

```
!pip install bitsandbytes
```

```

Downloading nvidia_cuda_nvrtc_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl (24.6 MB)
24.6/24.6 MB 76.7 MB/s eta 0:00:00
Downloading nvidia_cuda_runtime_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl (883 kB)
883.7/883.7 kB 51.8 MB/s eta 0:00:00
Downloading nvidia_cudnn_cu12-9.1.0.70-py3-none-manylinux2014_x86_64.whl (664.8 MB)
664.8/664.8 MB 2.2 MB/s eta 0:00:00
Downloading nvidia_cufft_cu12-11.2.1.3-py3-none-manylinux2014_x86_64.whl (211.5 MB)
211.5/211.5 MB 4.9 MB/s eta 0:00:00
Downloading nvidia_curand_cu12-10.3.5.147-py3-none-manylinux2014_x86_64.whl (56.3 MB)
56.3/56.3 MB 44.3 MB/s eta 0:00:00
Downloading nvidia_cusolver_cu12-11.6.1.9-py3-none-manylinux2014_x86_64.whl (127.9 MB)
127.9/127.9 MB 20.3 MB/s eta 0:00:00
Downloading nvidia_cusparses_cu12-12.3.1.170-py3-none-manylinux2014_x86_64.whl (207.5 MB)
207.5/207.5 MB 5.1 MB/s eta 0:00:00
Downloading nvidia_nvjitlink_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl (21.1 MB)
21.1/21.1 MB 96.6 MB/s eta 0:00:00
Installing collected packages: nvidia-nvjitlink-cu12, nvidia-curand-cu12, nvidia-cufft-cu12, nvidia-cuda-runtime-cu12, nvidia-cuda-nvrtc-cu12, nvidia-cuda-cupti-cu12, nvidia-cublas-cu12, nvidia-cudnn-cu12, nvidia-cusparses-cu12, nvidia-cusolver-cu12
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-cusparses-cu12
Found existing installation: nvidia-cusparses-cu12 12.5.1.3
Uninstalling nvidia-cusparses-cu12-12.5.1.3:
Successfully uninstalled nvidia-cusparses-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 bitsandbytes-0.45.5 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-cusparses-cu12-12.3.1.170 nvidia-cusolver-cu12-11.6.1.9

```

```
!pip install accelerate==0.21.0 bitsandbytes transformers==4.31.0 scipy tensorboard
```

```

Downloading tokenizers-0.13.3-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (6.7 kB)

