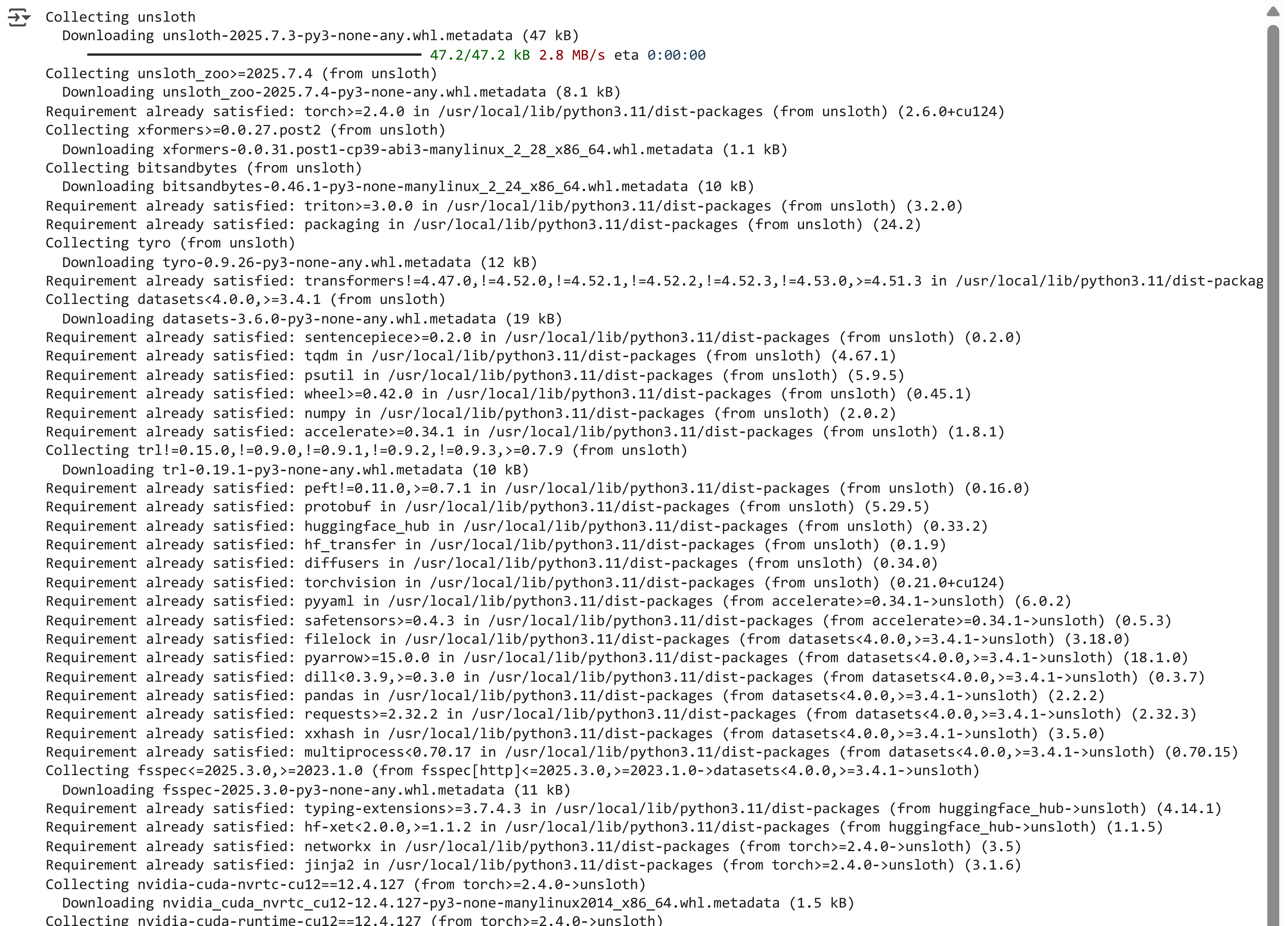
# NER task - Product Attribute Extraction

```
!pip install unsloth
```

