Breakdown of Each Library

1️⃣ accelerate – Hugging Face ka library jo multi-GPU, TPU & distributed training ko optimize karta hai.

Agar FSDP, DeepSpeed use kar rahe ho toh must-have hai.

2️⃣ peft (Parameter Efficient Fine-Tuning) –

LoRA, QLoRA, Adapters jaise low-memory tuning methods ke liye use hota hai. Full fine-tuning ki jagah lightweight & efficient tuning karne me madad karta hai.

3️⃣ bitsandbytes –

8-bit aur 4-bit quantization support karta hai. QLoRA fine-tuning me VRAM kaafi save hota hai.

4️⃣ git+https://github.com/huggingface/transformers –

Hugging Face ke transformers ka latest GitHub version install karta hai. Ye zaroori hai agar koi naye models ya features chahiye ho jo PyPI version me nahi mile.

5️⃣ trl (Transformer Reinforcement Learning) –

RLHF (Reinforcement Learning from Human Feedback) ke liye. Agar ChatGPT-like models banana hai toh `trl`` ka use hota hai.

6️⃣ py7zr –

7z format wali compressed files ko unzip karne ke liye. Agar Hugging Face ya kisi aur se compressed dataset mila toh ye useful hoga.

7️⃣ auto-gptq –

GPTQ-based quantization ke liye. Faster inference aur VRAM efficiency improve karta hai.

8️⃣ optimum –

Hugging Face ka library jo ONNX, TensorRT, Habana Gaudi, NeuronX jaise hardware optimizations provide karta hai.

Accelerated inference aur optimized training ke liye best hai.

 Summary

Agar low-VRAM GPUs (24GB ya less) par fine-tuning kar rahe ho toh bitsandbytes + peft + QLoRA combo best hai.

Agar multi-GPU/TPU cluster pe train kar rahe ho toh accelerate + optimum zaroori hai.