```

```
Requirement already satisfied: nvidia-cublas-cu12==12.4.5.8 in /usr/local/lib/python3.11/dist-packages (from torch>=1.10.0->a
Requirement already satisfied: nvidia-cufft-cu12==11.2.1.3 in /usr/local/lib/python3.11/dist-packages (from torch>=1.10.0->a
Requirement already satisfied: nvidia-curand-cu12==10.3.5.147 in /usr/local/lib/python3.11/dist-packages (from torch>=1.10.0
Requirement already satisfied: nvidia-cusolver-cu12==11.6.1.9 in /usr/local/lib/python3.11/dist-packages (from torch>=1.10.0
Requirement already satisfied: nvidia-cusparse-cu12==12.3.1.170 in /usr/local/lib/python3.11/dist-packages (from torch>=1.10
Requirement already satisfied: nvidia-cusparselt-cu12==0.6.2 in /usr/local/lib/python3.11/dist-packages (from torch>=1.10.0-
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in /usr/local/lib/python3.11/dist-packages (from torch>=1.10.0->acce
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.10.0->ac
Requirement already satisfied: nvidia-nvjitlink-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.10.
Requirement already satisfied: triton==3.2.0 in /usr/local/lib/python3.11/dist-packages (from torch>=1.10.0->accelerate==0.2
Requirement already satisfied: sympy==1.13.1 in /usr/local/lib/python3.11/dist-packages (from torch>=1.10.0->accelerate==0.2
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.11/dist-packages (from sympy==1.13.1->torch>=1.1
Requirement already satisfied: MarkupSafe>=2.1.1 in /usr/local/lib/python3.11/dist-packages (from werkzeug>=1.0.1->tensorboa
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests->transform
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests->transformers==4.31.0)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests->transformers==4
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests->transformers==4
Downloading accelerate-0.21.0-py3-none-any.whl (244 kB)
244.2/244.2 kB 10.7 MB/s eta 0:00:00
Downloading transformers-4.31.0-py3-none-any.whl (7.4 MB)
7.4/7.4 MB 80.9 MB/s eta 0:00:00
Downloading tokenizers-0.13.3-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (7.8 MB)
7.8/7.8 MB 105.9 MB/s eta 0:00:00
Installing collected packages: tokenizers, transformers, accelerate
  Attempting uninstall: tokenizers
    Found existing installation: tokenizers 0.21.1
    Uninstalling tokenizers-0.21.1:
      Successfully uninstalled tokenizers-0.21.1
  Attempting uninstall: transformers
    Found existing installation: transformers 4.51.3
    Uninstalling transformers-4.51.3:
      Successfully uninstalled transformers-4.51.3
  Attempting uninstall: accelerate
    Found existing installation: accelerate 1.5.2
    Uninstalling accelerate-1.5.2:
      Successfully uninstalled accelerate-1.5.2
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is
sentence-transformers 3.4.1 requires transformers<5.0.0,>=4.41.0, but you have transformers 4.31.0 which is incompatible.
Successfully installed accelerate-0.21.0 tokenizers-0.13.3 transformers-4.31.0
```

```
!pip install peft==0.3.0
```

```
Collecting peft==0.3.0
  Downloading peft-0.3.0-py3-none-any.whl.metadata (21 kB)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-packages (from peft==0.3.0) (2.0.2)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from peft==0.3.0) (24.2)
Requirement already satisfied: psutil in /usr/local/lib/python3.11/dist-packages (from peft==0.3.0) (5.9.5)
Requirement already satisfied: pyyaml in /usr/local/lib/python3.11/dist-packages (from peft==0.3.0) (6.0.2)
Requirement already satisfied: torch>=1.13.0 in /usr/local/lib/python3.11/dist-packages (from peft==0.3.0) (2.6.0+cu124)
Requirement already satisfied: transformers in /usr/local/lib/python3.11/dist-packages (from peft==0.3.0) (4.31.0)
Requirement already satisfied: accelerate in /usr/local/lib/python3.11/dist-packages (from peft==0.3.0) (0.21.0)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->peft==0.3.0) (3.18.0)
Requirement already satisfied: typing-extensions>=4.10.0 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->pef
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->peft==0.3.0) (3.4.2)
Requirement already satisfied: Jinja2 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->peft==0.3.0) (3.1.6)
Requirement already satisfied: fsspec in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->peft==0.3.0) (2025.3.2)
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13
Requirement already satisfied: nvidia-cudnn-cu12==9.1.0.70 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->p
Requirement already satisfied: nvidia-cublas-cu12==12.4.5.8 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->p
Requirement already satisfied: nvidia-cufft-cu12==11.2.1.3 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->p
Requirement already satisfied: nvidia-curand-cu12==10.3.5.147 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0
Requirement already satisfied: nvidia-cusolver-cu12==11.6.1.9 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0
Requirement already satisfied: nvidia-cusparse-cu12==12.3.1.170 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13
Requirement already satisfied: nvidia-cusparselt-cu12==0.6.2 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0-
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->peft
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->pe
Requirement already satisfied: nvidia-nvjitlink-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.
Requirement already satisfied: triton==3.2.0 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->peft==0.3.0) (3
Requirement already satisfied: sympy==1.13.1 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->peft==0.3.0) (1
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.11/dist-packages (from sympy==1.13.1->torch>=1.1
Requirement already satisfied: huggingface-hub<1.0,>=0.14.1 in /usr/local/lib/python3.11/dist-packages (from transformers->p
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.11/dist-packages (from transformers->peft==0.3.0)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages (from transformers->peft==0.3.0) (2.32.3)
Requirement already satisfied: tokenizers!=0.11.3,<0.14,>=0.11.1 in /usr/local/lib/python3.11/dist-packages (from transforme
Requirement already satisfied: safetensors>=0.3.1 in /usr/local/lib/python3.11/dist-packages (from transformers->peft==0.3.0
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.11/dist-packages (from transformers->peft==0.3.0) (4.67)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from Jinja2->torch>=1.13.0->peft=
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests->transform
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests->transformers->peft==0
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests->transformers->p
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests->transformers->p
Downloading peft-0.3.0-py3-none-any.whl (56 kB)
```

```
56.8/56.8 kB 2.3 MB/s eta 0:00:00
Installing collected packages: peft
  Attempting uninstall: peft
    Found existing installation: peft 0.14.0
    Uninstalling peft-0.14.0:
      Successfully uninstalled peft-0.14.0
  Successfully installed peft-0.3.0

!pip install trl

Requirement already satisfied: nvidia-cuda-runtime-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=2.0.0)
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=2.0.0)
Requirement already satisfied: nvidia-cudnn-cu12==9.1.0.70 in /usr/local/lib/python3.11/dist-packages (from torch>=2.0.0->accelerate)
Requirement already satisfied: nvidia-cublas-cu12==12.4.5.8 in /usr/local/lib/python3.11/dist-packages (from torch>=2.0.0->accelerate)
Requirement already satisfied: nvidia-cufft-cu12==11.2.1.3 in /usr/local/lib/python3.11/dist-packages (from torch>=2.0.0->accelerate)
Requirement already satisfied: nvidia-curand-cu12==10.3.5.147 in /usr/local/lib/python3.11/dist-packages (from torch>=2.0.0->accelerate)
Requirement already satisfied: nvidia-cusolver-cu12==11.6.1.9 in /usr/local/lib/python3.11/dist-packages (from torch>=2.0.0->accelerate)
Requirement already satisfied: nvidia-cusparse-cu12==12.3.1.170 in /usr/local/lib/python3.11/dist-packages (from torch>=2.0.0->accelerate)
Requirement already satisfied: nvidia-cusparselt-cu12==0.6.2 in /usr/local/lib/python3.11/dist-packages (from torch>=2.0.0->accelerate)
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in /usr/local/lib/python3.11/dist-packages (from torch>=2.0.0->accelerate)
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=2.0.0->accelerate)
Requirement already satisfied: nvidia-nvjitlink-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=2.0.0->accelerate)
Requirement already satisfied: triton==3.2.0 in /usr/local/lib/python3.11/dist-packages (from torch>=2.0.0->accelerate)
Requirement already satisfied: sympy==1.13.1 in /usr/local/lib/python3.11/dist-packages (from torch>=2.0.0->accelerate)
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)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas->datasets>=3.0.0)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from pandas->datasets>=3.0.0->trl)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas->datasets>=3.0.0->trl)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->pandas->datasets)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from jinja2->torch>=2.0.0->accelerate)
Downloading trl-0.17.0-py3-none-any.whl (348 kB)
348.0/348.0 kB 10.3 MB/s eta 0:00:00
Downloading accelerate-1.6.0-py3-none-any.whl (354 kB)
354.7/354.7 kB 32.2 MB/s eta 0:00:00
Downloading datasets-3.5.0-py3-none-any.whl (491 kB)
491.2/491.2 kB 43.6 MB/s eta 0:00:00
Downloading transformers-4.51.3-py3-none-any.whl (10.4 MB)
10.4/10.4 MB 126.0 MB/s eta 0:00:00
Downloading dill-0.3.8-py3-none-any.whl (116 kB)
116.3/116.3 kB 13.6 MB/s eta 0:00:00
Downloading fsspec-2024.12.0-py3-none-any.whl (183 kB)
183.9/183.9 kB 21.9 MB/s eta 0:00:00
Downloading multiprocessing-0.70.16-py311-none-any.whl (143 kB)
143.5/143.5 kB 17.0 MB/s eta 0:00:00
Downloading tokenizers-0.21.1-cp39-abi3-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (3.0 MB)
3.0/3.0 MB 107.8 MB/s eta 0:00:00
Downloading xxhash-3.5.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (194 kB)
194.8/194.8 kB 23.9 MB/s eta 0:00:00
Installing collected packages: xxhash, fsspec, dill, multiprocessing, tokenizers, transformers, datasets, accelerate, trl
  Attempting uninstall: fsspec
    Found existing installation: fsspec 2025.3.2
    Uninstalling fsspec-2025.3.2:
      Successfully uninstalled fsspec-2025.3.2
  Attempting uninstall: tokenizers
    Found existing installation: tokenizers 0.13.3
    Uninstalling tokenizers-0.13.3:
      Successfully uninstalled tokenizers-0.13.3
  Attempting uninstall: transformers
    Found existing installation: transformers 4.31.0
    Uninstalling transformers-4.31.0:
      Successfully uninstalled transformers-4.31.0
  Attempting uninstall: accelerate
    Found existing installation: accelerate 0.21.0
    Uninstalling accelerate-0.21.0:
      Successfully uninstalled accelerate-0.21.0
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is
  gcsfs 2025.3.2 requires fsspec==2025.3.2, but you have fsspec 2024.12.0 which is incompatible.
  Successfully installed accelerate-1.6.0 datasets-3.5.0 dill-0.3.8 fsspec-2024.12.0 multiprocessing-0.70.16 tokenizers-0.21.1 trl-0.17.0

!export LD_LIBRARY_PATH=/usr/local/cuda/lib64:$LD_LIBRARY_PATH
!export PATH=/usr/local/cuda/bin:$PATH

!nvidia-smi
```

Mon Apr 28 01:18:51 2025

NVIDIA-SMI 550.54.15				Driver Version: 550.54.15		CUDA Version: 12.4	
GPU Name		Persistence-M	Bus-Id	Disp.A	Volatile Uncorr. ECC		
Fan	Temp	Perf	Pwr:Usage/Cap	Memory-Usage	GPU-Util	Compute M.	MIG M.

