```
!pip install bitsandbytes>=0.39.0
!pip install --upgrade accelerate transformers
!pip install --upgrade transformers
!pip install --force-reinstall flash-attn --no-cache-dir --no-build-
isolation
```

```
Requirement already satisfied: accelerate in
/usr/local/lib/python3.11/dist-packages (1.3.0)
Collecting accelerate
  Downloading accelerate-1.4.0-py3-none-any.whl.metadata (19 kB)
Requirement already satisfied: transformers in
/usr/local/lib/python3.11/dist-packages (4.48.3)
Collecting transformers
  Downloading transformers-4.49.0-py3-none-any.whl.metadata (44 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 44.0/44.0 kB 2.9 MB/s eta
0:00:00
ent already satisfied: numpy<3.0.0,>=1.17 in
/usr/local/lib/python3.11/dist-packages (from accelerate) (1.26.4)
Requirement already satisfied: packaging>=20.0 in
/usr/local/lib/python3.11/dist-packages (from accelerate) (24.2)
Requirement already satisfied: psutil in
/usr/local/lib/python3.11/dist-packages (from accelerate) (5.9.5)
Requirement already satisfied: pyyaml in
/usr/local/lib/python3.11/dist-packages (from accelerate) (6.0.2)
Requirement already satisfied: torch>=2.0.0 in
/usr/local/lib/python3.11/dist-packages (from accelerate)
(2.5.1+cu124)
Requirement already satisfied: huggingface-hub>=0.21.0 in
/usr/local/lib/python3.11/dist-packages (from accelerate) (0.28.1)
Requirement already satisfied: safetensors>=0.4.3 in
/usr/local/lib/python3.11/dist-packages (from accelerate) (0.5.3)
Requirement already satisfied: filelock in
/usr/local/lib/python3.11/dist-packages (from transformers) (3.17.0)
Requirement already satisfied: regex!=2019.12.17 in
/usr/local/lib/python3.11/dist-packages (from transformers)
(2024.11.6)
Requirement already satisfied: requests in
/usr/local/lib/python3.11/dist-packages (from transformers) (2.32.3)
Requirement already satisfied: tokenizers<0.22,>=0.21 in
/usr/local/lib/python3.11/dist-packages (from transformers) (0.21.0)
Requirement already satisfied: tqdm>=4.27 in
/usr/local/lib/python3.11/dist-packages (from transformers) (4.67.1)
Requirement already satisfied: fsspec>=2023.5.0 in
/usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.21.0-
>accelerate) (2024.10.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.21.0-
>accelerate) (4.12.2)
Requirement already satisfied: networkx in
/usr/local/lib/python3.11/dist-packages (from torch>=2.0.0-
>accelerate) (3.4.2)
Requirement already satisfied: jinja2 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.0.0-
>accelerate) (3.1.5)
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.4.127 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.0.0-
>accelerate) (12.4.127)
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.4.127
in /usr/local/lib/python3.11/dist-packages (from torch>=2.0.0-
>accelerate) (12.4.127)
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.4.127 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.0.0-
>accelerate) (12.4.127)
Requirement already satisfied: nvidia-cudnn-cu12==9.1.0.70 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.0.0-
>accelerate) (9.1.0.70)
Requirement already satisfied: nvidia-cublas-cu12==12.4.5.8 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.0.0-
>accelerate) (12.4.5.8)
Requirement already satisfied: nvidia-cufft-cu12==11.2.1.3 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.0.0-
>accelerate) (11.2.1.3)
Requirement already satisfied: nvidia-curand-cu12==10.3.5.147 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.0.0-
>accelerate) (10.3.5.147)
Requirement already satisfied: nvidia-cusolver-cu12==11.6.1.9 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.0.0-
>accelerate) (11.6.1.9)
Requirement already satisfied: nvidia-cusparse-cu12==12.3.1.170 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.0.0-
>accelerate) (12.3.1.170)
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.0.0-
>accelerate) (2.21.5)
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.0.0-
>accelerate) (12.4.127)
Requirement already satisfied: nvidia-nvjitlink-cu12==12.4.127 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.0.0-
>accelerate) (12.4.127)
Requirement already satisfied: triton==3.1.0 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.0.0-
>accelerate) (3.1.0)
Requirement already satisfied: sympy==1.13.1 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.0.0-
>accelerate) (1.13.1)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in
/usr/local/lib/python3.11/dist-packages (from sympy==1.13.1-
>torch>=2.0.0->accelerate) (1.3.0)
```

```
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.11/dist-packages (from requests->transformers)
(3.4.1)
Requirement already satisfied: idna<4,>=2.5 in
/usr/local/lib/python3.11/dist-packages (from requests->transformers)
(3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.11/dist-packages (from requests->transformers)
