```
!pip install transformers torch datasets soundfile librosa

Requirement already satisfied: transformers in
/usr/local/lib/python3.11/dist-packages (4.48.3)
Requirement already satisfied: torch in
/usr/local/lib/python3.11/dist-packages (2.5.1+cu124)
Collecting datasets
  Downloading datasets-3.3.2-py3-none-any.whl.metadata (19 kB)
Requirement already satisfied: soundfile in
/usr/local/lib/python3.11/dist-packages (0.13.1)
Requirement already satisfied: librosa in
/usr/local/lib/python3.11/dist-packages (0.10.2.post1)
Requirement already satisfied: filelock in
/usr/local/lib/python3.11/dist-packages (from transformers) (3.17.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.24.0 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: typing-extensions>=4.8.0 in
/usr/local/lib/python3.11/dist-packages (from torch) (4.12.2)
Requirement already satisfied: networkx in
/usr/local/lib/python3.11/dist-packages (from torch) (3.4.2)
Requirement already satisfied: jinja2 in
/usr/local/lib/python3.11/dist-packages (from torch) (3.1.5)
Requirement already satisfied: fsspec in
/usr/local/lib/python3.11/dist-packages (from torch) (2024.10.0)
Collecting nvidia-cuda-nvrtc-cu12==12.4.127 (from torch)
  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)
  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)
  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)
  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)
  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)
  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)
  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)
  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)
  Downloading nvidia_cusparse_cu12-12.3.1.170-py3-none-
manylinux2014_x86_64.whl.metadata (1.6 kB)
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in
/usr/local/lib/python3.11/dist-packages (from torch) (2.21.5)
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in
/usr/local/lib/python3.11/dist-packages (from torch) (12.4.127)
Collecting nvidia-nvjitlink-cu12==12.4.127 (from torch)
  Downloading nvidia_nvjitlink_cu12-12.4.127-py3-none-
manylinux2014_x86_64.whl.metadata (1.5 kB)
Requirement already satisfied: triton==3.1.0 in
/usr/local/lib/python3.11/dist-packages (from torch) (3.1.0)
Requirement already satisfied: sympy==1.13.1 in
/usr/local/lib/python3.11/dist-packages (from torch) (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)
(1.3.0)
Requirement already satisfied: pyarrow>=15.0.0 in
/usr/local/lib/python3.11/dist-packages (from datasets) (18.1.0)
Collecting dill<0.3.9,>=0.3.0 (from datasets)
  Downloading dill-0.3.8-py3-none-any.whl.metadata (10 kB)
Requirement already satisfied: pandas in
/usr/local/lib/python3.11/dist-packages (from datasets) (2.2.2)
Collecting xxhash (from datasets)
  Downloading xxhash-3.5.0-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (12 kB)
Collecting multiprocess<0.70.17 (from datasets)
  Downloading multiprocess-0.70.16-py311-none-any.whl.metadata (7.2
kB)
Requirement already satisfied: aiohttp in
/usr/local/lib/python3.11/dist-packages (from datasets) (3.11.13)
Requirement already satisfied: cffi>=1.0 in
/usr/local/lib/python3.11/dist-packages (from soundfile) (1.17.1)
Requirement already satisfied: audioread>=2.1.9 in
```

```
/usr/local/lib/python3.11/dist-packages (from librosa) (3.0.1)
Requirement already satisfied: scipy>=1.2.0 in
/usr/local/lib/python3.11/dist-packages (from librosa) (1.13.1)
Requirement already satisfied: scikit-learn>=0.20.0 in
/usr/local/lib/python3.11/dist-packages (from librosa) (1.6.1)
Requirement already satisfied: joblib>=0.14 in
/usr/local/lib/python3.11/dist-packages (from librosa) (1.4.2)
Requirement already satisfied: decorator>=4.3.0 in
/usr/local/lib/python3.11/dist-packages (from librosa) (4.4.2)
Requirement already satisfied: numba>=0.51.0 in
/usr/local/lib/python3.11/dist-packages (from librosa) (0.60.0)
Requirement already satisfied: pooch>=1.1 in
/usr/local/lib/python3.11/dist-packages (from librosa) (1.8.2)
Requirement already satisfied: soxr>=0.3.2 in
/usr/local/lib/python3.11/dist-packages (from librosa) (0.5.0.post1)
Requirement already satisfied: lazy-loader>=0.1 in
/usr/local/lib/python3.11/dist-packages (from librosa) (0.4)
Requirement already satisfied: msgpack>=1.0 in
/usr/local/lib/python3.11/dist-packages (from librosa) (1.1.0)
Requirement already satisfied: pycparser in
/usr/local/lib/python3.11/dist-packages (from cffi>=1.0->soundfile)
(2.22)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets)
(2.4.6)
Requirement already satisfied: aiosignal>=1.1.2 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets)
(1.3.2)
Requirement already satisfied: attrs>=17.3.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets)
(25.1.0)
Requirement already satisfied: frozenlist>=1.1.1 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets)
(1.5.0)
Requirement already satisfied: multidict<7.0,>=4.5 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets)
(6.1.0)
Requirement already satisfied: propcache>=0.2.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets)
(0.3.0)
Requirement already satisfied: yarl<2.0,>=1.17.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets)
(1.18.3)
Requirement already satisfied: llvmlite<0.44,>=0.43.0dev0 in
/usr/local/lib/python3.11/dist-packages (from numba>=0.51.0->librosa)
(0.43.0)
Requirement already satisfied: platformdirs>=2.5.0 in
/usr/local/lib/python3.11/dist-packages (from pooch>=1.1->librosa)
(4.3.6)
```

```
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: threadpoolctl>=3.1.0 in
/usr/local/lib/python3.11/dist-packages (from scikit-learn>=0.20.0-
>librosa) (3.5.0)
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.11/dist-packages (from jinja2->torch) (3.0.2)
Requirement already satisfied: python-dateutil>=2.8.2 in
/usr/local/lib/python3.11/dist-packages (from pandas->datasets)
(2.8.2)
Requirement already satisfied: pytz>=2020.1 in
/usr/local/lib/python3.11/dist-packages (from pandas->datasets)
(2025.1)
Requirement already satisfied: tzdata>=2022.7 in
/usr/local/lib/python3.11/dist-packages (from pandas->datasets)
(2025.1)
Requirement already satisfied: six>=1.5 in
/usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2-
>pandas->datasets) (1.17.0)
Downloading nvidia_cublas_cu12-12.4.5.8-py3-none-
manylinux2014_x86_64.whl (363.4 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 363.4/363.4 MB 3.8 MB/s eta
0:00:00
anylinux2014_x86_64.whl (13.8 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 13.8/13.8 MB 43.7 MB/s eta
0:00:00
anylinux2014_x86_64.whl (24.6 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 24.6/24.6 MB 67.5 MB/s eta
0:00:00
e_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl (883 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 883.7/883.7 kB 42.4 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 6.2 MB/s eta
0:00:00
anylinux2014_x86_64.whl (56.3 MB)
```

```
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 56.3/56.3 MB 12.1 MB/s eta
0:00:00
anylinux2014_x86_64.whl (127.9 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 127.9/127.9 MB 7.4 MB/s eta
0:00:00
anylinux2014_x86_64.whl (207.5 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 207.5/207.5 MB 5.3 MB/s eta
0:00:00
anylinux2014_x86_64.whl (21.1 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 21.1/21.1 MB 100.5 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 485.4/485.4 kB 35.7 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 116.3/116.3 kB 11.8 MB/s eta
0:00:00
ultiprocess-0.70.16-py311-none-any.whl (143 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 143.5/143.5 kB 14.5 MB/s eta
0:00:00
anylinux_2_17_x86_64.manylinux2014_x86_64.whl (194 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 194.8/194.8 kB 18.4 MB/s eta
0:00:00
e-cu12, nvidia-cuda-nvrtc-cu12, nvidia-cuda-cupti-cu12, nvidia-cublas-
cu12, dill, nvidia-cusparse-cu12, nvidia-cudnn-cu12, multiprocess,
nvidia-cusolver-cu12, datasets
  Attempting uninstall: nvidia-nvjitlink-cu12
    Found existing installation: nvidia-nvjitlink-cu12 12.5.82
    Uninstalling nvidia-nvjitlink-cu12-12.5.82:
      Successfully uninstalled nvidia-nvjitlink-cu12-12.5.82
  Attempting uninstall: nvidia-curand-cu12
    Found existing installation: nvidia-curand-cu12 10.3.6.82
    Uninstalling nvidia-curand-cu12-10.3.6.82:
      Successfully uninstalled nvidia-curand-cu12-10.3.6.82
  Attempting uninstall: nvidia-cufft-cu12
    Found existing installation: nvidia-cufft-cu12 11.2.3.61
    Uninstalling nvidia-cufft-cu12-11.2.3.61:
      Successfully uninstalled nvidia-cufft-cu12-11.2.3.61
  Attempting uninstall: nvidia-cuda-runtime-cu12
    Found existing installation: nvidia-cuda-runtime-cu12 12.5.82
    Uninstalling nvidia-cuda-runtime-cu12-12.5.82:
      Successfully uninstalled nvidia-cuda-runtime-cu12-12.5.82
  Attempting uninstall: nvidia-cuda-nvrtc-cu12
    Found existing installation: nvidia-cuda-nvrtc-cu12 12.5.82
    Uninstalling nvidia-cuda-nvrtc-cu12-12.5.82:
      Successfully uninstalled nvidia-cuda-nvrtc-cu12-12.5.82
  Attempting uninstall: nvidia-cuda-cupti-cu12
    Found existing installation: nvidia-cuda-cupti-cu12 12.5.82
    Uninstalling nvidia-cuda-cupti-cu12-12.5.82:
      Successfully uninstalled nvidia-cuda-cupti-cu12-12.5.82
  Attempting uninstall: nvidia-cublas-cu12
```

```
    Found existing installation: nvidia-cublas-cu12 12.5.3.2
    Uninstalling nvidia-cublas-cu12-12.5.3.2:
      Successfully uninstalled nvidia-cublas-cu12-12.5.3.2
  Attempting uninstall: nvidia-cusparse-cu12
    Found existing installation: nvidia-cusparse-cu12 12.5.1.3
    Uninstalling nvidia-cusparse-cu12-12.5.1.3:
      Successfully uninstalled nvidia-cusparse-cu12-12.5.1.3
  Attempting uninstall: nvidia-cudnn-cu12
    Found existing installation: nvidia-cudnn-cu12 9.3.0.75
    Uninstalling nvidia-cudnn-cu12-9.3.0.75:
      Successfully uninstalled nvidia-cudnn-cu12-9.3.0.75
  Attempting uninstall: nvidia-cusolver-cu12
    Found existing installation: nvidia-cusolver-cu12 11.6.3.83
    Uninstalling nvidia-cusolver-cu12-11.6.3.83:
      Successfully uninstalled nvidia-cusolver-cu12-11.6.3.83
Successfully installed datasets-3.3.2 dill-0.3.8 multiprocess-0.70.16
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-nvjitlink-cu12-12.4.127 xxhash-3.5.0

!pip install -U transformers accelerate bitsandbytes

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 2.7 MB/s eta
0:00:00
ent 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)
Collecting bitsandbytes
  Downloading bitsandbytes-0.45.3-py3-none-
manylinux_2_24_x86_64.whl.metadata (5.0 kB)
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: psutil in
/usr/local/lib/python3.11/dist-packages (from accelerate) (5.9.5)
Requirement already satisfied: torch>=2.0.0 in
/usr/local/lib/python3.11/dist-packages (from accelerate)
(2.5.1+cu124)
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: 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 transformers-4.49.0-py3-none-any.whl (10.0 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 10.0/10.0 MB 52.6 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 342.1/342.1 kB 14.9 MB/s eta
0:00:00
anylinux_2_24_x86_64.whl (76.1 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 76.1/76.1 MB 10.3 MB/s eta
0:00:00
ers, bitsandbytes, 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 bitsandbytes-0.45.3
transformers-4.49.0
```

```
import torch
from transformers import WhisperProcessor,
WhisperForConditionalGeneration, BitsAndBytesConfig
import time
import psutil
import os
from datasets import load_dataset

device = "cuda" if torch.cuda.is_available() else "cpu"
model_size = "medium"
processor = WhisperProcessor.from_pretrained(f"openai/whisper-
{model_size}")
model =
WhisperForConditionalGeneration.from_pretrained(f"openai/whisper-
{model_size}").to(device)
```

```
/usr/local/lib/python3.11/dist-packages/huggingface_hub/utils/
_auth.py:94: UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your
settings tab (https://huggingface.co/settings/tokens), set it as
secret in your Google Colab and restart your session.
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to
access public models or datasets.
  warnings.warn(
```


```
{"model_id":"fca3063fef2f4104bc0559e6bdd26847","version_major":2,"vers
ion_minor":0}

{"model_id":"d5673dfda25a4f878da90799b905122c","version_major":2,"vers
ion_minor":0}

{"model_id":"8139845dfdf64578965a6a21f6e35fd4","version_major":2,"vers
ion_minor":0}

{"model_id":"3ea85c0fcb8b4284b9c5b748aaa7db38","version_major":2,"vers
ion_minor":0}

{"model_id":"c08f105363324ecba017320676679700","version_major":2,"vers
ion_minor":0}

{"model_id":"27c13459bed848119cf8fdfb075c9f52","version_major":2,"vers
ion_minor":0}
```