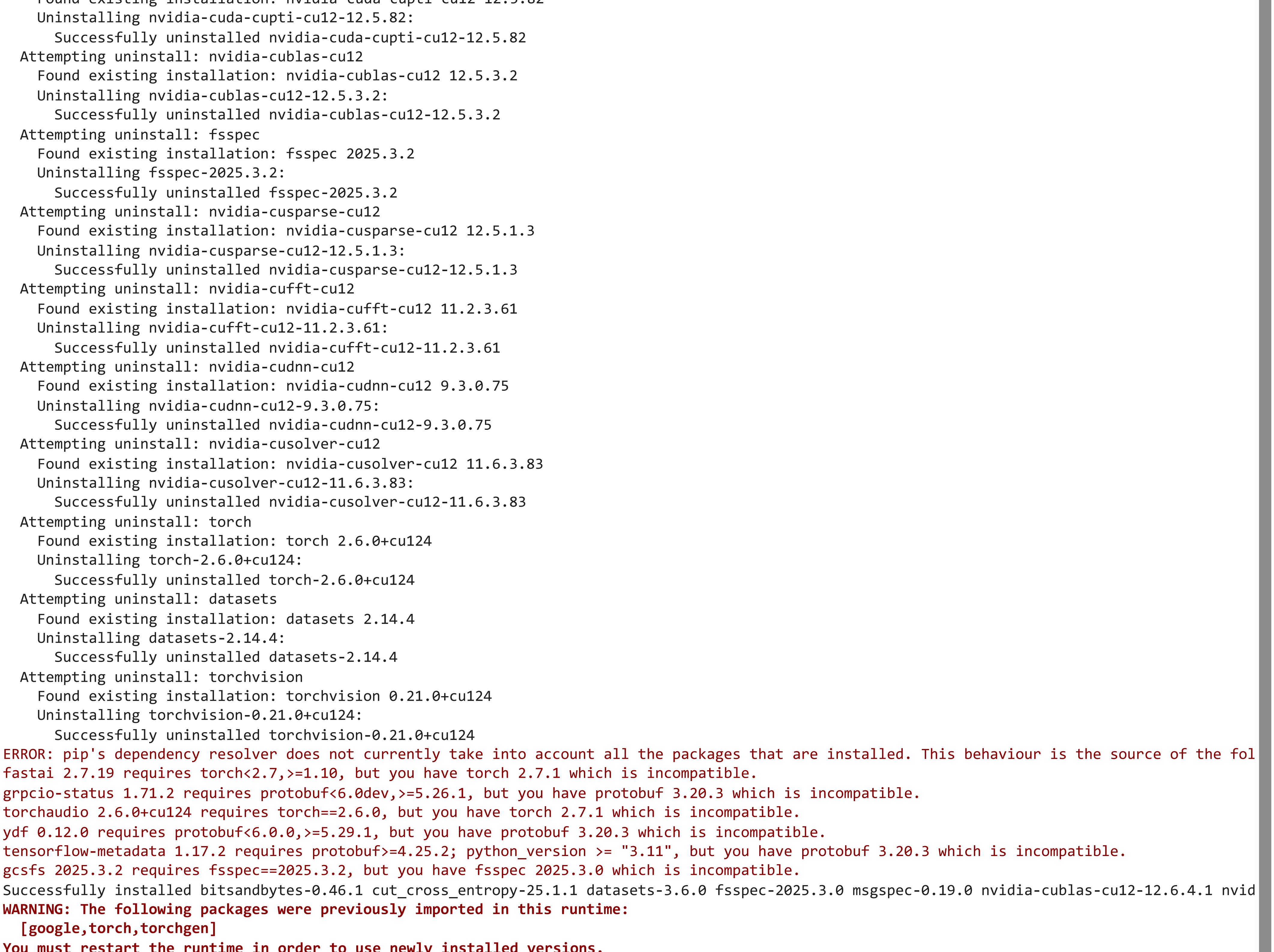


```
Collecting unsloth
  Downloading unsloth-2025.7.3-py3-none-any.whl.metadata (47 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 47.2/47.2 kB 2.8 MB/s eta 0:00:00
Collecting unsloth_zoo>=2025.7.4 (from unsloth)
  Downloading unsloth_zoo-2025.7.4-py3-none-any.whl.metadata (8.1 kB)
Requirement already satisfied: torch>=2.4.0 in /usr/local/lib/python3.11/dist-packages (from unsloth) (2.6.0+cu124)
Collecting xformers>=0.0.27.post2 (from unsloth)
  Downloading xformers-0.0.31.post1-cp39-abi3-manylinux_2_28_x86_64.whl.metadata (1.1 kB)
Collecting bitsandbytes (from unsloth)
  Downloading bitsandbytes-0.46.1-py3-none-manylinux_2_24_x86_64.whl.metadata (10 kB)
Requirement already satisfied: triton>=3.0.0 in /usr/local/lib/python3.11/dist-packages (from unsloth) (3.2.0)
Requirement already satisfied: packaging in /usr/local/lib/python3.11/dist-packages (from unsloth) (24.2)
Collecting tyro (from unsloth)
  Downloading tyro-0.9.26-py3-none-any.whl.metadata (12 kB)
Requirement already satisfied: transformers!=4.47.0,!=4.52.0,!=4.52.1,!=4.52.2,!=4.52.3,!=4.53.0,>=4.51.3 in /usr/local/lib/python3.11/dist-packag
Collecting datasets<4.0.0,>=3.4.1 (from unsloth)
  Downloading datasets-3.6.0-py3-none-any.whl.metadata (19 kB)
Requirement already satisfied: sentencepiece>=0.2.0 in /usr/local/lib/python3.11/dist-packages (from unsloth) (0.2.0)
Requirement already satisfied: tqdm in /usr/local/lib/python3.11/dist-packages (from unsloth) (4.67.1)
Requirement already satisfied: psutil in /usr/local/lib/python3.11/dist-packages (from unsloth) (5.9.5)
Requirement already satisfied: wheel>=0.42.0 in /usr/local/lib/python3.11/dist-packages (from unsloth) (0.45.1)
Requirement already satisfied: numpy in /usr/local/lib/python3.11/dist-packages (from unsloth) (2.0.2)
Requirement already satisfied: accelerate>=0.34.1 in /usr/local/lib/python3.11/dist-packages (from unsloth) (1.8.1)
Collecting trl!=0.15.0,!=0.9.0,!=0.9.1,!=0.9.2,!=0.9.3,>=0.7.9 (from unsloth)
  Downloading trl-0.19.1-py3-none-any.whl.metadata (10 kB)
Requirement already satisfied: peft!=0.11.0,>=0.7.1 in /usr/local/lib/python3.11/dist-packages (from unsloth) (0.16.0)
Requirement already satisfied: protobuf in /usr/local/lib/python3.11/dist-packages (from unsloth) (5.29.5)
Requirement already satisfied: huggingface_hub in /usr/local/lib/python3.11/dist-packages (from unsloth) (0.33.2)
Requirement already satisfied: hf_transfer in /usr/local/lib/python3.11/dist-packages (from unsloth) (0.1.9)
Requirement already satisfied: diffusers in /usr/local/lib/python3.11/dist-packages (from unsloth) (0.34.0)
Requirement already satisfied: torchvision in /usr/local/lib/python3.11/dist-packages (from unsloth) (0.21.0+cu124)
Requirement already satisfied: pyyaml in /usr/local/lib/python3.11/dist-packages (from accelerate>=0.34.1->unsloth) (6.0.2)
Requirement already satisfied: safetensors>=0.4.3 in /usr/local/lib/python3.11/dist-packages (from accelerate>=0.34.1->unsloth) (0.5.3)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from datasets<4.0.0,>=3.4.1->unsloth) (3.18.0)
Requirement already satisfied: pyarrow>=15.0.0 in /usr/local/lib/python3.11/dist-packages (from datasets<4.0.0,>=3.4.1->unsloth) (18.1.0)
Requirement already satisfied: dill<0.3.9,>=0.3.0 in /usr/local/lib/python3.11/dist-packages (from datasets<4.0.0,>=3.4.1->unsloth) (0.3.7)
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages (from datasets<4.0.0,>=3.4.1->unsloth) (2.2.2)
Requirement already satisfied: requests>=2.32.2 in /usr/local/lib/python3.11/dist-packages (from datasets<4.0.0,>=3.4.1->unsloth) (2.32.3)
Requirement already satisfied: xxhash in /usr/local/lib/python3.11/dist-packages (from datasets<4.0.0,>=3.4.1->unsloth) (3.5.0)
Requirement already satisfied: multiprocess<0.70.17 in /usr/local/lib/python3.11/dist-packages (from datasets<4.0.0,>=3.4.1->unsloth) (0.70.15)
Collecting fsspec<=2025.3.0,>=2023.1.0 (from fsspec[http]<=2025.3.0,>=2023.1.0->datasets<4.0.0,>=3.4.1->unsloth)
  Downloading fsspec-2025.3.0-py3-none-any.whl.metadata (11 kB)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.11/dist-packages (from huggingface_hub->unsloth) (4.14.1)
Requirement already satisfied: hf-xet<2.0.0,>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from huggingface_hub->unsloth) (1.1.5)
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-packages (from torch>=2.4.0->unsloth) (3.5)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.11/dist-packages (from torch>=2.4.0->unsloth) (3.1.6)
Collecting nvidia-cuda-nvrtc-cu12==12.4.127 (from torch>=2.4.0->unsloth)
  Downloading nvidia_cuda_nvrtc_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cuda-runtime-cu12==12.4.127 (from torch>=2.4.0->unsloth)
```

```
  Downloading nvidia_cuda_runtime_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cuda-cupti-cu12==12.4.127 (from torch>=2.4.0->unsloth)
  Downloading nvidia_cuda_cupti_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cudnn-cu12==9.1.0.70 (from torch>=2.4.0->unsloth)
  Downloading nvidia_cudnn_cu12-9.1.0.70-py3-none-manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cublas-cu12==12.4.5.8 (from torch>=2.4.0->unsloth)
  Downloading nvidia_cublas_cu12-12.4.5.8-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cufft-cu12==11.2.1.3 (from torch>=2.4.0->unsloth)
  Downloading nvidia_cufft_cu12-11.2.1.3-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-curand-cu12==10.3.5.147 (from torch>=2.4.0->unsloth)
  Downloading nvidia_curand_cu12-10.3.5.147-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cusolver-cu12==11.6.1.9 (from torch>=2.4.0->unsloth)
  Downloading nvidia_cusolver_cu12-11.6.1.9-py3-none-manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cusparse-cu12==12.3.1.170 (from torch>=2.4.0->unsloth)
  Downloading nvidia_cusparse_cu12-12.3.1.170-py3-none-manylinux2014_x86_64.whl.metadata (1.6 kB)
Requirement already satisfied: nvidia-cusparselt-cu12==0.6.2 in /usr/local/lib/python3.11/dist-packages (from torch>=2.4.0->unsloth) (0.6.2)
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in /usr/local/lib/python3.11/dist-packages (from torch>=2.4.0->unsloth) (2.21.5)
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=2.4.0->unsloth) (12.4.127)
Collecting nvidia-nvjitlink-cu12==12.4.127 (from torch>=2.4.0->unsloth)
  Downloading nvidia_nvjitlink_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Requirement already satisfied: sympy==1.13.1 in /usr/local/lib/python3.11/dist-packages (from torch>=2.4.0->unsloth) (1.13.1)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.11/dist-packages (from sympy==1.13.1->torch>=2.4.0->unsloth) (1.3.0)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.11/dist-packages (from transformers!=4.47.0,!=4.52.0,!=4.52.1,!=4.52.2,
Requirement already satisfied: tokenizers<0.22,>=0.21 in /usr/local/lib/python3.11/dist-packages (from transformers!=4.47.0,!=4.52.0,!=4.52.1,!=4.
Collecting protobuf (from unsloth)
  Downloading protobuf-3.20.3-py2.py3-none-any.whl.metadata (720 bytes)
Collecting cut_cross_entropy (from unsloth_zoo>=2025.7.4->unsloth)
  Downloading cut_cross_entropy-25.1.1-py3-none-any.whl.metadata (9.3 kB)
Requirement already satisfied: pillow in /usr/local/lib/python3.11/dist-packages (from unsloth_zoo>=2025.7.4->unsloth) (11.2.1)
Collecting msgspec (from unsloth_zoo>=2025.7.4->unsloth)
  Downloading msgspec-0.19.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (6.9 kB)
Collecting torch>=2.4.0 (from unsloth)
  Downloading torch-2.7.1-cp311-cp311-manylinux_2_28_x86_64.whl.metadata (29 kB)
