.ubuntu.com/ubuntu jammy-backports/main amd64 Packages [68.8 kB]
Get:19 http://us-east-1.ec2.archive.ubuntu.com/ubuntu jammy-backports/universe amd64 Packages [30.0 kB]
Get:20 http://us-east-1.ec2.archive.ubuntu.com/ubuntu jammy-backports/universe Translation-en [16.5 kB]
```

Python & Virtual Environment Setup

GPT2 and TensorRT-LLM Setup

```
(trt-llm-env) ubuntu@ip-172-31-72-27:~$ git clone https://huggingface.co/gpt2 /home/ubuntu/models_src/gpt2
Cloning into '/home/ubuntu/models_src/gpt2'...
remote: Enumerating objects: 87, done.
remote: Counting objects: 100% (3/3), done.
remote: Compressing objects: 100% (2/2), done.
remote: Total 87 (delta 0), reused 0 (delta 0), pack-reused 84 (from 1)
Unpacking objects: 100% (87/87), 1.65 MiB | 8.50 MiB/s, done.
Filtering content: 100% (11/11), 5.23 GiB | 77.46 MiB/s, done.
(trt-llm-env) ubuntu@ip-172-31-72-27:~$ cd /home/ubuntu
(trt-llm-env) ubuntu@ip-172-31-72-27:~$ git clone https://github.com/NVIDIA/TensorRT-LLM.git
Cloning into 'TensorRT-LLM'...
remote: Enumerating objects: 95583, done.
remote: Counting objects: 100% (227/227), done.
remote: Compressing objects: 100% (146/146), done.
remote: Total 95583 (delta 145), reused 82 (delta 81), pack-reused 95356 (from 2)
Receiving objects: 100% (95583/95583), 1.59 GiB | 46.42 MiB/s, done.
Resolving deltas: 100% (69921/69921), done.
Updating files: 100% (5877/5877), done.
Filtering content: 100% (2377/2377), 1.79 GiB | 60.28 MiB/s, done.
(trt-llm-env) ubuntu@ip-172-31-72-27:~$ sudo apt-get -y install libopenmpi-dev && pip3 install --upgrade pip setuptools && pip3 install tensorrt_llm

Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
libopenmpi-dev is already the newest version (4.1.2-2ubuntu1).
0 upgraded, 0 newly installed, 0 to remove and 49 not upgraded.
Requirement already satisfied: pip in ./trt-llm-env/lib/python3.10/site-packages (25.1.1)
Requirement already satisfied: setuptools in ./trt-llm-env/lib/python3.10/site-packages (79.0.1)
Collecting setuptools
  Using cached setuptools-80.9.0-py3-none-any.whl.metadata (6.6 kB)
Using cached setuptools-80.9.0-py3-none-any.whl (1.2 MB)
Installing collected packages: setuptools
  Attempting uninstall: setuptools
    Found existing installation: setuptools 79.0.1
    Uninstalling setuptools-79.0.1:
      Successfully uninstalled setuptools-79.0.1


Successfully installed setuptools-80.9.0
Requirement already satisfied: tensorrt_llm in ./trt-llm-env/lib/python3.10/site-packages (0.20.0)
Requirement already satisfied: accelerate>=0.25.0 in ./trt-llm-env/lib/python3.10/site-packages (from tensorrt_llm) (1.8.1)
Requirement already satisfied: build in ./trt-llm-env/lib/python3.10/site-packages (from tensorrt_llm) (1.2.2.post1)
Requirement already satisfied: colored in ./trt-llm-env/lib/python3.10/site-packages (from tensorrt_llm) (2.3.0)
Requirement already satisfied: cuda-python in ./trt-llm-env/lib/python3.10/site-packages (from tensorrt_llm) (12.9.0)
Requirement already satisfied: diffusers>=0.27.0 in ./trt-llm-env/lib/python3.10/site-packages (from tensorrt_llm) (0.34.0)
Requirement already satisfied: lark in ./trt-llm-env/lib/python3.10/site-packages (from tensorrt_llm) (1.2.2)
Requirement already satisfied: mpi4py in ./trt-llm-env/lib/python3.10/site-packages (from tensorrt_llm) (4.1.0)
Requirement already satisfied: numpy<2 in ./trt-llm-env/lib/python3.10/site-packages (from tensorrt_llm) (1.26.4)
```

Model Checkpoint Conversion and Engine Build

```
(trt-llm-env) ubuntu@ip-172-31-68-64:~$ python3 /home/ubuntu/TensorRT-LLM/examples/models/core/gpt/convert_checkpoint.py --model_dir
 /home/ubuntu/models_src/gpt2 --output_dir /home/ubuntu/models_trt/gpt2/tllm_checkpoint --dtype float16
(trt-llm-env) ubuntu@ip-172-31-68-64:~$
(trt-llm-env) ubuntu@ip-172-31-68-64:~$
(trt-llm-env) ubuntu@ip-172-31-68-64:~$ tree /home/ubuntu/models_trt/gpt2/tllm_checkpoint
/home/ubuntu/models_trt/gpt2/tllm_checkpoint
├── config.json
└── rank0.safetensors