0	Tesla T4	Off	00000000:00:04.0 Off	0
N/A	47C	P8	10W / 70W	0MiB / 15360MiB
				0% Default
				N/A

Processes:							GPU Memory Usage
GPU	GI	CI	PID	Type	Process name		
ID	ID	ID					
No running processes found							

```
!pip uninstall -y peft
!pip install peft==0.10.0
```

```

Found existing installation: peft 0.3.0
Uninstalling peft-0.3.0:
  Successfully uninstalled peft-0.3.0
Collecting peft==0.10.0
  Downloading peft-0.10.0-py3-none-any.whl.metadata (13 kB)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-packages (from peft==0.10.0) (2.0.2)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from peft==0.10.0) (24.2)
Requirement already satisfied: psutil in /usr/local/lib/python3.11/dist-packages (from peft==0.10.0) (5.9.5)
Requirement already satisfied: pyyaml in /usr/local/lib/python3.11/dist-packages (from peft==0.10.0) (6.0.2)
Requirement already satisfied: torch>=1.13.0 in /usr/local/lib/python3.11/dist-packages (from peft==0.10.0) (2.6.0+cu124)
Requirement already satisfied: transformers in /usr/local/lib/python3.11/dist-packages (from peft==0.10.0) (4.51.3)
Requirement already satisfied: tqdm in /usr/local/lib/python3.11/dist-packages (from peft==0.10.0) (4.67.1)
Requirement already satisfied: accelerate>=0.21.0 in /usr/local/lib/python3.11/dist-packages (from peft==0.10.0) (1.6.0)
Requirement already satisfied: safetensors in /usr/local/lib/python3.11/dist-packages (from peft==0.10.0) (0.5.3)
Requirement already satisfied: huggingface-hub>=0.17.0 in /usr/local/lib/python3.11/dist-packages (from peft==0.10.0) (0.30.0)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.17.0->peft==0.10.0) (3.16.0)
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.17.0->peft==0.10.0) (2025.5.0)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.17.0->peft==0.10.0) (2.32.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.17.0->peft==0.10.0) (4.12.0)
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->peft==0.10.0) (3.4.2)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->peft==0.10.0) (3.1.6)
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->peft==0.10.0) (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->peft==0.10.0) (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->peft==0.10.0) (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->peft==0.10.0) (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->peft==0.10.0) (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->peft==0.10.0) (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->peft==0.10.0) (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->peft==0.10.0) (11.6.1.9)
Requirement already satisfied: nvidia-cusparselt-cu12==0.6.2 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->peft==0.10.0) (0.6.2)
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->peft==0.10.0) (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->peft==0.10.0) (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->peft==0.10.0) (12.4.127)
Requirement already satisfied: triton==3.2.0 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->peft==0.10.0) (3.2.0)
Requirement already satisfied: sympy==1.13.1 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->peft==0.10.0) (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->peft==0.10.0) (1.3.0)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.11/dist-packages (from transformers->peft==0.10.0) (2024.11.6)
Requirement already satisfied: tokenizers<0.22, >=0.21 in /usr/local/lib/python3.11/dist-packages (from transformers->peft==0.10.0) (0.21.0)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from jinja2->torch>=1.13.0->peft==0.10.0) (3.0.2)
Requirement already satisfied: charset-normalizer<4, >=2 in /usr/local/lib/python3.11/dist-packages (from requests->huggingface-hub>=0.17.0->peft==0.10.0) (3.4.1)
Requirement already satisfied: idna<4, >=2.5 in /usr/local/lib/python3.11/dist-packages (from requests->huggingface-hub>=0.17.0->peft==0.10.0) (3.10.1)
Requirement already satisfied: urllib3<3, >=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests->huggingface-hub>=0.17.0->peft==0.10.0) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests->huggingface-hub>=0.17.0->peft==0.10.0) (2025.11.12)
Downloading peft-0.10.0-py3-none-any.whl (199 kB)
199.1/199.1 kB 4.1 MB/s eta 0:00:00
Installing collected packages: peft
Successfully installed peft-0.10.0

```

```

import numpy as np
import pandas as pd

import os
from tqdm import tqdm
import bitsandbytes as bnb

```

```

import torch
import torch.nn as nn
import transformers
from datasets import Dataset
from peft import LoraConfig, PeftConfig
from trl import SFTTrainer
from trl import setup_chat_format
from transformers import (AutoModelForCausalLM,

```

```

        AutoTokenizer,
        BitsAndBytesConfig,
        TrainingArguments,
        pipeline,
        logging)
from sklearn.metrics import (accuracy_score,
                             classification_report,
                             confusion_matrix)
from sklearn.model_selection import train_test_split

!nvcc --version

nvcc: NVIDIA (R) Cuda compiler driver
Copyright (c) 2005–2024 NVIDIA Corporation
Built on Thu_Jun__6_02:18:23_PDT_2024
Cuda compilation tools, release 12.5, V12.5.82
Build cuda_12.5.r12.5/compiler.34385749_0

import torch
print(f"pytorch version {torch.__version__}")

pytorch version 2.6.0+cu124

device = torch.device("cuda:0" if torch.cuda.is_available() else "cpu")
print(f"working on {device}")

working on cuda:0

filename = "/content/all-data.csv"

df = pd.read_csv(filename,
                  names=["sentiment", "text"],
                  encoding="utf-8", encoding_errors="replace")

X_train = list()
X_test = list()
for sentiment in ["positive", "neutral", "negative"]:
    train, test = train_test_split(df[df.sentiment==sentiment],
                                   train_size=300,
                                   test_size=300,
                                   random_state=42)

    X_train.append(train)
    X_test.append(test)

X_train = pd.concat(X_train).sample(frac=1, random_state=10)
X_test = pd.concat(X_test)

eval_idx = [idx for idx in df.index if idx not in list(X_train.index) + list(X_test.index)]
X_eval = df[df.index.isin(eval_idx)]
X_eval = (X_eval
          .groupby('sentiment', group_keys=False)
          .apply(lambda x: x.sample(n=50, random_state=10, replace=True)))
X_train = X_train.reset_index(drop=True)

def generate_prompt(data_point):
    return f"""
        Analyze the sentiment of the news headline enclosed in square brackets,
        determine if it is positive, neutral, or negative, and return the answer as
        the corresponding sentiment label "positive" or "neutral" or "negative".

        [{data_point["text"]} = {data_point["sentiment"]}
        """.strip()

def generate_test_prompt(data_point):
    return f"""
        Analyze the sentiment of the news headline enclosed in square brackets,
        determine if it is positive, neutral, or negative, and return the answer as
        the corresponding sentiment label "positive" or "neutral" or "negative".

        [{data_point["text"]} = """.strip()

X_train = pd.DataFrame(X_train.apply(generate_prompt, axis=1),
                       columns=["text"])
X_eval = pd.DataFrame(X_eval.apply(generate_prompt, axis=1),