{"model_id":"7b5c055d39e64928ab236835a0b89a23","version_major":2,"version_minor":0}

{"model_id":"fc1d3ae184d0424bb13a5559169583e9","version_major":2,"version_minor":0}

{"model_id":"91fc8fcedec64c3ab88af374d3798620","version_major":2,"version_minor":0}

{"model_id":"3bb9eb3ea352447980ad79aaff39a00d","version_major":2,"version_minor":0}

{"model_id":"d1523b15cb134d5fabcb2fe0f19d098f","version_major":2,"version_minor":0}

Downloading the entire dataset takes around 30 mins, so we use a single audio file for benchmarking

```
from datasets import load_dataset #datasets is a library in the huggingface ecosystem

dataset = load_dataset(
    "librispeech_asr",
    "clean",
    split="validation",
    streaming=True,#we don wanna access the whole dataset at once ,but rather access it piece by piece when needed
    trust_remote_code=True, #this is a flag, that we acknowledge the risk and willing to run the coee
    #Why Code? Datasets need special instructions to load correctly, like how to read their files or organize their data.
)
sample = next(iter(dataset))
sample_audio = sample["audio"]["array"]
sample_rate = sample["audio"]["sampling_rate"]
```

{"model_id":"e7a7c7da69ce4b98a2df1285992a60ad","version_major":2,"version_minor":0}

{"model_id":"706102a6b40a43eab061ec6368e69a3f","version_major":2,"version_minor":0}

the audio files chosen are random, that is each time a different file will be present at position 0

```
def transcribe(audio, processor, model, device):
    input_features = processor(audio, sampling_rate=sample_rate, return_tensors="pt").input_features.to(device)
    input_features = input_features.to(torch.float16)
    start_time = time.time()
    predicted_ids = model.generate(input_features)
```

```
    transcription = processor.batch_decode(predicted_ids,
skip_special_tokens=True)
    end_time = time.time()
    latency = end_time - start_time
    return transcription, latency

def get_cpu_memory_usage():
    process = psutil.Process()
    memory_info = process.memory_info()
    return memory_info.rss / (1024 * 1024)
```

QUANTIZATION

```
# INT8 Quantization (CPU)
if device == "cpu":
    quantized_model_int8 = torch.quantization.quantize_dynamic(
        model, {torch.nn.Linear}, dtype=torch.qint8
    )
    transcription_int8, latency_int8 = transcribe(sample_audio,
processor, quantized_model_int8, device)
    cpu_memory_int8 = get_cpu_memory_usage()
    print(f"INT8 Quantized CPU Transcription:
{transcription_int8[0]}")
    print(f"INT8 Quantized CPU Latency: {latency_int8:.4f} seconds")
    print(f"INT8 Quantized CPU Memory: {cpu_memory_int8:.4f} MB")

# 4-bit Quantization (GPU)
if device == "cuda":
    quantization_config = BitsAndBytesConfig(
        load_in_4bit=True,
        bnb_4bit_compute_dtype=torch.float16,
        bnb_4bit_quant_type="nf4",
    )
    quantized_model_4bit =
WhisperForConditionalGeneration.from_pretrained(
        f"openai/whisper-{model_size}",
quantization_config=quantization_config
    ).to(device)
    transcription_4bit, latency_4bit = transcribe(sample_audio,
processor, quantized_model_4bit, device)
    gpu_peak_memory = torch.cuda.max_memory_allocated(device=device)
    print(f"4-bit Quantized GPU Transcription:
{transcription_4bit[0]}")
    print(f"4-bit Quantized GPU Latency: {latency_4bit:.4f} seconds")
    print(f"4-bit Quantized GPU Peak Memory: {gpu_peak_memory / (1024
* 1024):.2f} MB")

`low_cpu_mem_usage` was None, now default to True since model is
quantized.
```

```
4-bit Quantized GPU Transcription:  He was in a fevered state of mind, owing to the blight his wife's action threatened to cast upon his entire future.
4-bit Quantized GPU Latency: 2.6260 seconds
4-bit Quantized GPU Peak Memory: 4301.86 MB
```

```
num_runs = 10
latencies = []
cpu_memory_usages = []

if device == "cuda":
    torch.cuda.reset_peak_memory_stats(device=device)
    start_memory = torch.cuda.memory_allocated(device=device)

    for _ in range(num_runs):
        transcription, latency = transcribe(sample_audio, processor, model, device)
        latencies.append(latency)

    end_memory = torch.cuda.memory_allocated(device=device)
    peak_memory = torch.cuda.max_memory_allocated(device=device)
    memory_usage = end_memory - start_memory

    avg_latency = sum(latencies) / num_runs
    print(f"GPU Transcription: {transcription[0]}")
    print(f"GPU Average Latency: {avg_latency:.4f} seconds")
    print(f"GPU Memory Usage: {memory_usage / (1024 * 1024):.2f} MB")
    print(f"GPU Peak Memory Usage: {peak_memory / (1024 * 1024):.2f} MB")

else:
    for _ in range(num_runs):
        start_cpu_memory = get_cpu_memory_usage()
        transcription, latency = transcribe(sample_audio, processor, model, device)
        end_cpu_memory = get_cpu_memory_usage()
        cpu_memory_usages.append(end_cpu_memory - start_cpu_memory)
        latencies.append(latency)

    avg_latency = sum(latencies) / num_runs
    avg_cpu_memory = sum(cpu_memory_usages) / num_runs

    print(f"CPU Transcription: {transcription[0]}")
    print(f"CPU Average Latency: {avg_latency:.4f} seconds")
    print(f"CPU Average Memory Usage: {avg_cpu_memory:.2f} MB")
```

```
GPU Transcription:  He was in a fevered state of mind, owing to the blight his wife's action threatened to cast upon his entire future.
GPU Average Latency: 1.3950 seconds
```

```
GPU Memory Usage: 0.00 MB
GPU Peak Memory Usage: 4103.10 MB
```

