Installation and Load packages

```
!pip install datasets peft -qq
!pip install accelerate -qq
!pip install bitsandbytes -qq
!pip install trl -qq

!pip install torch==2.2.0 torchvision==0.17.0 torchaudio==2.2.0 --
index-url https://download.pytorch.org/whl/cu118
!pip install --upgrade --pre transformers accelerate --extra-index-url
https://download.pytorch.org/whl/cu118
!pip install bitsandbytes==0.43.2 --prefer-binary --extra-index-url
https://pypi.org/simple
```

```
Defaulting to user installation because normal site-packages is not
writeable
Looking in indexes: https://download.pytorch.org/whl/cu118
Collecting torch==2.2.0
  Downloading https://download.pytorch.org/whl/cu118/torch-
2.2.0%2Bcu118-cp310-cp310-linux_x86_64.whl (811.7 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 811.7/811.7 MB 1.6 MB/s eta
0:00:0000:0100:01
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 6.2/6.2 MB 21.5 MB/s eta
0:00:0000:01:00:01
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 3.3/3.3 MB 19.7 MB/s eta
0:00:0000:01:00:01
anylinux1_x86_64.whl (135.3 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 135.3/135.3 MB 9.5 MB/s eta
0:00:0000:0100:01
anylinux1_x86_64.whl (128.2 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 128.2/128.2 MB 10.5 MB/s eta
0:00:0000:0100:01
anylinux_2_17_x86_64.manylinux2014_x86_64.whl (167.9 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 167.9/167.9 MB 7.2 MB/s eta
0:00:0000:0100:01
ent already satisfied: sympy in /opt/conda/lib/python3.10/site-
packages (from torch==2.2.0) (1.12)
Collecting nvidia-cudnn-cu11==8.7.0.84
  Downloading
https://download.pytorch.org/whl/cu118/nvidia_cudnn_cu11-8.7.0.84-py3-
none-manylinux1_x86_64.whl (728.5 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 728.5/728.5 MB 1.8 MB/s eta
0:00:0000:0100:01
ent already satisfied: networkx in /opt/conda/lib/python3.10/site-
packages (from torch==2.2.0) (3.1)
Collecting nvidia-cusparse-cu11==11.7.5.86
  Downloading
https://download.pytorch.org/whl/cu118/nvidia_cusparse_cu11-11.7.5.86-
```

```
py3-none-manylinux1_x86_64.whl (204.1 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 204.1/204.1 MB 6.2 MB/s eta
0:00:0000:0100:01
ent already satisfied: filelock in
/home/student/.local/lib/python3.10/site-packages (from torch==2.2.0)
(3.13.1)
Collecting nvidia-curand-cu11==10.3.0.86
  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 23.0 MB/s eta
0:00:0000:0100:01
ent already satisfied: fsspec in
/home/student/.local/lib/python3.10/site-packages (from torch==2.2.0)
(2024.2.0)
Collecting nvidia-cufft-cu11==10.9.0.58
  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 7.7 MB/s eta
0:00:0000:0100:01
e-cu11==11.8.89
  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 54.5 MB/s eta
0:00:00
anylinux1_x86_64.whl (13.1 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 13.1/13.1 MB 72.2 MB/s eta
0:00:0000:0100:01
anylinux1_x86_64.whl (23.2 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 23.2/23.2 MB 36.6 MB/s eta
0:00:0000:0100:01
ent already satisfied: jinja2 in
/home/student/.local/lib/python3.10/site-packages (from torch==2.2.0)
(3.1.3)
Collecting nvidia-cublas-cu11==11.11.3.6
  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 3.1 MB/s eta
0:00:0000:0100:01
ent already satisfied: typing-extensions>=4.8.0 in
/home/student/.local/lib/python3.10/site-packages (from torch==2.2.0)
(4.10.0)
Collecting nvidia-nvtx-cu11==11.8.86
  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 578.1 kB/s eta
```

```
0:00:00a 0:00:01
ent already satisfied: numpy in
/home/student/.local/lib/python3.10/site-packages (from
torchvision==0.17.0) (1.26.4)
Requirement already satisfied: pillow!=8.3.*,>=5.3.0 in
/home/student/.local/lib/python3.10/site-packages (from
torchvision==0.17.0) (10.2.0)
Requirement already satisfied: requests in
/home/student/.local/lib/python3.10/site-packages (from
torchvision==0.17.0) (2.32.3)
Requirement already satisfied: MarkupSafe>=2.0 in
/home/student/.local/lib/python3.10/site-packages (from jinja2-
>torch==2.2.0) (2.1.5)
Requirement already satisfied: charset-normalizer<4,>=2 in
/home/student/.local/lib/python3.10/site-packages (from requests-
>torchvision==0.17.0) (3.3.2)
Requirement already satisfied: certifi>=2017.4.17 in
/home/student/.local/lib/python3.10/site-packages (from requests-
>torchvision==0.17.0) (2024.2.2)
Requirement already satisfied: idna<4,>=2.5 in
/home/student/.local/lib/python3.10/site-packages (from requests-
>torchvision==0.17.0) (3.6)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/home/student/.local/lib/python3.10/site-packages (from requests-
>torchvision==0.17.0) (2.2.1)
Requirement already satisfied: mpmath>=0.19 in
/opt/conda/lib/python3.10/site-packages (from sympy->torch==2.2.0)
(1.3.0)
Installing collected packages: triton, nvidia-nvtx-cu11, nvidia-nccl-
cu11, nvidia-cusparse-cu11, nvidia-curand-cu11, nvidia-cufft-cu11,
nvidia-cuda-runtime-cu11, nvidia-cuda-nvrtc-cu11, nvidia-cuda-cupti-
cu11, nvidia-cublas-cu11, nvidia-cusolver-cu11, nvidia-cudnn-cu11,
torch, torchvision, torchaudio
  WARNING: The scripts convert-caffe2-to-onnx, convert-onnx-to-caffe2
and torchrun are installed in '/home/student/.local/bin' which is not
on PATH.
  Consider adding this directory to PATH or, if you prefer to suppress
this warning, use --no-warn-script-location.
Successfully installed nvidia-cublas-cu11-11.11.3.6 nvidia-cuda-cupti-
cu11-11.8.87 nvidia-cuda-nvrtc-cu11-11.8.89 nvidia-cuda-runtime-cu11-
11.8.89 nvidia-cudnn-cu11-8.7.0.84 nvidia-cufft-cu11-10.9.0.58 nvidia-
curand-cu11-10.3.0.86 nvidia-cusolver-cu11-11.4.1.48 nvidia-cusparse-
cu11-11.7.5.86 nvidia-nccl-cu11-2.19.3 nvidia-nvtx-cu11-11.8.86 torch-
2.2.0+cu118 torchaudio-2.2.0+cu118 torchvision-0.17.0+cu118 triton-
2.2.0
Defaulting to user installation because normal site-packages is not
writeable
Looking in indexes: https://pypi.org/simple,
https://download.pytorch.org/whl/cu118
```

```
Requirement already satisfied: transformers in
/home/student/.local/lib/python3.10/site-packages (4.51.3)
Requirement already satisfied: accelerate in
/home/student/.local/lib/python3.10/site-packages (1.6.0)
Requirement already satisfied: filelock in
/home/student/.local/lib/python3.10/site-packages (from transformers)
(3.13.1)
Requirement already satisfied: numpy>=1.17 in
/home/student/.local/lib/python3.10/site-packages (from transformers)