Agar RLHF (like ChatGPT) fine-tune karna hai toh TRL package kaam aayega.

```
!pip install accelerate peft bitsandbytes
git+https://github.com/huggingface/transformers trl py7zr auto-gptq
optimum
```

```
Collecting git+https://github.com/huggingface/transformers
  Cloning https://github.com/huggingface/transformers to /tmp/pip-req-
build-g7yzgny2
  Running command git clone --filter=blob:none --quiet
https://github.com/huggingface/transformers /tmp/pip-req-build-
g7yzgny2
  Resolved https://github.com/huggingface/transformers to commit
92c5ca9dd70de3ade2af2eb835c96215cc50e815
  Installing build dependencies ... ents to build wheel ... etadata
(pyproject.toml) ... ent already satisfied: accelerate in
/usr/local/lib/python3.11/dist-packages (1.3.0)
Requirement already satisfied: peft in /usr/local/lib/python3.11/dist-
packages (0.14.0)
Collecting bitsandbytes
  Using cached bitsandbytes-0.45.2-py3-none-
manylinux_2_24_x86_64.whl.metadata (5.8 kB)
Collecting trl
  Downloading trl-0.15.1-py3-none-any.whl.metadata (11 kB)
Collecting py7zr
  Downloading py7zr-0.22.0-py3-none-any.whl.metadata (16 kB)
Collecting auto-gptq
  Downloading auto_gptq-0.7.1-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (18 kB)
Collecting optimum
  Downloading optimum-1.24.0-py3-none-any.whl.metadata (21 kB)
Requirement already satisfied: numpy<3.0.0,>=1.17 in
/usr/local/lib/python3.11/dist-packages (from accelerate) (1.26.4)
Requirement already satisfied: packaging>=20.0 in
/usr/local/lib/python3.11/dist-packages (from accelerate) (24.2)
Requirement already satisfied: psutil in
/usr/local/lib/python3.11/dist-packages (from accelerate) (5.9.5)
Requirement already satisfied: pyyaml in
/usr/local/lib/python3.11/dist-packages (from accelerate) (6.0.2)
Requirement already satisfied: torch>=2.0.0 in
/usr/local/lib/python3.11/dist-packages (from accelerate)
(2.5.1+cu124)
Requirement already satisfied: huggingface-hub>=0.21.0 in
/usr/local/lib/python3.11/dist-packages (from accelerate) (0.28.1)
Requirement already satisfied: safetensors>=0.4.3 in
/usr/local/lib/python3.11/dist-packages (from accelerate) (0.5.2)
Requirement already satisfied: tqdm in /usr/local/lib/python3.11/dist-
packages (from peft) (4.67.1)
Requirement already satisfied: filelock in
/usr/local/lib/python3.11/dist-packages (from
transformers==4.50.0.dev0) (3.17.0)
Requirement already satisfied: regex!=2019.12.17 in
/usr/local/lib/python3.11/dist-packages (from
transformers==4.50.0.dev0) (2024.11.6)
Requirement already satisfied: requests in
```

```
/usr/local/lib/python3.11/dist-packages (from
transformers==4.50.0.dev0) (2.32.3)
Requirement already satisfied: tokenizers<0.22,>=0.21 in
/usr/local/lib/python3.11/dist-packages (from
transformers==4.50.0.dev0) (0.21.0)
Collecting datasets>=2.21.0 (from trl)
  Downloading datasets-3.3.2-py3-none-any.whl.metadata (19 kB)
Requirement already satisfied: rich in /usr/local/lib/python3.11/dist-
packages (from trl) (13.9.4)
Collecting texttable (from py7zr)
  Downloading texttable-1.7.0-py2.py3-none-any.whl.metadata (9.8 kB)
Collecting pycryptodomex>=3.16.0 (from py7zr)
  Downloading pycryptodomex-3.21.0-cp36-abi3-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (3.4 kB)
Collecting pyzstd>=0.15.9 (from py7zr)
  Downloading pyzstd-0.16.2-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (2.4 kB)
Collecting pyppmd<1.2.0,>=1.1.0 (from py7zr)
  Downloading pyppmd-1.1.1-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (5.5 kB)
Collecting pybcj<1.1.0,>=1.0.0 (from py7zr)
  Downloading pybcj-1.0.3-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (3.9 kB)
Collecting multivolumefile>=0.2.3 (from py7zr)
  Downloading multivolumefile-0.2.3-py3-none-any.whl.metadata (6.3 kB)
Collecting inflate64<1.1.0,>=1.0.0 (from py7zr)
  Downloading inflate64-1.0.1-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (3.8 kB)
Collecting brotli>=1.1.0 (from py7zr)
  Downloading Brotli-1.1.0-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (5.5 kB)
Requirement already satisfied: sentencepiece in
/usr/local/lib/python3.11/dist-packages (from auto-gptq) (0.2.0)
Collecting rouge (from auto-gptq)
  Downloading rouge-1.0.1-py3-none-any.whl.metadata (4.1 kB)
Collecting gekko (from auto-gptq)
  Downloading gekko-1.2.1-py3-none-any.whl.metadata (3.0 kB)
Requirement already satisfied: pyarrow>=15.0.0 in
/usr/local/lib/python3.11/dist-packages (from datasets>=2.21.0->trl)
(17.0.0)
Collecting dill<0.3.9,>=0.3.0 (from datasets>=2.21.0->trl)
  Downloading dill-0.3.8-py3-none-any.whl.metadata (10 kB)
Requirement already satisfied: pandas in
/usr/local/lib/python3.11/dist-packages (from datasets>=2.21.0->trl)
(2.2.2)
Collecting xxhash (from datasets>=2.21.0->trl)
  Downloading xxhash-3.5.0-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (12 kB)
Collecting multiprocess<0.70.17 (from datasets>=2.21.0->trl)
```

```
  Downloading multiprocess-0.70.16-py311-none-any.whl.metadata (7.2
kB)
Requirement already satisfied: fsspec<=2024.12.0,>=2023.1.0 in
/usr/local/lib/python3.11/dist-packages (from
fsspec[http]<=2024.12.0,>=2023.1.0->datasets>=2.21.0->trl) (2024.10.0)
Requirement already satisfied: aiohttp in
/usr/local/lib/python3.11/dist-packages (from datasets>=2.21.0->trl)
(3.11.12)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.21.0-
>accelerate) (4.12.2)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.11/dist-packages (from requests-
>transformers==4.50.0.dev0) (3.4.1)
Requirement already satisfied: idna<4,>=2.5 in
/usr/local/lib/python3.11/dist-packages (from requests-
>transformers==4.50.0.dev0) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.11/dist-packages (from requests-
>transformers==4.50.0.dev0) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.11/dist-packages (from requests-
>transformers==4.50.0.dev0) (2025.1.31)
Requirement already satisfied: networkx in
/usr/local/lib/python3.11/dist-packages (from torch>=2.0.0-
>accelerate) (3.4.2)
Requirement already satisfied: jinja2 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.0.0-
>accelerate) (3.1.5)
Collecting nvidia-cuda-nvrtc-cu12==12.4.127 (from torch>=2.0.0-
>accelerate)
  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>=2.0.0-
>accelerate)
  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>=2.0.0-
>accelerate)
  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>=2.0.0->accelerate)
  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>=2.0.0-
>accelerate)
  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>=2.0.0->accelerate)
```

```
  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>=2.0.0-
>accelerate)
  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>=2.0.0-
>accelerate)
  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>=2.0.0-
>accelerate)
  Downloading nvidia_cusparse_cu12-12.3.1.170-py3-none-
manylinux2014_x86_64.whl.metadata (1.6 kB)
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.0.0-
>accelerate) (2.21.5)
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.0.0-
>accelerate) (12.4.127)
Collecting nvidia-nvjitlink-cu12==12.4.127 (from torch>=2.0.0-
>accelerate)
  Downloading nvidia_nvjitlink_cu12-12.4.127-py3-none-
manylinux2014_x86_64.whl.metadata (1.5 kB)
Requirement already satisfied: triton==3.1.0 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.0.0-
>accelerate) (3.1.0)
Requirement already satisfied: sympy==1.13.1 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.0.0-
>accelerate) (1.13.1)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in
/usr/local/lib/python3.11/dist-packages (from sympy==1.13.1-
>torch>=2.0.0->accelerate) (1.3.0)
Requirement already satisfied: markdown-it-py>=2.2.0 in
/usr/local/lib/python3.11/dist-packages (from rich->trl) (3.0.0)
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in
/usr/local/lib/python3.11/dist-packages (from rich->trl) (2.18.0)
Requirement already satisfied: six in /usr/local/lib/python3.11/dist-
packages (from rouge->auto-gptq) (1.17.0)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp-
>datasets>=2.21.0->trl) (2.4.6)
Requirement already satisfied: aiosignal>=1.1.2 in
/usr/local/lib/python3.11/dist-packages (from aiohttp-
>datasets>=2.21.0->trl) (1.3.2)
Requirement already satisfied: attrs>=17.3.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp-
>datasets>=2.21.0->trl) (25.1.0)
Requirement already satisfied: frozenlist>=1.1.1 in
```

```
/usr/local/lib/python3.11/dist-packages (from aiohttp-
>datasets>=2.21.0->trl) (1.5.0)
Requirement already satisfied: multidict<7.0,>=4.5 in
/usr/local/lib/python3.11/dist-packages (from aiohttp-
>datasets>=2.21.0->trl) (6.1.0)
Requirement already satisfied: propcache>=0.2.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp-
>datasets>=2.21.0->trl) (0.2.1)
Requirement already satisfied: yarl<2.0,>=1.17.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp-
>datasets>=2.21.0->trl) (1.18.3)
Requirement already satisfied: mdurl~=0.1 in
/usr/local/lib/python3.11/dist-packages (from markdown-it-py>=2.2.0-
>rich->trl) (0.1.2)
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.11/dist-packages (from jinja2->torch>=2.0.0-
>accelerate) (3.0.2)
Requirement already satisfied: python-dateutil>=2.8.2 in
/usr/local/lib/python3.11/dist-packages (from pandas-
>datasets>=2.21.0->trl) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in
/usr/local/lib/python3.11/dist-packages (from pandas-
>datasets>=2.21.0->trl) (2025.1)
Requirement already satisfied: tzdata>=2022.7 in
/usr/local/lib/python3.11/dist-packages (from pandas-
>datasets>=2.21.0->trl) (2025.1)
Downloading bitsandbytes-0.45.2-py3-none-manylinux_2_24_x86_64.whl
(69.7 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 69.7/69.7 MB 10.3 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 318.9/318.9 kB 22.3 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 67.9/67.9 kB 5.1 MB/s eta
0:00:00
anylinux_2_17_x86_64.manylinux2014_x86_64.whl (23.5 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 23.5/23.5 MB 79.5 MB/s eta
0:00:00
um-1.24.0-py3-none-any.whl (433 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 433.6/433.6 kB 28.3 MB/s eta
0:00:00
anylinux_2_17_x86_64.manylinux2014_x86_64.whl (2.9 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2.9/2.9 MB 72.6 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 485.4/485.4 kB 28.7 MB/s eta
0:00:00
anylinux_2_17_x86_64.manylinux2014_x86_64.whl (96 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 96.2/96.2 kB 6.9 MB/s eta
0:00:00
ultivolumefile-0.2.3-py3-none-any.whl (17 kB)
```

```
Downloading pybcj-1.0.3-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl (50 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 50.6/50.6 kB 3.9 MB/s eta
0:00:00
ex-3.21.0-cp36-abi3-manylinux_2_17_x86_64.manylinux2014_x86_64.whl
(2.3 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2.3/2.3 MB 78.8 MB/s eta
0:00:00
d-1.1.1-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl
(141 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 141.3/141.3 kB 10.8 MB/s eta
0:00:00
anylinux_2_17_x86_64.manylinux2014_x86_64.whl (413 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 413.7/413.7 kB 27.0 MB/s eta
0:00:00
anylinux2014_x86_64.whl (363.4 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 363.4/363.4 MB 4.4 MB/s eta
0:00:00
anylinux2014_x86_64.whl (13.8 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 13.8/13.8 MB 85.5 MB/s eta
0:00:00
anylinux2014_x86_64.whl (24.6 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 24.6/24.6 MB 74.6 MB/s eta
0:00:00
e_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl (883 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 883.7/883.7 kB 44.2 MB/s eta
0:00:00
anylinux2014_x86_64.whl (664.8 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 664.8/664.8 MB 2.8 MB/s eta
0:00:00
anylinux2014_x86_64.whl (211.5 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 211.5/211.5 MB 5.3 MB/s eta
0:00:00
anylinux2014_x86_64.whl (56.3 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 56.3/56.3 MB 11.6 MB/s eta
0:00:00
anylinux2014_x86_64.whl (127.9 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 127.9/127.9 MB 7.9 MB/s eta
0:00:00
anylinux2014_x86_64.whl (207.5 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 207.5/207.5 MB 6.3 MB/s eta
0:00:00
anylinux2014_x86_64.whl (21.1 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 21.1/21.1 MB 78.9 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 13.2/13.2 MB 92.9 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 116.3/116.3 kB 8.7 MB/s eta
0:00:00
```

```
ultiprocess-0.70.16-py311-none-any.whl (143 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 143.5/143.5 kB 10.8 MB/s eta
0:00:00
anylinux_2_17_x86_64.manylinux2014_x86_64.whl (194 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 194.8/194.8 kB 15.0 MB/s eta
0:00:00
ers
  Building wheel for transformers (pyproject.toml) ... ers:
filename=transformers-4.50.0.dev0-py3-none-any.whl size=10849200
sha256=acf0b514d1c4e80f18d7ce69d3727450c4f80a0817352a0a0ad0856f72c289f
a
  Stored in directory:
/tmp/pip-ephem-wheel-cache-7nyt0lwr/wheels/04/a3/f1/b88775f8e166582752
5b19ac7590250f1038d947067beba9fb
Successfully built transformers
Installing collected packages: texttable, brotli, xxhash, rouge,
pyzstd, pyppmd, pycryptodomex, pybcj, nvidia-nvjitlink-cu12, nvidia-
curand-cu12, nvidia-cufft-cu12, nvidia-cuda-runtime-cu12, nvidia-cuda-
nvrtc-cu12, nvidia-cuda-cupti-cu12, nvidia-cublas-cu12,
multivolumefile, inflate64, gekko, dill, py7zr, nvidia-cusparse-cu12,
nvidia-cudnn-cu12, multiprocess, nvidia-cusolver-cu12, transformers,
datasets, optimum, bitsandbytes, trl, auto-gptq
  Attempting uninstall: nvidia-nvjitlink-cu12
    Found existing installation: nvidia-nvjitlink-cu12 12.5.82
    Uninstalling nvidia-nvjitlink-cu12-12.5.82:
      Successfully uninstalled nvidia-nvjitlink-cu12-12.5.82
  Attempting uninstall: nvidia-curand-cu12
    Found existing installation: nvidia-curand-cu12 10.3.6.82
    Uninstalling nvidia-curand-cu12-10.3.6.82:
      Successfully uninstalled nvidia-curand-cu12-10.3.6.82
  Attempting uninstall: nvidia-cufft-cu12
    Found existing installation: nvidia-cufft-cu12 11.2.3.61
    Uninstalling nvidia-cufft-cu12-11.2.3.61:
      Successfully uninstalled nvidia-cufft-cu12-11.2.3.61
  Attempting uninstall: nvidia-cuda-runtime-cu12
    Found existing installation: nvidia-cuda-runtime-cu12 12.5.82
    Uninstalling nvidia-cuda-runtime-cu12-12.5.82:
      Successfully uninstalled nvidia-cuda-runtime-cu12-12.5.82
  Attempting uninstall: nvidia-cuda-nvrtc-cu12
    Found existing installation: nvidia-cuda-nvrtc-cu12 12.5.82
    Uninstalling nvidia-cuda-nvrtc-cu12-12.5.82:
      Successfully uninstalled nvidia-cuda-nvrtc-cu12-12.5.82
  Attempting uninstall: nvidia-cuda-cupti-cu12
    Found existing installation: nvidia-cuda-cupti-cu12 12.5.82
    Uninstalling nvidia-cuda-cupti-cu12-12.5.82:
      Successfully uninstalled nvidia-cuda-cupti-cu12-12.5.82
  Attempting uninstall: nvidia-cublas-cu12
    Found existing installation: nvidia-cublas-cu12 12.5.3.2
    Uninstalling nvidia-cublas-cu12-12.5.3.2:
```

```
      Successfully uninstalled nvidia-cublas-cu12-12.5.3.2
  Attempting uninstall: nvidia-cusparse-cu12
    Found existing installation: nvidia-cusparse-cu12 12.5.1.3
    Uninstalling nvidia-cusparse-cu12-12.5.1.3:
      Successfully uninstalled nvidia-cusparse-cu12-12.5.1.3
  Attempting uninstall: nvidia-cudnn-cu12
    Found existing installation: nvidia-cudnn-cu12 9.3.0.75
    Uninstalling nvidia-cudnn-cu12-9.3.0.75:
      Successfully uninstalled nvidia-cudnn-cu12-9.3.0.75
  Attempting uninstall: nvidia-cusolver-cu12
    Found existing installation: nvidia-cusolver-cu12 11.6.3.83
    Uninstalling nvidia-cusolver-cu12-11.6.3.83:
      Successfully uninstalled nvidia-cusolver-cu12-11.6.3.83
  Attempting uninstall: transformers
    Found existing installation: transformers 4.48.3
    Uninstalling transformers-4.48.3:
      Successfully uninstalled transformers-4.48.3
Successfully installed auto-gptq-0.7.1 bitsandbytes-0.45.2 brotli-
1.1.0 datasets-3.3.2 dill-0.3.8 gekko-1.2.1 inflate64-1.0.1
multiprocess-0.70.16 multivolumefile-0.2.3 nvidia-cublas-cu12-12.4.5.8
nvidia-cuda-cupti-cu12-12.4.127 nvidia-cuda-nvrtc-cu12-12.4.127
nvidia-cuda-runtime-cu12-12.4.127 nvidia-cudnn-cu12-9.1.0.70 nvidia-
cufft-cu12-11.2.1.3 nvidia-curand-cu12-10.3.5.147 nvidia-cusolver-
cu12-11.6.1.9 nvidia-cusparse-cu12-12.3.1.170 nvidia-nvjitlink-cu12-
12.4.127 optimum-1.24.0 py7zr-0.22.0 pybcj-1.0.3 pycryptodomex-3.21.0
pyppmd-1.1.1 pyzstd-0.16.2 rouge-1.0.1 texttable-1.7.0 transformers-
4.50.0.dev0 trl-0.15.1 xxhash-3.5.0
```

