

```
# 🗑️ Uninstall all cached Hugging Face components
!pip uninstall -y transformers accelerate peft trl
```

```
🔄 Found existing installation: transformers 4.52.4
Uninstalling transformers-4.52.4:
  Successfully uninstalled transformers-4.52.4
Found existing installation: accelerate 1.7.0
Uninstalling accelerate-1.7.0:
  Successfully uninstalled accelerate-1.7.0
Found existing installation: peft 0.15.2
Uninstalling peft-0.15.2:
  Successfully uninstalled peft-0.15.2
WARNING: Skipping trl as it is not installed.
```

```
# 📦 1. Install Required Libraries - all with correct versions
!pip install -U "transformers>=4.38.0" "datasets" "accelerate" "trl>=0.7.9" "peft>=0.7.1" "bitsandbytes"
```

```
🔄 Collecting transformers>=4.38.0
  Downloading transformers-4.52.4-py3-none-any.whl.metadata (38 kB)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from transformers>=4.38.0) (3.18.0)
Requirement already satisfied: datasets in /usr/local/lib/python3.11/dist-packages (from transformers>=4.38.0) (2.14.4)
Collecting datasets
  Downloading datasets-3.6.0-py3-none-any.whl.metadata (19 kB)
Collecting accelerate
  Downloading accelerate-1.8.1-py3-none-any.whl.metadata (19 kB)
Collecting trl>=0.7.9
  Downloading trl-0.19.0-py3-none-any.whl.metadata (10 kB)
Collecting peft>=0.7.1
  Downloading peft-0.15.2-py3-none-any.whl.metadata (13 kB)
Collecting bitsandbytes
  Downloading bitsandbytes-0.46.0-py3-none-manylinux_2_24_x86_64.whl.metadata (10 kB)
Requirement already satisfied: huggingface-hub<1.0,>=0.30.0 in /usr/local/lib/python3.11/dist-packages (from transformers>=4.38.0) (3.18.0)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-packages (from transformers>=4.38.0) (2.0.2)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from transformers>=4.38.0) (24.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-packages (from transformers>=4.38.0) (6.0.2)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.11/dist-packages (from transformers>=4.38.0) (2024.1)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages (from transformers>=4.38.0) (2.32.3)
Requirement already satisfied: tokenizers<0.22,>=0.21 in /usr/local/lib/python3.11/dist-packages (from transformers>=4.38.0) (0.20.3)
Requirement already satisfied: safetensors>=0.4.3 in /usr/local/lib/python3.11/dist-packages (from transformers>=4.38.0) (0.5.3)
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.11/dist-packages (from transformers>=4.38.0) (4.67.1)
Requirement already satisfied: pyarrow>=15.0.0 in /usr/local/lib/python3.11/dist-packages (from datasets) (18.1.0)
Requirement already satisfied: dill<0.3.9,>=0.3.0 in /usr/local/lib/python3.11/dist-packages (from datasets) (0.3.7)
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages (from datasets) (2.2.2)
Requirement already satisfied: xxhash in /usr/local/lib/python3.11/dist-packages (from datasets) (3.5.0)
Requirement already satisfied: multiprocess<0.70.17 in /usr/local/lib/python3.11/dist-packages (from datasets) (0.70.15)
Collecting fsspec<=2025.3.0,>=2023.1.0 (from fsspec[http]<=2025.3.0,>=2023.1.0->datasets)
  Downloading fsspec-2025.3.0-py3-none-any.whl.metadata (11 kB)
Requirement already satisfied: psutil in /usr/local/lib/python3.11/dist-packages (from accelerate) (5.9.5)
Requirement already satisfied: torch>=2.0.0 in /usr/local/lib/python3.11/dist-packages (from accelerate) (2.6.0+cu124)
Requirement already satisfied: aiohttp!=4.0.0a0,!4.0.0a1 in /usr/local/lib/python3.11/dist-packages (from fsspec[http]<=2025.3.0->datasets) (3.11.10)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0,>=0.30.0->transformers) (4.12.2)
Requirement already satisfied: hf-xet<2.0.0,>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0,>=0.30.0->transformers) (1.2.0)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (3.3.0)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (3.10.1)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (2025.1.1)
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-packages (from torch>=2.0.0->accelerate) (3.5)
Requirement already satisfied: Jinja2 in /usr/local/lib/python3.11/dist-packages (from torch>=2.0.0->accelerate) (3.1.6)
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)
```

```
# Sanity check
import transformers
import trl
import peft

print(transformers.__version__) # should be 4.38.0 or newer
print(trl.__version__)         # should be 0.7.9 or newer
print(peft.__version__)        # should be 0.7.1 or newer
```

```
# 🗝️ 2. Login to Hugging Face (you'll need a token from https://huggingface.co/settings/tokens)
from huggingface_hub import login
login() # Enter your HF token here (with write access)
```

```
import torch
from datasets import load_dataset
from transformers import AutoTokenizer, AutoModelForCausalLM, BitsAndBytesConfig
from peft import LoraConfig, get_peft_model, TaskType
from trl import SFTConfig, SFTTrainer
```

```
bnb = BitsAndBytesConfig(
    load_in_4bit=True,
    bnb_4bit_quant_type="nf4",
    bnb_4bit_use_double_quant=True,
    bnb_4bit_compute_dtype=torch.float16
)
```

```
model = AutoModelForCausalLM.from_pretrained(
    model_id,
    device_map="auto",
    quantization_config=bnb,
    trust_remote_code=True
)
```

```
# LoRA adapters
peft_cfg = LoraConfig(
    r=8, lora_alpha=16,
    target_modules=["q_proj", "v_proj"],
    lora_dropout=0.05,
    bias="none",
    task_type=TaskType.CAUSAL_LM
)
```

```
model = get_peft_model(model, peft_cfg)
```

```
# 4. Load dataset
raw = load_dataset("yahma/alpaca-cleaned")
```

```
def formatting_func(example):
    prompt = f"### Instruction:\n{example['instruction']}\n"
    if example['input']:
        prompt += f"### Input:\n{example['input']}\n"
    return prompt + f"### Response:\n{example['output']}"

# 🚀 Subset to speed it up – 5k train, 500 eval
raw["train"] = raw["train"].shuffle(seed=42).select(range(5500))
ds = raw["train"].train_test_split(test_size=500 / 5500)
```