(2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.11/dist-packages (from requests->transformers)
(2025.1.31)
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.11/dist-packages (from jinja2->torch>=2.0.0-
>accelerate) (3.0.2)
Downloading accelerate-1.4.0-py3-none-any.whl (342 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 342.1/342.1 kB 13.6 MB/s eta
0:00:00
ers-4.49.0-py3-none-any.whl (10.0 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 10.0/10.0 MB 62.9 MB/s eta
0:00:00
ers, accelerate
  Attempting uninstall: transformers
    Found existing installation: transformers 4.48.3
    Uninstalling transformers-4.48.3:
      Successfully uninstalled transformers-4.48.3
  Attempting uninstall: accelerate
    Found existing installation: accelerate 1.3.0
    Uninstalling accelerate-1.3.0:
      Successfully uninstalled accelerate-1.3.0
Successfully installed accelerate-1.4.0 transformers-4.49.0
Requirement already satisfied: transformers in
/usr/local/lib/python3.11/dist-packages (4.49.0)
Requirement already satisfied: filelock in
/usr/local/lib/python3.11/dist-packages (from transformers) (3.17.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.26.0 in
/usr/local/lib/python3.11/dist-packages (from transformers) (0.28.1)
Requirement already satisfied: numpy>=1.17 in
/usr/local/lib/python3.11/dist-packages (from transformers) (1.26.4)
Requirement already satisfied: packaging>=20.0 in
/usr/local/lib/python3.11/dist-packages (from transformers) (24.2)
Requirement already satisfied: pyyaml>=5.1 in
/usr/local/lib/python3.11/dist-packages (from transformers) (6.0.2)
Requirement already satisfied: regex!=2019.12.17 in
/usr/local/lib/python3.11/dist-packages (from transformers)
(2024.11.6)
Requirement already satisfied: requests in
/usr/local/lib/python3.11/dist-packages (from transformers) (2.32.3)
Requirement already satisfied: tokenizers<0.22,>=0.21 in
```

```
/usr/local/lib/python3.11/dist-packages (from transformers) (0.21.0)
Requirement already satisfied: safetensors>=0.4.1 in
/usr/local/lib/python3.11/dist-packages (from transformers) (0.5.3)
Requirement already satisfied: tqdm>=4.27 in
/usr/local/lib/python3.11/dist-packages (from transformers) (4.67.1)
Requirement already satisfied: fsspec>=2023.5.0 in
/usr/local/lib/python3.11/dist-packages (from huggingface-
hub<1.0,>=0.26.0->transformers) (2024.10.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.11/dist-packages (from huggingface-
hub<1.0,>=0.26.0->transformers) (4.12.2)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.11/dist-packages (from requests->transformers)
(3.4.1)
Requirement already satisfied: idna<4,>=2.5 in
/usr/local/lib/python3.11/dist-packages (from requests->transformers)
(3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.11/dist-packages (from requests->transformers)
(2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.11/dist-packages (from requests->transformers)
(2025.1.31)
Collecting flash-attn
  Downloading flash_attn-2.7.4.post1.tar.gz (6.0 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 6.0/6.0 MB 74.8 MB/s eta
0:00:00
etadata (setup.py) ...  flash-attn)
  Downloading torch-2.6.0-cp311-cp311-manylinux1_x86_64.whl.metadata
(28 kB)
Collecting einops (from flash-attn)
  Downloading einops-0.8.1-py3-none-any.whl.metadata (13 kB)
Collecting filelock (from torch->flash-attn)
  Downloading filelock-3.17.0-py3-none-any.whl.metadata (2.9 kB)
Collecting typing-extensions>=4.10.0 (from torch->flash-attn)
  Downloading typing_extensions-4.12.2-py3-none-any.whl.metadata (3.0
kB)
Collecting networkx (from torch->flash-attn)
  Downloading networkx-3.4.2-py3-none-any.whl.metadata (6.3 kB)
Collecting jinja2 (from torch->flash-attn)
  Downloading jinja2-3.1.6-py3-none-any.whl.metadata (2.9 kB)
Collecting fsspec (from torch->flash-attn)
  Downloading fsspec-2025.2.0-py3-none-any.whl.metadata (11 kB)
Collecting nvidia-cuda-nvrtc-cu12==12.4.127 (from torch->flash-attn)
  Downloading nvidia_cuda_nvrtc_cu12-12.4.127-py3-none-
manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cuda-runtime-cu12==12.4.127 (from torch->flash-attn)
  Downloading nvidia_cuda_runtime_cu12-12.4.127-py3-none-
manylinux2014_x86_64.whl.metadata (1.5 kB)
```

```
Collecting nvidia-cuda-cupti-cu12==12.4.127 (from torch->flash-attn)
  Downloading nvidia_cuda_cupti_cu12-12.4.127-py3-none-
manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cudnn-cu12==9.1.0.70 (from torch->flash-attn)
  Downloading nvidia_cudnn_cu12-9.1.0.70-py3-none-
manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cublas-cu12==12.4.5.8 (from torch->flash-attn)
  Downloading nvidia_cublas_cu12-12.4.5.8-py3-none-
manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cufft-cu12==11.2.1.3 (from torch->flash-attn)
  Downloading nvidia_cufft_cu12-11.2.1.3-py3-none-
manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-curand-cu12==10.3.5.147 (from torch->flash-attn)