FOR BASELINE TRANSRIPTION INPUT SHOULD BE float16 FOR 4 BIT GPU QUANTIZATION INPUT SHOULD BE float32

```
!pip install auto-gptq
```

```
Collecting auto-gptq
  Downloading auto_gptq-0.7.1-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (18 kB)
Requirement already satisfied: accelerate>=0.26.0 in
/usr/local/lib/python3.11/dist-packages (from auto-gptq) (1.4.0)
Requirement already satisfied: datasets in
/usr/local/lib/python3.11/dist-packages (from auto-gptq) (3.3.2)
Requirement already satisfied: sentencepiece in
/usr/local/lib/python3.11/dist-packages (from auto-gptq) (0.2.0)
Requirement already satisfied: numpy in
/usr/local/lib/python3.11/dist-packages (from auto-gptq) (1.26.4)
Collecting rouge (from auto-gptq)
  Downloading rouge-1.0.1-py3-none-any.whl.metadata (4.1 kB)
Collecting gekko (from auto-gptq)
  Downloading gekko-1.2.1-py3-none-any.whl.metadata (3.0 kB)
Requirement already satisfied: torch>=1.13.0 in
/usr/local/lib/python3.11/dist-packages (from auto-gptq) (2.5.1+cu124)
Requirement already satisfied: safetensors in
/usr/local/lib/python3.11/dist-packages (from auto-gptq) (0.5.3)
Requirement already satisfied: transformers>=4.31.0 in
/usr/local/lib/python3.11/dist-packages (from auto-gptq) (4.49.0)
Requirement already satisfied: peft>=0.5.0 in
/usr/local/lib/python3.11/dist-packages (from auto-gptq) (0.14.0)
Requirement already satisfied: tqdm in /usr/local/lib/python3.11/dist-
packages (from auto-gptq) (4.67.1)
Requirement already satisfied: packaging>=20.0 in
/usr/local/lib/python3.11/dist-packages (from accelerate>=0.26.0-
>auto-gptq) (24.2)
Requirement already satisfied: psutil in
/usr/local/lib/python3.11/dist-packages (from accelerate>=0.26.0-
>auto-gptq) (5.9.5)
Requirement already satisfied: pyyaml in
/usr/local/lib/python3.11/dist-packages (from accelerate>=0.26.0-
>auto-gptq) (6.0.2)
Requirement already satisfied: huggingface-hub>=0.21.0 in
/usr/local/lib/python3.11/dist-packages (from accelerate>=0.26.0-
>auto-gptq) (0.28.1)
Requirement already satisfied: filelock in
/usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->auto-
gptq) (3.17.0)
Requirement already satisfied: typing-extensions>=4.8.0 in
```

```
/usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->auto-
gptq) (4.12.2)
Requirement already satisfied: networkx in
/usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->auto-
gptq) (3.4.2)
Requirement already satisfied: jinja2 in
/usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->auto-
gptq) (3.1.5)
Requirement already satisfied: fsspec in
/usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->auto-
gptq) (2024.10.0)
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.4.127 in
/usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->auto-
gptq) (12.4.127)
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.4.127
in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->auto-
gptq) (12.4.127)
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.4.127 in
/usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->auto-
gptq) (12.4.127)
Requirement already satisfied: nvidia-cudnn-cu12==9.1.0.70 in
/usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->auto-
gptq) (9.1.0.70)
Requirement already satisfied: nvidia-cublas-cu12==12.4.5.8 in
/usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->auto-
gptq) (12.4.5.8)
Requirement already satisfied: nvidia-cufft-cu12==11.2.1.3 in
/usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->auto-
gptq) (11.2.1.3)
Requirement already satisfied: nvidia-curand-cu12==10.3.5.147 in
/usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->auto-
gptq) (10.3.5.147)
Requirement already satisfied: nvidia-cusolver-cu12==11.6.1.9 in
/usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->auto-
gptq) (11.6.1.9)
Requirement already satisfied: nvidia-cusparse-cu12==12.3.1.170 in
/usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->auto-
gptq) (12.3.1.170)
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in
/usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->auto-
gptq) (2.21.5)
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in
/usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->auto-
gptq) (12.4.127)
Requirement already satisfied: nvidia-nvjitlink-cu12==12.4.127 in
/usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->auto-
gptq) (12.4.127)
Requirement already satisfied: triton==3.1.0 in
/usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->auto-
```

```
gptq) (3.1.0)
Requirement already satisfied: sympy==1.13.1 in
/usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->auto-
gptq) (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>=1.13.0->auto-gptq) (1.3.0)
Requirement already satisfied: regex!=2019.12.17 in
/usr/local/lib/python3.11/dist-packages (from transformers>=4.31.0-
>auto-gptq) (2024.11.6)
Requirement already satisfied: requests in
/usr/local/lib/python3.11/dist-packages (from transformers>=4.31.0-
>auto-gptq) (2.32.3)
Requirement already satisfied: tokenizers<0.22,>=0.21 in
/usr/local/lib/python3.11/dist-packages (from transformers>=4.31.0-
>auto-gptq) (0.21.0)
Requirement already satisfied: pyarrow>=15.0.0 in
/usr/local/lib/python3.11/dist-packages (from datasets->auto-gptq)
(18.1.0)
Requirement already satisfied: dill<0.3.9,>=0.3.0 in
/usr/local/lib/python3.11/dist-packages (from datasets->auto-gptq)
(0.3.8)
Requirement already satisfied: pandas in
/usr/local/lib/python3.11/dist-packages (from datasets->auto-gptq)
(2.2.2)
Requirement already satisfied: xxhash in
/usr/local/lib/python3.11/dist-packages (from datasets->auto-gptq)
(3.5.0)
Requirement already satisfied: multiprocess<0.70.17 in
/usr/local/lib/python3.11/dist-packages (from datasets->auto-gptq)
(0.70.16)
Requirement already satisfied: aiohttp in
/usr/local/lib/python3.11/dist-packages (from datasets->auto-gptq)
(3.11.13)
Requirement already satisfied: six in /usr/local/lib/python3.11/dist-
packages (from rouge->auto-gptq) (1.17.0)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets->auto-
gptq) (2.4.6)
Requirement already satisfied: aiosignal>=1.1.2 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets->auto-
gptq) (1.3.2)
Requirement already satisfied: attrs>=17.3.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets->auto-
gptq) (25.1.0)
Requirement already satisfied: frozenlist>=1.1.1 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets->auto-
gptq) (1.5.0)
Requirement already satisfied: multidict<7.0,>=4.5 in
```

```
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets->auto-
gptq) (6.1.0)
Requirement already satisfied: propcache>=0.2.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets->auto-
gptq) (0.3.0)
Requirement already satisfied: yarl<2.0,>=1.17.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets->auto-
gptq) (1.18.3)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.11/dist-packages (from requests-
>transformers>=4.31.0->auto-gptq) (3.4.1)
Requirement already satisfied: idna<4,>=2.5 in
/usr/local/lib/python3.11/dist-packages (from requests-
>transformers>=4.31.0->auto-gptq) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.11/dist-packages (from requests-
>transformers>=4.31.0->auto-gptq) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.11/dist-packages (from requests-
>transformers>=4.31.0->auto-gptq) (2025.1.31)
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.11/dist-packages (from jinja2->torch>=1.13.0-
>auto-gptq) (3.0.2)
Requirement already satisfied: python-dateutil>=2.8.2 in
/usr/local/lib/python3.11/dist-packages (from pandas->datasets->auto-
gptq) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in
/usr/local/lib/python3.11/dist-packages (from pandas->datasets->auto-
gptq) (2025.1)
Requirement already satisfied: tzdata>=2022.7 in
/usr/local/lib/python3.11/dist-packages (from pandas->datasets->auto-
gptq) (2025.1)
Downloading auto_gptq-0.7.1-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl (23.5 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 23.5/23.5 MB 84.6 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 13.2/13.2 MB 103.7 MB/s eta
0:00:00

from auto_gptq import AutoGPTQForCausalLM, BaseQuantizeConfig

quantize_config = BaseQuantizeConfig(bits=4, group_size=128)  #
Example settings

# Load the model and quantize it
model_gptq = AutoGPTQForCausalLM.from_pretrained(
    f"openai/whisper-{model_size}", quantize_config
)

#
```

```
/usr/local/lib/python3.11/dist-packages/auto_gptq/nn_modules/
triton_utils/kernels.py:410: FutureWarning:
`torch.cuda.amp.custom_fwd(args...)` is deprecated. Please use
`torch.amp.custom_fwd(args..., device_type='cuda')` instead.
  @custom_fwd
/usr/local/lib/python3.11/dist-packages/auto_gptq/nn_modules/triton_ut
ils/kernels.py:418: FutureWarning:
`torch.cuda.amp.custom_bwd(args...)` is deprecated. Please use
`torch.amp.custom_bwd(args..., device_type='cuda')` instead.
  @custom_bwd
/usr/local/lib/python3.11/dist-packages/auto_gptq/nn_modules/triton_ut
ils/kernels.py:461: FutureWarning:
`torch.cuda.amp.custom_fwd(args...)` is deprecated. Please use
`torch.amp.custom_fwd(args..., device_type='cuda')` instead.
  @custom_fwd(cast_inputs=torch.float16)
WARNING:auto_gptq.nn_modules.qlinear.qlinear_cuda:CUDA extension not
installed.
WARNING:auto_gptq.nn_modules.qlinear.qlinear_cuda_old:CUDA extension
not installed.

---------------------------------------------------------------------
-----
TypeError                                 Traceback (most recent call
last)
<ipython-input-18-9c91091317d4> in <cell line: 0>()
      4
      5 # Load the model and quantize it
----> 6 model_gptq = AutoGPTQForCausalLM.from_pretrained(
      7     f"openai/whisper-{model_size}", quantize_config
      8 )

/usr/local/lib/python3.11/dist-packages/auto_gptq/modeling/auto.py in
from_pretrained(cls, pretrained_model_name_or_path, quantize_config,
max_memory, trust_remote_code, **model_init_kwargs)
     73         **model_init_kwargs,
     74     ) -> BaseGPTQForCausalLM:
---> 75         model_type =
check_and_get_model_type(pretrained_model_name_or_path,
trust_remote_code)
     76         return
GPTQ_CAUSAL_LM_MODEL_MAP[model_type].from_pretrained(
     77
pretrained_model_name_or_path=pretrained_model_name_or_path,

/usr/local/lib/python3.11/dist-packages/auto_gptq/modeling/_utils.py
in check_and_get_model_type(model_dir, trust_remote_code)
    303     config = AutoConfig.from_pretrained(model_dir,
trust_remote_code=trust_remote_code)
    304     if config.model_type not in SUPPORTED_MODELS:
--> 305         raise TypeError(f"{config.model_type} isn't supported
```

```
yet.")
    306     model_type = config.model_type
    307     return model_type

TypeError: whisper isn't supported yet.
```