Collecting sympy>=1.13.3 (from torch>=2.4.0->unsloth)
  Downloading sympy-1.14.0-py3-none-any.whl.metadata (12 kB)
Collecting nvidia-cuda-nvrtc-cu12==12.6.77 (from torch>=2.4.0->unsloth)
  Downloading nvidia_cuda_nvrtc_cu12-12.6.77-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cuda-runtime-cu12==12.6.77 (from torch>=2.4.0->unsloth)
  Downloading nvidia_cuda_runtime_cu12-12.6.77-py3-none-manylinux2014_x86_64.manylinux_2_17_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cuda-cupti-cu12==12.6.80 (from torch>=2.4.0->unsloth)
  Downloading nvidia_cuda_cupti_cu12-12.6.80-py3-none-manylinux2014_x86_64.manylinux_2_17_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cudnn-cu12==9.5.1.17 (from torch>=2.4.0->unsloth)
  Downloading nvidia_cudnn_cu12-9.5.1.17-py3-none-manylinux_2_28_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cublas-cu12 (from nvidia-cudnn-cu12==9.1.0.70->torch>=2.4.0->unsloth)
  Downloading nvidia_cublas_cu12-12.6.4.1-py3-none-manylinux2014_x86_64.manylinux_2_17_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cufft-cu12==11.3.0.4 (from torch>=2.4.0->unsloth)
  Downloading nvidia_cufft_cu12-11.3.0.4-py3-none-manylinux2014_x86_64.manylinux_2_17_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-curand-cu12==10.3.7.77 (from torch>=2.4.0->unsloth)
  Downloading nvidia_curand_cu12-10.3.7.77-py3-none-manylinux2014_x86_64.manylinux_2_17_x86_64.whl.metadata (1.5 kB)
```

```
Collecting nvidia-cusolver-cu12==11.7.1.2 (from torch>=2.4.0->unsloth)
  Downloading nvidia_cusolver_cu12-11.7.1.2-py3-none-manylinux2014_x86_64.manylinux_2_17_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cusparse-cu12 (from nvidia-cusolver-cu12==11.6.1.9->torch>=2.4.0->unsloth)
  Downloading nvidia_cusparse_cu12-12.5.4.2-py3-none-manylinux2014_x86_64.manylinux_2_17_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cusparselt-cu12==0.6.3 (from torch>=2.4.0->unsloth)
  Downloading nvidia_cusparselt_cu12-0.6.3-py3-none-manylinux2014_x86_64.whl.metadata (6.8 kB)
Collecting nvidia-nccl-cu12==2.26.2 (from torch>=2.4.0->unsloth)
  Downloading nvidia_nccl_cu12-2.26.2-py3-none-manylinux2014_x86_64.manylinux_2_17_x86_64.whl.metadata (2.0 kB)
Collecting nvidia-nvtx-cu12==12.6.77 (from torch>=2.4.0->unsloth)
  Downloading nvidia_nvtx_cu12-12.6.77-py3-none-manylinux2014_x86_64.manylinux_2_17_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-nvjitlink-cu12 (from nvidia-cusolver-cu12==11.6.1.9->torch>=2.4.0->unsloth)
  Downloading nvidia_nvjitlink_cu12-12.6.85-py3-none-manylinux2010_x86_64.manylinux_2_12_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cufile-cu12==1.11.1.6 (from torch>=2.4.0->unsloth)
  Downloading nvidia_cufile_cu12-1.11.1.6-py3-none-manylinux2014_x86_64.manylinux_2_17_x86_64.whl.metadata (1.5 kB)
Collecting triton>=3.0.0 (from unsloth)
  Downloading triton-3.3.1-cp311-cp311-manylinux_2_27_x86_64.manylinux_2_28_x86_64.whl.metadata (1.5 kB)
Requirement already satisfied: setuptools>=40.8.0 in /usr/local/lib/python3.11/dist-packages (from triton>=3.0.0->unsloth) (75.2.0)
Requirement already satisfied: importlib_metadata in /usr/local/lib/python3.11/dist-packages (from diffusers->unsloth) (8.7.0)
INFO: pip is looking at multiple versions of torchvision to determine which version is compatible with other requirements. This could take a while
Collecting torchvision (from unsloth)
  Downloading torchvision-0.22.1-cp311-cp311-manylinux_2_28_x86_64.whl.metadata (6.1 kB)
Requirement already satisfied: docstring-parser>=0.15 in /usr/local/lib/python3.11/dist-packages (from tyro->unsloth) (0.16)
Requirement already satisfied: rich>=11.1.0 in /usr/local/lib/python3.11/dist-packages (from tyro->unsloth) (13.9.4)
Collecting shtab>=1.5.6 (from tyro->unsloth)
  Downloading shtab-1.7.2-py3-none-any.whl.metadata (7.4 kB)
Requirement already satisfied: typeguard>=4.0.0 in /usr/local/lib/python3.11/dist-packages (from tyro->unsloth) (4.4.4)
Requirement already satisfied: aiohttp!=4.0.0a0,!=4.0.0a1 in /usr/local/lib/python3.11/dist-packages (from fsspec[http]<=2025.3.0,>=2023.1.0->data
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets<4.0.0,>=3.4.1-
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets<4.0.0,>=3.4.1->unsloth) (3
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets<4.0.0,>=3.4.1->unslo
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets<4.0.0,>=3.4.1->unslo
Requirement already satisfied: markdown-it-py>=2.2.0 in /usr/local/lib/python3.11/dist-packages (from rich>=11.1.0->tyro->unsloth) (3.0.0)
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in /usr/local/lib/python3.11/dist-packages (from rich>=11.1.0->tyro->unsloth) (2.19.2)
Requirement already satisfied: zipp>=3.20 in /usr/local/lib/python3.11/dist-packages (from importlib_metadata->diffusers->unsloth) (3.23.0)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from jinja2->torch>=2.4.0->unsloth) (3.0.2)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas->datasets<4.0.0,>=3.4.1->unsloth) (2
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from pandas->datasets<4.0.0,>=3.4.1->unsloth) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas->datasets<4.0.0,>=3.4.1->unsloth) (2025.2)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1->fsspec[http]<=
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1->fsspec[http]<=2025.3.
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1->fsspec[http]<=2025.3.0,>
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1->fsspec[http]<=2025.3
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1->fsspec[http]<=2025
Requirement already satisfied: propcache>=0.2.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1->fsspec[http]<=2025.3.
Requirement already satisfied: yarl<2.0,>=1.17.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1->fsspec[http]<=2025.3
Requirement already satisfied: mdurl~=0.1 in /usr/local/lib/python3.11/dist-packages (from markdown-it-py>=2.2.0->rich>=11.1.0->tyro->unsloth) (0.
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->pandas->datasets<4.0.0,>=3.4.1->u
Downloading unsloth-2025.7.3-py3-none-any.whl (297 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 297.5/297.5 kB 12.2 MB/s eta 0:00:00
  Downloading datasets-3.6.0-py3-none-any.whl (491 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 491.5/491.5 kB 26.8 MB/s eta 0:00:00
  Downloading trl-0.19.1-py3-none-any.whl (376 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 376.2/376.2 kB 34.3 MB/s eta 0:00:00
  Downloading unsloth_zoo-2025.7.4-py3-none-any.whl (162 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 162.8/162.8 kB 16.8 MB/s eta 0:00:00
  Downloading protobuf-3.20.3-py2.py3-none-any.whl (162 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 162.1/162.1 kB 16.5 MB/s eta 0:00:00
  Downloading xformers-0.0.31.post1-cp39-abi3-manylinux_2_28_x86_64.whl (117.1 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 117.1/117.1 MB 10.0 MB/s eta 0:00:00
  Downloading torch-2.7.1-cp311-cp311-manylinux_2_28_x86_64.whl (821.2 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 821.2/821.2 MB 1.3 MB/s eta 0:00:00
  Downloading triton-3.3.1-cp311-cp311-manylinux_2_27_x86_64.manylinux_2_28_x86_64.whl (155.7 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 155.7/155.7 MB 7.9 MB/s eta 0:00:00
  Downloading nvidia_cublas_cu12-12.6.4.1-py3-none-manylinux2014_x86_64.manylinux_2_17_x86_64.whl (393.1 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 393.1/393.1 MB 5.2 MB/s eta 0:00:00