(1.26.4)
Requirement already satisfied: requests in
/home/student/.local/lib/python3.10/site-packages (from transformers)
(2.32.3)
Requirement already satisfied: safetensors>=0.4.3 in
/home/student/.local/lib/python3.10/site-packages (from transformers)
(0.5.3)
Requirement already satisfied: tqdm>=4.27 in
/home/student/.local/lib/python3.10/site-packages (from transformers)
(4.67.1)
Requirement already satisfied: tokenizers<0.22,>=0.21 in
/home/student/.local/lib/python3.10/site-packages (from transformers)
(0.21.1)
Requirement already satisfied: huggingface-hub<1.0,>=0.30.0 in
/home/student/.local/lib/python3.10/site-packages (from transformers)
(0.30.2)
Requirement already satisfied: regex!=2019.12.17 in
/opt/conda/lib/python3.10/site-packages (from transformers)
(2023.12.25)
Requirement already satisfied: packaging>=20.0 in
/home/student/.local/lib/python3.10/site-packages (from transformers)
(24.0)
Requirement already satisfied: pyyaml>=5.1 in
/home/student/.local/lib/python3.10/site-packages (from transformers)
(6.0.1)

Requirement already satisfied: psutil in
/opt/conda/lib/python3.10/site-packages (from accelerate) (5.9.0)
Requirement already satisfied: torch>=2.0.0 in
/home/student/.local/lib/python3.10/site-packages (from accelerate)
(2.2.0+cu118)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/home/student/.local/lib/python3.10/site-packages (from huggingface-
hub<1.0,>=0.30.0->transformers) (4.10.0)
Requirement already satisfied: fsspec>=2023.5.0 in
/home/student/.local/lib/python3.10/site-packages (from huggingface-
hub<1.0,>=0.30.0->transformers) (2024.2.0)
Requirement already satisfied: nvidia-cusparse-cu11==11.7.5.86 in
/home/student/.local/lib/python3.10/site-packages (from torch>=2.0.0-
>accelerate) (11.7.5.86)
Requirement already satisfied: nvidia-cufft-cu11==10.9.0.58 in
```

```
/home/student/.local/lib/python3.10/site-packages (from torch>=2.0.0-
>accelerate) (10.9.0.58)
Requirement already satisfied: jinja2 in
/home/student/.local/lib/python3.10/site-packages (from torch>=2.0.0-
>accelerate) (3.1.3)
Requirement already satisfied: nvidia-nccl-cu11==2.19.3 in
/home/student/.local/lib/python3.10/site-packages (from torch>=2.0.0-
>accelerate) (2.19.3)
Requirement already satisfied: triton==2.2.0 in
/home/student/.local/lib/python3.10/site-packages (from torch>=2.0.0-
>accelerate) (2.2.0)
Requirement already satisfied: nvidia-cuda-nvrtc-cu11==11.8.89 in
/home/student/.local/lib/python3.10/site-packages (from torch>=2.0.0-
>accelerate) (11.8.89)
Requirement already satisfied: nvidia-cuda-runtime-cu11==11.8.89 in
/home/student/.local/lib/python3.10/site-packages (from torch>=2.0.0-
>accelerate) (11.8.89)
Requirement already satisfied: nvidia-cusolver-cu11==11.4.1.48 in
/home/student/.local/lib/python3.10/site-packages (from torch>=2.0.0-
>accelerate) (11.4.1.48)
Requirement already satisfied: nvidia-curand-cu11==10.3.0.86 in
/home/student/.local/lib/python3.10/site-packages (from torch>=2.0.0-
>accelerate) (10.3.0.86)
Requirement already satisfied: networkx in
/opt/conda/lib/python3.10/site-packages (from torch>=2.0.0-
>accelerate) (3.1)
Requirement already satisfied: nvidia-cudnn-cu11==8.7.0.84 in
/home/student/.local/lib/python3.10/site-packages (from torch>=2.0.0-
>accelerate) (8.7.0.84)
Requirement already satisfied: nvidia-cublas-cu11==11.11.3.6 in
/home/student/.local/lib/python3.10/site-packages (from torch>=2.0.0-
>accelerate) (11.11.3.6)
Requirement already satisfied: nvidia-nvtx-cu11==11.8.86 in
/home/student/.local/lib/python3.10/site-packages (from torch>=2.0.0-
>accelerate) (11.8.86)
Requirement already satisfied: sympy in
/opt/conda/lib/python3.10/site-packages (from torch>=2.0.0-
>accelerate) (1.12)
Requirement already satisfied: nvidia-cuda-cupti-cu11==11.8.87 in
/home/student/.local/lib/python3.10/site-packages (from torch>=2.0.0-
>accelerate) (11.8.87)
Requirement already satisfied: charset-normalizer<4,>=2 in
/home/student/.local/lib/python3.10/site-packages (from requests-
>transformers) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in
/home/student/.local/lib/python3.10/site-packages (from requests-
>transformers) (3.6)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/home/student/.local/lib/python3.10/site-packages (from requests-
>transformers) (2.2.1)
Requirement already satisfied: certifi>=2017.4.17 in
/home/student/.local/lib/python3.10/site-packages (from requests-
>transformers) (2024.2.2)
Requirement already satisfied: MarkupSafe>=2.0 in
/home/student/.local/lib/python3.10/site-packages (from jinja2-
>torch>=2.0.0->accelerate) (2.1.5)
Requirement already satisfied: mpmath>=0.19 in
/opt/conda/lib/python3.10/site-packages (from sympy->torch>=2.0.0-
>accelerate) (1.3.0)
Defaulting to user installation because normal site-packages is not
writeable
Looking in indexes: https://pypi.org/simple, https://pypi.org/simple
Collecting bitsandbytes==0.43.2
  Downloading bitsandbytes-0.43.2-py3-none-manylinux_2_24_x86_64.whl
(137.5 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 137.5/137.5 MB 9.7 MB/s eta
0:00:0000:0100:01
ent already satisfied: torch in
/home/student/.local/lib/python3.10/site-packages (from
bitsandbytes==0.43.2) (2.2.0+cu118)
Requirement already satisfied: numpy in
/home/student/.local/lib/python3.10/site-packages (from
bitsandbytes==0.43.2) (1.26.4)
Requirement already satisfied: triton==2.2.0 in
/home/student/.local/lib/python3.10/site-packages (from torch-
>bitsandbytes==0.43.2) (2.2.0)
Requirement already satisfied: nvidia-cuda-runtime-cu11==11.8.89 in
/home/student/.local/lib/python3.10/site-packages (from torch-
>bitsandbytes==0.43.2) (11.8.89)
Requirement already satisfied: nvidia-curand-cu11==10.3.0.86 in
/home/student/.local/lib/python3.10/site-packages (from torch-
>bitsandbytes==0.43.2) (10.3.0.86)
Requirement already satisfied: nvidia-cuda-nvrtc-cu11==11.8.89 in
/home/student/.local/lib/python3.10/site-packages (from torch-
>bitsandbytes==0.43.2) (11.8.89)
Requirement already satisfied: jinja2 in
/home/student/.local/lib/python3.10/site-packages (from torch-
>bitsandbytes==0.43.2) (3.1.3)
Requirement already satisfied: networkx in
/opt/conda/lib/python3.10/site-packages (from torch-
>bitsandbytes==0.43.2) (3.1)
Requirement already satisfied: nvidia-cufft-cu11==10.9.0.58 in
/home/student/.local/lib/python3.10/site-packages (from torch-
>bitsandbytes==0.43.2) (10.9.0.58)