Not all quantization libraries support all model architectures. GPTQ:auto-gptq library Does not support Whisper yet

AWSQ

```
!pip install autoawq

from awq import AutoAWQForCausalLM

Collecting autoawq
  Downloading autoawq-0.2.8.tar.gz (71 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0.0/71.6 kB ? eta -:--:--
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 61.4/71.6 kB 78.3 MB/s eta
0:00:01 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 71.6/71.6 kB 1.5 MB/s
eta 0:00:00
etadata (setup.py) ... ent already satisfied: torch>=2.5.1 in
/usr/local/lib/python3.11/dist-packages (from autoawq) (2.5.1+cu124)
Requirement already satisfied: triton in
/usr/local/lib/python3.11/dist-packages (from autoawq) (3.1.0)
Collecting transformers<=4.47.1,>=4.45.0 (from autoawq)
  Downloading transformers-4.47.1-py3-none-any.whl.metadata (44 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 44.1/44.1 kB 1.7 MB/s eta
0:00:00
ent already satisfied: tokenizers>=0.12.1 in
/usr/local/lib/python3.11/dist-packages (from autoawq) (0.21.0)
Requirement already satisfied: typing_extensions>=4.8.0 in
/usr/local/lib/python3.11/dist-packages (from autoawq) (4.12.2)
Requirement already satisfied: accelerate in
/usr/local/lib/python3.11/dist-packages (from autoawq) (1.3.0)
Collecting datasets>=2.20 (from autoawq)
  Downloading datasets-3.3.2-py3-none-any.whl.metadata (19 kB)
Requirement already satisfied: zstandard in
/usr/local/lib/python3.11/dist-packages (from autoawq) (0.23.0)
Requirement already satisfied: huggingface_hub>=0.26.5 in
/usr/local/lib/python3.11/dist-packages (from autoawq) (0.28.1)
Requirement already satisfied: filelock in
/usr/local/lib/python3.11/dist-packages (from datasets>=2.20->autoawq)
(3.17.0)
Requirement already satisfied: numpy>=1.17 in
/usr/local/lib/python3.11/dist-packages (from datasets>=2.20->autoawq)
(1.26.4)
Requirement already satisfied: pyarrow>=15.0.0 in
```

```
/usr/local/lib/python3.11/dist-packages (from datasets>=2.20->autoawq)
(18.1.0)
Collecting dill<0.3.9,>=0.3.0 (from datasets>=2.20->autoawq)
  Downloading dill-0.3.8-py3-none-any.whl.metadata (10 kB)
Requirement already satisfied: pandas in
/usr/local/lib/python3.11/dist-packages (from datasets>=2.20->autoawq)
(2.2.2)
Requirement already satisfied: requests>=2.32.2 in
/usr/local/lib/python3.11/dist-packages (from datasets>=2.20->autoawq)
(2.32.3)
Requirement already satisfied: tqdm>=4.66.3 in
/usr/local/lib/python3.11/dist-packages (from datasets>=2.20->autoawq)
(4.67.1)
Collecting xxhash (from datasets>=2.20->autoawq)
  Downloading xxhash-3.5.0-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (12 kB)
Collecting multiprocess<0.70.17 (from datasets>=2.20->autoawq)
  Downloading multiprocess-0.70.16-py311-none-any.whl.metadata (7.2
kB)
Requirement already satisfied: fsspec<=2024.12.0,>=2023.1.0 in
/usr/local/lib/python3.11/dist-packages (from
fsspec[http]<=2024.12.0,>=2023.1.0->datasets>=2.20->autoawq)
(2024.10.0)
Requirement already satisfied: aiohttp in
/usr/local/lib/python3.11/dist-packages (from datasets>=2.20->autoawq)
(3.11.13)
Requirement already satisfied: packaging in
/usr/local/lib/python3.11/dist-packages (from datasets>=2.20->autoawq)
(24.2)
Requirement already satisfied: pyyaml>=5.1 in
/usr/local/lib/python3.11/dist-packages (from datasets>=2.20->autoawq)
(6.0.2)
Requirement already satisfied: networkx in
/usr/local/lib/python3.11/dist-packages (from torch>=2.5.1->autoawq)
(3.4.2)
Requirement already satisfied: jinja2 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.5.1->autoawq)
(3.1.5)
Collecting nvidia-cuda-nvrtc-cu12==12.4.127 (from torch>=2.5.1-
>autoawq)
  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>=2.5.1-
>autoawq)
  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>=2.5.1-
>autoawq)
  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>=2.5.1->autoawq)
  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>=2.5.1->autoawq)
  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>=2.5.1->autoawq)
  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>=2.5.1->autoawq)
  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>=2.5.1->autoawq)
  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>=2.5.1-
>autoawq)
  Downloading nvidia_cusparse_cu12-12.3.1.170-py3-none-
manylinux2014_x86_64.whl.metadata (1.6 kB)
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.5.1->autoawq)
(2.21.5)
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.5.1->autoawq)
(12.4.127)
Collecting nvidia-nvjitlink-cu12==12.4.127 (from torch>=2.5.1-
>autoawq)
  Downloading nvidia_nvjitlink_cu12-12.4.127-py3-none-
manylinux2014_x86_64.whl.metadata (1.5 kB)
Requirement already satisfied: sympy==1.13.1 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.5.1->autoawq)
(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>=2.5.1->autoawq) (1.3.0)
Requirement already satisfied: regex!=2019.12.17 in
/usr/local/lib/python3.11/dist-packages (from
transformers<=4.47.1,>=4.45.0->autoawq) (2024.11.6)
Requirement already satisfied: safetensors>=0.4.1 in
/usr/local/lib/python3.11/dist-packages (from
transformers<=4.47.1,>=4.45.0->autoawq) (0.5.3)
Requirement already satisfied: psutil in
/usr/local/lib/python3.11/dist-packages (from accelerate->autoawq)
(5.9.5)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets>=2.20-
>autoawq) (2.4.6)
Requirement already satisfied: aiosignal>=1.1.2 in
```

```
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets>=2.20->autoawq) (1.3.2)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets>=2.20->autoawq) (25.1.0)
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets>=2.20->autoawq) (1.5.0)
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets>=2.20->autoawq) (6.1.0)
Requirement already satisfied: propcache>=0.2.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets>=2.20->autoawq) (0.3.0)
Requirement already satisfied: yarl<2.0,>=1.17.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets>=2.20->autoawq) (1.18.3)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets>=2.20->autoawq) (3.4.1)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets>=2.20->autoawq) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets>=2.20->autoawq) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets>=2.20->autoawq) (2025.1.31)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from jinja2->torch>=2.5.1->autoawq) (3.0.2)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas->datasets>=2.20->autoawq) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from pandas->datasets>=2.20->autoawq) (2025.1)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas->datasets>=2.20->autoawq) (2025.1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->pandas->datasets>=2.20->autoawq) (1.17.0)