Breakdown 1 from huggingface_hub import notebook_login

This imports the notebook_login function, which is used for authentication inside Jupyter Notebooks or Google Colab.

2 notebook_login()

This will prompt you to enter your Hugging Face access token. You can get the token from Hugging Face website.

 Why is this important?

If you are downloading a private model or dataset, authentication is required.

If you want to upload your fine-tuned model back to Hugging Face, you need to log in first.

 Alternative for Script-Based Login

If you are running a script (not in a notebook), use:

from huggingface_hub import login

login(token="your_huggingface_token")

```
from huggingface_hub import notebook_login
notebook_login()
```

{"model_id":"fcb74cff6fb44065ac447aa7bdc5de56","version_major":2,"version_minor":0}

Breakdown of Each Import

1 import torch

PyTorch is the core deep-learning library used for training models. It helps in tensor operations, GPU acceleration, and model training.

2 from datasets import load_dataset, Dataset

load_dataset: Used to load datasets from Hugging Face Hub or local files. Dataset: Helps in creating a dataset manually from Python objects (like a list or dictionary).

3 from peft import LoraConfig, AutoPeftModelForCausalLM, prepare_model_for_kbit_training, get_peft_model

LoraConfig: Configuration for LoRA (Low-Rank Adaptation), which makes fine-tuning more memory efficient.

AutoPeftModelForCausalLM: Loads a causal language model with PEFT (Parameter-Efficient Fine-Tuning).

prepare_model_for_kbit_training: Optimizes the model for low-bit training (8-bit/4-bit with QLoRA).

get_peft_model: Converts a standard model into a LoRA-optimized model.

4 from transformers import AutoModelForCausalLM, AutoTokenizer, GPTQConfig, TrainingArguments

AutoModelForCausalLM: Loads a pre-trained causal language model (like LLaMA, Mistral).

AutoTokenizer: Tokenizer for preprocessing text input.

GPTQConfig: Configures GPTQ (Quantized GPT) for efficient inference.

TrainingArguments: Defines training settings like epochs, batch size, optimizer, learning rate, etc.

5 from trl import SFTTrainer

SFTTrainer: Trainer from the trl library used for Supervised Fine-Tuning (SFT).

6 simplifies LoRA-based fine-tuning and integrates well with Hugging Face. 6 import os

Used for handling file paths and system settings, like saving models, loading datasets, etc.

 What is this setup used for?

 Fine-tuning large language models (LLMs) efficiently using LoRA and QLoRA.

 Using Hugging Face datasets and models.

 Training a model with low-bit precision (4-bit/8-bit) for better memory efficiency.

```
import torch
from datasets import load_dataset, Dataset
from peft import LoraConfig, AutoPeftModelForCausalLM,
prepare_model_for_kbit_training, get_peft_model
from transformers import AutoModelForCausalLM, AutoTokenizer,
GPTQConfig, TrainingArguments
from trl import SFTTrainer
import os
```

Understanding the Code: Loading and Preparing the Dataset for Fine-Tuning This code is loading, processing, and converting a dataset into a format suitable for fine-tuning an LLM (like Mistral or LLaMA-2) for text summarization. Let's break it down step by step.

1 Loading the Dataset

data = load_dataset("samsum", split="train")

 load_dataset("samsum", split="train") loads the Samsum dataset, which contains dialogues and their summaries.

 split="train" ensures that we load only the training set.

 Samsum Dataset Overview

dialogue: A conversation between people.

summary: A short summary of that conversation.

 Example from the dataset:

dialogue summary

Alice: Hey, how are you? Bob: I'm good, you? Alice and Bob greet each other.

2 Converting Dataset to Pandas DataFrame

data_df = data.to_pandas()

 This converts the dataset into a Pandas DataFrame for easier processing.

3 Formatting Data for LLM Fine-Tuning

data_df["text"] = data_df[["dialogue", "summary"]].apply(lambda x: "###Human: Summarize this following dialogue: " + x["dialogue"] + "\n###Assistant: " + x["summary"], axis=1)

 Purpose: It formats the data into a ChatML-style prompt to fine-tune LLaMA or Mistral.

 How It Works:

It takes the dialogue and summary columns.

It transforms them into a prompt-response format for LLM training.

 Example Output:

###Human: Summarize this following dialogue:

Alice: Hey, how are you?

Bob: I'm good, you?

###Assistant: Alice and Bob greet each other.

 This format mimics human-AI interactions, making it suitable for instruction-tuned models like Mistral or LLaMA-2-Chat.

4 Checking the First Example

print(data_df.iloc[0])

 data_df.iloc[0] prints the first row of the dataset after formatting.

5 Converting Back to Hugging Face Dataset

data = Dataset.from_pandas(data_df)

 Why? Since fine-tuning with Transformers & PEFT requires a Hugging Face dataset, we convert it back after processing.

 Summary

 Loads the Samsum dataset (dialogue → summary).

 Formats it into a prompt-response structure for LLM fine-tuning.

 Converts it back into a Hugging Face Dataset for training.

 Next: Do you want to tokenize this dataset for Mistral/LLaMA-2 fine-tuning?

```
data = load_dataset("samsum", split="train")
```

{"model_id":"3efd2912ac0b4669bdde0228c10130f9","version_major":2,"version_minor":0}

{"model_id":"c6d70535c1f544fbb4a9886a81f4eca3","version_major":2,"version_minor":0}

```
The repository for samsum contains custom code which must be executed
to correctly load the dataset. You can inspect the repository content
at https://hf.co/datasets/samsum.
You can avoid this prompt in future by passing the argument
`trust_remote_code=True`.

Do you wish to run the custom code? [y/N] y
```

{"model_id":"ebc8674835a5411bafdb9c7cab47f4ad","version_major":2,"version_minor":0}

{"model_id":"40bc7809f6814f6382dd64571bf2c2a8","version_major":2,"version_minor":0}

{"model_id":"a28aba8b9aae47caad4628718ef2d148","version_major":2,"version_minor":0}

{"model_id":"4bec36d7cbfe49efbe150c4c997f91c2","version_major":2,"version_minor":0}

```
data_df = data.to_pandas()

data_df
```