  Downloading nvidia_curand_cu12-10.3.5.147-py3-none-
manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cusolver-cu12==11.6.1.9 (from torch->flash-attn)
  Downloading nvidia_cusolver_cu12-11.6.1.9-py3-none-
manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cusparse-cu12==12.3.1.170 (from torch->flash-attn)
  Downloading nvidia_cusparse_cu12-12.3.1.170-py3-none-
manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cusparselt-cu12==0.6.2 (from torch->flash-attn)
  Downloading nvidia_cusparselt_cu12-0.6.2-py3-none-
manylinux2014_x86_64.whl.metadata (6.8 kB)
Collecting nvidia-nccl-cu12==2.21.5 (from torch->flash-attn)
  Downloading nvidia_nccl_cu12-2.21.5-py3-none-
manylinux2014_x86_64.whl.metadata (1.8 kB)
Collecting nvidia-nvtx-cu12==12.4.127 (from torch->flash-attn)
  Downloading nvidia_nvtx_cu12-12.4.127-py3-none-
manylinux2014_x86_64.whl.metadata (1.7 kB)
Collecting nvidia-nvjitlink-cu12==12.4.127 (from torch->flash-attn)
  Downloading nvidia_nvjitlink_cu12-12.4.127-py3-none-
manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting triton==3.2.0 (from torch->flash-attn)
  Downloading triton-3.2.0-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (1.4 kB)
Collecting sympy==1.13.1 (from torch->flash-attn)
  Downloading sympy-1.13.1-py3-none-any.whl.metadata (12 kB)
Collecting mpmath<1.4,>=1.1.0 (from sympy==1.13.1->torch->flash-attn)
  Downloading mpmath-1.3.0-py3-none-any.whl.metadata (8.6 kB)
Collecting MarkupSafe>=2.0 (from jinja2->torch->flash-attn)
  Downloading MarkupSafe-3.0.2-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (4.0 kB)
Downloading einops-0.8.1-py3-none-any.whl (64 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 64.4/64.4 kB 187.7 MB/s eta
0:00:00
anylinux1_x86_64.whl (766.7 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 766.7/766.7 MB 201.1 MB/s eta
0:00:00
```

```
anylinux2014_x86_64.whl (363.4 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 363.4/363.4 MB 98.0 MB/s eta
0:00:00
anylinux2014_x86_64.whl (13.8 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 13.8/13.8 MB 162.5 MB/s eta
0:00:00
anylinux2014_x86_64.whl (24.6 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 24.6/24.6 MB 158.2 MB/s eta
0:00:00
e_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl (883 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 883.7/883.7 kB 213.9 MB/s eta
0:00:00
anylinux2014_x86_64.whl (664.8 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 664.8/664.8 MB 201.1 MB/s eta
0:00:00
anylinux2014_x86_64.whl (211.5 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 211.5/211.5 MB 230.6 MB/s eta
0:00:00
anylinux2014_x86_64.whl (56.3 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 56.3/56.3 MB 261.9 MB/s eta
0:00:00
anylinux2014_x86_64.whl (127.9 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 127.9/127.9 MB 209.6 MB/s eta
0:00:00
anylinux2014_x86_64.whl (207.5 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 207.5/207.5 MB 151.9 MB/s eta
0:00:00
anylinux2014_x86_64.whl (150.1 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 150.1/150.1 MB 236.1 MB/s eta
0:00:00
anylinux2014_x86_64.whl (188.7 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 188.7/188.7 MB 177.3 MB/s eta
0:00:00
anylinux2014_x86_64.whl (21.1 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 21.1/21.1 MB 262.5 MB/s eta
0:00:00
anylinux2014_x86_64.whl (99 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 99.1/99.1 kB 287.8 MB/s eta
0:00:00
py-1.13.1-py3-none-any.whl (6.2 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 6.2/6.2 MB 127.6 MB/s eta
0:00:00
anylinux_2_17_x86_64.manylinux2014_x86_64.whl (253.2 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 253.2/253.2 MB 274.4 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 184.5/184.5 kB 356.2 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 134.9/134.9 kB 332.1 MB/s eta
0:00:00
```

```
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.7/1.7 MB 334.6 MB/s eta
0:00:00
anylinux_2_17_x86_64.manylinux2014_x86_64.whl (23 kB)
Downloading mpmath-1.3.0-py3-none-any.whl (536 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 536.2/536.2 kB 368.1 MB/s eta
0:00:00
e=flash_attn-2.7.4.post1-cp311-cp311-linux_x86_64.whl size=187815463
sha256=d944fc7d2f962bce83fc4708c2fc0c21eaf8255962a0b350ae919362a51b7ef
2
  Stored in directory:
/tmp/pip-ephem-wheel-cache-3dfnh5xp/wheels/3d/88/d8/284b89f56af7d5bf36
6b10d6b8e251ac8a7c7bf3f04203fb4f
Successfully built flash-attn
Installing collected packages: triton, nvidia-cusparselt-cu12, mpmath,
typing-extensions, sympy, nvidia-nvtx-cu12, nvidia-nvjitlink-cu12,
nvidia-nccl-cu12, nvidia-curand-cu12, nvidia-cufft-cu12, nvidia-cuda-
runtime-cu12, nvidia-cuda-nvrtc-cu12, nvidia-cuda-cupti-cu12, nvidia-
cublas-cu12, networkx, MarkupSafe, fsspec, filelock, einops, nvidia-
cusparse-cu12, nvidia-cudnn-cu12, jinja2, nvidia-cusolver-cu12, torch,
flash-attn
  Attempting uninstall: triton