Downloading datasets-3.3.2-py3-none-any.whl (485 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 485.4/485.4 kB 21.2 MB/s eta 0:00:00
anylinux2014_x86_64.whl (363.4 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 363.4/363.4 MB 4.3 MB/s eta
```

```
0:00:00
anylinux2014_x86_64.whl (13.8 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 13.8/13.8 MB 82.9 MB/s eta
0:00:00
anylinux2014_x86_64.whl (24.6 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 24.6/24.6 MB 82.3 MB/s eta
0:00:00
e_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl (883 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 883.7/883.7 kB 48.9 MB/s eta
0:00:00
anylinux2014_x86_64.whl (664.8 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 664.8/664.8 MB 2.1 MB/s eta
0:00:00
anylinux2014_x86_64.whl (211.5 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 211.5/211.5 MB 5.5 MB/s eta
0:00:00
anylinux2014_x86_64.whl (56.3 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 56.3/56.3 MB 19.9 MB/s eta
0:00:00
anylinux2014_x86_64.whl (127.9 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 127.9/127.9 MB 7.5 MB/s eta
0:00:00
anylinux2014_x86_64.whl (207.5 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 207.5/207.5 MB 5.6 MB/s eta
0:00:00
anylinux2014_x86_64.whl (21.1 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 21.1/21.1 MB 91.1 MB/s eta
0:00:00
ers-4.47.1-py3-none-any.whl (10.1 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 10.1/10.1 MB 91.7 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 116.3/116.3 kB 10.6 MB/s eta
0:00:00
ultiprocess-0.70.16-py311-none-any.whl (143 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 143.5/143.5 kB 12.7 MB/s eta
0:00:00
anylinux_2_17_x86_64.manylinux2014_x86_64.whl (194 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 194.8/194.8 kB 17.4 MB/s eta
0:00:00
e=autoawq-0.2.8-py3-none-any.whl size=108744
sha256=9b9c98f42cc06b8ccd9969e5b9b197f082d3cdb5aa3ada3bcaab0ae610f204c
5
  Stored in directory:
/root/.cache/pip/wheels/fd/03/fe/99c1c678bfe8aca712186466969ed866f52fe
da95ae1dcd1b1
Successfully built autoawq
Installing collected packages: xxhash, nvidia-nvjitlink-cu12, nvidia-
curand-cu12, nvidia-cufft-cu12, nvidia-cuda-runtime-cu12, nvidia-cuda-
nvrtc-cu12, nvidia-cuda-cupti-cu12, nvidia-cublas-cu12, dill, nvidia-
```

```
cusparse-cu12, nvidia-cudnn-cu12, multiprocess, nvidia-cusolver-cu12,
transformers, datasets, autoawq
  Attempting uninstall: nvidia-nvjitlink-cu12
    Found existing installation: nvidia-nvjitlink-cu12 12.5.82
    Uninstalling nvidia-nvjitlink-cu12-12.5.82:
      Successfully uninstalled nvidia-nvjitlink-cu12-12.5.82
  Attempting uninstall: nvidia-curand-cu12
    Found existing installation: nvidia-curand-cu12 10.3.6.82
    Uninstalling nvidia-curand-cu12-10.3.6.82:
      Successfully uninstalled nvidia-curand-cu12-10.3.6.82
  Attempting uninstall: nvidia-cufft-cu12
    Found existing installation: nvidia-cufft-cu12 11.2.3.61
    Uninstalling nvidia-cufft-cu12-11.2.3.61:
      Successfully uninstalled nvidia-cufft-cu12-11.2.3.61
  Attempting uninstall: nvidia-cuda-runtime-cu12
    Found existing installation: nvidia-cuda-runtime-cu12 12.5.82
    Uninstalling nvidia-cuda-runtime-cu12-12.5.82:
      Successfully uninstalled nvidia-cuda-runtime-cu12-12.5.82
  Attempting uninstall: nvidia-cuda-nvrtc-cu12
    Found existing installation: nvidia-cuda-nvrtc-cu12 12.5.82
    Uninstalling nvidia-cuda-nvrtc-cu12-12.5.82:
      Successfully uninstalled nvidia-cuda-nvrtc-cu12-12.5.82
  Attempting uninstall: nvidia-cuda-cupti-cu12
    Found existing installation: nvidia-cuda-cupti-cu12 12.5.82
    Uninstalling nvidia-cuda-cupti-cu12-12.5.82:
      Successfully uninstalled nvidia-cuda-cupti-cu12-12.5.82
  Attempting uninstall: nvidia-cublas-cu12
    Found existing installation: nvidia-cublas-cu12 12.5.3.2
    Uninstalling nvidia-cublas-cu12-12.5.3.2:
      Successfully uninstalled nvidia-cublas-cu12-12.5.3.2
  Attempting uninstall: nvidia-cusparse-cu12
    Found existing installation: nvidia-cusparse-cu12 12.5.1.3
    Uninstalling nvidia-cusparse-cu12-12.5.1.3:
      Successfully uninstalled nvidia-cusparse-cu12-12.5.1.3
  Attempting uninstall: nvidia-cudnn-cu12
    Found existing installation: nvidia-cudnn-cu12 9.3.0.75
    Uninstalling nvidia-cudnn-cu12-9.3.0.75:
      Successfully uninstalled nvidia-cudnn-cu12-9.3.0.75
  Attempting uninstall: nvidia-cusolver-cu12
    Found existing installation: nvidia-cusolver-cu12 11.6.3.83
    Uninstalling nvidia-cusolver-cu12-11.6.3.83:
      Successfully uninstalled nvidia-cusolver-cu12-11.6.3.83
  Attempting uninstall: transformers
    Found existing installation: transformers 4.48.3
    Uninstalling transformers-4.48.3:
      Successfully uninstalled transformers-4.48.3
Successfully installed autoawq-0.2.8 datasets-3.3.2 dill-0.3.8
multiprocess-0.70.16 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-nvjitlink-cu12-12.4.127 transformers-
4.47.1 xxhash-3.5.0

!git clone https://github.com/mit-han-lab/awq.git
%cd awq
!pip install -r requirements.txt
!pip install .

Cloning into 'awq'...
fatal: could not read Username for 'https://github.com': No such
device or address
[Errno 2] No such file or directory: 'awq'
/content
ERROR: Could not open requirements file: [Errno 2] No such file or
directory: 'requirements.txt'
ERROR: Directory '.' is not installable. Neither 'setup.py' nor
'pyproject.toml' found.

!git clone https://github.com/mit-han-lab/awq.git

Cloning into 'awq'...
fatal: could not read Username for 'https://github.com': No such
device or address

%cd awq
!pip install -r requirements.txt
!pip install .

[Errno 2] No such file or directory: 'awq'
/content
ERROR: Could not open requirements file: [Errno 2] No such file or
directory: 'requirements.txt'
ERROR: Directory '.' is not installable. Neither 'setup.py' nor
'pyproject.toml' found.
```