{"summary":"{\n \"name\": \"data_df\",\n \"rows\": 14732,\n \"fields\": [\n   {\n   \"column\": \"id\",\n   \"properties\": {\n   \"dtype\": \"string\",\n   \"num_unique_values\": 14732,\n   \"samples\": [\n   \"13811908\",\n   \"13716431\",\n   \"13810214\"\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"dialogue\",\n   \"properties\": {\n   \"dtype\": \"string\",\n   \"num_unique_values\": 14265,\n   \"samples\": [\n   \"Charles: <file_other>\\r\\nCharles: It seems that the govt decided to fuck us even harder\\r\\nCharles: Every year prices go up\\r\\nCharles: Maybe raising taxes is a source of pleasure for those so called 'politicians'?\\r\\nMike: Dude...\\r\\nMike: I don't mind people interested in kinky stuff\\r\\nMike: But have never expected such political perversion in my own country\\r\\nCharles: Neither have I. It makes me think\\r\\nCharles: That wouldn't be a big deal if only the earnings would go up as well\\r\\nCharles: No wonder people resort to going abroad\\r\\nMike: Actually I've been considering working somewhere abroad lately\\r\\nMike: I've been looking for a job since May here and there's a few offers available\\r\\nMike: Mainly physical, hard work...\\r\\nCharles: This system sucks!\\r\\nCharles: You're wasting the best years of your life on education, then it turns out there's no job for you in your profession...\\r\\nMike: Yeah, it's true. But complaining rarely changes anything\\r\\nMike: It's better to take some action\\r\\nMike: Going abroad is not that bad, I think I'll give it a try. Wanna join me?\\r\\nCharles: And where would you like to go?\\r\\nMike: The Netherlands for starters, the perspectives are decent and it'd be a chance to visit Amsterdam, the city of freedom! :D\\r\\nCharles: Damn, you're pulling me in!\\r\\nCharles: And actually... besides earning some money it can serve as a good adventure\\r\\nMike: So? R U in?\\r\\nCharles: Give me a few days, I'll ask around and let you know soon\\r\\nMike: Fine, remember that I'm going anyway :p\\r\\nCharles: Sure, I feel I've already decided too! :D\\r\\nCharles: Talk to you soon\\r\\nMike: All right\",\n   \"Finn: I heard that Britney got expelled.\\r\\nTerry: why?\\r\\nOswald: what for?\\r\\

nFinn: you know her general behavior\\r\\nOswald: and she's absent a lot\\r\\nFinn: so yeah, that accumulated kinda, but now she's been accused of making that blue graffiti that popped up 3 days ago\\r\\nTerry: I was wondering if this might have anything to do with that graffiti\\r\\nFinn: apparently it does\\r\\nOswald: you think she did it?\\r\\nFinn: who knows? She certainly likes to draw\\r\\nTerry: maybe they just needed an excuse to finally kick her out\\r\\nFinn: maybe\\r\\nOswald: too bad, though, I kinda liked her\\r\\nFinn: liked her liked her? ;)\\r\\nOswald: no, ofc not, but she's pretty cool\\r\\nFinn: true that\",\n \"Darren: Look! 10am!\\r\\nFrank: Whoa! They're awesome knockers.\\r\\nDarren: Stop staring or she'll notice us! LOL\"\n],\n \"semantic_type\": \"\",\n \"description\": \"\"\n }\n },\n {\n \"column\": \"summary\",\n \"properties\": {\n \"dtype\": \"string\",\n \"num_unique_values\": 14730,\n \"samples\": [\n \"Violet sent Claire Austin's article.\",\n \"Michael, Tom and Chris tease Mark because he has a new girlfriend.\",\n \"Leo just had a very good time with his parents in Spain. Granny will be happy if he came round and told her about his holidays. He could also see flowers blooming in her and Daddy's garden.\"\n],\n \"semantic_type\": \"\",\n \"description\": \"\"\n }\n }\n]\n}","type":"dataframe","variable_name":"data_df"}

```
data_df["text"] = data_df[["dialogue", "summary"]].apply(lambda x: "###Human: Summarize this following dialogue: " + x["dialogue"] + "\n###Assistant: " +x["summary"], axis=1)
```

```
data_df
```

{"summary":"{\n \"name\": \"data_df\",\n \"rows\": 14732,\n \"fields\": [\n {\n \"column\": \"id\",\n \"properties\": {\n \"dtype\": \"string\",\n \"num_unique_values\": 14732,\n \"samples\": [\n \"13811908\",\n \"13716431\",\n \"13810214\"\n],\n \"semantic_type\": \"\",\n \"description\": \"\"\n }\n },\n {\n \"column\": \"dialogue\",\n \"properties\": {\n \"dtype\": \"string\",\n \"num_unique_values\": 14265,\n \"samples\": [\n \"Charles: <file_other>\\r\\nCharles: It seems that the govt decided to fuck us even harder\\r\\nCharles: Every year prices go up\\r\\nCharles: Maybe raising taxes is a source of pleasure for those so called 'politicians'?\\r\\nMike: Dude...\\r\\nMike: I don't mind people interested in kinky stuff\\r\\nMike: But have never expected such political perversion in my own country\\r\\nCharles: Neither have I. It makes me think\\r\\nCharles: That wouldn't be a big deal if only the earnings would go up as well\\r\\nCharles: No wonder people resort to going abroad\\r\\nMike: Actually I've been considering working somewhere abroad lately\\r\\nMike: I've been looking for a job since May here and there's a few offers available\\r\\nMike: Mainly physical, hard work...\\r\\nCharles: This system sucks!\\r\\nCharles: You're wasting the best

```
years of your life on education, then it turns out there's no job for
you in your profession...\\r\\nMike: Yeah, it's true. But complaining
rarely changes anything\\r\\nMike: It's better to take some action\\
r\\nMike: Going abroad is not that bad, I think I'll give it a try.
Wanna join me?\\r\\nCharles: And where would you like to go?\\r\\
nMike: The Netherlands for starters, the perspectives are decent and
it'd be a chance to visit Amsterdam, the city of freedom! :D\\r\\
nCharles: Damn, you're pulling me in!\\r\\nCharles: And actually...
besides earning some money it can serve as a good adventure\\r\\nMike:
So? R U in?\\r\\nCharles: Give me a few days, I'll ask around and let
you know soon\\r\\nMike: Fine, remember that I'm going anyway :p\\r\\
nCharles: Sure, I feel I've already decided too! :D\\r\\nCharles: Talk
to you soon\\r\\nMike: All right\",\n          \"Finn: I heard that
Britney got expelled.\\r\\nTerry: why?\\r\\nOswald: what for?\\r\\
nFinn: you know her general behavior\\r\\nOswald: and she's absent a
lot\\r\\nFinn: so yeah, that accumulated kinda, but now she's been
accused of making that blue graffiti that popped up 3 days ago\\r\\
nTerry: I was wondering if this might have anything to do with that
graffiti\\r\\nFinn: apparently it does\\r\\nOswald: you think she did
it?\\r\\nFinn: who knows? She certainly likes to draw\\r\\nTerry:
maybe they just needed an excuse to finally kick her out\\r\\nFinn:
maybe\\r\\nOswald: too bad, though, I kinda liked her\\r\\nFinn: liked
her liked her? ;)\\r\\nOswald: no, ofc not, but she's pretty cool\\r\\
nFinn: true that\",\n          \"Darren: Look! 10am!\\r\\nFrank: Whoa!
They're awesome knockers.\\r\\nDarren: Stop staring or she'll notice
us! LOL\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"summary\",\n      \"properties\": {\n        \"dtype\": \"string\",\
n        \"num_unique_values\": 14730,\n        \"samples\": [\n
\"Violet sent Claire Austin's article.\",\n          \"Michael, Tom
and Chris tease Mark because he has a new girlfriend.\",\n
\"Leo just had a very good time with his parents in Spain. Granny will
be happy if he came round and told her about his holidays. He could
also see flowers blooming in her and Daddy's garden.\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"text\",\n      \"properties\": {\n
\"dtype\": \"string\",\n        \"num_unique_values\": 14731,\n
\"samples\": [\n          \"###Human: Summarize this following
dialogue: Violet: hi! i came across this Austin's article and i
thought that you might find it interesting\\r\\nViolet: <file_other>\\
r\\nClaire: Hi! :) Thanks, but I've already read it. :)\\r\\nClaire:
But thanks for thinking about me :)\\n###Assistant: Violet sent Claire
Austin's article.\",\n          \"###Human: Summarize this following
dialogue: Pat: So does anyone know when the stream is going to
happen?\\r\\nLou: Unfortunately, no, but would really like to.\\r\\
nKevin: I don't think I'd be interested in this.\\r\\nPat: Y?\\r\\
nKevin: Seeing all the blood and internal organs makes me dizzy.\\r\\
nLou: So you're so gentle?\\r\\nPat: C'mon! Srsly?\\r\\nKevin: Yup.
Had the same thing since I was a child.\\r\\nLou: Maybe it's time to
```

```
change it?\\r\\nPat: Yeah! Give it a try!\\n###Assistant: Pat and Lou
are waiting for The stream but Kevin is not interested as it makes him
dizzy.  \",\n          \"###Human: Summarize this following dialogue:
Jane: <gif_file>\\r\\nJane: Whaddya think? \\r\\nShona: This ur tinder
profile thing?\\r\\nJane: Yeah, I'm updating my profile tonite. Kinda
nervoous though... :( \\r\\nJane: What if i get another guy like John?
o.O\\r\\nShona: John was a dickhead\\r\\nJane: preach sistah!\\r\\
nShona: anyhoo - this time I've got u :D No slimeballs for you \\r\\
nJane: Not again *shudders*\\r\\nJane: You know he forgot my
birthday??!!\\r\\nShona: wanker\\n###Assistant: Jane is updating her
Tinder profile tonight and together with Shona they don't want to find
another guy like John, who forgot Jane's birthday.\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    }\n  ]\n}","type":"dataframe","variable_name":"data_df"}
```


###Human: Summarize this following dialogue:

Amanda: I baked cookies. Do you want some?
Jerry: Yes, I'd love some.