    Found existing installation: triton 3.1.0
    Uninstalling triton-3.1.0:
      Successfully uninstalled triton-3.1.0
  Attempting uninstall: mpmath
    Found existing installation: mpmath 1.3.0
    Uninstalling mpmath-1.3.0:
      Successfully uninstalled mpmath-1.3.0
  Attempting uninstall: typing-extensions
    Found existing installation: typing_extensions 4.12.2
    Uninstalling typing_extensions-4.12.2:
      Successfully uninstalled typing_extensions-4.12.2
  Attempting uninstall: sympy
    Found existing installation: sympy 1.13.1
    Uninstalling sympy-1.13.1:
      Successfully uninstalled sympy-1.13.1
  Attempting uninstall: nvidia-nvtx-cu12
    Found existing installation: nvidia-nvtx-cu12 12.4.127
    Uninstalling nvidia-nvtx-cu12-12.4.127:
      Successfully uninstalled nvidia-nvtx-cu12-12.4.127
  Attempting uninstall: nvidia-nvjitlink-cu12
    Found existing installation: nvidia-nvjitlink-cu12 12.4.127
    Uninstalling nvidia-nvjitlink-cu12-12.4.127:
      Successfully uninstalled nvidia-nvjitlink-cu12-12.4.127
  Attempting uninstall: nvidia-nccl-cu12
    Found existing installation: nvidia-nccl-cu12 2.21.5
    Uninstalling nvidia-nccl-cu12-2.21.5:
      Successfully uninstalled nvidia-nccl-cu12-2.21.5
  Attempting uninstall: nvidia-curand-cu12
```

```
    Found existing installation: nvidia-curand-cu12 10.3.5.147
    Uninstalling nvidia-curand-cu12-10.3.5.147:
      Successfully uninstalled nvidia-curand-cu12-10.3.5.147
  Attempting uninstall: nvidia-cufft-cu12
    Found existing installation: nvidia-cufft-cu12 11.2.1.3
    Uninstalling nvidia-cufft-cu12-11.2.1.3:
      Successfully uninstalled nvidia-cufft-cu12-11.2.1.3
  Attempting uninstall: nvidia-cuda-runtime-cu12
    Found existing installation: nvidia-cuda-runtime-cu12 12.4.127
    Uninstalling nvidia-cuda-runtime-cu12-12.4.127:
      Successfully uninstalled nvidia-cuda-runtime-cu12-12.4.127
  Attempting uninstall: nvidia-cuda-nvrtc-cu12
    Found existing installation: nvidia-cuda-nvrtc-cu12 12.4.127
    Uninstalling nvidia-cuda-nvrtc-cu12-12.4.127:
      Successfully uninstalled nvidia-cuda-nvrtc-cu12-12.4.127
  Attempting uninstall: nvidia-cuda-cupti-cu12
    Found existing installation: nvidia-cuda-cupti-cu12 12.4.127
    Uninstalling nvidia-cuda-cupti-cu12-12.4.127:
      Successfully uninstalled nvidia-cuda-cupti-cu12-12.4.127
  Attempting uninstall: nvidia-cublas-cu12
    Found existing installation: nvidia-cublas-cu12 12.4.5.8
    Uninstalling nvidia-cublas-cu12-12.4.5.8:
      Successfully uninstalled nvidia-cublas-cu12-12.4.5.8
  Attempting uninstall: networkx
    Found existing installation: networkx 3.4.2
    Uninstalling networkx-3.4.2:
      Successfully uninstalled networkx-3.4.2
  Attempting uninstall: MarkupSafe
    Found existing installation: MarkupSafe 3.0.2
    Uninstalling MarkupSafe-3.0.2:
      Successfully uninstalled MarkupSafe-3.0.2
  Attempting uninstall: fsspec
    Found existing installation: fsspec 2024.10.0
    Uninstalling fsspec-2024.10.0:
      Successfully uninstalled fsspec-2024.10.0
  Attempting uninstall: filelock
    Found existing installation: filelock 3.17.0
    Uninstalling filelock-3.17.0:
      Successfully uninstalled filelock-3.17.0
  Attempting uninstall: einops
    Found existing installation: einops 0.8.1
    Uninstalling einops-0.8.1:
      Successfully uninstalled einops-0.8.1
  Attempting uninstall: nvidia-cusparse-cu12
    Found existing installation: nvidia-cusparse-cu12 12.3.1.170
    Uninstalling nvidia-cusparse-cu12-12.3.1.170:
      Successfully uninstalled nvidia-cusparse-cu12-12.3.1.170
  Attempting uninstall: nvidia-cudnn-cu12
    Found existing installation: nvidia-cudnn-cu12 9.1.0.70
    Uninstalling nvidia-cudnn-cu12-9.1.0.70:
      Successfully uninstalled nvidia-cudnn-cu12-9.1.0.70
  Attempting uninstall: jinja2
    Found existing installation: Jinja2 3.1.5
    Uninstalling Jinja2-3.1.5:
      Successfully uninstalled Jinja2-3.1.5
  Attempting uninstall: nvidia-cusolver-cu12
    Found existing installation: nvidia-cusolver-cu12 11.6.1.9
    Uninstalling nvidia-cusolver-cu12-11.6.1.9:
      Successfully uninstalled nvidia-cusolver-cu12-11.6.1.9
  Attempting uninstall: torch
    Found existing installation: torch 2.5.1+cu124
    Uninstalling torch-2.5.1+cu124:
      Successfully uninstalled torch-2.5.1+cu124
ERROR: pip's dependency resolver does not currently take into account
all the packages that are installed. This behaviour is the source of
the following dependency conflicts.
fastai 2.7.18 requires torch<2.6,>=1.10, but you have torch 2.6.0
which is incompatible.
gcsfs 2024.10.0 requires fsspec==2024.10.0, but you have fsspec
2025.2.0 which is incompatible.
torchaudio 2.5.1+cu124 requires torch==2.5.1, but you have torch 2.6.0
which is incompatible.
torchvision 0.20.1+cu124 requires torch==2.5.1, but you have torch
2.6.0 which is incompatible.