Could not find the awq library

```
```

VLLM

```
!pip install vllm

Collecting vllm
  Downloading vllm-0.7.3-cp38-abi3-manylinux1_x86_64.whl.metadata (25
kB)
Requirement already satisfied: psutil in
/usr/local/lib/python3.11/dist-packages (from vllm) (5.9.5)
Requirement already satisfied: sentencepiece in
```

```
/usr/local/lib/python3.11/dist-packages (from vllm) (0.2.0)
Requirement already satisfied: numpy<2.0.0 in
/usr/local/lib/python3.11/dist-packages (from vllm) (1.26.4)
Requirement already satisfied: numba==0.60.0 in
/usr/local/lib/python3.11/dist-packages (from vllm) (0.60.0)
Requirement already satisfied: requests>=2.26.0 in
/usr/local/lib/python3.11/dist-packages (from vllm) (2.32.3)
Requirement already satisfied: tqdm in /usr/local/lib/python3.11/dist-
packages (from vllm) (4.67.1)
Collecting blake3 (from vllm)
  Downloading blake3-1.0.4-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (4.2 kB)
Requirement already satisfied: py-cpuinfo in
/usr/local/lib/python3.11/dist-packages (from vllm) (9.0.0)
Requirement already satisfied: transformers>=4.48.2 in
/usr/local/lib/python3.11/dist-packages (from vllm) (4.49.0)
Requirement already satisfied: tokenizers>=0.19.1 in
/usr/local/lib/python3.11/dist-packages (from vllm) (0.21.0)
Requirement already satisfied: protobuf in
/usr/local/lib/python3.11/dist-packages (from vllm) (4.25.6)
Collecting fastapi!=0.113.*,!=0.114.0,>=0.107.0 (from
fastapi[standard]!=0.113.*,!=0.114.0,>=0.107.0; python_version >=
"3.9"->vllm)
  Downloading fastapi-0.115.11-py3-none-any.whl.metadata (27 kB)
Requirement already satisfied: aiohttp in
/usr/local/lib/python3.11/dist-packages (from vllm) (3.11.13)
Requirement already satisfied: openai>=1.52.0 in
/usr/local/lib/python3.11/dist-packages (from vllm) (1.61.1)
Requirement already satisfied: pydantic>=2.9 in
/usr/local/lib/python3.11/dist-packages (from vllm) (2.10.6)
Requirement already satisfied: prometheus_client>=0.18.0 in
/usr/local/lib/python3.11/dist-packages (from vllm) (0.21.1)
Requirement already satisfied: pillow in
/usr/local/lib/python3.11/dist-packages (from vllm) (11.1.0)
Collecting prometheus-fastapi-instrumentator>=7.0.0 (from vllm)
  Downloading prometheus_fastapi_instrumentator-7.0.2-py3-none-
any.whl.metadata (13 kB)
Collecting tiktoken>=0.6.0 (from vllm)
  Downloading tiktoken-0.9.0-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (6.7 kB)
Collecting lm-format-enforcer<0.11,>=0.10.9 (from vllm)
  Downloading lm_format_enforcer-0.10.11-py3-none-any.whl.metadata (17
kB)
Collecting outlines==0.1.11 (from vllm)
  Downloading outlines-0.1.11-py3-none-any.whl.metadata (17 kB)
Collecting lark==1.2.2 (from vllm)
  Downloading lark-1.2.2-py3-none-any.whl.metadata (1.8 kB)
Collecting xgrammar==0.1.11 (from vllm)
  Downloading xgrammar-0.1.11-cp311-cp311-
```

```
manylinux_2_27_x86_64.manylinux_2_28_x86_64.whl.metadata (2.0 kB)
Requirement already satisfied: typing_extensions>=4.10 in
/usr/local/lib/python3.11/dist-packages (from vllm) (4.12.2)
Requirement already satisfied: filelock>=3.16.1 in
/usr/local/lib/python3.11/dist-packages (from vllm) (3.17.0)
Collecting partial-json-parser (from vllm)
  Downloading partial_json_parser-0.2.1.1.post5-py3-none-
any.whl.metadata (6.1 kB)
Requirement already satisfied: pyzmq in
/usr/local/lib/python3.11/dist-packages (from vllm) (24.0.1)
Collecting msgspec (from vllm)
  Downloading msgspec-0.19.0-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (6.9 kB)
Collecting gguf==0.10.0 (from vllm)
  Downloading gguf-0.10.0-py3-none-any.whl.metadata (3.5 kB)
Requirement already satisfied: importlib_metadata in
/usr/local/lib/python3.11/dist-packages (from vllm) (8.6.1)
Collecting mistral_common>=1.5.0 (from mistral_common[opencv]>=1.5.0-
>vllm)
  Downloading mistral_common-1.5.3-py3-none-any.whl.metadata (4.5 kB)
Requirement already satisfied: pyyaml in
/usr/local/lib/python3.11/dist-packages (from vllm) (6.0.2)
Requirement already satisfied: einops in
/usr/local/lib/python3.11/dist-packages (from vllm) (0.8.1)
Collecting compressed-tensors==0.9.1 (from vllm)
  Downloading compressed_tensors-0.9.1-py3-none-any.whl.metadata (6.8
kB)
Collecting depyf==0.18.0 (from vllm)
  Downloading depyf-0.18.0-py3-none-any.whl.metadata (7.1 kB)
Requirement already satisfied: cloudpickle in
/usr/local/lib/python3.11/dist-packages (from vllm) (3.1.1)
Collecting ray==2.40.0 (from ray[adag]==2.40.0->vllm)
  Downloading ray-2.40.0-cp311-cp311-manylinux2014_x86_64.whl.metadata
(17 kB)
Requirement already satisfied: torch==2.5.1 in
/usr/local/lib/python3.11/dist-packages (from vllm) (2.5.1+cu124)
Requirement already satisfied: torchaudio==2.5.1 in
/usr/local/lib/python3.11/dist-packages (from vllm) (2.5.1+cu124)
Requirement already satisfied: torchvision==0.20.1 in
/usr/local/lib/python3.11/dist-packages (from vllm) (0.20.1+cu124)
Collecting xformers==0.0.28.post3 (from vllm)
  Downloading xformers-0.0.28.post3-cp311-cp311-
manylinux_2_28_x86_64.whl.metadata (1.0 kB)
Collecting astor (from depyf==0.18.0->vllm)
  Downloading astor-0.8.1-py2.py3-none-any.whl.metadata (4.2 kB)
Requirement already satisfied: dill in /usr/local/lib/python3.11/dist-
packages (from depyf==0.18.0->vllm) (0.3.8)
Requirement already satisfied: llvmlite<0.44,>=0.43.0dev0 in
/usr/local/lib/python3.11/dist-packages (from numba==0.60.0->vllm)
```

```
(0.43.0)
Collecting interegular (from outlines==0.1.11->vllm)
  Downloading interegular-0.3.3-py37-none-any.whl.metadata (3.0 kB)
Requirement already satisfied: jinja2 in
/usr/local/lib/python3.11/dist-packages (from outlines==0.1.11->vllm)
(3.1.5)
Requirement already satisfied: nest_asyncio in
/usr/local/lib/python3.11/dist-packages (from outlines==0.1.11->vllm)
(1.6.0)
Collecting diskcache (from outlines==0.1.11->vllm)
  Downloading diskcache-5.6.3-py3-none-any.whl.metadata (20 kB)
Requirement already satisfied: referencing in
/usr/local/lib/python3.11/dist-packages (from outlines==0.1.11->vllm)
(0.36.2)
Requirement already satisfied: jsonschema in
/usr/local/lib/python3.11/dist-packages (from outlines==0.1.11->vllm)
(4.23.0)
Collecting pycountry (from outlines==0.1.11->vllm)
  Downloading pycountry-24.6.1-py3-none-any.whl.metadata (12 kB)
Collecting airportsdata (from outlines==0.1.11->vllm)
  Downloading airportsdata-20250224-py3-none-any.whl.metadata (9.0 kB)
Collecting outlines_core==0.1.26 (from outlines==0.1.11->vllm)
  Downloading outlines_core-0.1.26-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (3.8 kB)
Requirement already satisfied: click>=7.0 in
/usr/local/lib/python3.11/dist-packages (from ray==2.40.0-
>ray[adag]==2.40.0->vllm) (8.1.8)
Requirement already satisfied: msgpack<2.0.0,>=1.0.0 in
/usr/local/lib/python3.11/dist-packages (from ray==2.40.0-
>ray[adag]==2.40.0->vllm) (1.1.0)
Requirement already satisfied: packaging in
/usr/local/lib/python3.11/dist-packages (from ray==2.40.0-
>ray[adag]==2.40.0->vllm) (24.2)
Requirement already satisfied: aiosignal in
/usr/local/lib/python3.11/dist-packages (from ray==2.40.0-
>ray[adag]==2.40.0->vllm) (1.3.2)
Requirement already satisfied: frozenlist in
/usr/local/lib/python3.11/dist-packages (from ray==2.40.0-
>ray[adag]==2.40.0->vllm) (1.5.0)
Requirement already satisfied: cupy-cuda12x in
/usr/local/lib/python3.11/dist-packages (from ray[adag]==2.40.0->vllm)
(13.3.0)
Requirement already satisfied: networkx in
/usr/local/lib/python3.11/dist-packages (from torch==2.5.1->vllm)
(3.4.2)
Requirement already satisfied: fsspec in
/usr/local/lib/python3.11/dist-packages (from torch==2.5.1->vllm)