###Assistant:

Amanda baked cookies and will bring Jerry some.

```
print(data_df.iloc[0])
```

```
id                                                  13818513
dialogue    Amanda: I baked  cookies. Do you want some?\r\...
summary     Amanda baked cookies and will bring Jerry some...
text        ###Human: Summarize this following dialogue: A...
Name: 0, dtype: object
```

```
data = Dataset.from_pandas(data_df)
```

```
data
```

```
Dataset({
    features: ['id', 'dialogue', 'summary', 'text'],
    num_rows: 14732
})
```

```
tokenizer = AutoTokenizer.from_pretrained("TheBloke/Mistral-7B-
Instruct-v0.1-GPTQ")
```


```
{"model_id":"b24c0099f6a24d718bc19c0d22c165e4","version_major":2,"vers
ion_minor":0}
```

```
{"model_id":"f07c0ad86ade433f9dd910dea675a7cf","version_major":2,"vers
ion_minor":0}
```

{"model_id":"cd38a7972cc44908abba119469bcee4e","version_major":2,"version_minor":0}

{"model_id":"c91772fa220d4a88820657972087193f","version_major":2,"version_minor":0}

```
tokenizer.eos_token
```

{"type":"string"}

```
tokenizer.eos_token_id
```

```
2
```

```
tokenizer.pad_token
```

```
tokenizer.pad_token = tokenizer.eos_token
```

```
quantization_config_loading = GPTQConfig(bits=4, disable_exllama=True,
tokenizer=tokenizer)
```

```
Using `disable_exllama` is deprecated and will be removed in version
4.37. Use `use_exllama` instead and specify the version with
`exllama_config`.The value of `use_exllama` will be overwritten by
`disable_exllama` passed in `GPTQConfig` or stored in your config
file.
```

```
model = AutoModelForCausalLM.from_pretrained(
                          "TheBloke/Mistral-7B-Instruct-v0.1-GPTQ",

quantization_config=quantization_config_loading,
                          device_map="auto")
```

{"model_id":"01da0a8e53c54adda9461f9ab7701e4a","version_major":2,"version_minor":0}

```
/usr/local/lib/python3.11/dist-packages/transformers/quantizers/
auto.py:207: UserWarning: You passed `quantization_config` or
equivalent parameters to `from_pretrained` but the model you're
loading already has a `quantization_config` attribute. The
`quantization_config` from the model will be used.However, loading
attributes (e.g. ['backend', 'use_cuda_fp16', 'use_exllama',
'max_input_length', 'exllama_config', 'disable_exllama']) will be
overwritten with the one you passed to `from_pretrained`. The rest
will be ignored.
  warnings.warn(warning_msg)
/usr/local/lib/python3.11/dist-packages/auto_gptq/nn_modules/triton_ut
ils/kernels.py:410: FutureWarning:
`torch.cuda.amp.custom_fwd(args...)` is deprecated. Please use
`torch.amp.custom_fwd(args..., device_type='cuda')` instead.
  @custom_fwd
/usr/local/lib/python3.11/dist-packages/auto_gptq/nn_modules/triton_ut
```

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

{"model_id":"6e3788b34105481997e2c8b0345727f1","version_major":2,"vers
ion_minor":0}

`loss_type=None` was set in the config but it is unrecognised.Using
the default loss: `ForCausalLMLoss`.
Some weights of the model checkpoint at TheBloke/Mistral-7B-Instruct-
v0.1-GPTQ were not used when initializing MistralForCausalLM:
{'model.layers.15.self_attn.q_proj.bias',
'model.layers.23.self_attn.q_proj.bias',
'model.layers.21.self_attn.q_proj.bias',
'model.layers.4.self_attn.q_proj.bias',
'model.layers.27.mlp.gate_proj.bias',
'model.layers.21.self_attn.v_proj.bias',
'model.layers.27.mlp.up_proj.bias',
'model.layers.1.self_attn.q_proj.bias',
'model.layers.12.mlp.gate_proj.bias',
'model.layers.2.self_attn.q_proj.bias',
'model.layers.10.self_attn.q_proj.bias',
'model.layers.10.mlp.down_proj.bias',
'model.layers.1.mlp.gate_proj.bias',
'model.layers.16.self_attn.q_proj.bias',
'model.layers.26.self_attn.k_proj.bias',
'model.layers.11.mlp.gate_proj.bias',
'model.layers.30.self_attn.v_proj.bias',
'model.layers.4.self_attn.o_proj.bias',
'model.layers.16.mlp.down_proj.bias',
'model.layers.6.self_attn.q_proj.bias',
'model.layers.13.self_attn.k_proj.bias',
'model.layers.3.self_attn.q_proj.bias',
'model.layers.12.self_attn.q_proj.bias',
'model.layers.11.self_attn.k_proj.bias',
'model.layers.18.self_attn.k_proj.bias',
'model.layers.14.mlp.down_proj.bias',
'model.layers.17.self_attn.v_proj.bias',
'model.layers.27.self_attn.q_proj.bias',
'model.layers.26.mlp.gate_proj.bias',
```

```
'model.layers.8.mlp.up_proj.bias',
'model.layers.20.self_attn.o_proj.bias',
'model.layers.6.mlp.gate_proj.bias',
'model.layers.1.self_attn.v_proj.bias',
'model.layers.17.self_attn.q_proj.bias',
'model.layers.3.mlp.up_proj.bias',
'model.layers.15.self_attn.o_proj.bias',
'model.layers.15.mlp.down_proj.bias',
'model.layers.17.self_attn.o_proj.bias',
'model.layers.30.self_attn.q_proj.bias',
'model.layers.5.mlp.gate_proj.bias',
'model.layers.8.self_attn.k_proj.bias',
'model.layers.21.mlp.down_proj.bias',
'model.layers.23.mlp.gate_proj.bias',
'model.layers.24.self_attn.k_proj.bias',
'model.layers.17.self_attn.k_proj.bias',
'model.layers.1.mlp.up_proj.bias',
'model.layers.6.self_attn.o_proj.bias',
'model.layers.5.self_attn.k_proj.bias',
'model.layers.14.self_attn.q_proj.bias',
'model.layers.12.self_attn.o_proj.bias',
'model.layers.25.mlp.up_proj.bias',
'model.layers.0.self_attn.q_proj.bias',
'model.layers.27.mlp.down_proj.bias',
'model.layers.7.mlp.up_proj.bias', 'model.layers.11.mlp.up_proj.bias',
'model.layers.12.mlp.down_proj.bias',
'model.layers.5.mlp.up_proj.bias',
'model.layers.23.mlp.down_proj.bias',
'model.layers.0.mlp.gate_proj.bias',
'model.layers.9.self_attn.q_proj.bias',
'model.layers.13.mlp.down_proj.bias',
'model.layers.5.self_attn.v_proj.bias',
'model.layers.26.self_attn.q_proj.bias',
'model.layers.27.self_attn.o_proj.bias',
'model.layers.17.mlp.up_proj.bias',
'model.layers.31.mlp.down_proj.bias',
'model.layers.12.self_attn.k_proj.bias',
'model.layers.3.mlp.down_proj.bias',
'model.layers.31.mlp.up_proj.bias',
'model.layers.5.self_attn.q_proj.bias',
'model.layers.6.mlp.down_proj.bias',
'model.layers.24.mlp.down_proj.bias',
'model.layers.3.mlp.gate_proj.bias',
'model.layers.6.mlp.up_proj.bias',
'model.layers.22.mlp.gate_proj.bias',
'model.layers.7.mlp.down_proj.bias',
'model.layers.18.mlp.down_proj.bias',
'model.layers.22.mlp.down_proj.bias',
'model.layers.29.mlp.up_proj.bias',
```

```
'model.layers.0.self_attn.k_proj.bias',
'model.layers.2.mlp.down_proj.bias',
'model.layers.19.mlp.gate_proj.bias',
'model.layers.2.mlp.up_proj.bias',
'model.layers.29.mlp.gate_proj.bias',
'model.layers.0.mlp.up_proj.bias', 'model.layers.24.mlp.up_proj.bias',
'model.layers.21.self_attn.k_proj.bias',
'model.layers.18.mlp.up_proj.bias',
'model.layers.28.self_attn.o_proj.bias',
'model.layers.28.mlp.gate_proj.bias',
'model.layers.15.self_attn.k_proj.bias',
'model.layers.4.mlp.up_proj.bias',
'model.layers.31.self_attn.k_proj.bias',
'model.layers.7.self_attn.q_proj.bias',
'model.layers.18.self_attn.q_proj.bias',
'model.layers.28.self_attn.q_proj.bias',
'model.layers.20.mlp.down_proj.bias',
'model.layers.0.self_attn.v_proj.bias',
'model.layers.12.self_attn.v_proj.bias',
'model.layers.9.self_attn.o_proj.bias',
'model.layers.16.self_attn.v_proj.bias',
'model.layers.18.self_attn.o_proj.bias',
'model.layers.13.mlp.up_proj.bias',
'model.layers.15.mlp.gate_proj.bias',
'model.layers.25.self_attn.v_proj.bias',
'model.layers.22.self_attn.o_proj.bias',
'model.layers.24.self_attn.v_proj.bias',
'model.layers.9.self_attn.k_proj.bias',
'model.layers.29.self_attn.o_proj.bias',
'model.layers.17.mlp.down_proj.bias',
'model.layers.21.mlp.up_proj.bias',
'model.layers.10.self_attn.k_proj.bias',
'model.layers.11.self_attn.o_proj.bias',
'model.layers.20.self_attn.v_proj.bias',
'model.layers.22.self_attn.v_proj.bias',
'model.layers.31.self_attn.v_proj.bias',
'model.layers.31.self_attn.o_proj.bias',
'model.layers.22.self_attn.k_proj.bias',
'model.layers.22.mlp.up_proj.bias',
'model.layers.4.self_attn.k_proj.bias',
'model.layers.27.self_attn.v_proj.bias',
'model.layers.29.self_attn.k_proj.bias',
'model.layers.11.mlp.down_proj.bias',
'model.layers.14.self_attn.o_proj.bias',
'model.layers.31.self_attn.q_proj.bias',
'model.layers.14.mlp.gate_proj.bias',
'model.layers.14.self_attn.v_proj.bias',
'model.layers.4.mlp.gate_proj.bias',
'model.layers.6.self_attn.v_proj.bias',
```