0 directories, 2 files
```

```
(trt-llm-env) ubuntu@ip-172-31-68-64:~$ trtllm-build --checkpoint_dir /home/ubuntu/models_trt/gpt2/tllm_checkpoint   --output_dir /
home/ubuntu/models_trt/gpt2/trt_engine_fp16_t4   --gemm_plugin float16   --remove_input_padding enable   --logits_dtype float16   -
-max_batch_size 1   --max_seq_len 1024   --workers 1   --log_level info   --context_fmha disable
(trt-llm-env) ubuntu@ip-172-31-68-64:~$ tree /home/ubuntu/models_trt/gpt2/trt_engine_fp16_t4
/home/ubuntu/models_trt/gpt2/trt_engine_fp16_t4
├── config.json
└── rank0.engine

0 directories, 2 files
```

```
(trt-llm-env) ubuntu@ip-172-31-68-64:~$ cp TensorRT-LLM/triton_backend/all_models/gpt/tensorrt_llm/config.pbtxt /home/ubuntu/triton_model_repo/gpt2/config.pbtxt
(trt-llm-env) ubuntu@ip-172-31-68-64:~$ MODEL_FOLDER=/home/ubuntu/triton_model_repo
FILL_TEMPLATE_SCRIPT=/home/ubuntu/TensorRT-LLM/triton_backend/tools/fill_template.py
ENGINE_DIR=/home/ubuntu/triton_model_repo/gpt2/1
(trt-llm-env) ubuntu@ip-172-31-68-64:~$
(trt-llm-env) ubuntu@ip-172-31-68-64:~$ python3 ${FILL_TEMPLATE_SCRIPT} ${MODEL_FOLDER}/gpt2/config.pbtxt engine_dir:${ENGINE_DIR},name:gpt2
name: "tensorrt_llm"
backend: "python"
max_batch_size: 1024

# # Uncomment this for dynamic_batching
# dynamic_batching {
#     max_queue_delay_microseconds: 50000
# }

input [
  {
    name: "input_ids"
    data_type: TYPE_INT32
    dims: [ -1 ]
  },
  {
    name: "input_lengths"
    data_type: TYPE_INT32
    dims: [ 1 ]
    reshape: { shape: [ ] }
  },
  {
    name: "request_output_len"
    data_type: TYPE_INT32
    dims: [ -1 ]
  },
  {
    name: "end_id"
    data_type: TYPE_INT32
    dims: [ 1 ]
    reshape: { shape: [ ] }
  },
  {
    name: "pad_id"
    data_type: TYPE_INT32
    dims: [ 1 ]
    reshape: { shape: [ ] }
  },
  {
    name: "beam_width"
    data_type: TYPE_INT32
    dims: [ 1 ]
    reshape: { shape: [ ] }
    optional: true
  },
  {
    name: "temperature"
    data_type: TYPE_FP32
```

Config.pbtxt Creation

```
cp /home/ubuntu/models_trt/gpt2/trt_engine_fp16_t4  /rank0.engine
/home/ubuntu/triton_model_repo/gpt2/ 1/rank0.engine

cp /home/ubuntu/models_trt/gpt2/trt_engine_fp16_t4  /config.json /home/ubuntu/triton_model_repo/gpt2/  1/config.json

cp TensorRT-LLM/triton_backend/all_models/gpt/tensorrt_llm/1/model.py
/home/ubuntu/triton_model_repo/gpt2/1/model.py

cp /home/ubuntu/models_src/gpt2/tokenizer_config.json
/home/ubuntu/triton_model_repo/gpt2/1/tokenizer/tokenizer_config.json

cp /home/ubuntu/models_src/gpt2/tokenizer.json /home/ubuntu/triton_model_repo/gpt2/1/tokenizer/tokenizer.json

cp /home/ubuntu/models_src/gpt2/vocab.json /home/ubuntu/triton_model_repo/gpt2/1/tokenizer/vocab.json

cp /home/ubuntu/models_src/gpt2/merges.txt /home/ubuntu/triton_model_repo/gpt2/1/tokenizer/merges.txt
```

Copy Model Files & Check Repo Structure

```
(trt-llm-env) ubuntu@ip-172-31-68-64:~$ tree /home/ubuntu/triton_model_repo
/home/ubuntu/triton_model_repo
└── gpt2
    ├── 1
    │   ├── __pycache__
    │   │   └── model.cpython-312.pyc
    │   ├── config.json
    │   ├── model.py
    │   ├── rank0.engine
    │   └── tokenizer
    │       ├── merges.txt
    │       ├── tokenizer.json
    │       ├── tokenizer_config.json
    │       └── vocab.json
    └── config.pbtxt