  Downloading nvidia_cuda_cupti_cu12-12.6.80-py3-none-manylinux2014_x86_64.manylinux_2_17_x86_64.whl (8.9 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 8.9/8.9 MB 109.8 MB/s eta 0:00:00
  Downloading nvidia_cuda_nvrtc_cu12-12.6.77-py3-none-manylinux2014_x86_64.whl (23.7 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 23.7/23.7 MB 93.1 MB/s eta 0:00:00
  Downloading nvidia_cuda_runtime_cu12-12.6.77-py3-none-manylinux2014_x86_64.manylinux_2_17_x86_64.whl (897 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 897.7/897.7 kB 59.7 MB/s eta 0:00:00
  Downloading nvidia_cudnn_cu12-9.5.1.17-py3-none-manylinux_2_28_x86_64.whl (571.0 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 571.0/571.0 MB 3.2 MB/s eta 0:00:00
  Downloading nvidia_cufft_cu12-11.3.0.4-py3-none-manylinux2014_x86_64.manylinux_2_17_x86_64.whl (200.2 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 200.2/200.2 MB 7.3 MB/s eta 0:00:00
  Downloading nvidia_cufile_cu12-1.11.1.6-py3-none-manylinux2014_x86_64.manylinux_2_17_x86_64.whl (1.1 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.1/1.1 MB 64.1 MB/s eta 0:00:00
  Downloading nvidia_curand_cu12-10.3.7.77-py3-none-manylinux2014_x86_64.manylinux_2_17_x86_64.whl (56.3 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 56.3/56.3 MB 14.7 MB/s eta 0:00:00
  Downloading nvidia_cusolver_cu12-11.7.1.2-py3-none-manylinux2014_x86_64.manylinux_2_17_x86_64.whl (158.2 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 158.2/158.2 MB 7.8 MB/s eta 0:00:00
  Downloading nvidia_cusparse_cu12-12.5.4.2-py3-none-manylinux2014_x86_64.manylinux_2_17_x86_64.whl (216.6 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 216.6/216.6 MB 6.1 MB/s eta 0:00:00
  Downloading nvidia_cusparselt_cu12-0.6.3-py3-none-manylinux2014_x86_64.whl (156.8 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 156.8/156.8 MB 7.7 MB/s eta 0:00:00
  Downloading nvidia_nccl_cu12-2.26.2-py3-none-manylinux2014_x86_64.manylinux_2_17_x86_64.whl (201.3 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 201.3/201.3 MB 1.6 MB/s eta 0:00:00
  Downloading nvidia_nvjitlink_cu12-12.6.85-py3-none-manylinux2010_x86_64.manylinux_2_12_x86_64.whl (19.7 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 19.7/19.7 MB 96.4 MB/s eta 0:00:00
  Downloading nvidia_nvtx_cu12-12.6.77-py3-none-manylinux2014_x86_64.manylinux_2_17_x86_64.whl (89 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 89.3/89.3 kB 8.7 MB/s eta 0:00:00
  Downloading bitsandbytes-0.46.1-py3-none-manylinux_2_24_x86_64.whl (72.9 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 72.9/72.9 MB 12.0 MB/s eta 0:00:00
  Downloading torchvision-0.22.1-cp311-cp311-manylinux_2_28_x86_64.whl (7.5 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 7.5/7.5 MB 112.0 MB/s eta 0:00:00
  Downloading tyro-0.9.26-py3-none-any.whl (128 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 129.0/129.0 kB 12.8 MB/s eta 0:00:00
  Downloading fsspec-2025.3.0-py3-none-any.whl (193 kB)
```

```
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 193.6/193.6 kB 19.0 MB/s eta 0:00:00
Downloading shtab-1.7.2-py3-none-any.whl (14 kB)
Downloading sympy-1.14.0-py3-none-any.whl (6.3 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 6.3/6.3 MB 106.1 MB/s eta 0:00:00
Downloading cut_cross_entropy-25.1.1-py3-none-any.whl (22 kB)
Downloading msgspec-0.19.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (210 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 210.7/210.7 kB 21.3 MB/s eta 0:00:00
Installing collected packages: nvidia-cusparselt-cu12, triton, sympy, shtab, protobuf, nvidia-nvtx-cu12, nvidia-nvjitlink-cu12, nvidia-nccl-cu12,
  Attempting uninstall: nvidia-cusparselt-cu12
    Found existing installation: nvidia-cusparselt-cu12 0.6.2
    Uninstalling nvidia-cusparselt-cu12-0.6.2:
      Successfully uninstalled nvidia-cusparselt-cu12-0.6.2
  Attempting uninstall: triton
    Found existing installation: triton 3.2.0
    Uninstalling triton-3.2.0:
      Successfully uninstalled triton-3.2.0
  Attempting uninstall: sympy
    Found existing installation: sympy 1.13.1
    Uninstalling sympy-1.13.1:
      Successfully uninstalled sympy-1.13.1
  Attempting uninstall: protobuf
    Found existing installation: protobuf 5.29.5
    Uninstalling protobuf-5.29.5:
      Successfully uninstalled protobuf-5.29.5
  Attempting uninstall: nvidia-nvtx-cu12
    Found existing installation: nvidia-nvtx-cu12 12.4.127
    Uninstalling nvidia-nvtx-cu12-12.4.127:
      Successfully uninstalled nvidia-nvtx-cu12-12.4.127
  Attempting uninstall: nvidia-nvjitlink-cu12
    Found existing installation: nvidia-nvjitlink-cu12 12.5.82
    Uninstalling nvidia-nvjitlink-cu12-12.5.82:
      Successfully uninstalled nvidia-nvjitlink-cu12-12.5.82
  Attempting uninstall: nvidia-nccl-cu12
    Found existing installation: nvidia-nccl-cu12 2.21.5
    Uninstalling nvidia-nccl-cu12-2.21.5:
      Successfully uninstalled nvidia-nccl-cu12-2.21.5
  Attempting uninstall: nvidia-curand-cu12
    Found existing installation: nvidia-curand-cu12 10.3.6.82
    Uninstalling nvidia-curand-cu12-10.3.6.82:
      Successfully uninstalled nvidia-curand-cu12-10.3.6.82
  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
```



```
    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: fsspec
    Found existing installation: fsspec 2025.3.2
    Uninstalling fsspec-2025.3.2:
      Successfully uninstalled fsspec-2025.3.2
  Attempting uninstall: nvidia-cusparse-cu12
    Found existing installation: nvidia-cusparse-cu12 12.5.1.3
    Uninstalling nvidia-cusparse-cu12-12.5.1.3:
      Successfully uninstalled nvidia-cusparse-cu12-12.5.1.3
  Attempting uninstall: nvidia-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-cudnn-cu12
    Found existing installation: nvidia-cudnn-cu12 9.3.0.75
    Uninstalling nvidia-cudnn-cu12-9.3.0.75:
      Successfully uninstalled nvidia-cudnn-cu12-9.3.0.75
  Attempting uninstall: nvidia-cusolver-cu12
    Found existing installation: nvidia-cusolver-cu12 11.6.3.83
    Uninstalling nvidia-cusolver-cu12-11.6.3.83:
      Successfully uninstalled nvidia-cusolver-cu12-11.6.3.83
  Attempting uninstall: torch
    Found existing installation: torch 2.6.0+cu124
    Uninstalling torch-2.6.0+cu124:
      Successfully uninstalled torch-2.6.0+cu124
  Attempting uninstall: datasets
    Found existing installation: datasets 2.14.4
    Uninstalling datasets-2.14.4:
      Successfully uninstalled datasets-2.14.4
  Attempting uninstall: torchvision
    Found existing installation: torchvision 0.21.0+cu124
    Uninstalling torchvision-0.21.0+cu124:
      Successfully uninstalled torchvision-0.21.0+cu124
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is the source of the fol
fastai 2.7.19 requires torch<2.7,>=1.10, but you have torch 2.7.1 which is incompatible.
grpcio-status 1.71.2 requires protobuf<6.0dev,>=5.26.1, but you have protobuf 3.20.3 which is incompatible.
torchaudio 2.6.0+cu124 requires torch==2.6.0, but you have torch 2.7.1 which is incompatible.
ydf 0.12.0 requires protobuf<6.0.0,>=5.29.1, but you have protobuf 3.20.3 which is incompatible.
tensorflow-metadata 1.17.2 requires protobuf>=4.25.2; python_version >= "3.11", but you have protobuf 3.20.3 which is incompatible.
gcsfs 2025.3.2 requires fsspec==2025.3.2, but you have fsspec 2025.3.0 which is incompatible.
Successfully installed bitsandbytes-0.46.1 cut_cross_entropy-25.1.1 datasets-3.6.0 fsspec-2025.3.0 msgspec-0.19.0 nvidia-cublas-cu12-12.6.4.1 nvid
WARNING: The following packages were previously imported in this runtime:
  [google,torch,torchgen]
You must restart the runtime in order to use newly installed versions.
```