```

```

        columns=["text"])

y_true = X_test.sentiment
X_test = pd.DataFrame(X_test.apply(generate_test_prompt, axis=1), columns=["text"])

train_data = Dataset.from_pandas(X_train)
eval_data = Dataset.from_pandas(X_eval)

↳ <ipython-input-12-cf78b53f0b42>:24: DeprecationWarning: DataFrameGroupBy.apply operated on the grouping columns. This behavior
    .apply(lambda x: x.sample(n=50, random_state=10, replace=True)))

def evaluate(y_true, y_pred):
    labels = ['positive', 'neutral', 'negative']
    mapping = {'positive': 2, 'neutral': 1, 'none':1, 'negative': 0}
    def map_func(x):
        return mapping.get(x, 1)

    y_true = np.vectorize(map_func)(y_true)
    y_pred = np.vectorize(map_func)(y_pred)

    # Calculate accuracy
    accuracy = accuracy_score(y_true=y_true, y_pred=y_pred)
    print(f'Accuracy: {accuracy:.3f}')

    # Generate accuracy report
    unique_labels = set(y_true) # Get unique labels

    for label in unique_labels:
        label_indices = [i for i in range(len(y_true))
                        if y_true[i] == label]
        label_y_true = [y_true[i] for i in label_indices]
        label_y_pred = [y_pred[i] for i in label_indices]
        accuracy = accuracy_score(label_y_true, label_y_pred)
        print(f'Accuracy for label {label}: {accuracy:.3f}')

    # Generate classification report
    class_report = classification_report(y_true=y_true, y_pred=y_pred)
    print('\nClassification Report:')
    print(class_report)

    # Generate confusion matrix
    conf_matrix = confusion_matrix(y_true=y_true, y_pred=y_pred, labels=[0, 1, 2])
    print('\nConfusion Matrix:')
    print(conf_matrix)

# The model that you want to train from the Hugging Face hub
model_name = "NousResearch/llama-2-7b-chat-hf"

# The instruction dataset to use
dataset_name = "mlabonne/guanaco-llama2-1k"

# Fine-tuned model name
new_model = "llama-2-7b-miniguanaco"

#####
# QLoRA parameters
#####

# LoRA attention dimension
lora_r = 64

# Alpha parameter for LoRA scaling
lora_alpha = 16

# Dropout probability for LoRA layers
lora_dropout = 0.1

#####
# bitsandbytes parameters
#####

# Activate 4-bit precision base model loading
use_4bit = True

# Compute dtype for 4-bit base models
bnb_4bit_compute_dtype = "float16"

```



```

# Quantization type (fp4 or nf4)
bnb_4bit_quant_type = "nf4"

# Activate nested quantization for 4-bit base models (double quantization)
use_nested_quant = False

#####
# TrainingArguments parameters
#####

# Output directory where the model predictions and checkpoints will be stored
output_dir = "./results"

# Number of training epochs
num_train_epochs = 1

# Enable fp16/bf16 training (set bf16 to True with an A100)
fp16 = False
bf16 = False

# Batch size per GPU for training
per_device_train_batch_size = 4

# Batch size per GPU for evaluation
per_device_eval_batch_size = 4

# Number of update steps to accumulate the gradients for
gradient_accumulation_steps = 1

# Enable gradient checkpointing
gradient_checkpointing = True

# Maximum gradient normal (gradient clipping)
max_grad_norm = 0.3

# Initial learning rate (AdamW optimizer)
learning_rate = 2e-4

# Weight decay to apply to all layers except bias/LayerNorm weights
weight_decay = 0.001

# Optimizer to use
optim = "paged_adamw_32bit"

# Learning rate schedule (constant a bit better than cosine)
lr_scheduler_type = "constant"

# Number of training steps (overrides num_train_epochs)
max_steps = -1

# Ratio of steps for a linear warmup (from 0 to learning rate)
warmup_ratio = 0.03

# Group sequences into batches with same length
# Saves memory and speeds up training considerably
group_by_length = True

# Save checkpoint every X updates steps
save_steps = 25

# Log every X updates steps
logging_steps = 25

#####
# SFT parameters
#####

# Maximum sequence length to use
max_seq_length = None

# Pack multiple short examples in the same input sequence to increase efficiency
packing = False

# Load the entire model on the GPU 0
device_map = {"": 0}

```

```
!pip uninstall torch -y  
!pip install torch torchvision torchaudio --index-url https://download.pytorch.org/whl/cu118
```



```

Found existing installation: torch 2.6.0+cu124
Uninstalling torch-2.6.0+cu124:
Traceback (most recent call last):
  File "/usr/lib/python3.11/shutil.py", line 853, in move
    os.rename(src, real_dst)
OSError: [Errno 18] Invalid cross-device link: '/usr/local/lib/python3.11/dist-packages/torch/' -> '/usr/local/lib/python3.1

```

During handling of the above exception, another exception occurred:

```

Traceback (most recent call last):
  File "/usr/local/lib/python3.11/dist-packages/pip/_internal/cli/base_command.py", line 179, in exc_logging_wrapper
    status = run_func(*args)
             ~~~~~
  File "/usr/local/lib/python3.11/dist-packages/pip/_internal/commands/uninstall.py", line 106, in run
    uninstall_pathset = req.uninstall(
                       ~~~~~
  File "/usr/local/lib/python3.11/dist-packages/pip/_internal/req/req_install.py", line 722, in uninstall
    uninstalled_pathset.remove(auto_confirm, verbose)
  File "/usr/local/lib/python3.11/dist-packages/pip/_internal/req/req_uninstall.py", line 370, in remove
    moved.stash(path)
  File "/usr/local/lib/python3.11/dist-packages/pip/_internal/req/req_uninstall.py", line 261, in stash
    renames(path, new_path)
  File "/usr/local/lib/python3.11/dist-packages/pip/_internal/utils/misc.py", line 356, in renames
    shutil.move(old, new)
  File "/usr/lib/python3.11/shutil.py", line 869, in move
    copytree(src, real_dst, copy_function=copy_function,
  File "/usr/lib/python3.11/shutil.py", line 573, in copytree
    return _copytree(entries=entries, src=src, dst=dst, symlinks=symlinks,
    ~~~~~
  File "/usr/lib/python3.11/shutil.py", line 509, in _copytree
    copytree(srcobj, dstname, symlinks, ignore, copy_function,
  File "/usr/lib/python3.11/shutil.py", line 573, in copytree
    return _copytree(entries=entries, src=src, dst=dst, symlinks=symlinks,
    ~~~~~
  File "/usr/lib/python3.11/shutil.py", line 513, in _copytree
    copy_function(srcobj, dstname)
  File "/usr/lib/python3.11/shutil.py", line 448, in copy2
    copyfile(src, dst, follow_symlinks=follow_symlinks)
  File "/usr/lib/python3.11/shutil.py", line 269, in copyfile
    _fastcopy_sendfile(fsrc, fdst)
  File "/usr/lib/python3.11/shutil.py", line 144, in _fastcopy_sendfile
    sent = os.sendfile(outfd, infd, offset, blocksize)
    ~~~~~