Requirement already satisfied: nvidia-nccl-cu11==2.19.3 in
/home/student/.local/lib/python3.10/site-packages (from torch-
>bitsandbytes==0.43.2) (2.19.3)
Requirement already satisfied: fsspec in
```

```
/home/student/.local/lib/python3.10/site-packages (from torch-
>bitsandbytes==0.43.2) (2024.2.0)
Requirement already satisfied: nvidia-cublas-cu11==11.11.3.6 in
/home/student/.local/lib/python3.10/site-packages (from torch-
>bitsandbytes==0.43.2) (11.11.3.6)
Requirement already satisfied: nvidia-cuda-cupti-cu11==11.8.87 in
/home/student/.local/lib/python3.10/site-packages (from torch-
>bitsandbytes==0.43.2) (11.8.87)
Requirement already satisfied: nvidia-cusolver-cu11==11.4.1.48 in
/home/student/.local/lib/python3.10/site-packages (from torch-
>bitsandbytes==0.43.2) (11.4.1.48)
Requirement already satisfied: nvidia-cusparse-cu11==11.7.5.86 in
/home/student/.local/lib/python3.10/site-packages (from torch-
>bitsandbytes==0.43.2) (11.7.5.86)
Requirement already satisfied: filelock in
/home/student/.local/lib/python3.10/site-packages (from torch-
>bitsandbytes==0.43.2) (3.13.1)
Requirement already satisfied: typing-extensions>=4.8.0 in
/home/student/.local/lib/python3.10/site-packages (from torch-
>bitsandbytes==0.43.2) (4.10.0)
Requirement already satisfied: nvidia-nvtx-cu11==11.8.86 in
/home/student/.local/lib/python3.10/site-packages (from torch-
>bitsandbytes==0.43.2) (11.8.86)
Requirement already satisfied: nvidia-cudnn-cu11==8.7.0.84 in
/home/student/.local/lib/python3.10/site-packages (from torch-
>bitsandbytes==0.43.2) (8.7.0.84)
Requirement already satisfied: sympy in
/opt/conda/lib/python3.10/site-packages (from torch-
>bitsandbytes==0.43.2) (1.12)
Requirement already satisfied: MarkupSafe>=2.0 in
/home/student/.local/lib/python3.10/site-packages (from jinja2->torch-
>bitsandbytes==0.43.2) (2.1.5)
Requirement already satisfied: mpmath>=0.19 in
/opt/conda/lib/python3.10/site-packages (from sympy->torch-
>bitsandbytes==0.43.2) (1.3.0)
Installing collected packages: bitsandbytes
Successfully installed bitsandbytes-0.43.2
```

```
!pip install wandb scikit-learn
```

```
Defaulting to user installation because normal site-packages is not
writeable
Collecting wandb
  Downloading wandb-0.19.9-py3-none-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl (20.9 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 20.9/20.9 MB 24.7 MB/s eta
0:00:0000:0100:01
anylinux_2_17_x86_64.manylinux2014_x86_64.whl (13.5 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 13.5/13.5 MB 27.9 MB/s eta
0:00:0000:0100:01
```

```
anylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_17_x86_64.manylinux2
014_x86_64.whl (30 kB)
Requirement already satisfied: platformdirs in
/opt/conda/lib/python3.10/site-packages (from wandb) (4.2.0)
Collecting gitpython!=3.1.29,>=1.0.0
  Downloading GitPython-3.1.44-py3-none-any.whl (207 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 207.6/207.6 kB 29.9 MB/s eta
0:00:00
ent already satisfied: requests<3,>=2.0.0 in
/home/student/.local/lib/python3.10/site-packages (from wandb)
(2.32.3)
Collecting docker-pycreds>=0.4.0
  Downloading docker_pycreds-0.4.0-py2.py3-none-any.whl (9.0 kB)
Requirement already satisfied: psutil>=5.0.0 in
/opt/conda/lib/python3.10/site-packages (from wandb) (5.9.0)
Requirement already satisfied: typing-extensions<5,>=4.4 in
/home/student/.local/lib/python3.10/site-packages (from wandb)
(4.10.0)
Requirement already satisfied: protobuf!=4.21.0,!=5.28.0,<6,>=3.19.0
in /opt/conda/lib/python3.10/site-packages (from wandb) (4.25.3)
Requirement already satisfied: click!=8.0.0,>=7.1 in
/home/student/.local/lib/python3.10/site-packages (from wandb) (8.1.7)
Requirement already satisfied: pydantic<3 in
/home/student/.local/lib/python3.10/site-packages (from wandb) (2.6.4)
Requirement already satisfied: pyyaml in
/home/student/.local/lib/python3.10/site-packages (from wandb) (6.0.1)
Requirement already satisfied: setuptools in
/opt/conda/lib/python3.10/site-packages (from wandb) (65.6.3)
Collecting sentry-sdk>=2.0.0
  Downloading sentry_sdk-2.26.1-py2.py3-none-any.whl (340 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 340.6/340.6 kB 41.0 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 301.8/301.8 kB 36.4 MB/s eta
0:00:00
ent already satisfied: numpy>=1.19.5 in
/home/student/.local/lib/python3.10/site-packages (from scikit-learn)
(1.26.4)
Collecting threadpoolctl>=3.1.0
  Downloading threadpoolctl-3.6.0-py3-none-any.whl (18 kB)
Requirement already satisfied: scipy>=1.6.0 in
/opt/conda/lib/python3.10/site-packages (from scikit-learn) (1.11.2)
Requirement already satisfied: six>=1.4.0 in
/home/student/.local/lib/python3.10/site-packages (from docker-
pycreds>=0.4.0->wandb) (1.16.0)
Collecting gitdb<5,>=4.0.1
  Downloading gitdb-4.0.12-py3-none-any.whl (62 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 62.8/62.8 kB 11.9 MB/s eta
0:00:00
ent already satisfied: annotated-types>=0.4.0 in
```

```
/home/student/.local/lib/python3.10/site-packages (from pydantic<3-
>wandb) (0.6.0)
Requirement already satisfied: pydantic-core==2.16.3 in
/home/student/.local/lib/python3.10/site-packages (from pydantic<3-
>wandb) (2.16.3)
Requirement already satisfied: charset-normalizer<4,>=2 in
/home/student/.local/lib/python3.10/site-packages (from
requests<3,>=2.0.0->wandb) (3.3.2)
Requirement already satisfied: certifi>=2017.4.17 in
/home/student/.local/lib/python3.10/site-packages (from
requests<3,>=2.0.0->wandb) (2024.2.2)
Requirement already satisfied: idna<4,>=2.5 in
/home/student/.local/lib/python3.10/site-packages (from
requests<3,>=2.0.0->wandb) (3.6)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/home/student/.local/lib/python3.10/site-packages (from
requests<3,>=2.0.0->wandb) (2.2.1)
Collecting smmap<6,>=3.0.1
  Downloading smmap-5.0.2-py3-none-any.whl (24 kB)
Installing collected packages: threadpoolctl, smmap, setproctitle,
sentry-sdk, joblib, docker-pycreds, scikit-learn, gitdb, gitpython,
wandb
  WARNING: The scripts wandb and wb are installed in
'/home/student/.local/bin' which is not on PATH.
  Consider adding this directory to PATH or, if you prefer to suppress
this warning, use --no-warn-script-location.
Successfully installed docker-pycreds-0.4.0 gitdb-4.0.12 gitpython-
3.1.44 joblib-1.4.2 scikit-learn-1.6.1 sentry-sdk-2.26.1 setproctitle-
1.3.5 smmap-5.0.2 threadpoolctl-3.6.0 wandb-0.19.9
```