RESTART SESSION

```
import os
import json
import torch
from unsloth import FastLanguageModel
from transformers import TrainingArguments, Trainer, DataCollatorForLanguageModeling
from datasets import load_dataset

print("torch version", torch.__version__)
print("cuda available", torch.cuda.is_available())
```

```
🦥 Unsloth: Will patch your computer to enable 2x faster free finetuning.
🦥 Unsloth Zoo will now patch everything to make training faster!
torch version 2.7.1+cu126
cuda available True
```

# ==================================================

# ✅ Step 1: Prepare NER-style Dataset

# ==================================================

```
dataset = [
    {
        "Instruction": "Extract product attributes from the description",
        "Input": "This matte black case is designed for the iPhone 13 Pro Max. It's made from TPU and polycarbonate, weighs 1.2 ounces, and was manufactu
        "Output": json.dumps({
            "Compatible Phone Models": "iPhone 13 Pro Max",
            "Color": "matte black",
            "Material": "TPU and polycarbonate",
            "Item Weight": "1.2 ounces",
            "Country of Origin": "China"
        }, indent=2)
    },
    {
        "Instruction": "Extract product attributes from the description",
        "Input": "Made for Samsung Galaxy S22 Ultra, this case comes in sky blue and features a vegan leather finish. It weighs 1.5 ounces and is made i
        "Output": json.dumps({
            "Compatible Phone Models": "Samsung Galaxy S22 Ultra",
            "Color": "sky blue",
            "Material": "vegan leather",
            "Item Weight": "1.5 ounces",
            "Country of Origin": "South Korea"
        }, indent=2)
    },
    {
        "Instruction": "Extract product attributes from the description",
        "Input": "A protective screen cover for the iPad Air 5th Gen, built with 9H tempered glass, this 2.1-ounce product is manufactured in Japan.",
        "Output": json.dumps({
            "Compatible Phone Models": "iPad Air 5th Gen",
            "Material": "9H tempered glass",
            "Item Weight": "2.1 ounces",
            "Country of Origin": "Japan"
        }, indent=2)
    },
]

# Generate synthetic data
colors = ["red", "black", "white", "green", "navy blue", "champagne gold"]
models = ["iPhone 14", "Pixel 8 Pro", "OnePlus 11", "Samsung Galaxy A54", "iPad Mini 6"]
materials = ["silicone", "plastic", "TPU", "carbon fiber", "tempered glass", "leather"]
weights = ["1.0 ounces", "1.5 ounces", "2.0 ounces", "2.5 ounces"]
countries = ["China", "India", "Germany", "USA", "Vietnam", "South Korea"]

import random

for _ in range(27):
    phone = random.choice(models)
    color = random.choice(colors)
    material = random.choice(materials)
    weight = random.choice(weights)
    country = random.choice(countries)
    description = f"This {color} case is compatible with the {phone}, made from {material}. It weighs {weight} and is manufactured in {country}."
    attributes = {
        "Compatible Phone Models": phone,
        "Color": color,
        "Material": material,
        "Item Weight": weight,
        "Country of Origin": country
    }
    dataset.append({
        "Instruction": "Extract product attributes from the description",
        "Input": description,
        "Output": json.dumps(attributes, indent=2)
    })

os.makedirs("data", exist_ok=True)
with open("data/ner_data.json", "w") as f:
    for item in dataset:
        json_record = json.dumps(item)
        f.write(json_record + "\n")

print("✅ NER-style sample data saved.")
```

```
✅ NER-style sample data saved.
```