```

KeyboardInterrupt

During handling of the above exception, another exception occurred:

```

Traceback (most recent call last):
  File "/usr/local/bin/pip3", line 10, in <module>
    sys.exit(main())
    ~~~~~
  File "/usr/local/lib/python3.11/dist-packages/pip/_internal/cli/main.py", line 80, in main
    return command.main(cmd_args)
    ~~~~~
  File "/usr/local/lib/python3.11/dist-packages/pip/_internal/cli/base_command.py", line 100, in main
    return self._main(args)
    ~~~~~
  File "/usr/local/lib/python3.11/dist-packages/pip/_internal/cli/base_command.py", line 232, in _main
    return run(options, args)
    ~~~~~
  File "/usr/local/lib/python3.11/dist-packages/pip/_internal/cli/base_command.py", line 215, in exc_logging_wrapper
    logger.critical("Operation cancelled by user")
  File "/usr/lib/python3.11/logging/__init__.py", line 1536, in critical
    self._log(CRITICAL, msg, args, **kwargs)
  File "/usr/lib/python3.11/logging/__init__.py", line 1622, in _log
    fn, lno, func, sinfo = self.findCaller(stack_info, stacklevel)
    ~~~~~
  File "/usr/lib/python3.11/logging/__init__.py", line 1582, in findCaller
    if not _is_internal_frame(f):
    ~~~~~
  File "/usr/lib/python3.11/logging/__init__.py", line 196, in _is_internal_frame
    filename = os.path.normcase(frame.f_code.co_filename)
    ~~~~~
  File "<frozen posixpath>", line 52, in normcase

```

KeyboardInterrupt

^C

WARNING: Ignoring invalid distribution ~orch (/usr/local/lib/python3.11/dist-packages)

WARNING: Ignoring invalid distribution ~orch (/usr/local/lib/python3.11/dist-packages)

Looking in indexes: <https://download.pytorch.org/whl/cu118>

Collecting torch

Downloading https://download.pytorch.org/whl/cu118/torch-2.7.0%2Bcu118-cp311-cp311-manylinux_2_28_x86_64.whl.metadata (28

Requirement already satisfied: torchvision in /usr/local/lib/python3.11/dist-packages (0.21.0+cu124)

Requirement already satisfied: torchaudio in /usr/local/lib/python3.11/dist-packages (2.6.0+cu124)

Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from torch) (3.18.0)

Requirement already satisfied: typing_extensions=4.10.0 in /usr/local/lib/python3.11/dist-packages (from torch) (4.10.0)

```

Requirement already satisfied: typing-extensions>=4.10.0 in /usr/local/lib/python3.11/dist-packages (from torch) (4.13.2)
Collecting sympy>=1.13.3 (from torch)
  Downloading https://download.pytorch.org/whl/sympy-1.13.3-py3-none-any.whl.metadata (12 kB)
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) (3.1.6)
Requirement already satisfied: fsspec in /usr/local/lib/python3.11/dist-packages (from torch) (2024.12.0)
Collecting nvidia-cuda-nvrtc-cu11==11.8.89 (from torch)
  Downloading https://download.pytorch.org/whl/cu118/nvidia_cuda_nvrtc_cu11-11.8.89-py3-none-manylinux1_x86_64.whl (23.2 MB)
    23.2/23.2 MB 104.9 MB/s eta 0:00:00
Collecting nvidia-cuda-runtime-cu11==11.8.89 (from torch)
  Downloading https://download.pytorch.org/whl/cu118/nvidia_cuda_runtime_cu11-11.8.89-py3-none-manylinux1_x86_64.whl (875 kB)
    875.6/875.6 kB 61.8 MB/s eta 0:00:00
Collecting nvidia-cuda-cupti-cu11==11.8.87 (from torch)
  Downloading https://download.pytorch.org/whl/cu118/nvidia_cuda_cupti_cu11-11.8.87-py3-none-manylinux1_x86_64.whl (13.1 MB)
    13.1/13.1 MB 128.3 MB/s eta 0:00:00
Collecting nvidia-cudnn-cu11==9.1.0.70 (from torch)
  Downloading https://download.pytorch.org/whl/cu118/nvidia_cudnn_cu11-9.1.0.70-py3-none-manylinux2014_x86_64.whl (663.9 MB)
    663.9/663.9 MB 2.3 MB/s eta 0:00:00
Collecting nvidia-cublas-cu11==11.11.3.6 (from torch)
  Downloading https://download.pytorch.org/whl/cu118/nvidia_cublas_cu11-11.11.3.6-py3-none-manylinux1_x86_64.whl (417.9 MB)
    417.9/417.9 MB 2.8 MB/s eta 0:00:00
Collecting nvidia-cufft-cu11==10.9.0.58 (from torch)
  Downloading https://download.pytorch.org/whl/cu118/nvidia_cufft_cu11-10.9.0.58-py3-none-manylinux1_x86_64.whl (168.4 MB)
    168.4/168.4 MB 7.3 MB/s eta 0:00:00
Collecting nvidia-curand-cu11==10.3.0.86 (from torch)
  Downloading https://download.pytorch.org/whl/cu118/nvidia_curand_cu11-10.3.0.86-py3-none-manylinux1_x86_64.whl (58.1 MB)
    58.1/58.1 MB 40.4 MB/s eta 0:00:00
Collecting nvidia-cusolver-cu11==11.4.1.48 (from torch)
  Downloading https://download.pytorch.org/whl/cu118/nvidia_cusolver_cu11-11.4.1.48-py3-none-manylinux1_x86_64.whl (128.2 MB)
    128.2/128.2 MB 17.1 MB/s eta 0:00:00
Collecting nvidia-cuspars-cu11==11.7.5.86 (from torch)
  Downloading https://download.pytorch.org/whl/cu118/nvidia_cuspars-cu11-11.7.5.86-py3-none-manylinux1_x86_64.whl (204.1 MB)
    204.1/204.1 MB 4.4 MB/s eta 0:00:00
Collecting nvidia-nccl-cu11==2.12.5 (from torch)
  Downloading https://download.pytorch.org/whl/cu118/nvidia_nccl_cu11-2.12.5-py3-none-manylinux2014_x86_64.whl (147.8 MB)
    147.8/147.8 MB 7.1 MB/s eta 0:00:00
Collecting nvidia-nvtx-cu11==11.8.86 (from torch)
  Downloading https://download.pytorch.org/whl/cu118/nvidia_nvtx_cu11-11.8.86-py3-none-manylinux1_x86_64.whl (99 kB)
    99.1/99.1 kB 5.2 MB/s eta 0:00:00
Collecting triton==3.3.0 (from torch)
  Downloading https://download.pytorch.org/whl/triton-3.3.0-cp311-cp311-manylinux_2_27_x86_64_manylinux_2_28_x86_64.whl.metadata (27 kB)
Requirement already satisfied: setuptools>=40.8.0 in /usr/local/lib/python3.11/dist-packages (from triton==3.3.0->torch) (75)
Requirement already satisfied: numpy in /usr/local/lib/python3.11/dist-packages (from torchvision) (2.0.2)
Collecting torch
  Downloading https://download.pytorch.org/whl/cu118/torch-2.6.0%2Bcu118-cp311-cp311-linux_x86_64.whl.metadata (27 kB)
Requirement already satisfied: pillow!=8.3.*,>=5.3.0 in /usr/local/lib/python3.11/dist-packages (from torchvision) (11.1.0)
Requirement already satisfied: triton==3.2.0 in /usr/local/lib/python3.11/dist-packages (from torch) (3.2.0)
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.
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from jinja2->torch) (3.0.2)
  Downloading https://download.pytorch.org/whl/cu118/torch-2.6.0%2Bcu118-cp311-cp311-linux_x86_64.whl (848.7 MB)
    848.7/848.7 MB 2.3 MB/s eta 0:00:00
WARNING: Ignoring invalid distribution ~orch (/usr/local/lib/python3.11/dist-packages)
Installing collected packages: nvidia-nvtx-cu11, nvidia-nccl-cu11, nvidia-cuspars-cu11, nvidia-curand-cu11, nvidia-cufft-cu
WARNING: Ignoring invalid distribution ~orch (/usr/local/lib/python3.11/dist-packages)
Successfully installed nvidia-cublas-cu11-11.11.3.6 nvidia-cuda-cupti-cu11-11.8.87 nvidia-cuda-nvrtc-cu11-11.8.89 nvidia-cud
WARNING: The following packages were previously imported in this runtime:
[nvidia]
You must restart the runtime in order to use newly installed versions.