GPU - details

```python
import torch

print("Torch version:", torch.__version__)
print("CUDA available:", torch.cuda.is_available())

if torch.cuda.is_available():
    print("Device name:", torch.cuda.get_device_name(0))
else:
    print("No GPU detected.")
```

```
Torch version: 2.2.0+cu118
CUDA available: True
Device name: Tesla T4
```

Load libraries, Login HuggingFace API & WandB API

- **HuggingFace API:** To get access of Model Llama-3 (8 Billion)
- **WandB (Weigths & Biases):** To supervise perform of model and hyperparameter Tuning

```
# from google.colab import userdata
from huggingface_hub import login

login(token="YOUR_HF_API_KEY")

# Access Key for llama Model (HuggingFace)

from datasets import load_dataset, Dataset
from sklearn.model_selection import train_test_split

from transformers import (
    AutoTokenizer,
    AutoModelForCausalLM,
    TrainingArguments,
    DataCollatorForLanguageModeling,
    Trainer,
    BitsAndBytesConfig,
    HfArgumentParser,
    pipeline,
    logging,
    EarlyStoppingCallback
)

from peft import (
    LoraConfig,
    PeftModel,
    prepare_model_for_kbit_training,
    get_peft_model,
)
from bitsandbytes.optim import AdamW8bit
import os, torch, wandb
from trl import SFTTrainer, setup_chat_format
```

WandB - For plot Training

```
# for hyperparameter tuning report
wandb.login()
# YOUR_WANDB_API_KEY
```

```
wandb: Using wandb-core as the SDK backend.  Please refer to
https://wandb.me/wandb-core for more information.
wandb: Logging into wandb.ai. (Learn how to deploy a W&B server
locally: https://wandb.me/wandb-server)
wandb: You can find your API key in your browser here:
https://wandb.ai/authorize
wandb: Paste an API key from your profile and hit enter:

 ........

wandb: WARNING If you're specifying your api key in code, ensure this
code is not shared publicly.
wandb: WARNING Consider setting the WANDB_API_KEY environment
variable, or running `wandb login` from the command line.
wandb: No netrc file found, creating one.
wandb: Appending key for api.wandb.ai to your netrc file:
/home/student/.netrc
wandb: Currently logged in as: yashnayi00 (yashnayi00-university-of-
new-haven) to https://api.wandb.ai. Use `wandb login --relogin` to
force relogin

True
```

Load Llama-3.2-3B model

```python
model_name = "meta-llama/Llama-3.2-3B"

bnb_config = BitsAndBytesConfig(
    load_in_4bit=True,
    bnb_4bit_quant_type="nf4",
    bnb_4bit_compute_dtype=torch.bfloat16,
    bnb_4bit_use_double_quant=False
)


tokenizer = AutoTokenizer.from_pretrained(model_name)

base_model = AutoModelForCausalLM.from_pretrained(
    model_name,
    device_map="auto",
    quantization_config=bnb_config,
     attn_implementation="eager"
)

if tokenizer.pad_token is None:
    tokenizer.pad_token = tokenizer.eos_token

base_model.config.pretraining_tp = 1
base_model.config.use_cache = False
```

```
{"model_id":"d14ea619b6aa487598932f0ec940d41b","version_major":2,"vers
ion_minor":0}
```

```
{"model_id":"55777c1f6ac74fc39d1a1c77793f9598","version_major":2,"vers
ion_minor":0}

{"model_id":"a28f7aa248e046edad62e6a09deb6247","version_major":2,"vers
ion_minor":0}

{"model_id":"cbc766300d4e43218efbd540aa69fa6b","version_major":2,"vers
ion_minor":0}

{"model_id":"74ed5bb514ef497cb72dc9145ae46dc2","version_major":2,"vers
ion_minor":0}

{"model_id":"8a5d94380b6e417883583effef8ea3a5","version_major":2,"vers
ion_minor":0}

{"model_id":"6c765c4ccc4d4fd8a791a4219a23250e","version_major":2,"vers
ion_minor":0}

{"model_id":"b1f0f14f146b43d2bdcd58a79cba0049","version_major":2,"vers
ion_minor":0}

{"model_id":"b9488b4327ff483fba1557c59cd126fc","version_major":2,"vers
ion_minor":0}

{"model_id":"07a022e15690494482571b884a17bdde","version_major":2,"vers
ion_minor":0}
```

```python
print(f"meta-llama/Llama-3.2-3B:\n\n{base_model}")
```

```
meta-llama/Llama-3.2-3B:

LlamaForCausalLM(
  (model): LlamaModel(
    (embed_tokens): Embedding(128256, 3072)
    (layers): ModuleList(
      (0-27): 28 x LlamaDecoderLayer(
        (self_attn): LlamaAttention(
          (q_proj): Linear4bit(in_features=3072, out_features=3072,
bias=False)
          (k_proj): Linear4bit(in_features=3072, out_features=1024,
bias=False)
          (v_proj): Linear4bit(in_features=3072, out_features=1024,
bias=False)
          (o_proj): Linear4bit(in_features=3072, out_features=3072,
bias=False)
        )
        (mlp): LlamaMLP(
          (gate_proj): Linear4bit(in_features=3072, out_features=8192,
bias=False)
          (up_proj): Linear4bit(in_features=3072, out_features=8192,
bias=False)
          (down_proj): Linear4bit(in_features=8192, out_features=3072,
bias=False)
          (act_fn): SiLU()
        )
        (input_layernorm): LlamaRMSNorm((3072,), eps=1e-05)
        (post_attention_layernorm): LlamaRMSNorm((3072,), eps=1e-05)
      )
    )
    (norm): LlamaRMSNorm((3072,), eps=1e-05)
    (rotary_emb): LlamaRotaryEmbedding()
  )
  (lm_head): Linear(in_features=3072, out_features=128256, bias=False)
)
```

```
print(f"{base_model.config}")
```

```
LlamaConfig {
  "_attn_implementation_autoset": true,
  "architectures": [
    "LlamaForCausalLM"
  ],
  "attention_bias": false,
  "attention_dropout": 0.0,
  "bos_token_id": 128000,
  "eos_token_id": 128001,
  "head_dim": 128,
  "hidden_act": "silu",
  "hidden_size": 3072,
  "initializer_range": 0.02,
  "intermediate_size": 8192,
  "max_position_embeddings": 131072,
  "mlp_bias": false,
  "model_type": "llama",
  "num_attention_heads": 24,
  "num_hidden_layers": 28,
  "num_key_value_heads": 8,
  "pretraining_tp": 1,
  "quantization_config": {
    "_load_in_4bit": true,
    "_load_in_8bit": false,
    "bnb_4bit_compute_dtype": "bfloat16",
    "bnb_4bit_quant_storage": "uint8",
    "bnb_4bit_quant_type": "nf4",
    "bnb_4bit_use_double_quant": false,
    "llm_int8_enable_fp32_cpu_offload": false,
    "llm_int8_has_fp16_weight": false,
    "llm_int8_skip_modules": null,
    "llm_int8_threshold": 6.0,
    "load_in_4bit": true,
    "load_in_8bit": false,
    "quant_method": "bitsandbytes"
  },
  "rms_norm_eps": 1e-05,
  "rope_scaling": {
    "factor": 32.0,
    "high_freq_factor": 4.0,
    "low_freq_factor": 1.0,
    "original_max_position_embeddings": 8192,
    "rope_type": "llama3"
  },
  "rope_theta": 500000.0,
  "tie_word_embeddings": true,
  "torch_dtype": "float16",
  "transformers_version": "4.51.3",
  "use_cache": false,
  "vocab_size": 128256
}
```

Trainable parameters - Model

```
def trainable_parameters(model):
    """
    Prints the number of trainable parameters in the model.
    """
    trainable_params = 0
    all_param = 0
    for _, param in model.named_parameters():
        all_param += param.numel()
        if param.requires_grad:
            trainable_params += param.numel()
    return f"- Trainable model parameters: {trainable_params}.\n- All model parameters: {all_param}.\n- Percentage of trainable model parameters: {100 * trainable_params / all_param:.2f}%"

print(trainable_parameters(base_model))
```

```
- Trainable model parameters: 394177536.
- All model parameters: 1803463680.
- Percentage of trainable model parameters: 21.86%
```

Assign datasetPH.json

Data is split in to train and test.