4 directories, 9 files
(trt-llm-env) ubuntu@ip-172-31-68-64:~$
```

Start Triton Server

```
(trt-llm-env) ubuntu@ip-172-31-68-64:~$ docker run --rm -it --net host --gpus all \
  --shm-size=2g --ulimit memlock=-1 --ulimit stack=67108864 \
  -v /home/ubuntu/triton_model_repo:/home/ubuntu/triton_model_repo \
  nvcr.io/nvidia/tritonserver:25.06-trtllm-python-py3 \
  tritonserver --model-repository=/home/ubuntu/triton_model_repo
```

```
==============================
== Triton Inference Server ==
==============================

NVIDIA Release 25.06 (build 179868725)
Triton Server Version 2.59.0

Copyright (c) 2018-2025, NVIDIA CORPORATION & AFFILIATES.  All rights reserved.

Various files include modifications (c) NVIDIA CORPORATION & AFFILIATES.  All rights reserved.

GOVERNING TERMS: The software and materials are governed by the NVIDIA Software License Agreement
(found at https://www.nvidia.com/en-us/agreements/enterprise-software/nvidia-software-license-agreement/)
and the Product-Specific Terms for NVIDIA AI Products
(found at https://www.nvidia.com/en-us/agreements/enterprise-software/product-specific-terms-for-ai-products/).

NOTE: CUDA Forward Compatibility mode ENABLED.
  Using CUDA 12.9 driver version 575.51.02 with kernel driver version 535.247.01.
  See https://docs.nvidia.com/deploy/cuda-compatibility/ for details.
```

Triton Server Status

```
[TensorRT-LLM] TensorRT-LLM version: 0.20.0
I0714 22:24:39.321789 1 model_lifecycle.cc:849] "successfully loaded 'gpt2'"
I0714 22:24:39.321920 1 server.cc:611]
+------------------+------+
| Repository Agent | Path |
+------------------+------+
+------------------+------+

I0714 22:24:39.321974 1 server.cc:638]
+---------+-------------------------------------------------------+------------------------------------------------------------------------------------------------------------------------------------------------------------+
| Backend | Path                                                  | Config                                                                                                                                                     |
+---------+-------------------------------------------------------+------------------------------------------------------------------------------------------------------------------------------------------------------------+
| python  | /opt/tritonserver/backends/python/libtriton_python.so | {"cmdline":{"auto-complete-config":"true","backend-directory":"/opt/tritonserver/backends","min-compute-capability":"6.000000","default-max-batch-size":"4"}} |
+---------+-------------------------------------------------------+------------------------------------------------------------------------------------------------------------------------------------------------------------+

I0714 22:24:39.322074 1 server.cc:681]
+-------+---------+--------+
| Model | Version | Status |
+-------+---------+--------+
| gpt2  | 1       | READY  |
+-------+---------+--------+

I0714 22:24:39.401656 1 metrics.cc:890] "Collecting metrics for GPU 0: Tesla T4"
I0714 22:24:39.406242 1 metrics.cc:783] "Collecting CPU metrics"
I0714 22:24:39.406431 1 tritonserver.cc:2598]
+----------------------------------+------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
| Option                           | Value                                                                                                                                                                                    |
+----------------------------------+------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
| server_id                        | triton                                                                                                                                                                                   |
| server_version                   | 2.59.0                                                                                                                                                                                   |
| server_extensions                | classification sequence model_repository model_repository(unload_dependents) schedule_policy model_configuration system_shared_memory cuda_shared_memory binary_tensor_data parameters statistics trace logging |
| model_repository_path[0]         | /home/ubuntu/triton_model_repo                                                                                                                                                           |
| model_control_mode               | MODE_NONE                                                                                                                                                                                |
| strict_model_config              | 0                                                                                                                                                                                        |
| model_config_name                |                                                                                                                                                                                          |
| rate_limit                       | OFF                                                                                                                                                                                      |
| pinned_memory_pool_byte_size     | 268435456                                                                                                                                                                                |
| cuda_memory_pool_byte_size{0}    | 67108864                                                                                                                                                                                 |
| min_supported_compute_capability | 6.0                                                                                                                                                                                      |
| strict_readiness                 | 1                                                                                                                                                                                        |
| exit_timeout                     | 30                                                                                                                                                                                       |
| cache_enabled                    | 0                                                                                                                                                                                        |
+----------------------------------+------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+

I0714 22:24:39.445471 1 grpc_server.cc:2562] "Started GRPCInferenceService at 0.0.0.0:8001"
I0714 22:24:39.446462 1 http_server.cc:4832] "Started HTTPService at 0.0.0.0:8000"
I0714 22:24:39.489409 1 http_server.cc:358] "Started Metrics Service at 0.0.0.0:8002"
```