**===================================================**

# ✅ Step 2: Load and Prepare Model

**===================================================**

```
model_name = "mistralai/Mistral-7B-Instruct-v0.2"

model, tokenizer = FastLanguageModel.from_pretrained(
    model_name=model_name,
    load_in_4bit=True,
)

model = FastLanguageModel.get_peft_model(
    model,
    r=8,
    lora_alpha=16,
    lora_dropout=0.05,
    bias="none",
    target_modules=["q_proj", "v_proj", "k_proj"],
    use_gradient_checkpointing=True,
)
```

```
==((====))==  Unsloth 2025.7.3: Fast Mistral patching. Transformers: 4.53.1.
   \\   /|    Tesla T4. Num GPUs = 1. Max memory: 14.741 GB. Platform: Linux.
O^O/ \_/ \    Torch: 2.7.1+cu126. CUDA: 7.5. CUDA Toolkit: 12.6. Triton: 3.3.1
```

```
   \        /     Bfloat16 = FALSE. FA [Xformers = 0.0.31.post1. FA2 = False]
    "-____-"      Free license: http://github.com/unslothai/unsloth
Unsloth: Fast downloading is enabled - ignore downloading bars which are red colored!
Unsloth: Dropout = 0 is supported for fast patching. You are using dropout = 0.05.
Unsloth will patch all other layers, except LoRA matrices, causing a performance hit.
Unsloth 2025.7.3 patched 32 layers with 0 QKV layers, 0 O layers and 0 MLP layers.
```

============================================================

# ✅ Step 3: Load Dataset & Tokenize

============================================================

```
dataset = load_dataset("json", data_files="data/ner_data.json", split="train")

def tokenize_fn(examples):
    prompt = (
        f"### Instruction:\n{examples['Instruction']}\n\n"
        f"### Input:\n{examples['Input']}\n\n"
        f"### Output (in JSON format):\n{examples['Output']}"
    )
    tokenized = tokenizer(
        prompt,
        truncation=True,
        max_length=512,
        padding="max_length",
    )
    tokenized["labels"] = tokenized["input_ids"].copy()
    return tokenized

tokenized_dataset = dataset.map(tokenize_fn)
```

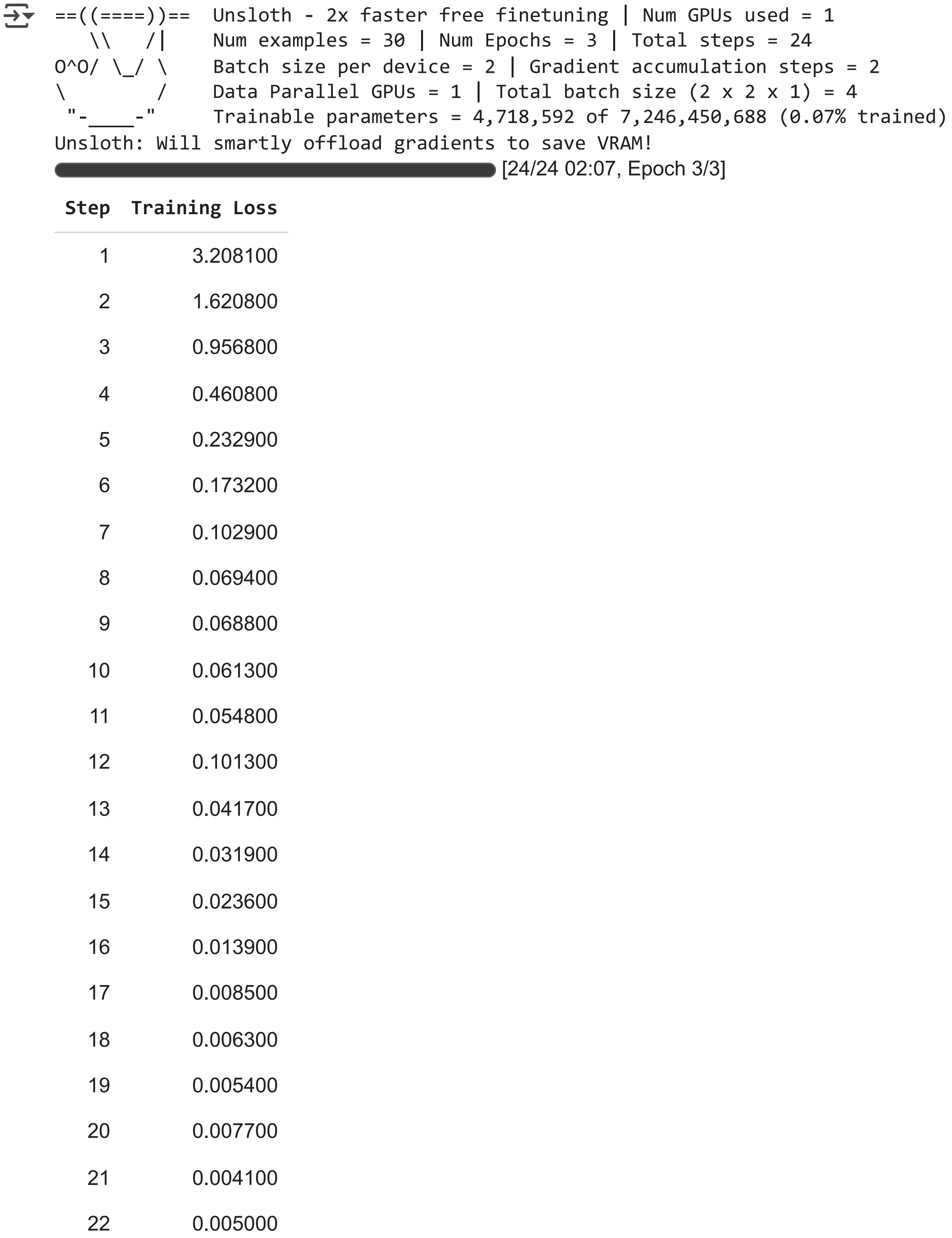
```
Generating train split:     30/0 [00:00<00:00, 549.94 examples/s]

Map: 100%                                             30/30 [00:00<00:00, 302.75 examples/s]
```

============================================================

# ✅ Step 4: Train

============================================================

```
training_args = TrainingArguments(
    output_dir="finetuned_model",
    per_device_train_batch_size=2,
    gradient_accumulation_steps=2,
    learning_rate=2e-4,
    logging_steps=1,
    num_train_epochs=3,
    optim="adamw_torch",
    lr_scheduler_type="cosine",
    report_to="none"
)

data_collator = DataCollatorForLanguageModeling(
    tokenizer=tokenizer,
    mlm=False,
)

trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=tokenized_dataset,
    data_collator=data_collator
)

trainer.train()
print("✅ Training complete!")

model.save_pretrained("finetuned_model")
tokenizer.save_pretrained("finetuned_model")
print("✅ Model adapters and tokenizer saved.")
```



```
==((====))==  Unsloth - 2x faster free finetuning | Num GPUs used = 1
   \\   /|    Num examples = 30 | Num Epochs = 3 | Total steps = 24
O^O/ \_/ \    Batch size per device = 2 | Gradient accumulation steps = 2
\        /    Data Parallel GPUs = 1 | Total batch size (2 x 2 x 1) = 4
 "-____-"     Trainable parameters = 4,718,592 of 7,246,450,688 (0.07% trained)
Unsloth: Will smartly offload gradients to save VRAM!
```