Successfully installed MarkupSafe-3.0.2 einops-0.8.1 filelock-3.17.0
flash-attn-2.7.4.post1 fsspec-2025.2.0 jinja2-3.1.6 mpmath-1.3.0
networkx-3.4.2 nvidia-cublas-cu12-12.4.5.8 nvidia-cuda-cupti-cu12-
12.4.127 nvidia-cuda-nvrtc-cu12-12.4.127 nvidia-cuda-runtime-cu12-
12.4.127 nvidia-cudnn-cu12-9.1.0.70 nvidia-cufft-cu12-11.2.1.3 nvidia-
curand-cu12-10.3.5.147 nvidia-cusolver-cu12-11.6.1.9 nvidia-cusparse-
cu12-12.3.1.170 nvidia-cusparselt-cu12-0.6.2 nvidia-nccl-cu12-2.21.5
nvidia-nvjitlink-cu12-12.4.127 nvidia-nvtx-cu12-12.4.127 sympy-1.13.1
torch-2.6.0 triton-3.2.0 typing-extensions-4.12.2
```

```
!pip cache purge # Clearing the cache to remove any problematic files
!pip install --upgrade --force-reinstall torchvision # Reinstalling
torchvision to ensure a clean install
```

```
Files removed: 84
Collecting torchvision
  Downloading torchvision-0.21.0-cp311-cp311-
manylinux1_x86_64.whl.metadata (6.1 kB)
Collecting numpy (from torchvision)
  Downloading numpy-2.2.3-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (62 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 62.0/62.0 kB 2.9 MB/s eta
0:00:00
 torchvision)
  Downloading torch-2.6.0-cp311-cp311-manylinux1_x86_64.whl.metadata
```

```
(28 kB)
Collecting pillow!=8.3.*,>=5.3.0 (from torchvision)
  Downloading pillow-11.1.0-cp311-cp311-
manylinux_2_28_x86_64.whl.metadata (9.1 kB)
Collecting filelock (from torch==2.6.0->torchvision)
  Downloading filelock-3.17.0-py3-none-any.whl.metadata (2.9 kB)
Collecting typing-extensions>=4.10.0 (from torch==2.6.0->torchvision)
  Downloading typing_extensions-4.12.2-py3-none-any.whl.metadata (3.0
kB)
Collecting networkx (from torch==2.6.0->torchvision)
  Downloading networkx-3.4.2-py3-none-any.whl.metadata (6.3 kB)
Collecting jinja2 (from torch==2.6.0->torchvision)
  Downloading jinja2-3.1.6-py3-none-any.whl.metadata (2.9 kB)
Collecting fsspec (from torch==2.6.0->torchvision)
  Downloading fsspec-2025.2.0-py3-none-any.whl.metadata (11 kB)
Collecting nvidia-cuda-nvrtc-cu12==12.4.127 (from torch==2.6.0-
>torchvision)
  Downloading nvidia_cuda_nvrtc_cu12-12.4.127-py3-none-
manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cuda-runtime-cu12==12.4.127 (from torch==2.6.0-
>torchvision)
  Downloading nvidia_cuda_runtime_cu12-12.4.127-py3-none-
manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cuda-cupti-cu12==12.4.127 (from torch==2.6.0-
>torchvision)
  Downloading nvidia_cuda_cupti_cu12-12.4.127-py3-none-
manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cudnn-cu12==9.1.0.70 (from torch==2.6.0-
>torchvision)
  Downloading nvidia_cudnn_cu12-9.1.0.70-py3-none-
manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cublas-cu12==12.4.5.8 (from torch==2.6.0-
>torchvision)
  Downloading nvidia_cublas_cu12-12.4.5.8-py3-none-
manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cufft-cu12==11.2.1.3 (from torch==2.6.0-
>torchvision)
  Downloading nvidia_cufft_cu12-11.2.1.3-py3-none-
manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-curand-cu12==10.3.5.147 (from torch==2.6.0-
>torchvision)
  Downloading nvidia_curand_cu12-10.3.5.147-py3-none-
manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cusolver-cu12==11.6.1.9 (from torch==2.6.0-
>torchvision)
  Downloading nvidia_cusolver_cu12-11.6.1.9-py3-none-
manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cusparse-cu12==12.3.1.170 (from torch==2.6.0-
>torchvision)
```

```
  Downloading nvidia_cusparse_cu12-12.3.1.170-py3-none-
manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cusparselt-cu12==0.6.2 (from torch==2.6.0-
>torchvision)
  Downloading nvidia_cusparselt_cu12-0.6.2-py3-none-
manylinux2014_x86_64.whl.metadata (6.8 kB)
Collecting nvidia-nccl-cu12==2.21.5 (from torch==2.6.0->torchvision)
  Downloading nvidia_nccl_cu12-2.21.5-py3-none-
manylinux2014_x86_64.whl.metadata (1.8 kB)
Collecting nvidia-nvtx-cu12==12.4.127 (from torch==2.6.0->torchvision)
  Downloading nvidia_nvtx_cu12-12.4.127-py3-none-
manylinux2014_x86_64.whl.metadata (1.7 kB)
Collecting nvidia-nvjitlink-cu12==12.4.127 (from torch==2.6.0-
>torchvision)
  Downloading nvidia_nvjitlink_cu12-12.4.127-py3-none-
manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting triton==3.2.0 (from torch==2.6.0->torchvision)
  Downloading triton-3.2.0-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (1.4 kB)
Collecting sympy==1.13.1 (from torch==2.6.0->torchvision)
  Downloading sympy-1.13.1-py3-none-any.whl.metadata (12 kB)
Collecting mpmath<1.4,>=1.1.0 (from sympy==1.13.1->torch==2.6.0-
>torchvision)
  Downloading mpmath-1.3.0-py3-none-any.whl.metadata (8.6 kB)
Collecting MarkupSafe>=2.0 (from jinja2->torch==2.6.0->torchvision)