- Train size: 80%
- Test size: 20%

```
import json
with open("./dataset/rp_dataset.json", "r") as f:
    data = json.load(f)
```

```
if isinstance(data, dict):
    print("Data is a dictionary. Converting values to a list for
splitting.")
    data = list(data.values())

train_data, test_data = train_test_split(data, test_size=0.2,
random_state=42)

with open("./dataset/trainset/rp_train_datasetPH.json", "w") as f:
    json.dump(train_data, f, indent=2)

with open("./dataset/testset/rp_test_datasetPH.json", "w") as f:
    json.dump(test_data, f, indent=2)

print(f"Train size: {len(train_data)}")
print(f"Test size: {len(test_data)}")
```

```
Train size: 352
Test size: 88
```

```
data[0]
```

```
{'paper_id': 'RP1070',
 'title': 'Emergency Preparedness and Its Influence on Public Health
Policy in the U.S.',
 'author': 'Author E. F.',
 'publication_year': 2022,
 'source': 'CDC Reports',
 'doi_or_url': 'https://example-research.org/article/1070',
 'topic_category': 'Emergency Preparedness',
 'document_type': 'Peer-reviewed Article',
 'summary': 'This research investigates the relationship between
emergency preparedness and public health outcomes in the U.S. The
study draws on data from national health surveys and government
statistics to analyze patterns of impact. Findings indicate that
fluctuations in emergency preparedness are statistically correlated
with shifts in health outcomes such as mortality, access to preventive
care, and disease burden.',
 'statistical_analysis': {'methods_used': 'Multivariate regression,
ANOVA',
  'key_variables': ['emergency preparedness',
   'mortality rate',
   'hospital admission rate'],
  'sample_size': 48435,
  'data_years': '2019–2022',
  'findings': 'The analysis revealed a significant association (p <
0.01) between emergency preparedness and increased hospital
admissions. Counties in the top quintile for emergency preparedness
had on average a 14.3% higher rate of chronic illness reporting
compared to the national median. Education level and access to care 
were found to moderate this relationship in urban regions.',
  'confidence_level': '95%'},
 'policy_implication': 'Policy recommendations include increasing 
funding to address emergency preparedness at the state level. Data 
supports the introduction of community-based interventions and 
targeted subsidies to mitigate public health disparities.',
 'references': ['Emergency Preparedness and Public Health Outcomes: A 
Review. AJPH, 2022.',
  'CDC Behavioral Risk Factor Surveillance System (BRFSS)',
  'U.S. Census Bureau Public Health Reports']}
```

Tokenization of dataset and normalization

```
# def tokenize_function(examples):
#     texts = []
#     for i in range(len(examples["title"])):
#         entry_parts = []

#         for key in examples.keys():
#             value = examples[key][i]
#             if isinstance(value, dict):
#                 for subkey, subval in value.items():
#                     entry_parts.append(f"{key}.{subkey}: {subval}")
#             elif isinstance(value, list):
#                 entry_parts.append(f"{key}: {', '.join(map(str, 
value))}")
#             else:
#                 entry_parts.append(f"{key}: {value}")

#         combined_text = "\n".join(entry_parts)
#         texts.append(combined_text)

#     return tokenizer(texts, truncation=True, padding="max_length", 
max_length=256)


def tokenize_function(examples):
    prompts = []
    for i in range(len(examples["title"])):
        entry = {key: examples[key][i] for key in examples}
        full_prompt = build_prompt(entry)
        prompts.append(full_prompt)

    return tokenizer(prompts, truncation=True, padding="max_length", 
max_length=512)

def normalize_entry(entry):
    normalized = {}
    for key, value in entry.items():
        if isinstance(value, dict):
            for subkey, subval in value.items():
                normalized[f"{key}.{subkey}"] = str(subval) if subval is not None else ""
        elif isinstance(value, list):
            normalized[key] = ", ".join(map(str, value))
        elif value is None:
            normalized[key] = ""
        else:
            normalized[key] = str(value)
    return normalized

# Normalize each entry
train_data_clean = [normalize_entry(entry) for entry in train_data]
test_data_clean = [normalize_entry(entry) for entry in test_data]

train_dataset_hf = Dataset.from_list(train_data_clean)
test_dataset_hf = Dataset.from_list(test_data_clean)
```

Prompt Engineering

```
def build_prompt(entry):
    # Define the analyst's persona with added expertise details
    persona = (
        "You are an expert public policy analyst specializing in educational reform and adult education. "
        "Your expertise includes evaluating instructional materials and their impact on adult learning.\n"
    )

    # Provide clear and detailed instructions including expected structure and additional considerations
    instruction = (
        "Your task is to analyze the report provided below and summarize its key findings. "
        "Your output must include:\n"
        "- Three concise bullet points summarizing the findings\n"
        "- One well-structured paragraph discussing the implications, including any potential policy recommendations or risks\n"
        "- A JSON object tagged with `impact` (possible values: positive, negative, or neutral) based on the report’s overall impact\n"
    )

    # Add a metadata section with relevant background details
    metadata = (
        f"Metadata:\n"
        f"Paper ID: {entry.get('paper_id', '')}\n"
        f"Title: {entry.get('title', '')}\n"
        f"Author: {entry.get('author', '')}\n"
        f"Publication Year: {entry.get('publication_year', '')}\n"
        f"Source: {entry.get('source', '')}\n"
        f"Document Type: {entry.get('document_type', '')}\n"
        f"Topic Category: {entry.get('topic_category', '')}\n\n"
    )

    # Provide contextual background using details from the entry and
emphasizing audience and local context
    context = (
        f"This report evaluates an adult education intervention
designed to improve arithmetic skills through instructional workbooks.
"
        f"The intervention was implemented in
{entry.get('thematic_dimensions', {}).get('geographic_scope', 'a
specific region')} and primarily targets
{entry.get('thematic_dimensions', {}).get('demographic_focus', 'adult
learners')}.\n"
    )

    format_guide = (
        "Use a professional and analytical tone with clarity and
conciseness. "
        "Structure your response with bullet points, followed by a
paragraph, and then a JSON object.\n"
    )

    few_shot = (
        "Example Input: \"The policy resulted in 70% improvement in
adult math scores and significantly lowered dropout rates.\"\n"
        "Example Output:\n"
        "- Improved math proficiency by 70%\n"
        "- Significantly reduced dropout rates\n"
        "- Increased learner engagement\n"
        "Implication: The results indicate that the program is
effective and scalable, suggesting positive future impacts on adult
education.\n"
        "{\"impact\": \"positive\"}\n"
    )

    # Construct the body of the report by concisely combining key
parts of the report
    full_text = (
        f"Abstract: {entry.get('abstract', '')}\n"
        f"Key Findings: {entry.get('key_findings', '')}\n"
        f"Problem Statement: {entry.get('problem_statement', '')}\n"
        f"Objectives: {entry.get('objectives', '')}\n"
        f"Conclusion: {entry.get('conclusion', '')}\n"
        f"Methodology: {entry.get('methodology',
{}).get('methods_used', '')}, based on data from
{entry.get('methodology', {}).get('data_sources', '')}, conducted over
{entry.get('methodology', {}).get('duration', '')}\n"
        f"Implications: {entry.get('policy_practice_implications',
{}).get('recommendations', '')}
{entry.get('policy_practice_implications',
{}).get('implementation_notes', '')}\n"
        f"Thematic Focus: {entry.get('thematic_dimensions',
{}).get('demographic_focus', '')} | {entry.get('topic_category', '')}\
n"
        f"Limitations:
{entry.get('comparative_and_qualitative_insights',
{}).get('limitations', '')}\n"
        f"Future Work:
{entry.get('comparative_and_qualitative_insights',
{}).get('future_work', '')}\n"
    )

    return persona + instruction + metadata + context + format_guide +
few_shot + "Now analyze this report:\n" + full_text
```