```
'model.layers.24.mlp.gate_proj.bias',
'model.layers.28.self_attn.k_proj.bias',
'model.layers.21.mlp.gate_proj.bias',
'model.layers.30.mlp.gate_proj.bias',
'model.layers.19.self_attn.v_proj.bias',
'model.layers.6.self_attn.k_proj.bias',
'model.layers.23.self_attn.k_proj.bias',
'model.layers.10.self_attn.v_proj.bias',
'model.layers.26.self_attn.v_proj.bias',
'model.layers.19.mlp.down_proj.bias',
'model.layers.23.self_attn.v_proj.bias',
'model.layers.7.self_attn.o_proj.bias',
'model.layers.10.self_attn.o_proj.bias',
'model.layers.26.mlp.down_proj.bias',
'model.layers.7.self_attn.v_proj.bias',
'model.layers.0.mlp.down_proj.bias',
'model.layers.20.self_attn.k_proj.bias',
'model.layers.3.self_attn.k_proj.bias',
'model.layers.26.self_attn.o_proj.bias',
'model.layers.28.self_attn.v_proj.bias',
'model.layers.5.self_attn.o_proj.bias',
'model.layers.14.self_attn.k_proj.bias',
'model.layers.28.mlp.up_proj.bias',
'model.layers.1.self_attn.o_proj.bias',
'model.layers.19.self_attn.o_proj.bias',
'model.layers.9.mlp.gate_proj.bias',
'model.layers.23.self_attn.o_proj.bias',
'model.layers.23.mlp.up_proj.bias',
'model.layers.27.self_attn.k_proj.bias',
'model.layers.15.mlp.up_proj.bias',
'model.layers.24.self_attn.o_proj.bias',
'model.layers.13.self_attn.q_proj.bias',
'model.layers.16.self_attn.o_proj.bias',
'model.layers.19.self_attn.k_proj.bias',
'model.layers.25.self_attn.q_proj.bias',
'model.layers.30.mlp.up_proj.bias',
'model.layers.3.self_attn.v_proj.bias',
'model.layers.4.self_attn.v_proj.bias',
'model.layers.13.mlp.gate_proj.bias',
'model.layers.16.mlp.gate_proj.bias',
'model.layers.8.mlp.down_proj.bias',
'model.layers.18.mlp.gate_proj.bias',
'model.layers.2.self_attn.v_proj.bias',
'model.layers.8.self_attn.v_proj.bias',
'model.layers.20.mlp.gate_proj.bias',
'model.layers.8.self_attn.o_proj.bias',
'model.layers.25.self_attn.o_proj.bias',
'model.layers.29.self_attn.v_proj.bias',
'model.layers.2.self_attn.o_proj.bias',
```

```
'model.layers.7.mlp.gate_proj.bias',
'model.layers.16.mlp.up_proj.bias',
'model.layers.29.self_attn.q_proj.bias',
'model.layers.24.self_attn.q_proj.bias',
'model.layers.16.self_attn.k_proj.bias',
'model.layers.13.self_attn.o_proj.bias',
'model.layers.21.self_attn.o_proj.bias',
'model.layers.18.self_attn.v_proj.bias',
'model.layers.25.mlp.gate_proj.bias',
'model.layers.3.self_attn.o_proj.bias',
'model.layers.10.mlp.up_proj.bias',
'model.layers.2.self_attn.k_proj.bias',
'model.layers.13.self_attn.v_proj.bias',
'model.layers.30.mlp.down_proj.bias',
'model.layers.8.self_attn.q_proj.bias',
'model.layers.22.self_attn.q_proj.bias',
'model.layers.9.mlp.down_proj.bias',
'model.layers.25.self_attn.k_proj.bias',
'model.layers.10.mlp.gate_proj.bias',
'model.layers.30.self_attn.k_proj.bias',
'model.layers.5.mlp.down_proj.bias',
'model.layers.7.self_attn.k_proj.bias',
'model.layers.29.mlp.down_proj.bias',
'model.layers.0.self_attn.o_proj.bias',
'model.layers.26.mlp.up_proj.bias',
'model.layers.17.mlp.gate_proj.bias',
'model.layers.11.self_attn.q_proj.bias',
'model.layers.2.mlp.gate_proj.bias',
'model.layers.15.self_attn.v_proj.bias',
'model.layers.11.self_attn.v_proj.bias',
'model.layers.12.mlp.up_proj.bias',
'model.layers.14.mlp.up_proj.bias',
'model.layers.19.mlp.up_proj.bias',
'model.layers.4.mlp.down_proj.bias',
'model.layers.1.mlp.down_proj.bias',
'model.layers.28.mlp.down_proj.bias',
'model.layers.20.mlp.up_proj.bias',
'model.layers.1.self_attn.k_proj.bias',
'model.layers.20.self_attn.q_proj.bias',
'model.layers.8.mlp.gate_proj.bias',
'model.layers.30.self_attn.o_proj.bias',
'model.layers.19.self_attn.q_proj.bias',
'model.layers.31.mlp.gate_proj.bias',
'model.layers.9.self_attn.v_proj.bias',
'model.layers.25.mlp.down_proj.bias',
'model.layers.9.mlp.up_proj.bias'}
- This IS expected if you are initializing MistralForCausalLM from the
checkpoint of a model trained on another task or with another
architecture (e.g. initializing a BertForSequenceClassification model
```

```
from a BertForPreTraining model).
- This IS NOT expected if you are initializing MistralForCausalLM from
the checkpoint of a model that you expect to be exactly identical
(initializing a BertForSequenceClassification model from a
BertForSequenceClassification model).

{"model_id":"86be6488e3254dab98704c3658de5284","version_major":2,"vers
ion_minor":0}

print(model)

MistralForCausalLM(
  (model): MistralModel(
    (embed_tokens): Embedding(32000, 4096, padding_idx=0)
    (layers): ModuleList(
      (0-31): 32 x MistralDecoderLayer(
        (self_attn): MistralAttention(
          (k_proj): QuantLinear()
          (o_proj): QuantLinear()
          (q_proj): QuantLinear()
          (v_proj): QuantLinear()
        )
        (mlp): MistralMLP(
          (act_fn): SiLU()
          (down_proj): QuantLinear()
          (gate_proj): QuantLinear()
          (up_proj): QuantLinear()
        )
        (input_layernorm): MistralRMSNorm((4096,), eps=1e-05)
        (post_attention_layernorm): MistralRMSNorm((4096,), eps=1e-05)
      )
    )
    (norm): MistralRMSNorm((4096,), eps=1e-05)
    (rotary_emb): MistralRotaryEmbedding()
  )
  (lm_head): Linear(in_features=4096, out_features=32000, bias=False)
)

# Load a 4-bit quantized model
quantization_config = BitsAndBytesConfig(
    load_in_4bit=True,    # Enable 4-bit quantization
    bnb_4bit_compute_dtype=torch.float16,  # Use fp16 for computation
    bnb_4bit_use_double_quant=True,  # Use double quantization for
memory efficiency
)

# Load model and tokenizer
model = AutoModelForCausalLM.from_pretrained(
    "mistralai/Mistral-7B-Instruct-v0.1",
    quantization_config=quantization_config,
```

```
    device_map="auto"  # Automatically assigns layers to available
GPUs
)

model.config.use_cache=False

model.config.pretraining_tp=1

model.gradient_checkpointing_enable()

model = prepare_model_for_kbit_training(model)
```

r=16 controls how much LoRA modifies the model (higher = more expressive).

 lora_alpha=16 scales LoRA's effect on training.

 lora_dropout=0.05 prevents overfitting (good for small datasets).

 target_modules=["q_proj", "v_proj"] makes LoRA memory-efficient.

 Great for fine-tuning LLaMA, Mistral, Falcon on low-VRAM GPUs.