(2024.10.0)
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.4.127 in
```

```
/usr/local/lib/python3.11/dist-packages (from torch==2.5.1->vllm)
(12.4.127)
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.4.127
in /usr/local/lib/python3.11/dist-packages (from torch==2.5.1->vllm)
(12.4.127)
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.4.127 in
/usr/local/lib/python3.11/dist-packages (from torch==2.5.1->vllm)
(12.4.127)
Requirement already satisfied: nvidia-cudnn-cu12==9.1.0.70 in
/usr/local/lib/python3.11/dist-packages (from torch==2.5.1->vllm)
(9.1.0.70)
Requirement already satisfied: nvidia-cublas-cu12==12.4.5.8 in
/usr/local/lib/python3.11/dist-packages (from torch==2.5.1->vllm)
(12.4.5.8)
Requirement already satisfied: nvidia-cufft-cu12==11.2.1.3 in
/usr/local/lib/python3.11/dist-packages (from torch==2.5.1->vllm)
(11.2.1.3)
Requirement already satisfied: nvidia-curand-cu12==10.3.5.147 in
/usr/local/lib/python3.11/dist-packages (from torch==2.5.1->vllm)
(10.3.5.147)
Requirement already satisfied: nvidia-cusolver-cu12==11.6.1.9 in
/usr/local/lib/python3.11/dist-packages (from torch==2.5.1->vllm)
(11.6.1.9)
Requirement already satisfied: nvidia-cusparse-cu12==12.3.1.170 in
/usr/local/lib/python3.11/dist-packages (from torch==2.5.1->vllm)
(12.3.1.170)
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in
/usr/local/lib/python3.11/dist-packages (from torch==2.5.1->vllm)
(2.21.5)
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in
/usr/local/lib/python3.11/dist-packages (from torch==2.5.1->vllm)
(12.4.127)
Requirement already satisfied: nvidia-nvjitlink-cu12==12.4.127 in
/usr/local/lib/python3.11/dist-packages (from torch==2.5.1->vllm)
(12.4.127)
Requirement already satisfied: triton==3.1.0 in
/usr/local/lib/python3.11/dist-packages (from torch==2.5.1->vllm)
(3.1.0)
Requirement already satisfied: sympy==1.13.1 in
/usr/local/lib/python3.11/dist-packages (from torch==2.5.1->vllm)
(1.13.1)
Collecting pybind11 (from xgrammar==0.1.11->vllm)
  Downloading pybind11-2.13.6-py3-none-any.whl.metadata (9.5 kB)
Requirement already satisfied: pytest in
/usr/local/lib/python3.11/dist-packages (from xgrammar==0.1.11->vllm)
(8.3.5)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in
/usr/local/lib/python3.11/dist-packages (from sympy==1.13.1-
>torch==2.5.1->vllm) (1.3.0)
```

```
Collecting starlette<0.47.0,>=0.40.0 (from fastapi!=0.113.*,!
=0.114.0,>=0.107.0->fastapi[standard]!=0.113.*,!=0.114.0,>=0.107.0;
python_version >= "3.9"->vllm)
  Downloading starlette-0.46.0-py3-none-any.whl.metadata (6.2 kB)
Collecting fastapi-cli>=0.0.5 (from fastapi-cli[standard]>=0.0.5;
extra == "standard"->fastapi[standard]!=0.113.*,!=0.114.0,>=0.107.0;
python_version >= "3.9"->vllm)
  Downloading fastapi_cli-0.0.7-py3-none-any.whl.metadata (6.2 kB)
Requirement already satisfied: httpx>=0.23.0 in
/usr/local/lib/python3.11/dist-packages (from fastapi[standard]!
=0.113.*,!=0.114.0,>=0.107.0; python_version >= "3.9"->vllm) (0.28.1)
Collecting python-multipart>=0.0.18 (from fastapi[standard]!=0.113.*,!
=0.114.0,>=0.107.0; python_version >= "3.9"->vllm)
  Downloading python_multipart-0.0.20-py3-none-any.whl.metadata (1.8
kB)
Collecting email-validator>=2.0.0 (from fastapi[standard]!=0.113.*,!
=0.114.0,>=0.107.0; python_version >= "3.9"->vllm)
  Downloading email_validator-2.2.0-py3-none-any.whl.metadata (25 kB)
Collecting uvicorn>=0.12.0 (from uvicorn[standard]>=0.12.0; extra ==
"standard"->fastapi[standard]!=0.113.*,!=0.114.0,>=0.107.0;
python_version >= "3.9"->vllm)
  Downloading uvicorn-0.34.0-py3-none-any.whl.metadata (6.5 kB)
Requirement already satisfied: opencv-python-headless>=4.0.0 in
/usr/local/lib/python3.11/dist-packages (from
mistral_common[opencv]>=1.5.0->vllm) (4.11.0.86)
Requirement already satisfied: anyio<5,>=3.5.0 in
/usr/local/lib/python3.11/dist-packages (from openai>=1.52.0->vllm)
(3.7.1)
Requirement already satisfied: distro<2,>=1.7.0 in
/usr/local/lib/python3.11/dist-packages (from openai>=1.52.0->vllm)
(1.9.0)
Requirement already satisfied: jiter<1,>=0.4.0 in
/usr/local/lib/python3.11/dist-packages (from openai>=1.52.0->vllm)
(0.8.2)
Requirement already satisfied: sniffio in
/usr/local/lib/python3.11/dist-packages (from openai>=1.52.0->vllm)
(1.3.1)
Requirement already satisfied: annotated-types>=0.6.0 in
/usr/local/lib/python3.11/dist-packages (from pydantic>=2.9->vllm)
(0.7.0)
Requirement already satisfied: pydantic-core==2.27.2 in
/usr/local/lib/python3.11/dist-packages (from pydantic>=2.9->vllm)
(2.27.2)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.11/dist-packages (from requests>=2.26.0->vllm)
(3.4.1)
Requirement already satisfied: idna<4,>=2.5 in
/usr/local/lib/python3.11/dist-packages (from requests>=2.26.0->vllm)
(3.10)
```

```
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.11/dist-packages (from requests>=2.26.0->vllm)
(2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.11/dist-packages (from requests>=2.26.0->vllm)
(2025.1.31)
Requirement already satisfied: regex>=2022.1.18 in
/usr/local/lib/python3.11/dist-packages (from tiktoken>=0.6.0->vllm)
(2024.11.6)
Requirement already satisfied: huggingface-hub<1.0,>=0.16.4 in
/usr/local/lib/python3.11/dist-packages (from tokenizers>=0.19.1-
>vllm) (0.28.1)
Requirement already satisfied: safetensors>=0.4.1 in
/usr/local/lib/python3.11/dist-packages (from transformers>=4.48.2-
>vllm) (0.5.3)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->vllm) (2.4.6)
Requirement already satisfied: attrs>=17.3.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->vllm) (25.1.0)
Requirement already satisfied: multidict<7.0,>=4.5 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->vllm) (6.1.0)
Requirement already satisfied: propcache>=0.2.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->vllm) (0.3.0)
Requirement already satisfied: yarl<2.0,>=1.17.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->vllm) (1.18.3)
Requirement already satisfied: zipp>=3.20 in
/usr/local/lib/python3.11/dist-packages (from importlib_metadata-
>vllm) (3.21.0)
Collecting dnspython>=2.0.0 (from email-validator>=2.0.0-
>fastapi[standard]!=0.113.*,!=0.114.0,>=0.107.0; python_version >=
"3.9"->vllm)
  Downloading dnspython-2.7.0-py3-none-any.whl.metadata (5.8 kB)
Requirement already satisfied: typer>=0.12.3 in
/usr/local/lib/python3.11/dist-packages (from fastapi-cli>=0.0.5-
>fastapi-cli[standard]>=0.0.5; extra == "standard"->fastapi[standard]!
=0.113.*,!=0.114.0,>=0.107.0; python_version >= "3.9"->vllm) (0.15.2)
Collecting rich-toolkit>=0.11.1 (from fastapi-cli>=0.0.5->fastapi-
cli[standard]>=0.0.5; extra == "standard"->fastapi[standard]!
=0.113.*,!=0.114.0,>=0.107.0; python_version >= "3.9"->vllm)
  Downloading rich_toolkit-0.13.2-py3-none-any.whl.metadata (999
bytes)
Requirement already satisfied: httpcore==1.* in
/usr/local/lib/python3.11/dist-packages (from httpx>=0.23.0-
>fastapi[standard]!=0.113.*,!=0.114.0,>=0.107.0; python_version >=
"3.9"->vllm) (1.0.7)
Requirement already satisfied: h11<0.15,>=0.13 in
/usr/local/lib/python3.11/dist-packages (from httpcore==1.*-
>httpx>=0.23.0->fastapi[standard]!=0.113.*,!=0.114.0,>=0.107.0;
python_version >= "3.9"->vllm) (0.14.0)
```

```
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.11/dist-packages (from jinja2-
>outlines==0.1.11->vllm) (3.0.2)
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in
/usr/local/lib/python3.11/dist-packages (from jsonschema-
>outlines==0.1.11->vllm) (2024.10.1)
Requirement already satisfied: rpds-py>=0.7.1 in
/usr/local/lib/python3.11/dist-packages (from jsonschema-