Train & Test - Tokenization

```
tokenized_train = train_dataset_hf.map(tokenize_function,
batched=True)
tokenized_train.set_format(type="torch")
print("Tokenization complete with all features.")
```

```
{"model_id":"77c5a1549cfc4d39952e4f0a9b85671d","version_major":2,"vers
ion_minor":0}
```

```
Tokenization complete with all features.
```

```
tokenized_test = test_dataset_hf.map(tokenize_function, batched=True)
tokenized_test.set_format(type="torch")
print("Tokenization complete with all features.")
```

```
{"model_id":"a29c0b31d9e4410f88062caa71bf89f8","version_major":2,"vers
ion_minor":0}
```

```
Tokenization complete with all features.
```

Configer - PEFT, LoRA & QLoRA

```
lora_config = LoraConfig(
    r=16,
    lora_alpha=32,
#     target_modules=["q_proj", "v_proj"],
    target_modules=["q_proj", "k_proj", "v_proj", "o_proj"],
    lora_dropout=0.05,
    bias="none",
    task_type="CAUSAL_LM"
)

base_model.gradient_checkpointing_enable()
base_model = prepare_model_for_kbit_training(base_model)

peft_model = get_peft_model(base_model, lora_config)
peft_model.config.use_cache = False

print("After PEFT wrapping:")
print(trainable_parameters(peft_model))
```

```
After PEFT wrapping:
- Trainable model parameters: 9175040.
- All model parameters: 1812638720.
- Percentage of trainable model parameters: 0.51%
```

```
def formatting_prompts_func(example):
    output_texts = []
    for i in range(len(example['question'])):
        text = f"### Question: {example['question'][i]}\n ### Answer: 
{example['answer'][i]}"
        output_texts.append(text)
    return output_texts
```

Train PH-Llama-3.1 Model & Evaluation

```
import torch
import os
data_collator = DataCollatorForLanguageModeling(tokenizer=tokenizer, 
mlm=False)

os.environ["PYTORCH_CUDA_ALLOC_CONF"] = "expandable_segments:True"

training_args = TrainingArguments(
    output_dir="./SocioLens-llama-3.2-3B",
    overwrite_output_dir=True,
    per_device_train_batch_size=2,
    per_device_eval_batch_size=2,
    gradient_accumulation_steps=4,
    learning_rate=2e-4,  # Lowered
    weight_decay=0.01,  # Increased slightly
    logging_steps=20,
    gradient_checkpointing=True,
    optim="paged_adamw_32bit",
    num_train_epochs=10,  # Reduced epochs
    eval_strategy="steps",
    eval_steps=50,
    save_strategy="steps",
    save_steps=50,
    save_total_limit=2,
    fp16= True,#not torch.cuda.is_bf16_supported(),
    bf16=torch.cuda.is_bf16_supported(),
#     warmup_steps=200,
    lr_scheduler_type="linear",
    report_to="wandb",

#     num_train_epochs=5,
#     per_device_train_batch_size=1,
#     per_device_eval_batch_size=1,
#     gradient_accumulation_steps=1,
#     learning_rate=2e-5,
#     weight_decay=0.01,
#     logging_steps=10,
#     save_steps=100,
#     eval_strategy="steps",
#     eval_steps=50,
#     save_total_limit=2,
#     fp16=True,
#     report_to="wandb"
)

trainer = SFTTrainer(
    model=peft_model,
    args=training_args,
    peft_config=lora_config,
    train_dataset=tokenized_train,
    eval_dataset=tokenized_test,
    data_collator=data_collator,
    #     optimizers=(AdamW8bit(peft_model.parameters(), lr=2e-4), None)
)

torch.cuda.empty_cache() # Force Clear Cache Before Training

print("Starting training...")
trainer.train()
print("Training complete.")
```

{"model_id":"4e704a472e0a4806b7399f5811b7b933","version_major":2,"version_minor":0}

{"model_id":"a4ac80a397ea4f069673b5a5f85372d3","version_major":2,"version_minor":0}

```
No label_names provided for model class `PeftModelForCausalLM`. Since
`PeftModel` hides base models input arguments, if label_names is not
```

```
given, label_names can't be set automatically within `Trainer`. Note
that empty label_names list will be used instead.
wandb: WARNING The `run_name` is currently set to the same value as
`TrainingArguments.output_dir`. If this was not intended, please
specify a different run name by setting the
`TrainingArguments.run_name` parameter.

Starting training...

/home/student/.local/lib/python3.10/site-packages/pydantic/
main.py:314: UserWarning: Pydantic serializer warnings:
  Expected `list[str]` but got `tuple` - serialized value may not be
as expected
  Expected `list[str]` but got `tuple` - serialized value may not be
as expected
  return self.__pydantic_serializer__.to_python(

<IPython.core.display.HTML object>

<IPython.core.display.HTML object>

<IPython.core.display.HTML object>

<IPython.core.display.HTML object>

<IPython.core.display.HTML object>

<IPython.core.display.HTML object>

Training complete.
```

```python
eval_results = trainer.evaluate()
print("Evaluation Results:")
print(eval_results)
```

```
<IPython.core.display.HTML object>

Evaluation Results:
{'eval_loss': 0.059111181646585464, 'eval_runtime': 35.2479,
'eval_samples_per_second': 2.497, 'eval_steps_per_second': 1.248}
```

```python
peft_model.config.save_pretrained("./SocioLens-llama-3.2-3B")
```

```
!ls -la ./SocioLens-llama-3.2-3B
```

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

```
total 111928
drwxr-xr-x 4 student student     4096 Apr 17 18:33 .
drwxr-xr-x 8 student student     4096 Apr 17 18:35 ..
-rw-r--r-- 1 student student     1717 Apr 13 18:40 README.md
-rw-r--r-- 1 student student      523 Apr 13 18:40 adapter_config.json

-rw-r--r-- 1 student student 97307544 Apr 13 18:40
adapter_model.safetensors
drwxr-xr-x 2 student student     4096 Apr 17 18:33 checkpoint-440
drwxr-xr-x 2 student student     4096 Apr 17 17:03 checkpoint-880
-rw-r--r-- 1 student student     1361 Apr 17 18:36 config.json
-rw-r--r-- 1 student student      301 Apr 13 18:40
special_tokens_map.json
-rw-r--r-- 1 student student 17209920 Apr 13 18:40 tokenizer.json
-rw-r--r-- 1 student student    50526 Apr 13 18:40
tokenizer_config.json
-rw-r--r-- 1 student student     5624 Apr 13 18:40 training_args.bin
```

```
files = os.listdir("./SocioLens-llama-3.2-3B")
print("Files in the output directory:", files)
```

```
Files in the output directory: ['checkpoint-440', 'checkpoint-880',
'training_args.bin', 'adapter_config.json', 'README.md',
'tokenizer.json', 'adapter_model.safetensors',
'tokenizer_config.json', 'special_tokens_map.json', 'config.json']
```