  Downloading MarkupSafe-3.0.2-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (4.0 kB)
Downloading torchvision-0.21.0-cp311-cp311-manylinux1_x86_64.whl (7.2
MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 7.2/7.2 MB 83.8 MB/s eta
0:00:00
anylinux1_x86_64.whl (766.7 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 766.7/766.7 MB 1.2 MB/s eta
0:00:00
anylinux2014_x86_64.whl (363.4 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 363.4/363.4 MB 2.7 MB/s eta
0:00:00
anylinux2014_x86_64.whl (13.8 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 13.8/13.8 MB 110.7 MB/s eta
0:00:00
anylinux2014_x86_64.whl (24.6 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 24.6/24.6 MB 85.6 MB/s eta
0:00:00
e_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl (883 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 883.7/883.7 kB 61.2 MB/s eta
0:00:00
anylinux2014_x86_64.whl (664.8 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 664.8/664.8 MB 2.6 MB/s eta
0:00:00
```

```
anylinux2014_x86_64.whl (211.5 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 211.5/211.5 MB 1.4 MB/s eta
0:00:00
anylinux2014_x86_64.whl (56.3 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 56.3/56.3 MB 14.9 MB/s eta
0:00:00
anylinux2014_x86_64.whl (127.9 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 127.9/127.9 MB 8.9 MB/s eta
0:00:00
anylinux2014_x86_64.whl (207.5 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 207.5/207.5 MB 6.3 MB/s eta
0:00:00
anylinux2014_x86_64.whl (150.1 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 150.1/150.1 MB 6.8 MB/s eta
0:00:00
anylinux2014_x86_64.whl (188.7 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 188.7/188.7 MB 6.4 MB/s eta
0:00:00
anylinux2014_x86_64.whl (21.1 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 21.1/21.1 MB 89.9 MB/s eta
0:00:00
anylinux2014_x86_64.whl (99 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 99.1/99.1 kB 8.4 MB/s eta
0:00:00
py-1.13.1-py3-none-any.whl (6.2 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 6.2/6.2 MB 111.8 MB/s eta
0:00:00
anylinux_2_17_x86_64.manylinux2014_x86_64.whl (253.2 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 253.2/253.2 MB 5.3 MB/s eta
0:00:00
anylinux_2_28_x86_64.whl (4.5 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 4.5/4.5 MB 48.3 MB/s eta
0:00:00
py-2.2.3-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl
(16.4 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 16.4/16.4 MB 60.2 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 184.5/184.5 kB 14.6 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 134.9/134.9 kB 13.4 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.7/1.7 MB 66.8 MB/s eta
0:00:00
anylinux_2_17_x86_64.manylinux2014_x86_64.whl (23 kB)
Downloading mpmath-1.3.0-py3-none-any.whl (536 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 536.2/536.2 kB 36.8 MB/s eta
0:00:00
pmath, typing-extensions, sympy, pillow, nvidia-nvtx-cu12, nvidia-
nvjitlink-cu12, nvidia-nccl-cu12, nvidia-curand-cu12, nvidia-cufft-
```

```
cu12, nvidia-cuda-runtime-cu12, nvidia-cuda-nvrtc-cu12, nvidia-cuda-
cupti-cu12, nvidia-cublas-cu12, numpy, networkx, MarkupSafe, fsspec,
filelock, nvidia-cusparse-cu12, nvidia-cudnn-cu12, jinja2, nvidia-
cusolver-cu12, torch, torchvision
  Attempting uninstall: triton
    Found existing installation: triton 3.2.0
    Uninstalling triton-3.2.0:
      Successfully uninstalled triton-3.2.0
  Attempting uninstall: nvidia-cusparselt-cu12
    Found existing installation: nvidia-cusparselt-cu12 0.6.2
    Uninstalling nvidia-cusparselt-cu12-0.6.2:
      Successfully uninstalled nvidia-cusparselt-cu12-0.6.2
  Attempting uninstall: mpmath
    Found existing installation: mpmath 1.3.0
    Uninstalling mpmath-1.3.0:
      Successfully uninstalled mpmath-1.3.0
  Attempting uninstall: typing-extensions
    Found existing installation: typing_extensions 4.12.2
    Uninstalling typing_extensions-4.12.2:
      Successfully uninstalled typing_extensions-4.12.2
  Attempting uninstall: sympy
    Found existing installation: sympy 1.13.1
    Uninstalling sympy-1.13.1:
      Successfully uninstalled sympy-1.13.1
  Attempting uninstall: pillow
    Found existing installation: pillow 11.1.0
    Uninstalling pillow-11.1.0:
      Successfully uninstalled pillow-11.1.0
  Attempting uninstall: nvidia-nvtx-cu12
    Found existing installation: nvidia-nvtx-cu12 12.4.127
    Uninstalling nvidia-nvtx-cu12-12.4.127:
      Successfully uninstalled nvidia-nvtx-cu12-12.4.127
  Attempting uninstall: nvidia-nvjitlink-cu12
    Found existing installation: nvidia-nvjitlink-cu12 12.4.127