>outlines==0.1.11->vllm) (0.23.1)
Collecting httptools>=0.6.3 (from uvicorn[standard]>=0.12.0; extra ==
"standard"->fastapi[standard]!=0.113.*,!=0.114.0,>=0.107.0;
python_version >= "3.9"->vllm)
  Downloading httptools-0.6.4-cp311-cp311-
manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_17_x86_64.manylinux
2014_x86_64.whl.metadata (3.6 kB)
Collecting python-dotenv>=0.13 (from uvicorn[standard]>=0.12.0; extra
== "standard"->fastapi[standard]!=0.113.*,!=0.114.0,>=0.107.0;
python_version >= "3.9"->vllm)
  Downloading python_dotenv-1.0.1-py3-none-any.whl.metadata (23 kB)
Collecting uvloop!=0.15.0,!=0.15.1,>=0.14.0 (from
uvicorn[standard]>=0.12.0; extra == "standard"->fastapi[standard]!
=0.113.*,!=0.114.0,>=0.107.0; python_version >= "3.9"->vllm)
  Downloading uvloop-0.21.0-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (4.9 kB)
Collecting watchfiles>=0.13 (from uvicorn[standard]>=0.12.0; extra ==
"standard"->fastapi[standard]!=0.113.*,!=0.114.0,>=0.107.0;
python_version >= "3.9"->vllm)
  Downloading watchfiles-1.0.4-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (4.9 kB)
Requirement already satisfied: websockets>=10.4 in
/usr/local/lib/python3.11/dist-packages (from
uvicorn[standard]>=0.12.0; extra == "standard"->fastapi[standard]!
=0.113.*,!=0.114.0,>=0.107.0; python_version >= "3.9"->vllm) (14.2)
Requirement already satisfied: fastrlock>=0.5 in
/usr/local/lib/python3.11/dist-packages (from cupy-cuda12x-
>ray[adag]==2.40.0->vllm) (0.8.3)
Requirement already satisfied: iniconfig in
/usr/local/lib/python3.11/dist-packages (from pytest-
>xgrammar==0.1.11->vllm) (2.0.0)
Requirement already satisfied: pluggy<2,>=1.5 in
/usr/local/lib/python3.11/dist-packages (from pytest-
>xgrammar==0.1.11->vllm) (1.5.0)
Requirement already satisfied: rich>=13.7.1 in
/usr/local/lib/python3.11/dist-packages (from rich-toolkit>=0.11.1-
>fastapi-cli>=0.0.5->fastapi-cli[standard]>=0.0.5; extra ==
"standard"->fastapi[standard]!=0.113.*,!=0.114.0,>=0.107.0;
python_version >= "3.9"->vllm) (13.9.4)
Requirement already satisfied: shellingham>=1.3.0 in
/usr/local/lib/python3.11/dist-packages (from typer>=0.12.3->fastapi-
```

```
cli>=0.0.5->fastapi-cli[standard]>=0.0.5; extra == "standard"-
>fastapi[standard]!=0.113.*,!=0.114.0,>=0.107.0; python_version >=
"3.9"->vllm) (1.5.4)
Requirement already satisfied: markdown-it-py>=2.2.0 in
/usr/local/lib/python3.11/dist-packages (from rich>=13.7.1->rich-
toolkit>=0.11.1->fastapi-cli>=0.0.5->fastapi-cli[standard]>=0.0.5;
extra == "standard"->fastapi[standard]!=0.113.*,!=0.114.0,>=0.107.0;
python_version >= "3.9"->vllm) (3.0.0)
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in
/usr/local/lib/python3.11/dist-packages (from rich>=13.7.1->rich-
toolkit>=0.11.1->fastapi-cli>=0.0.5->fastapi-cli[standard]>=0.0.5;
extra == "standard"->fastapi[standard]!=0.113.*,!=0.114.0,>=0.107.0;
python_version >= "3.9"->vllm) (2.18.0)
Requirement already satisfied: mdurl~=0.1 in
/usr/local/lib/python3.11/dist-packages (from markdown-it-py>=2.2.0-
>rich>=13.7.1->rich-toolkit>=0.11.1->fastapi-cli>=0.0.5->fastapi-
cli[standard]>=0.0.5; extra == "standard"->fastapi[standard]!
=0.113.*,!=0.114.0,>=0.107.0; python_version >= "3.9"->vllm) (0.1.2)
Downloading vllm-0.7.3-cp38-abi3-manylinux1_x86_64.whl (264.6 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 264.6/264.6 MB 5.3 MB/s eta
0:00:00
pressed_tensors-0.9.1-py3-none-any.whl (96 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 96.5/96.5 kB 8.9 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 71.6/71.6 kB 7.4 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 111.0/111.0 kB 10.9 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 87.6/87.6 kB 7.4 MB/s eta
0:00:00
anylinux2014_x86_64.whl (67.0 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 67.0/67.0 MB 13.2 MB/s eta
0:00:00
ers-0.0.28.post3-cp311-cp311-manylinux_2_28_x86_64.whl (16.7 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 16.7/16.7 MB 84.7 MB/s eta
0:00:00
mar-0.1.11-cp311-cp311-manylinux_2_27_x86_64.manylinux_2_28_x86_64.whl
(396 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 396.9/396.9 kB 34.0 MB/s eta
0:00:00
anylinux_2_17_x86_64.manylinux2014_x86_64.whl (343 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 343.3/343.3 kB 30.4 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 94.9/94.9 kB 9.9 MB/s eta
0:00:00
_format_enforcer-0.10.11-py3-none-any.whl (44 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 44.2/44.2 kB 4.0 MB/s eta
0:00:00
istral_common-1.5.3-py3-none-any.whl (6.5 MB)
```

```
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 6.5/6.5 MB 114.0 MB/s eta
0:00:00
etheus_fastapi_instrumentator-7.0.2-py3-none-any.whl (18 kB)
Downloading tiktoken-0.9.0-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl (1.2 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.2/1.2 MB 62.7 MB/s eta
0:00:00
anylinux_2_17_x86_64.manylinux2014_x86_64.whl (376 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 376.2/376.2 kB 32.5 MB/s eta
0:00:00
sgspec-0.19.0-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl (210 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 210.7/210.7 kB 19.9 MB/s eta
0:00:00
ail_validator-2.2.0-py3-none-any.whl (33 kB)
Downloading fastapi_cli-0.0.7-py3-none-any.whl (10 kB)
Downloading interegular-0.3.3-py37-none-any.whl (23 kB)
Downloading python_multipart-0.0.20-py3-none-any.whl (24 kB)
Downloading starlette-0.46.0-py3-none-any.whl (71 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 72.0/72.0 kB 7.7 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 62.3/62.3 kB 6.2 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 913.7/913.7 kB 57.2 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 45.5/45.5 kB 4.4 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 243.3/243.3 kB 23.6 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 6.3/6.3 MB 99.7 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 313.6/313.6 kB 29.5 MB/s eta
0:00:00
anylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_17_x86_64.manylinux2
014_x86_64.whl (459 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 459.8/459.8 kB 37.8 MB/s eta
0:00:00
anylinux_2_17_x86_64.manylinux2014_x86_64.whl (4.0 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 4.0/4.0 MB 92.2 MB/s eta
0:00:00
anylinux_2_17_x86_64.manylinux2014_x86_64.whl (452 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 452.6/452.6 kB 33.5 MB/s eta
0:00:00
ultipart, python-dotenv, pycountry, pybind11, partial-json-parser,
msgspec, lark, interegular, httptools, gguf, dnspython, diskcache,
astor, airportsdata, watchfiles, tiktoken, starlette, email-validator,
depyf, rich-toolkit, prometheus-fastapi-instrumentator, lm-format-
enforcer, fastapi, xformers, ray, outlines_core, mistral_common,
fastapi-cli, xgrammar, outlines, compressed-tensors, vllm
Successfully installed airportsdata-20250224 astor-0.8.1 blake3-1.0.4
```

```
compressed-tensors-0.9.1 depyf-0.18.0 diskcache-5.6.3 dnspython-2.7.0
email-validator-2.2.0 fastapi-0.115.11 fastapi-cli-0.0.7 gguf-0.10.0
httptools-0.6.4 interegular-0.3.3 lark-1.2.2 lm-format-enforcer-
0.10.11 mistral_common-1.5.3 msgspec-0.19.0 outlines-0.1.11
outlines_core-0.1.26 partial-json-parser-0.2.1.1.post5 prometheus-
fastapi-instrumentator-7.0.2 pybind11-2.13.6 pycountry-24.6.1 python-
dotenv-1.0.1 python-multipart-0.0.20 ray-2.40.0 rich-toolkit-0.13.2
starlette-0.46.0 tiktoken-0.9.0 uvicorn-0.34.0 uvloop-0.21.0 vllm-
0.7.3 watchfiles-1.0.4 xformers-0.0.28.post3 xgrammar-0.1.11
```