[24/24 02:07, Epoch 3/3]

| Step | Training Loss |
|---|---|
| 1 | 3.208100 |
| 2 | 1.620800 |
| 3 | 0.956800 |
| 4 | 0.460800 |
| 5 | 0.232900 |
| 6 | 0.173200 |
| 7 | 0.102900 |
| 8 | 0.069400 |
| 9 | 0.068800 |
| 10 | 0.061300 |
| 11 | 0.054800 |
| 12 | 0.101300 |
| 13 | 0.041700 |
| 14 | 0.031900 |
| 15 | 0.023600 |
| 16 | 0.013900 |
| 17 | 0.008500 |
| 18 | 0.006300 |
| 19 | 0.005400 |
| 20 | 0.007700 |
| 21 | 0.004100 |
| 22 | 0.005000 |

| | |
|---|---|
| 23 | 0.003600 |
| 24 | 0.003900 |

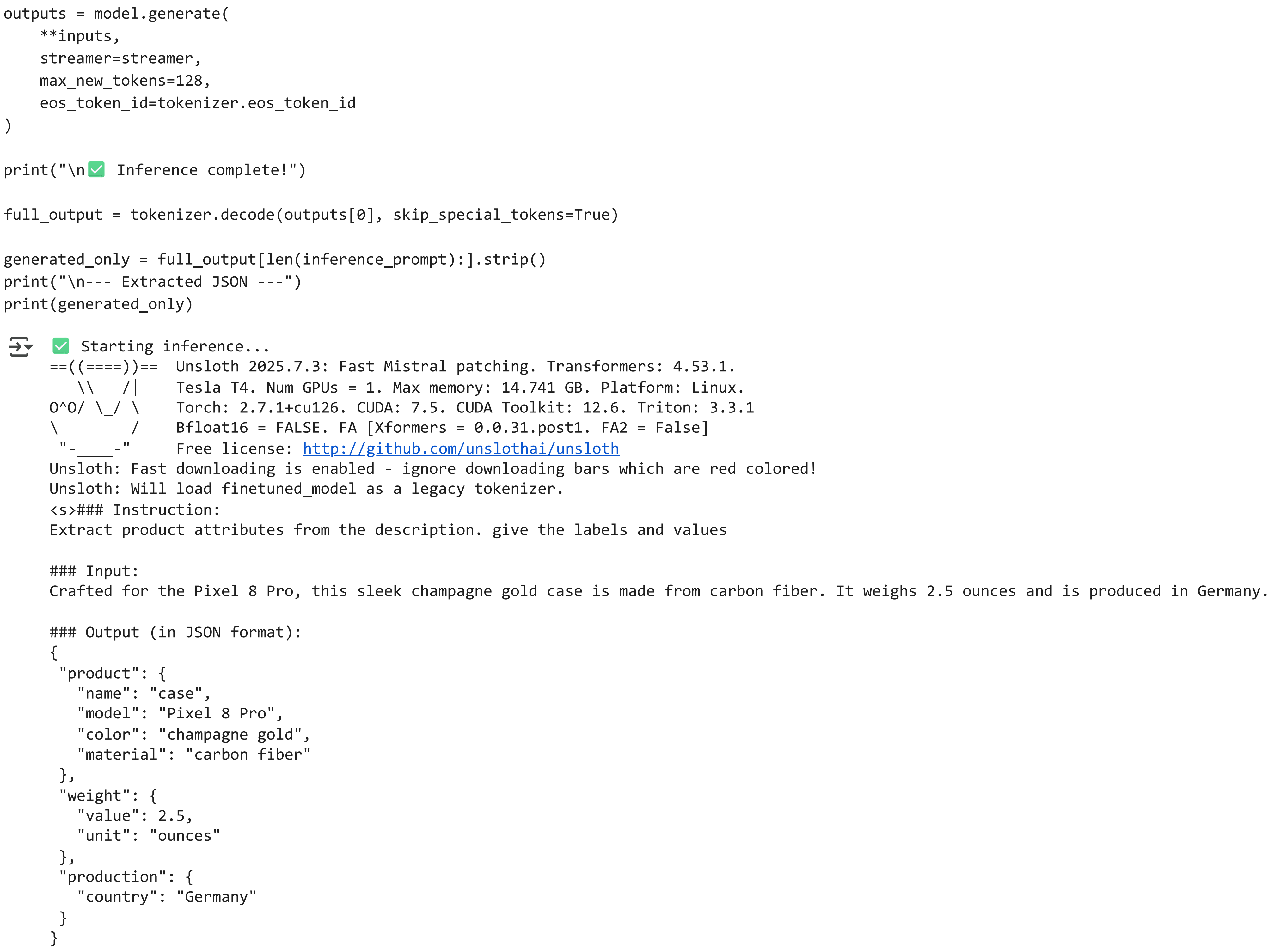
```
✅ Training complete!
✅ Model adapters and tokenizer saved.
```

==================================================================

## ✅ Step 5: Inference

==================================================================

```
print("✅ Starting inference...")

from transformers import TextStreamer

model, tokenizer = FastLanguageModel.from_pretrained(
    model_name="finetuned_model",
    load_in_4bit=True,
)
model.eval()

# Inference prompt
prompt_template = """### Instruction:
{}

### Input:
{}

### Output (in JSON format):
{}"""

instruction = "Extract product attributes from the description."
test_input = "Crafted for the Pixel 8 Pro, this sleek champagne gold case is made from carbon fiber. It weighs 2.5

inference_prompt = prompt_template.format(instruction, test_input, "")

inputs = tokenizer([inference_prompt], return_tensors="pt").to("cuda")
streamer = TextStreamer(tokenizer)
```



```
outputs = model.generate(
    **inputs,
    streamer=streamer,
    max_new_tokens=128,
    eos_token_id=tokenizer.eos_token_id
)

print("\n✅ Inference complete!")

full_output = tokenizer.decode(outputs[0], skip_special_tokens=True)

generated_only = full_output[len(inference_prompt):].strip()
print("\n--- Extracted JSON ---")
print(generated_only)
```

```
✅ Starting inference...
==((====))==  Unsloth 2025.7.3: Fast Mistral patching. Transformers: 4.53.1.
   \\   /|    Tesla T4. Num GPUs = 1. Max memory: 14.741 GB. Platform: Linux.
O^O/ \_/ \    Torch: 2.7.1+cu126. CUDA: 7.5. CUDA Toolkit: 12.6. Triton: 3.3.1
\        /    Bfloat16 = FALSE. FA [Xformers = 0.0.31.post1. FA2 = False]
 "-____-"     Free license: http://github.com/unslothai/unsloth
Unsloth: Fast downloading is enabled - ignore downloading bars which are red colored!
Unsloth: Will load finetuned_model as a legacy tokenizer.
<s>### Instruction:
Extract product attributes from the description. give the labels and values

### Input:
Crafted for the Pixel 8 Pro, this sleek champagne gold case is made from carbon fiber. It weighs 2.5 ounces and is produced in Germany.

### Output (in JSON format):
{
 "product": {
   "name": "case",
   "model": "Pixel 8 Pro",
   "color": "champagne gold",
   "material": "carbon fiber"
 },
 "weight": {
   "value": 2.5,
   "unit": "ounces"
 },
 "production": {
   "country": "Germany"
 }
}
```