```

[RESTART SESSION](#)

```
!pip install bitsandbytes-cuda118
```

```
ERROR: Could not find a version that satisfies the requirement bitsandbytes-cuda118 (from versions: none)
ERROR: No matching distribution found for bitsandbytes-cuda118
```

```
import torch
print(torch.cuda.is_available())
```

```
True
```

```
!pip install -U bitsandbytes
```

```
Requirement already satisfied: bitsandbytes in /usr/local/lib/python3.11/dist-packages (0.45.5)
Requirement already satisfied: torch<3,>=2.0 in /usr/local/lib/python3.11/dist-packages (from bitsandbytes) (2.6.0+cu118)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-packages (from bitsandbytes) (2.0.2)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.0->bitsandbytes) (3.18.0)
Requirement already satisfied: typing-extensions>=4.10.0 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.0->bitsandbytes) (4.12.0)
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.0->bitsandbytes) (3.4.2)
Requirement already satisfied: Jinja2 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.0->bitsandbytes) (3.1.6)
Requirement already satisfied: fsspec in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.0->bitsandbytes) (2024.12.0)
Requirement already satisfied: nvidia-cuda-nvrtc-cu11==11.8.89 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.0->bitsandbytes) (11.8.89)
Requirement already satisfied: nvidia-cuda-runtime-cu11==11.8.89 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.0->bitsandbytes) (11.8.89)
Requirement already satisfied: nvidia-cuda-cupti-cu11==11.8.87 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.0->bitsandbytes) (11.8.87)
Requirement already satisfied: nvidia-cudnn-cu11==9.1.0.70 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.0->bitsandbytes) (9.1.0.70)
Requirement already satisfied: nvidia-cublas-cu11==11.11.3.6 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.0->bitsandbytes) (11.11.3.6)
Requirement already satisfied: nvidia-cufft-cu11==10.9.0.58 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.0->bitsandbytes) (10.9.0.58)
Requirement already satisfied: nvidia-curand-cu11==10.3.0.86 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.0->bitsandbytes) (10.3.0.86)
Requirement already satisfied: nvidia-cusolver-cu11==11.4.1.48 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.0->bitsandbytes) (11.4.1.48)
Requirement already satisfied: nvidia-cusparse-cu11==11.7.5.86 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.0->bitsandbytes) (11.7.5.86)
Requirement already satisfied: nvidia-nccl-cu11==2.21.5 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.0->bitsandbytes) (2.21.5)
Requirement already satisfied: nvidia-nvtx-cu11==11.8.86 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.0->bitsandbytes) (11.8.86)
Requirement already satisfied: triton==3.2.0 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.0->bitsandbytes) (3.2.0)
Requirement already satisfied: sympy==1.13.1 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.0->bitsandbytes) (1.13.1)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from Jinja2->torch<3,>=2.0->bitsandbytes) (2.1.5)
Traceback (most recent call last):
  File "/usr/local/lib/python3.11/dist-packages/pip/_internal/cli/base_command.py", line 179, in exc_logging_wrapper
    status = run_func(*args)
             ~~~~~~
  File "/usr/local/lib/python3.11/dist-packages/pip/_internal/cli/req_command.py", line 67, in wrapper
    return func(self, options, args)
           ~~~~~
  File "/usr/local/lib/python3.11/dist-packages/pip/_internal/commands/install.py", line 447, in run
    conflicts = self._determine_conflicts(to_install)
               ~~~~~
  File "/usr/local/lib/python3.11/dist-packages/pip/_internal/commands/install.py", line 578, in _determine_conflicts
    return check_install_conflicts(to_install)
           ~~~~~
  File "/usr/local/lib/python3.11/dist-packages/pip/_internal/operations/check.py", line 101, in check_install_conflicts
    package_set, _ = create_package_set_from_installed()
                    ~~~~~
  File "/usr/local/lib/python3.11/dist-packages/pip/_internal/operations/check.py", line 43, in create_package_set_from_installed
    package_set[name] = PackageDetails(dist.version, dependencies)
                        ~~~~~
  File "/usr/local/lib/python3.11/dist-packages/pip/_internal/metadata/importlib/_dists.py", line 175, in version
    ^C
```

```
# Load dataset (you can process it here)
# dataset = load_dataset(dataset_name, split="train")
```

```
# Load tokenizer and model with QLoRA configuration
from transformers import BitsAndBytesConfig
compute_dtype = getattr(torch, bnb_4bit_compute_dtype)
```

```
bnb_config = BitsAndBytesConfig(
    load_in_4bit=use_4bit,
    bnb_4bit_quant_type=bnb_4bit_quant_type,
    bnb_4bit_compute_dtype=compute_dtype,
    bnb_4bit_use_double_quant=use_nested_quant,
)
```

```
# Check GPU compatibility with bfloat16
if compute_dtype == torch.float16 and use_4bit:
    major, _ = torch.cuda.get_device_capability()
    if major >= 8:
        print("=" * 80)
```


```

print("Your GPU supports bfloat16: accelerate training with bf16=True")
print("=" * 80)

# Load base model
model = AutoModelForCausalLM.from_pretrained(
    model_name,
    quantization_config=bnb_config,
    device_map=device_map
)
model.config.use_cache = False
model.config.pretraining_tp = 1

# Load LLaMA tokenizer
tokenizer = AutoTokenizer.from_pretrained(model_name, trust_remote_code=True)
tokenizer.pad_token = tokenizer.eos_token
tokenizer.padding_side = "right" # Fix weird overflow issue with fp16 training