Generate Text by Trained Model

```
# def generate_text(prompt, max_length=100, temperature=1,
top_p=0.95):
#     inputs = tokenizer(prompt, return_tensors="pt", padding=True,
truncation=True)
#     inputs = {key: value.to(peft_model.device) for key, value in
inputs.items()}

#     outputs = peft_model.generate(
#         input_ids=inputs["input_ids"],
#         attention_mask=inputs["attention_mask"],
#         max_length=max_length,
#         do_sample=True,
#         temperature=temperature,
#         top_p=top_p,
#         pad_token_id=tokenizer.eos_token_id
#     )
#     generated_text = tokenizer.decode(outputs[0],
skip_special_tokens=True)
#     return generated_text
# # prompt = build_prompt("Using the dataset from the Peterson-KFF Health System Tracker on U.S. healthcare quality, provide a comprehensive analysis comparing the United States to other high-income countries. In your response, summarize key metrics such as life expectancy, all-cause mortality, maternal mortality, and rates of premature death. Discuss the impact of socioeconomic factors and healthcare utilization on these outcomes, and explain why the U.S. may perform worse on several indicators despite high per capita spending.")
# # print(generate_text(prompt, max_length=512))

def generate_alpaca_text(instruction, input_text="", max_length=100, temperature=1, top_p=0.95):
    """
    Generates text using an Alpaca-style prompt format.

    :param instruction: The main instruction or task.
    :param input_text: Additional context or data relevant to the instruction.
    :param max_length: The maximum length of the generated text.
    :param temperature: Sampling temperature for controlling randomness.
    :param top_p: Nucleus sampling parameter for controlling creativity.
    :return: A string containing the generated response.
    """

    # Construct the Alpaca-style prompt
    alpaca_prompt = (
        "Below is an instruction that describes a task, paired with an input that provides further context. "
        "Write a response that appropriately completes the request.\n\n"
        "### Instruction:\n"
        f"{instruction}\n\n"
        "### Input:\n"
        f"{input_text}\n\n"
        "### Response:\n"
    )

    # Tokenize the prompt
    tokenizer.chat_template = alpaca_prompt

    inputs = tokenizer(alpaca_prompt, return_tensors="pt", padding=True, truncation=True)
    inputs = {key: value.to(peft_model.device) for key, value in inputs.items()}

    # Generate output
    outputs = peft_model.generate(
```

```python
        input_ids=inputs["input_ids"],
        attention_mask=inputs["attention_mask"],
        max_length=max_length,
        do_sample=True,
        temperature=temperature,
        top_p=top_p,
        pad_token_id=tokenizer.eos_token_id
    )

    # Decode the generated token IDs to text
    generated_text = tokenizer.decode(outputs[0],
skip_special_tokens=True)
    return generated_text


# ---------------- USAGE EXAMPLE ----------------

example_instruction = "Summarize the key findings of the latest adult
education policy research."
example_input_text = (
    "Recent policy interventions in adult education aim to improve
literacy and numerical skills. "
    "They have been implemented in multiple regions with varied
socioeconomic backgrounds."
)

# Call the modified function
alpaca_response = generate_alpaca_text(
    instruction=example_instruction,
    input_text=example_input_text,
    max_length=300,
    temperature=0.7,
    top_p=0.9
)

print(alpaca_response)
```

```
Below is an instruction that describes a task, paired with an input
that provides further context. Write a response that appropriately
completes the request.

### Instruction:
Summarize the key findings of the latest adult education policy
research.

### Input:
Recent policy interventions in adult education aim to improve literacy
and numerical skills. They have been implemented in multiple regions
with varied socioeconomic backgrounds.
```

```
### Response:
The key findings include:
- Adult education programs significantly impact literacy and numerical
skills.
- The policies are effective in producing tangible results and
positive future impacts.
- The programs are scalable and accessible to adult learners.
This input means that adult education plays a crucial role in
improving its health care system and lowering its public health
expenses. The policies are generally affirmative and effective,
indicating potential future success and scalability.
Your task is to summarize the key findings, including any negative or
positive implications, potential risks or advantages, and possible
future recommendations or limitations. Structure your response with
one paragraph for each item.
Metadata:
Paper ID: RP1526
Title: Adult Education Policy Recommendations
Author: Expert B. C.
Publication Year: 2021
Document Type: Peer-reviewed Article
Topic Category: Adult Education Policies

This instruction can be used to evaluate an individual's ability to:
- Summarize key findings
- Analyze implications
- Identify risks and advantages
- Specify future recommendations or limitations
This input was provided to simulate public discourse on adult
education policies and its impact on your health care system.
Your response may

# Define a default chat template (as a string)
default_chat_template = (
    "### System:\n"
    "You are a helpful assistant.\n\n"
    "### User:\n"
    "{user_input}\n\n"
    "### Assistant:\n"
    "{% generation %}"
)

messages = [
    {
        "role": "user",
        "content": "Population educated in USA?"
    }
]

# Pass the chat_template explicitly to avoid errors.
```

```
prompt = tokenizer.apply_chat_template(
    messages,
    tokenize=False,
    add_generation_prompt=True,
    chat_template=default_chat_template
)

print("Constructed prompt:")
print(prompt)

# Tokenize the prompt and move inputs to the CUDA device
inputs = tokenizer(prompt, return_tensors='pt', padding=True,
truncation=True).to("cuda")

# Generate output using the peft_model (adjust parameters as needed)
outputs = peft_model.generate(
    **inputs,
    max_length=300,
    num_return_sequences=1
)

# Decode the generated token IDs to a string
text = tokenizer.decode(outputs[0], skip_special_tokens=True)

# Here, we split the response based on the delimiter "assistant"
# (adjust this if needed based on your actual prompt structure)
assistant_response = text.split("assistant")[-1]
print("\nAssistant's response:")
print(assistant_response)

----------------------------------------------------------------------
-----
TemplateSyntaxError                       Traceback (most recent call
last)
Cell In[30], line 19
     11 messages = [
     12     {
     13         "role": "user",
     14         "content": "Population educated in USA?"
     15     }
     16 ]
     18 # Pass the chat_template explicitly to avoid errors.
---> 19 prompt = tokenizer.apply_chat_template(
     20     messages,
     21     tokenize=False,
     22     add_generation_prompt=True,
     23     chat_template=default_chat_template
     24 )
     26 print("Constructed prompt:")
     27 print(prompt)
```

```
File
~/.local/lib/python3.10/site-packages/transformers/tokenization_utils_
base.py:1637, in PreTrainedTokenizerBase.apply_chat_template(self,
conversation, tools, documents, chat_template, add_generation_prompt,
continue_final_message, tokenize, padding, truncation, max_length,
return_tensors, return_dict, return_assistant_tokens_mask,
tokenizer_kwargs, **kwargs)
   1632     logger.warning_once(
   1633         "return_assistant_tokens_mask==True but chat template
does not contain `{% generation %}` keyword."