    Uninstalling nvidia-nvjitlink-cu12-12.4.127:
      Successfully uninstalled nvidia-nvjitlink-cu12-12.4.127
  Attempting uninstall: nvidia-nccl-cu12
    Found existing installation: nvidia-nccl-cu12 2.21.5
    Uninstalling nvidia-nccl-cu12-2.21.5:
      Successfully uninstalled nvidia-nccl-cu12-2.21.5
  Attempting uninstall: nvidia-curand-cu12
    Found existing installation: nvidia-curand-cu12 10.3.5.147
    Uninstalling nvidia-curand-cu12-10.3.5.147:
      Successfully uninstalled nvidia-curand-cu12-10.3.5.147
  Attempting uninstall: nvidia-cufft-cu12
    Found existing installation: nvidia-cufft-cu12 11.2.1.3
    Uninstalling nvidia-cufft-cu12-11.2.1.3:
      Successfully uninstalled nvidia-cufft-cu12-11.2.1.3
  Attempting uninstall: nvidia-cuda-runtime-cu12
```

```
    Found existing installation: nvidia-cuda-runtime-cu12 12.4.127
    Uninstalling nvidia-cuda-runtime-cu12-12.4.127:
      Successfully uninstalled nvidia-cuda-runtime-cu12-12.4.127
  Attempting uninstall: nvidia-cuda-nvrtc-cu12
    Found existing installation: nvidia-cuda-nvrtc-cu12 12.4.127
    Uninstalling nvidia-cuda-nvrtc-cu12-12.4.127:
      Successfully uninstalled nvidia-cuda-nvrtc-cu12-12.4.127
  Attempting uninstall: nvidia-cuda-cupti-cu12
    Found existing installation: nvidia-cuda-cupti-cu12 12.4.127
    Uninstalling nvidia-cuda-cupti-cu12-12.4.127:
      Successfully uninstalled nvidia-cuda-cupti-cu12-12.4.127
  Attempting uninstall: nvidia-cublas-cu12
    Found existing installation: nvidia-cublas-cu12 12.4.5.8
    Uninstalling nvidia-cublas-cu12-12.4.5.8:
      Successfully uninstalled nvidia-cublas-cu12-12.4.5.8
  Attempting uninstall: numpy
    Found existing installation: numpy 1.26.4
    Uninstalling numpy-1.26.4:
      Successfully uninstalled numpy-1.26.4
  Attempting uninstall: networkx
    Found existing installation: networkx 3.4.2
    Uninstalling networkx-3.4.2:
      Successfully uninstalled networkx-3.4.2
  Attempting uninstall: MarkupSafe
    Found existing installation: MarkupSafe 3.0.2
    Uninstalling MarkupSafe-3.0.2:
      Successfully uninstalled MarkupSafe-3.0.2
  Attempting uninstall: fsspec
    Found existing installation: fsspec 2025.2.0
    Uninstalling fsspec-2025.2.0:
      Successfully uninstalled fsspec-2025.2.0
  Attempting uninstall: filelock
    Found existing installation: filelock 3.17.0
    Uninstalling filelock-3.17.0:
      Successfully uninstalled filelock-3.17.0
  Attempting uninstall: nvidia-cusparse-cu12
    Found existing installation: nvidia-cusparse-cu12 12.3.1.170
    Uninstalling nvidia-cusparse-cu12-12.3.1.170:
      Successfully uninstalled nvidia-cusparse-cu12-12.3.1.170
  Attempting uninstall: nvidia-cudnn-cu12
    Found existing installation: nvidia-cudnn-cu12 9.1.0.70
    Uninstalling nvidia-cudnn-cu12-9.1.0.70:
      Successfully uninstalled nvidia-cudnn-cu12-9.1.0.70
  Attempting uninstall: jinja2
    Found existing installation: Jinja2 3.1.6
    Uninstalling Jinja2-3.1.6:
      Successfully uninstalled Jinja2-3.1.6
  Attempting uninstall: nvidia-cusolver-cu12
    Found existing installation: nvidia-cusolver-cu12 11.6.1.9
    Uninstalling nvidia-cusolver-cu12-11.6.1.9:
      Successfully uninstalled nvidia-cusolver-cu12-11.6.1.9
  Attempting uninstall: torch
    Found existing installation: torch 2.6.0
    Uninstalling torch-2.6.0:
      Successfully uninstalled torch-2.6.0
  Attempting uninstall: torchvision
    Found existing installation: torchvision 0.20.1+cu124
    Uninstalling torchvision-0.20.1+cu124:
      Successfully uninstalled torchvision-0.20.1+cu124
ERROR: pip's dependency resolver does not currently take into account
all the packages that are installed. This behaviour is the source of
the following dependency conflicts.
tensorflow 2.18.0 requires numpy<2.1.0,>=1.26.0, but you have numpy
2.2.3 which is incompatible.
thinc 8.2.5 requires numpy<2.0.0,>=1.19.0; python_version >= "3.9",
but you have numpy 2.2.3 which is incompatible.
fastai 2.7.18 requires torch<2.6,>=1.10, but you have torch 2.6.0
which is incompatible.
pytensor 2.27.1 requires numpy<2,>=1.17.0, but you have numpy 2.2.3
which is incompatible.
langchain 0.3.19 requires numpy<2,>=1.26.4; python_version < "3.12",
but you have numpy 2.2.3 which is incompatible.
numba 0.61.0 requires numpy<2.2,>=1.24, but you have numpy 2.2.3 which
is incompatible.
gcsfs 2024.10.0 requires fsspec==2024.10.0, but you have fsspec
2025.2.0 which is incompatible.
gensim 4.3.3 requires numpy<2.0,>=1.18.5, but you have numpy 2.2.3
which is incompatible.
torchaudio 2.5.1+cu124 requires torch==2.5.1, but you have torch 2.6.0
which is incompatible.