```

 /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
 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(
config.json: 100%                               583/583 [00:00<00:00, 67.8kB/s]

model.safetensors.index.json: 100%                26.8k/26.8k [00:00<00:00, 3.24MB/s]

Fetching 2 files: 100%                          2/2 [00:41<00:00, 41.27s/it]

model-00001-of-00002.safetensors: 100%           9.98G/9.98G [00:40<00:00, 317MB/s]

model-00002-of-00002.safetensors: 100%           3.50G/3.50G [00:16<00:00, 312MB/s]

Loading checkpoint shards: 100%                  2/2 [00:16<00:00, 7.44s/it]

generation_config.json: 100%                     200/200 [00:00<00:00, 24.4kB/s]

tokenizer_config.json: 100%                      746/746 [00:00<00:00, 103kB/s]

tokenizer.model: 100%                            500k/500k [00:00<00:00, 12.1MB/s]

tokenizer.json: 100%                             1.84M/1.84M [00:00<00:00, 2.68MB/s]

added_tokens.json: 100%                          21.0/21.0 [00:00<00:00, 2.66kB/s]

special_tokens_map.json: 100%                    435/435 [00:00<00:00, 56.2kB/s]

```

```

def predict(test, model, tokenizer):
    y_pred = []
    for i in tqdm(range(len(X_test))):
        prompt = X_test.iloc[i]["text"]
        pipe = pipeline(task="text-generation",
                        model=model,
                        tokenizer=tokenizer,
                        max_new_tokens = 1,
                        # temperature = 0.0,
                        do_sample=False,
                        )
        result = pipe(prompt)
        answer = result[0]['generated_text'].split("=")[-1]
        if "positive" in answer:
            y_pred.append("positive")
        elif "negative" in answer:
            y_pred.append("negative")
        elif "neutral" in answer:
            y_pred.append("neutral")
        else:
            y_pred.append("none")
    return y_pred

```

```
y_pred = predict(test, model, tokenizer)
```



```

94%|██████████| 840/900 [03:31<00:13, 3.91it/s]Device set to use cuda:0
94%|██████████| 849/900 [03:31<00:13, 3.89it/s]Device set to use cuda:0
94%|██████████| 850/900 [03:31<00:12, 3.93it/s]Device set to use cuda:0
95%|██████████| 851/900 [03:32<00:12, 3.93it/s]Device set to use cuda:0
95%|██████████| 852/900 [03:32<00:12, 3.94it/s]Device set to use cuda:0
95%|██████████| 853/900 [03:32<00:12, 3.68it/s]Device set to use cuda:0
95%|██████████| 854/900 [03:33<00:12, 3.77it/s]Device set to use cuda:0
95%|██████████| 855/900 [03:33<00:11, 3.82it/s]Device set to use cuda:0
95%|██████████| 856/900 [03:33<00:11, 3.84it/s]Device set to use cuda:0
95%|██████████| 857/900 [03:33<00:11, 3.83it/s]Device set to use cuda:0
95%|██████████| 858/900 [03:34<00:10, 3.84it/s]Device set to use cuda:0
95%|██████████| 859/900 [03:34<00:10, 3.81it/s]Device set to use cuda:0
96%|██████████| 860/900 [03:34<00:10, 3.80it/s]Device set to use cuda:0
96%|██████████| 861/900 [03:34<00:10, 3.86it/s]Device set to use cuda:0
96%|██████████| 862/900 [03:35<00:09, 3.88it/s]Device set to use cuda:0
96%|██████████| 863/900 [03:35<00:09, 3.91it/s]Device set to use cuda:0
96%|██████████| 864/900 [03:35<00:09, 3.93it/s]Device set to use cuda:0
96%|██████████| 865/900 [03:35<00:08, 3.93it/s]Device set to use cuda:0
96%|██████████| 866/900 [03:36<00:08, 3.96it/s]Device set to use cuda:0
96%|██████████| 867/900 [03:36<00:08, 3.70it/s]Device set to use cuda:0
96%|██████████| 868/900 [03:36<00:08, 3.78it/s]Device set to use cuda:0
97%|██████████| 869/900 [03:36<00:08, 3.81it/s]Device set to use cuda:0
97%|██████████| 870/900 [03:37<00:07, 3.86it/s]Device set to use cuda:0
97%|██████████| 871/900 [03:37<00:07, 3.86it/s]Device set to use cuda:0
97%|██████████| 872/900 [03:37<00:07, 3.85it/s]Device set to use cuda:0
97%|██████████| 873/900 [03:37<00:07, 3.83it/s]Device set to use cuda:0
97%|██████████| 874/900 [03:38<00:06, 3.81it/s]Device set to use cuda:0
97%|██████████| 875/900 [03:38<00:06, 3.83it/s]Device set to use cuda:0
97%|██████████| 876/900 [03:38<00:06, 3.79it/s]Device set to use cuda:0
97%|██████████| 877/900 [03:38<00:06, 3.76it/s]Device set to use cuda:0
98%|██████████| 878/900 [03:39<00:05, 3.81it/s]Device set to use cuda:0
98%|██████████| 879/900 [03:39<00:05, 3.85it/s]Device set to use cuda:0
98%|██████████| 880/900 [03:39<00:05, 3.88it/s]Device set to use cuda:0
98%|██████████| 881/900 [03:40<00:04, 3.82it/s]Device set to use cuda:0
98%|██████████| 882/900 [03:40<00:04, 3.82it/s]Device set to use cuda:0
98%|██████████| 883/900 [03:40<00:04, 3.84it/s]Device set to use cuda:0
98%|██████████| 884/900 [03:40<00:04, 3.82it/s]Device set to use cuda:0
98%|██████████| 885/900 [03:41<00:03, 3.85it/s]Device set to use cuda:0
98%|██████████| 886/900 [03:41<00:03, 3.84it/s]Device set to use cuda:0
99%|██████████| 887/900 [03:41<00:03, 3.83it/s]Device set to use cuda:0
99%|██████████| 888/900 [03:41<00:03, 3.84it/s]Device set to use cuda:0
99%|██████████| 889/900 [03:42<00:02, 3.78it/s]Device set to use cuda:0
99%|██████████| 890/900 [03:42<00:02, 3.81it/s]Device set to use cuda:0
99%|██████████| 891/900 [03:42<00:02, 3.85it/s]Device set to use cuda:0
99%|██████████| 892/900 [03:42<00:02, 3.91it/s]Device set to use cuda:0
99%|██████████| 893/900 [03:43<00:01, 3.91it/s]Device set to use cuda:0
99%|██████████| 894/900 [03:43<00:01, 3.92it/s]Device set to use cuda:0
99%|██████████| 895/900 [03:43<00:01, 3.89it/s]Device set to use cuda:0
100%|██████████| 896/900 [03:43<00:01, 3.93it/s]Device set to use cuda:0
100%|██████████| 897/900 [03:44<00:00, 3.95it/s]Device set to use cuda:0
100%|██████████| 898/900 [03:44<00:00, 3.90it/s]Device set to use cuda:0
100%|██████████| 899/900 [03:44<00:00, 3.86it/s]Device set to use cuda:0
100%|██████████| 900/900 [03:44<00:00, 4.00it/s]

```