```
# ["q_proj", "v_proj", "k_proj"] → Adds key projection (more
expressive)
# ["q_proj", "v_proj", "out_proj"] → Also fine-tunes attention output

peft_config = LoraConfig(
        r=16, lora_alpha=16, lora_dropout=0.05, bias="none",
task_type="CAUSAL_LM", target_modules=["q_proj", "v_proj"]
    )

model = get_peft_model(model, peft_config)
```

```
/usr/local/lib/python3.11/dist-packages/peft/mapping.py:185:
UserWarning: The PEFT config's `base_model_name_or_path` was renamed
from 'TheBloke/Mistral-7B-Instruct-v0.1-GPTQ' to 'None'. Please ensure
that the correct base model is loaded when loading this checkpoint.
  warnings.warn(
```

```
print(model)
```

```
PeftModelForCausalLM(
  (base_model): LoraModel(
    (model): PeftModelForCausalLM(
      (base_model): LoraModel(
        (model): MistralForCausalLM(
          (model): MistralModel(
            (embed_tokens): Embedding(32000, 4096, padding_idx=0)
            (layers): ModuleList(
              (0-31): 32 x MistralDecoderLayer(
                (self_attn): MistralAttention(
                  (k_proj): QuantLinear()
```

```
                  (o_proj): QuantLinear()
                  (q_proj): lora.QuantLinear(
                    (base_layer): QuantLinear()
                    (lora_dropout): ModuleDict(
                      (default): Dropout(p=0.05, inplace=False)
                    )
                    (lora_A): ModuleDict(
                      (default): Linear(in_features=4096,
out_features=16, bias=False)
                    )
                    (lora_B): ModuleDict(
                      (default): Linear(in_features=16,
out_features=4096, bias=False)
                    )
                    (lora_embedding_A): ParameterDict()
                    (lora_embedding_B): ParameterDict()
                    (lora_magnitude_vector): ModuleDict()
                    (quant_linear_module): QuantLinear()
                  )
                  (v_proj): lora.QuantLinear(
                    (base_layer): QuantLinear()
                    (lora_dropout): ModuleDict(
                      (default): Dropout(p=0.05, inplace=False)
                    )
                    (lora_A): ModuleDict(
                      (default): Linear(in_features=4096,
out_features=16, bias=False)
                    )
                    (lora_B): ModuleDict(
                      (default): Linear(in_features=16,
out_features=1024, bias=False)
                    )
                    (lora_embedding_A): ParameterDict()
                    (lora_embedding_B): ParameterDict()
                    (lora_magnitude_vector): ModuleDict()
                    (quant_linear_module): QuantLinear()
                  )
                )
                (mlp): MistralMLP(
                  (act_fn): SiLU()
                  (down_proj): QuantLinear()
                  (gate_proj): QuantLinear()
                  (up_proj): QuantLinear()
                )
                (input_layernorm): MistralRMSNorm((4096,), eps=1e-05)
                (post_attention_layernorm): MistralRMSNorm((4096,),
eps=1e-05)
              )
            )
```

```
            (norm): MistralRMSNorm((4096,), eps=1e-05)
            (rotary_emb): MistralRotaryEmbedding()
          )
          (lm_head): Linear(in_features=4096, out_features=32000,
bias=False)
        )
      )
    )
  )
)

training_arguments = TrainingArguments(
        output_dir="mistral-finetuned-samsum",
        per_device_train_batch_size=8,
        gradient_accumulation_steps=1,
        optim="paged_adamw_32bit",
        learning_rate=2e-4,
        lr_scheduler_type="cosine",
        save_strategy="epoch",
        logging_steps=100,
        num_train_epochs=1,
        max_steps=250,
        fp16=True,
        push_to_hub=True,
        report_to="none"
  )

# Create the SFTTrainer
trainer = SFTTrainer(
        model=model,
        train_dataset=data,
        peft_config=peft_config,
        args=training_arguments,
        tokenizer=tokenizer,
    )

<ipython-input-35-92d616e9460c>:2: FutureWarning: `tokenizer` is
deprecated and removed starting from version 0.16.0 for
`SFTTrainer.__init__`. Use `processing_class` instead.
  trainer = SFTTrainer(

{"model_id":"74a4133dbb134847878a92ea38db93ee","version_major":2,"vers
ion_minor":0}

{"model_id":"8b0eb23227a14e6e877b44d2a9cacfef","version_major":2,"vers
ion_minor":0}

{"model_id":"1c5eb099a1b24bfc8916d2ffec613143","version_major":2,"vers
ion_minor":0}
```

{"model_id":"1b272d72aa5f4484bb6972064ca5e535","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.
```

```
trainer.train()
```

```
/usr/local/lib/python3.11/dist-packages/torch/_dynamo/eval_frame.py:632: UserWarning: torch.utils.checkpoint: the use_reentrant parameter should be passed explicitly. In version 2.5 we will raise an exception if use_reentrant is not passed. use_reentrant=False is recommended, but if you need to preserve the current default behavior, you can pass use_reentrant=True. Refer to docs for more details on the differences between the two variants.
  return fn(*args, **kwargs)
```