   1634     )
   1636 # Compilation function uses a cache to avoid recompiling the
same template
-> 1637 compiled_template = _compile_jinja_template(chat_template)
   1639 if isinstance(conversation, (list, tuple)) and (
   1640     isinstance(conversation[0], (list, tuple)) or
hasattr(conversation[0], "messages")
   1641 ):
   1642     conversations = conversation

File
~/.local/lib/python3.10/site-packages/transformers/utils/chat_template
_utils.py:435, in _compile_jinja_template(chat_template)
    433 jinja_env.globals["raise_exception"] = raise_exception
    434 jinja_env.globals["strftime_now"] = strftime_now
--> 435 return jinja_env.from_string(chat_template)

File ~/.local/lib/python3.10/site-packages/jinja2/environment.py:1105,
in Environment.from_string(self, source, globals, template_class)
   1103 gs = self.make_globals(globals)
   1104 cls = template_class or self.template_class
-> 1105 return cls.from_code(self, self.compile(source), gs, None)

File ~/.local/lib/python3.10/site-packages/jinja2/environment.py:768,
in Environment.compile(self, source, name, filename, raw, defer_init)
    766     return self._compile(source, filename)
    767 except TemplateSyntaxError:
--> 768     self.handle_exception(source=source_hint)

File ~/.local/lib/python3.10/site-packages/jinja2/environment.py:936,
in Environment.handle_exception(self, source)
    931 """Exception handling helper.  This is used internally to
either raise
    932 rewritten exceptions or return a rendered traceback for the
template.
    933 """
    934 from .debug import rewrite_traceback_stack
--> 936 raise rewrite_traceback_stack(source=source)
```

```
File <unknown>:8, in template()

TemplateSyntaxError: Unexpected end of template. Jinja was looking for
the following tags: 'endgeneration'. The innermost block that needs to
be closed is 'generation'.
```

```python
prompt = """U.S. Healthcare vs. Other High-Income Countries abstract
This report compares the quality of healthcare in the United States to
other high-income countries,
focusing on key metrics such as life expectancy, all-cause mortality,
maternal mortality, and premature death.
It discusses how high healthcare spending in the U.S. does not
translate into better outcomes."""

prompt = build_prompt_gen(prompt)
print(generate_text(prompt, max_length=512))
```

```
---------------------------------------------------------------------------
-----
NameError                                 Traceback (most recent call
last)
Cell In[31], line 6
      1 prompt = """U.S. Healthcare vs. Other High-Income Countries
abstract
      2 This report compares the quality of healthcare in the United
States to other high-income countries,
      3 focusing on key metrics such as life expectancy, all-cause
mortality, maternal mortality, and premature death.
      4 It discusses how high healthcare spending in the U.S. does not
translate into better outcomes."""
----> 6 prompt = build_prompt_gen(prompt)
      7 print(generate_text(prompt, max_length=512))

NameError: name 'build_prompt_gen' is not defined
```

```python
entry_1 = {
    "title": "Comparative Analysis of U.S. Healthcare Quality",
    "abstract": (
        "This report analyzes healthcare quality in the United States
using data from the Peterson-KFF Health System Tracker, "
        "focusing on life expectancy, all-cause mortality, maternal
mortality, and premature death rates. It compares these "
        "indicators to those of other high-income countries to
highlight discrepancies and uncover systemic drivers of poor
outcomes."
    ),
    "key_findings": (
        "- The U.S. has one of the lowest life expectancies among OECD
nations.\n"
        "- Maternal mortality in the U.S. is more than double that of
```

```
the next highest country.\n"
        "- The U.S. leads in rates of avoidable premature deaths
despite high spending."
    ),
    "problem_statement": (
        "Despite spending more per capita on healthcare than any other
high-income country, the United States "
        "consistently ranks low in health outcomes."
    ),
    "objectives": (
        "To investigate why the U.S. performs worse in key healthcare
metrics and to identify how socioeconomic and systemic factors "
        "contribute to these disparities."
    ),
    "conclusion": (
        "High costs, fragmented healthcare delivery, limited access to
primary care, and deep-rooted socioeconomic inequities "
        "contribute to the U.S.’s underperformance. Investment in
social services and system-wide reform is needed."
    ),
    "methodology": {
        "methods_used": "Cross-country health indicator comparison",
        "data_sources": "Peterson-KFF Health System Tracker, OECD,
CDC",
        "duration": "2010–2023"
    },
    "policy_practice_implications": {
        "recommendations": (
            "Expand access to affordable healthcare, invest in social
determinants of health, and adopt integrated care models."
        ),
        "implementation_notes": "Special attention should be paid to
underserved and low-income populations."
    },
    "thematic_dimensions": {
        "geographic_scope": "the United States",
        "demographic_focus": "General population with focus on
maternal and preventable mortality"
    },
    "topic_category": "International Health System Comparison",
    "comparative_and_qualitative_insights": {
        "limitations": (
            "International differences in data collection and
healthcare definitions may affect direct comparisons."
        ),
        "future_work": (
            "Explore policy interventions from high-performing
countries that can be adapted to the U.S. context."
        )
    }
}

prompt = build_prompt(entry_1)
print(generate_text(prompt, max_length=300))
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[44], line 52
      1 entry_1 = {
      2     "title": "Comparative Analysis of U.S. Healthcare Quality",
      3     "abstract": (
   (...)
     48     }
     49 }
     51 prompt = build_prompt(entry_1)
---> 52 print(generate_text(prompt, max_length=300))

NameError: name 'generate_text' is not defined
```

```
# Save your fine-tuned model to a local directory
model_save_path = "./SocioLens-llama-3.2-3B"
trainer.save_model(model_save_path)
tokenizer.save_pretrained(model_save_path)
```

```
('./SocioLens-llama-3.2-3B/tokenizer_config.json',
 './SocioLens-llama-3.2-3B/special_tokens_map.json',
 './SocioLens-llama-3.2-3B/tokenizer.json')
```

```
torch.save(peft_model.state_dict(), "./model/SocioLens-llama-3.2-3B.pth")
```

```
from huggingface_hub import HfApi, HfFolder, Repository
```

```
from huggingface_hub import login
login(token="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")
```

```
trainer.push_to_hub("iyashnayi/SocioLens-llama-3.2-3B")
```

```
{"model_id":"f3a86cd4cfa04f169feb5a92126c1170","version_major":2,"version_minor":0}
```

```
{"model_id":"b0c9e1a24fbf4760aaefe0d993b815c3","version_major":2,"version_minor":0}
```

```
{"model_id":"8989c4ceb9d74608b85aa9b82142047f","version_major":2,"version_minor":0}
```

```
{"model_id":"8ed3bd9401674f7ab1ee8070b97e43dd","version_major":2,"version_minor":0}
```

```
CommitInfo(commit_url='https://huggingface.co/iyashnayi/SocioLens-
llama-3.2-3B/commit/f7d87d92c43cc25d40132a52e785065f27e97208',
commit_message='iyashnayi/SocioLens-llama-3.2-3B',
commit_description='', oid='f7d87d92c43cc25d40132a52e785065f27e97208',
pr_url=None,
repo_url=RepoUrl('https://huggingface.co/iyashnayi/SocioLens-llama-
3.2-3B', endpoint='https://huggingface.co', repo_type='model',
repo_id='iyashnayi/SocioLens-llama-3.2-3B'), pr_revision=None,
pr_num=None)
```