```
evaluate(y_true, y_pred)
```

```

🔗 Accuracy: 0.722
Accuracy for label 0: 0.973
Accuracy for label 1: 0.303
Accuracy for label 2: 0.890

```

```
Classification Report:
```

	precision	recall	f1-score	support
0	0.81	0.97	0.88	300
1	0.80	0.30	0.44	300
2	0.63	0.89	0.74	300
accuracy			0.72	900
macro avg	0.75	0.72	0.69	900
weighted avg	0.75	0.72	0.69	900

```
Confusion Matrix:
```

```

[[292  7  1]
 [ 51 91 158]
 [ 17 16 267]]

```

```

# Load LoRA configuration
peft_config = LoraConfig( # Moved to SFTTrainer
    lora_alpha=lora_alpha,
    lora_dropout=lora_dropout,
    r=lora_r,
    bias="none".

```

```

        task_type="CAUSAL_LM",
    )

# Set training parameters
training_arguments = TrainingArguments(
    output_dir=output_dir,
    num_train_epochs=num_train_epochs,
    per_device_train_batch_size=per_device_train_batch_size,
    gradient_accumulation_steps=gradient_accumulation_steps,
    optim=optim,
    save_steps=save_steps,
    logging_steps=logging_steps,
    learning_rate=learning_rate,
    weight_decay=weight_decay,
    fp16=fp16,
    bf16=bf16,
    max_grad_norm=max_grad_norm,
    max_steps=max_steps,
    warmup_ratio=warmup_ratio,
    group_by_length=group_by_length,
    lr_scheduler_type=lr_scheduler_type,
    report_to="tensorboard"
)

# Set supervised fine-tuning parameters
trainer = SFTTrainer(
    model=model,
    train_dataset=train_data,
    peft_config=LoraConfig( # Declared here
        lora_alpha=lora_alpha,
        lora_dropout=lora_dropout,
        r=lora_r,
        bias="none",
        task_type="CAUSAL_LM",
    ),
    # dataset_text_field="text", # Removed
    # max_seq_length=max_seq_length, # Remove this line
    # tokenizer=tokenizer,
    args=training_arguments,
    # packing=packing,
)

# Train model
# trainer.train()

# # Save trained model
# trainer.model.save_pretrained(new_model)

```



```


-----
NameError                                Traceback (most recent call last)
<ipython-input-32-786002e7c8ed> in <cell line: 0>()
      1 # Load LoRA configuration
----> 2 peft_config = LoraConfig( # Moved to SFTTrainer
      3     lora_alpha=lora_alpha,
      4     lora_dropout=lora_dropout,
      5     r=lora_r,

NameError: name 'LoraConfig' is not defined

```

Next steps: [Explain error](#)

- Train model
trainer.train()


 [225/225 04:53, Epoch 1/1]

Step	Training Loss
------	---------------

25	1.816800
50	0.877400
75	0.954300
100	0.640700
125	0.901800
150	0.645900
175	0.912700
200	0.650800
225	0.820500

```
TrainOutput(global_step=225, training_loss=0.9134338039822049, metrics={'train_runtime': 295.7318,
'train_samples_per_second': 3.043, 'train_steps_per_second': 0.761, 'total_flos': 3619555072081920.0, 'train_loss':
0.9134338039822049})
```

```
# Save trained model and tokenizer
trainer.save_model()
trainer.model.save_pretrained('./trained')
tokenizer.save_pretrained('./trained')
```



```
('./trained/tokenizer_config.json',
 './trained/special_tokens_map.json',
 './trained/tokenizer.model',
 './trained/added_tokens.json',
 './trained/tokenizer.json')
```

```
%load_ext tensorboard
%tensorboard --logdir results/runs
```




Filter runs (regex)

☒ Run

☒ Apr28_01-24-35_905bc92ac654 ☐

Filter tags (regex)

All Scalars Image Histogram

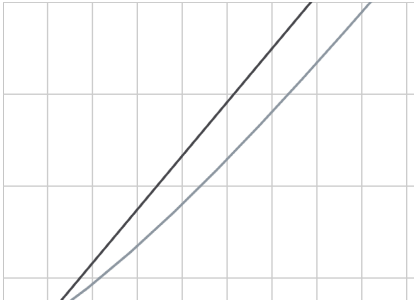
Settings

Pinned

Pin cards for a quick view and comparison

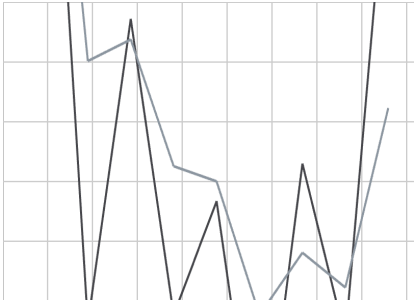
train 11 cards

train/epoch




Run	Smoothed	Value	Step	Rela
Apr28_01-24-35_905bc92ac654	0.8435	1	225	4.36

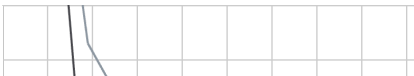
train/grad_norm



train/learning_rate



train/loss



Start coding or [generate](#) with AI.

```
import gc

del [model, tokenizer, peft_config, trainer, train_data, eval_data, bnb_config, training_arguments]
del [df, X_train, X_eval]
del [TrainingArguments, SFTTrainer, LoraConfig, BitsAndBytesConfig]

for _ in range(100):
    torch.cuda.empty_cache()
    gc.collect()

for _ in range(100):
    torch.cuda.empty_cache()
    gc.collect()

from peft import AutoPeftModelForCausalLM

finetuned_model = "./trained/"
compute_dtype = getattr(torch, "float16")
tokenizer = AutoTokenizer.from_pretrained(model_name )

model = AutoPeftModelForCausalLM.from_pretrained(
    finetuned_model,
    torch_dtype=torch.float16,
    return_dict=True,
    low_cpu_mem_usage=True,
    device_map=device,
)

merged_model = model.merge_and_unload()
```