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

```
TrainOutput(global_step=250, training_loss=1.8019019775390626, metrics={'train_runtime': 2763.5241, 'train_samples_per_second': 0.724, 'train_steps_per_second': 0.09, 'total_flos': 704127752404992.0, 'train_loss': 1.8019019775390626})
```

```
! cp -r /content/mistral-finetuned-samsum /content/drive/MyDrive/
```

```
trainer.push_to_hub()
```

```
from peft import AutoPeftModelForCausalLM
from transformers import GenerationConfig
from transformers import AutoTokenizer
import torch
tokenizer = AutoTokenizer.from_pretrained("/content/mistral-finetuned-samsum")

inputs = tokenizer("""
###Human: Summarize this following dialogue: Sunny: I'm at the railway station in Chennai Karthik: No problems so far? Sunny: no, everything's going smoothly Karthik: good. lets meet there soon!
###Assistant: """, return_tensors="pt").to("cuda")

model = AutoPeftModelForCausalLM.from_pretrained(
    "/content/mistral-finetuned-samsum",
    low_cpu_mem_usage=True,
    return_dict=True,
    torch_dtype=torch.float16,
    device_map="cuda")
```

```
Some weights of the model checkpoint at TheBloke/Mistral-7B-Instruct-v0.1-GPTQ were not used when initializing MistralForCausalLM:
```

```
{'model.layers.15.self_attn.q_proj.bias',
'model.layers.23.self_attn.q_proj.bias',
'model.layers.21.self_attn.q_proj.bias',
'model.layers.4.self_attn.q_proj.bias',
'model.layers.27.mlp.gate_proj.bias',
'model.layers.21.self_attn.v_proj.bias',
'model.layers.27.mlp.up_proj.bias',
'model.layers.1.self_attn.q_proj.bias',
'model.layers.12.mlp.gate_proj.bias',
'model.layers.2.self_attn.q_proj.bias',
'model.layers.10.self_attn.q_proj.bias',
'model.layers.10.mlp.down_proj.bias',
'model.layers.1.mlp.gate_proj.bias',
'model.layers.16.self_attn.q_proj.bias',
'model.layers.26.self_attn.k_proj.bias',
'model.layers.11.mlp.gate_proj.bias',
'model.layers.30.self_attn.v_proj.bias',
'model.layers.4.self_attn.o_proj.bias',
'model.layers.16.mlp.down_proj.bias',
'model.layers.6.self_attn.q_proj.bias',
'model.layers.13.self_attn.k_proj.bias',
'model.layers.3.self_attn.q_proj.bias',
'model.layers.12.self_attn.q_proj.bias',
'model.layers.11.self_attn.k_proj.bias',
'model.layers.18.self_attn.k_proj.bias',
'model.layers.14.mlp.down_proj.bias',
'model.layers.17.self_attn.v_proj.bias',
'model.layers.27.self_attn.q_proj.bias',
'model.layers.26.mlp.gate_proj.bias',
'model.layers.8.mlp.up_proj.bias',
'model.layers.20.self_attn.o_proj.bias',
'model.layers.6.mlp.gate_proj.bias',
'model.layers.1.self_attn.v_proj.bias',
'model.layers.17.self_attn.q_proj.bias',
'model.layers.3.mlp.up_proj.bias',
'model.layers.15.self_attn.o_proj.bias',
'model.layers.15.mlp.down_proj.bias',
'model.layers.17.self_attn.o_proj.bias',
'model.layers.30.self_attn.q_proj.bias',
'model.layers.5.mlp.gate_proj.bias',
'model.layers.8.self_attn.k_proj.bias',
'model.layers.21.mlp.down_proj.bias',
'model.layers.23.mlp.gate_proj.bias',
'model.layers.24.self_attn.k_proj.bias',
'model.layers.17.self_attn.k_proj.bias',
'model.layers.1.mlp.up_proj.bias',
'model.layers.6.self_attn.o_proj.bias',
'model.layers.5.self_attn.k_proj.bias',
'model.layers.14.self_attn.q_proj.bias',
```

```
'model.layers.12.self_attn.o_proj.bias',
'model.layers.25.mlp.up_proj.bias',
'model.layers.0.self_attn.q_proj.bias',
'model.layers.27.mlp.down_proj.bias',
'model.layers.7.mlp.up_proj.bias', 'model.layers.11.mlp.up_proj.bias',
'model.layers.12.mlp.down_proj.bias',
'model.layers.5.mlp.up_proj.bias',
'model.layers.23.mlp.down_proj.bias',
'model.layers.0.mlp.gate_proj.bias',
'model.layers.9.self_attn.q_proj.bias',
'model.layers.13.mlp.down_proj.bias',
'model.layers.5.self_attn.v_proj.bias',
'model.layers.26.self_attn.q_proj.bias',
'model.layers.27.self_attn.o_proj.bias',
'model.layers.17.mlp.up_proj.bias',
'model.layers.31.mlp.down_proj.bias',
'model.layers.12.self_attn.k_proj.bias',
'model.layers.3.mlp.down_proj.bias',
'model.layers.31.mlp.up_proj.bias',
'model.layers.5.self_attn.q_proj.bias',
'model.layers.6.mlp.down_proj.bias',
'model.layers.24.mlp.down_proj.bias',
'model.layers.3.mlp.gate_proj.bias',
'model.layers.6.mlp.up_proj.bias',
'model.layers.22.mlp.gate_proj.bias',
'model.layers.7.mlp.down_proj.bias',
'model.layers.18.mlp.down_proj.bias',
'model.layers.22.mlp.down_proj.bias',
'model.layers.29.mlp.up_proj.bias',
'model.layers.0.self_attn.k_proj.bias',
'model.layers.2.mlp.down_proj.bias',
'model.layers.19.mlp.gate_proj.bias',
'model.layers.2.mlp.up_proj.bias',
'model.layers.29.mlp.gate_proj.bias',
'model.layers.0.mlp.up_proj.bias', 'model.layers.24.mlp.up_proj.bias',
'model.layers.21.self_attn.k_proj.bias',
'model.layers.18.mlp.up_proj.bias',
'model.layers.28.self_attn.o_proj.bias',
'model.layers.28.mlp.gate_proj.bias',
'model.layers.15.self_attn.k_proj.bias',
'model.layers.4.mlp.up_proj.bias',
'model.layers.31.self_attn.k_proj.bias',
'model.layers.7.self_attn.q_proj.bias',
'model.layers.18.self_attn.q_proj.bias',
'model.layers.28.self_attn.q_proj.bias',
'model.layers.20.mlp.down_proj.bias',
'model.layers.0.self_attn.v_proj.bias',
'model.layers.12.self_attn.v_proj.bias',
'model.layers.9.self_attn.o_proj.bias',
```

```
'model.layers.16.self_attn.v_proj.bias',
'model.layers.18.self_attn.o_proj.bias',
'model.layers.13.mlp.up_proj.bias',
'model.layers.15.mlp.gate_proj.bias',
'model.layers.25.self_attn.v_proj.bias',
'model.layers.22.self_attn.o_proj.bias',
'model.layers.24.self_attn.v_proj.bias',
'model.layers.9.self_attn.k_proj.bias',
'model.layers.29.self_attn.o_proj.bias',
'model.layers.17.mlp.down_proj.bias',
'model.layers.21.mlp.up_proj.bias',
'model.layers.10.self_attn.k_proj.bias',
'model.layers.11.self_attn.o_proj.bias',
'model.layers.20.self_attn.v_proj.bias',
'model.layers.22.self_attn.v_proj.bias',
'model.layers.31.self_attn.v_proj.bias',
'model.layers.31.self_attn.o_proj.bias',
'model.layers.22.self_attn.k_proj.bias',
'model.layers.22.mlp.up_proj.bias',
'model.layers.4.self_attn.k_proj.bias',
'model.layers.27.self_attn.v_proj.bias',
'model.layers.29.self_attn.k_proj.bias',
'model.layers.11.mlp.down_proj.bias',
'model.layers.14.self_attn.o_proj.bias',
'model.layers.31.self_attn.q_proj.bias',
'model.layers.14.mlp.gate_proj.bias',
'model.layers.14.self_attn.v_proj.bias',
'model.layers.4.mlp.gate_proj.bias',
'model.layers.6.self_attn.v_proj.bias',
'model.layers.24.mlp.gate_proj.bias',
'model.layers.28.self_attn.k_proj.bias',
'model.layers.21.mlp.gate_proj.bias',
'model.layers.30.mlp.gate_proj.bias',
'model.layers.19.self_attn.v_proj.bias',
'model.layers.6.self_attn.k_proj.bias',
'model.layers.23.self_attn.k_proj.bias',
'model.layers.10.self_attn.v_proj.bias',
'model.layers.26.self_attn.v_proj.bias',
'model.layers.19.mlp.down_proj.bias',
'model.layers.23.self_attn.v_proj.bias',
'model.layers.7.self_attn.o_proj.bias',
'model.layers.10.self_attn.o_proj.bias',
'model.layers.26.mlp.down_proj.bias',
'model.layers.7.self_attn.v_proj.bias',
'model.layers.0.mlp.down_proj.bias',
'model.layers.20.self_attn.k_proj.bias',
'model.layers.3.self_attn.k_proj.bias',
'model.layers.26.self_attn.o_proj.bias',
'model.layers.28.self_attn.v_proj.bias',
```

```
'model.layers.5.self_attn.o_proj.bias',
'model.layers.14.self_attn.k_proj.bias',
'model.layers.28.mlp.up_proj.bias',
'model.layers.1.self_attn.o_proj.bias',
'model.layers.19.self_attn.o_proj.bias',
'model.layers.9.mlp.gate_proj.bias',
'model.layers.23.self_attn.o_proj.bias',
'model.layers.23.mlp.up_proj.bias',
'model.layers.27.self_attn.k_proj.bias',
'model.layers.15.mlp.up_proj.bias',
'model.layers.24.self_attn.o_proj.bias',
'model.layers.13.self_attn.q_proj.bias',
'model.layers.16.self_attn.o_proj.bias',
'model.layers.19.self_attn.k_proj.bias',
'model.layers.25.self_attn.q_proj.bias',
'model.layers.30.mlp.up_proj.bias',
'model.layers.3.self_attn.v_proj.bias',
'model.layers.4.self_attn.v_proj.bias',
'model.layers.13.mlp.gate_proj.bias',
'model.layers.16.mlp.gate_proj.bias',
'model.layers.8.mlp.down_proj.bias',
'model.layers.18.mlp.gate_proj.bias',
'model.layers.2.self_attn.v_proj.bias',
'model.layers.8.self_attn.v_proj.bias',
'model.layers.20.mlp.gate_proj.bias',
'model.layers.8.self_attn.o_proj.bias',
'model.layers.25.self_attn.o_proj.bias',
'model.layers.29.self_attn.v_proj.bias',
'model.layers.2.self_attn.o_proj.bias',
'model.layers.7.mlp.gate_proj.bias',
'model.layers.16.mlp.up_proj.bias',
'model.layers.29.self_attn.q_proj.bias',
'model.layers.24.self_attn.q_proj.bias',
'model.layers.16.self_attn.k_proj.bias',
'model.layers.13.self_attn.o_proj.bias',
'model.layers.21.self_attn.o_proj.bias',
'model.layers.18.self_attn.v_proj.bias',
'model.layers.25.mlp.gate_proj.bias',
'model.layers.3.self_attn.o_proj.bias',
'model.layers.10.mlp.up_proj.bias',
'model.layers.2.self_attn.k_proj.bias',
'model.layers.13.self_attn.v_proj.bias',
'model.layers.30.mlp.down_proj.bias',
'model.layers.8.self_attn.q_proj.bias',
'model.layers.22.self_attn.q_proj.bias',
'model.layers.9.mlp.down_proj.bias',
'model.layers.25.self_attn.k_proj.bias',
'model.layers.10.mlp.gate_proj.bias',
'model.layers.30.self_attn.k_proj.bias',
```

```
'model.layers.5.mlp.down_proj.bias',
'model.layers.7.self_attn.k_proj.bias',
'model.layers.29.mlp.down_proj.bias',
'model.layers.0.self_attn.o_proj.bias',
'model.layers.26.mlp.up_proj.bias',
'model.layers.17.mlp.gate_proj.bias',
'model.layers.11.self_attn.q_proj.bias',
'model.layers.2.mlp.gate_proj.bias',
'model.layers.15.self_attn.v_proj.bias',
'model.layers.11.self_attn.v_proj.bias',
'model.layers.12.mlp.up_proj.bias',
'model.layers.14.mlp.up_proj.bias',
'model.layers.19.mlp.up_proj.bias',
'model.layers.4.mlp.down_proj.bias',
'model.layers.1.mlp.down_proj.bias',
'model.layers.28.mlp.down_proj.bias',
'model.layers.20.mlp.up_proj.bias',
'model.layers.1.self_attn.k_proj.bias',
'model.layers.20.self_attn.q_proj.bias',
'model.layers.8.mlp.gate_proj.bias',
'model.layers.30.self_attn.o_proj.bias',
'model.layers.19.self_attn.q_proj.bias',
'model.layers.31.mlp.gate_proj.bias',
'model.layers.9.self_attn.v_proj.bias',
'model.layers.25.mlp.down_proj.bias',
'model.layers.9.mlp.up_proj.bias'}
- This IS expected if you are initializing MistralForCausalLM from the
checkpoint of a model trained on another task or with another
architecture (e.g. initializing a BertForSequenceClassification model
from a BertForPreTraining model).
- This IS NOT expected if you are initializing MistralForCausalLM from
the checkpoint of a model that you expect to be exactly identical
(initializing a BertForSequenceClassification model from a
BertForSequenceClassification model).

generation_config = GenerationConfig(
    do_sample=True,
    top_k=1,
    temperature=0.1,
    max_new_tokens=25,
    pad_token_id=tokenizer.eos_token_id
)

import time
st_time = time.time()
outputs = model.generate(**inputs,
generation_config=generation_config)
print(tokenizer.decode(outputs[0], skip_special_tokens=True))
print(time.time()-st_time)
```

```
###Human: Summarize this following dialogue: Vasanth: I'm at the railway station in Chennai Karthik: No problems so far? Vasanth: no, everything's going smoothly Karthik: good. lets meet there soon!
###Assistant:  Vasanth is at the railway station in Chennai. Everything is going smoothly. He will meet Karthik soon.
29.117424964904785
```