Successfully installed MarkupSafe-3.0.2 filelock-3.17.0 fsspec-
2025.2.0 jinja2-3.1.6 mpmath-1.3.0 networkx-3.4.2 numpy-2.2.3 nvidia-
cublas-cu12-12.4.5.8 nvidia-cuda-cupti-cu12-12.4.127 nvidia-cuda-
nvrtc-cu12-12.4.127 nvidia-cuda-runtime-cu12-12.4.127 nvidia-cudnn-
cu12-9.1.0.70 nvidia-cufft-cu12-11.2.1.3 nvidia-curand-cu12-10.3.5.147
nvidia-cusolver-cu12-11.6.1.9 nvidia-cusparse-cu12-12.3.1.170 nvidia-
cusparselt-cu12-0.6.2 nvidia-nccl-cu12-2.21.5 nvidia-nvjitlink-cu12-
12.4.127 nvidia-nvtx-cu12-12.4.127 pillow-11.1.0 sympy-1.13.1 torch-
2.6.0 torchvision-0.21.0 triton-3.2.0 typing-extensions-4.12.2
```

```
{"id":"c6f7c0d748804957a4993898ec24025c","pip_warning":{"packages":
["PIL","markupsafe","mpmath","sympy","torch","torchgen","triton"]}}
```

```python
from transformers import AutoTokenizer, AutoModelForCausalLM,
BitsAndBytesConfig
from accelerate.test_utils.testing import get_backend
import torch
import os
import time
import psutil
from huggingface_hub import login

# Hugging face token
YOUR_ACTUAL_TOKEN = "YOUR_TOKEN"
# Log in to Hugging Face Hub
login(token=YOUR_ACTUAL_TOKEN)

os.environ["TOKENIZERS_PARALLELISM"] = "false"  # To prevent long
warnings :)

# Load model with 4-bit quantization
quant_config = BitsAndBytesConfig(
    load_in_4bit=True,
    bnb_4bit_compute_dtype=torch.bfloat16,  # Ensures computation
stability
    bnb_4bit_use_double_quant=True,  # Improves efficiency
)

tokenizer = AutoTokenizer.from_pretrained("google/gemma-2b")

model = AutoModelForCausalLM.from_pretrained(
    "google/gemma-2b",
    quantization_config=quant_config,
    device_map="auto",
    torch_dtype=torch.bfloat16,
    attn_implementation="sdpa",
)
```

{"model_id":"aeda261bd2f64ad699a13b75b22aa377","version_major":2,"vers
ion_minor":0}

```python
device = model.device

# Get user input
input_text = input("Enter your prompt: ")
inputs = tokenizer(input_text, return_tensors="pt").to("cuda")

# Benchmarking Start
start_time = time.time()

# Text Generation
with torch.no_grad():
    outputs = model.generate(
        **inputs,
        do_sample=True,
        temperature=0.2,
        max_new_tokens=200,  # Increased response length
        top_p=0.9,
        top_k=20,
        use_cache=True,  # Enable KV Cache
        forced_bos_token_id=tokenizer.bos_token_id  # Forces English
output
    )

end_time = time.time()

# Decode and Print Output
generated_text = tokenizer.batch_decode(outputs,
skip_special_tokens=True)[0]
print("\nGenerated Text:\n", generated_text)

# Compute Latency
latency = (end_time - start_time) * 1000  # Convert to ms
print(f"\nLatency: {latency:.2f} ms")

# Compute Throughput
num_tokens = outputs.shape[1]  # Count generated tokens
throughput = num_tokens / (end_time - start_time)
print(f"Throughput: {throughput:.2f} tokens/sec")

# GPU Memory Usage
if torch.cuda.is_available():
    allocated_memory = torch.cuda.memory_allocated(device) / (1024 **
2)  # MB
    reserved_memory = torch.cuda.memory_reserved(device) / (1024 ** 2)
# MB
    print(f"Allocated GPU Memory: {allocated_memory:.2f} MB")
    print(f"Reserved GPU Memory: {reserved_memory:.2f} MB")

# CPU Memory Usage
memory_usage = psutil.Process().memory_info().rss / (1024 ** 2)  # MB
print(f"CPU Memory Usage: {memory_usage:.2f} MB")
```

```
Enter your prompt: what is machine learning

Generated Text:
 what is machine learning?

* Machine learning is the use of algorithms to allow computers to
learn without being explicitly programmed.

* Machine learning is the use of algorithms to allow computers to
learn without being explicitly programmed.

* Machine learning is the use of algorithms to allow computers to
learn without being explicitly programmed.

* Machine learning is the use of algorithms to allow computers to
learn without being explicitly programmed.
```

```
* Machine learning is the use of algorithms to allow computers to
learn without being explicitly programmed.

* Machine learning is the use of algorithms to allow computers to
learn without being explicitly programmed.

* Machine learning is the use of algorithms to allow computers to
learn without being explicitly programmed.

* Machine learning is the use of algorithms to allow computers to
learn without being explicitly programmed.

* Machine learning is the use of algorithms to allow computers to
learn without being explicitly programmed.

* Machine learning is the use of algorithms to allow computers to
learn without being explicitly programmed.

* Machine learning is the use of algorithms

Latency: 8388.83 ms
Throughput: 24.44 tokens/sec
Allocated GPU Memory: 2958.53 MB
Reserved GPU Memory: 3984.00 MB
CPU Memory Usage: 1588.25 MB
```

```python
from vllm import LLM, SamplingParams

# Load the LLM
llm = LLM(model="google/gemma-2b")

# Define parameters
sampling_params = SamplingParams(temperature=0.7, max_tokens=50)

# Run inference
prompt = "Explain AI inference optimization."
outputs = llm.generate(prompt, sampling_params)

# Print output
print("vLLM Output:", outputs[0].outputs[0].text)
from vllm import LLM, SamplingParams

# Load the LLM
llm = LLM(model="google/gemma-2b")

# Define parameters
sampling_params = SamplingParams(temperature=0.7, max_tokens=50)

# Run inference
prompt = "Explain AI inference optimization."
outputs = llm.generate(prompt, sampling_params)
```

```
# Print output
print("vLLM Output:", outputs[0].outputs[0].text)
```