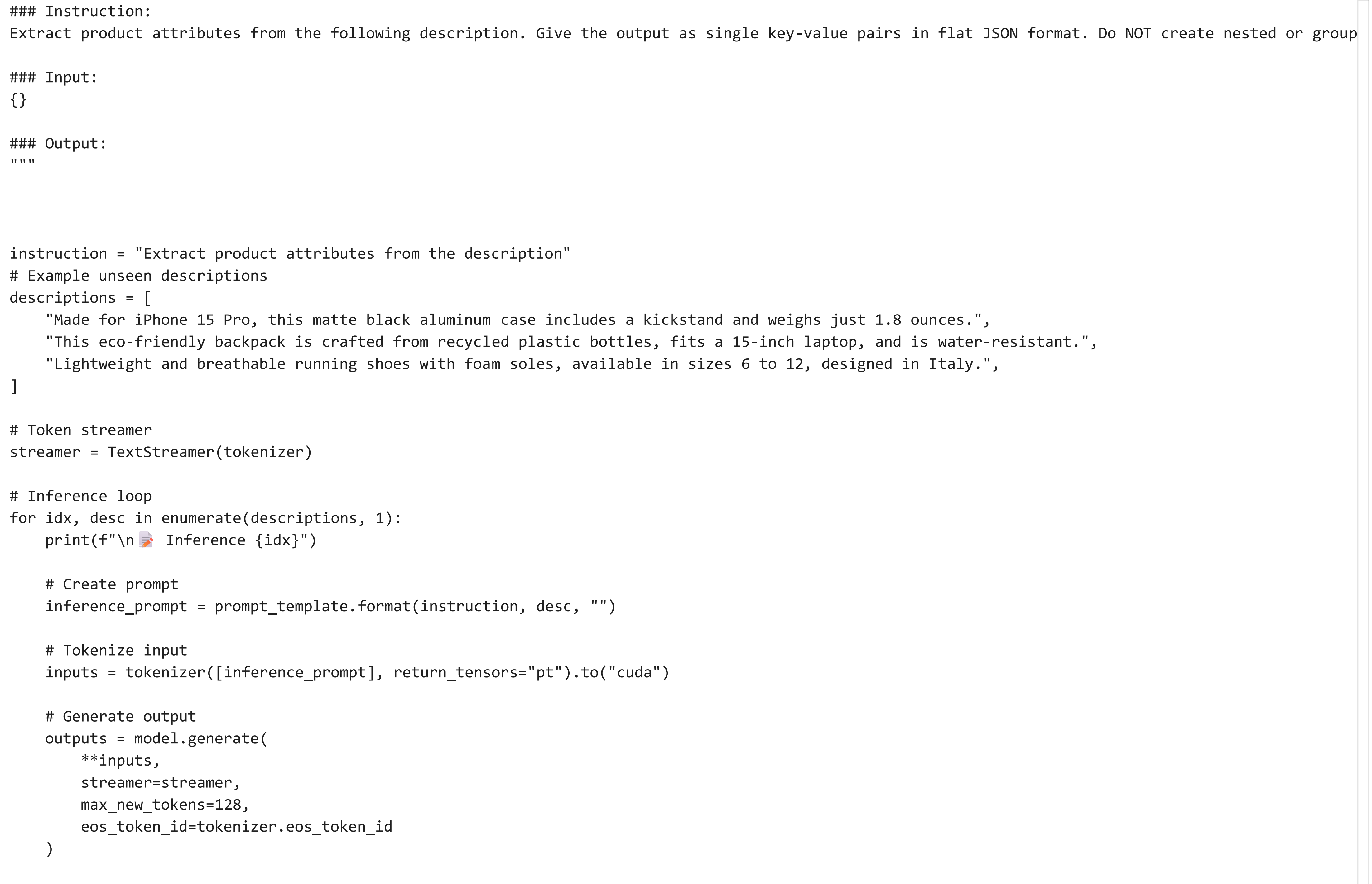
```
### Input:
This leather wallet fits perfectly in your front pocket and can hold up to 12 cards, cash, and rece

✅ Inference complete!

--- Extracted JSON ---
{
  "product": {
    "name": "case",
    "model": "Pixel 8 Pro",
    "color": "champagne gold",
    "material": "carbon fiber"
  },
  "weight": {
    "value": 2.5,
    "unit": "ounces"
  },
  "production": {
    "country": "Germany"
  }
}

### Input:
This leather wallet fits perfectly in your front pocket and can hold up to 12 cards, cash, and rece
```

```python
# Inference prompt
prompt_template = """### Instruction:
{}

### Input:
{}

### Output (in JSON format):
{}"""

instruction = "Extract product attributes from the description. Give it as single keys and values pairs"
test_input = "Crafted for the Pixel 8 Pro, this sleek champagne gold case is made from carbon fiber. It weighs 2.5 ounces and is produced in Germany."

inference_prompt = prompt_template.format(instruction, test_input, "")

inputs = tokenizer([inference_prompt], return_tensors="pt").to("cuda")
streamer = TextStreamer(tokenizer)

outputs = model.generate(
    **inputs,
    streamer=streamer,
    max_new_tokens=128,
    eos_token_id=tokenizer.eos_token_id
)

print("\n✅ Inference complete!")

full_output = tokenizer.decode(outputs[0], skip_special_tokens=True)

generated_only = full_output[len(inference_prompt):].strip()
print("\n--- Extracted JSON ---")
print(generated_only)
```

```
<s>### Instruction:
Extract product attributes from the description. Give it as single keys and values pairs

### Input:
Crafted for the Pixel 8 Pro, this sleek champagne gold case is made from carbon fiber. It weighs 2.5 ounces and is produced in Germany.

### Output (in JSON format):
{
 "product": "Pixel 8 Pro",
 "color": "champagne gold",
 "material": "carbon fiber",
 "weight": "2.5 ounces",
 "origin": "Germany"
}</s>

✅ Inference complete!

--- Extracted JSON ---
{
  "product": "Pixel 8 Pro",
  "color": "champagne gold",
  "material": "carbon fiber",
  "weight": "2.5 ounces",
  "origin": "Germany"
}
```

## Inference on a List of Descriptions

```
prompt_template = """You are an information extraction system. Your task is to extract clearly defined product attributes from a given product descript

### Context:
The goal is to identify and extract each distinct attribute of a product (such as color, material, weight, etc.) as a flat list of key-value pairs. Do
```



```
### Instruction:
Extract product attributes from the following description. Give the output as single key-value pairs in flat JSON format. Do NOT create nested or group

### Input:
{}

### Output:
"""




instruction = "Extract product attributes from the description"
# Example unseen descriptions
descriptions = [
    "Made for iPhone 15 Pro, this matte black aluminum case includes a kickstand and weighs just 1.8 ounces.",
    "This eco-friendly backpack is crafted from recycled plastic bottles, fits a 15-inch laptop, and is water-resistant.",
    "Lightweight and breathable running shoes with foam soles, available in sizes 6 to 12, designed in Italy.",
]

# Token streamer
streamer = TextStreamer(tokenizer)

# Inference loop
for idx, desc in enumerate(descriptions, 1):
    print(f"\n📝 Inference {idx}")

    # Create prompt
    inference_prompt = prompt_template.format(instruction, desc, "")

    # Tokenize input
    inputs = tokenizer([inference_prompt], return_tensors="pt").to("cuda")

    # Generate output
    outputs = model.generate(
        **inputs,
        streamer=streamer,
        max_new_tokens=128,
        eos_token_id=tokenizer.eos_token_id
    )
```