🔄 README.md: 100%  11.6k/11.6k [00:00<00:00, 477kB/s]


alpaca_data_cleaned.json: 100%  44.3M/44.3M [00:01<00:00, 34.7MB/s]


Generating train split: 100%  51760/51760 [00:00<00:00, 78446.79 examples/s]


```
# ⚙️ 5. Configure SFTTrainer w/ logging & autosave
sft_cfg = SFTConfig(
    output_dir="./phi2-alpaca-lora-4bit",
    per_device_train_batch_size=1,
    gradient_accumulation_steps=4,
    num_train_epochs=3,
    logging_steps=50,
    save_steps=500,
    save_total_limit=3,
    eval_strategy="steps",
    eval_steps=500,
    eval_packing=False,
    fp16=True,
    max_length=512,
    push_to_hub=True,
    hub_model_id="gauri-sharan/phi2-alpaca-lora-4bit",
    hub_private_repo=False,
    report_to="none"
)
```


🔄 average_tokens_across_devices is set to True but it is invalid when world size is 1. Turn it to False automatically.


```
# 🏠 6. Initialize Trainer
trainer = SFTTrainer(
    model=model,
    args=sft_cfg,
    train_dataset=ds["train"],
    eval_dataset=ds["test"],
    processing_class=tokenizer,
    formatting_func=formatting_func
)
```


🔄 Applying formatting function to train dataset: 100%  5000/5000 [00:01<00:00, 4681.93 examples/s]


Adding EOS to train dataset: 100%  5000/5000 [00:00<00:00, 7277.15 examples/s]


Tokenizing train dataset: 100%  5000/5000 [00:07<00:00, 756.93 examples/s]

Truncating train dataset: 100%  5000/5000 [00:00<00:00, 71614.26 examples/s]

Applying formatting function to eval dataset: 100%  500/500 [00:00<00:00, 3770.15 examples/s]

Adding EOS to eval dataset: 100%  500/500 [00:00<00:00, 5263.95 examples/s]

Tokenizing eval dataset: 100%  500/500 [00:00<00:00, 799.89 examples/s]

Truncating eval dataset: 100%  500/500 [00:00<00:00, 25578.77 examples/s]

No label_names provided for model class `PeftModelForCausalLM`. Since `PeftModel` hides base models input arguments, if label_name

```
# ✅ 7. Save a Model Card (README.md)
readme = """---
license: apache-2.0
tags:
- phi
- text-generation
- instruction-tuning
datasets:
- yahma/alpaca-cleaned
model-index:
- name: Phi2-Alpaca-LoRA
  results: []
```

```

---

# Phi2-Alpaca-LoRA

This is a LoRA finetuned version of [microsoft/phi-2](https://huggingface.co/microsoft/phi-2) using the [Stanford Alpaca dataset](https://stanfordnlp.github.io/alpaca/)

## 🧠 Training Details

- Base model: Phi-2
- Dataset: Alpaca (cleaned)
- Method: PEFT (LoRA) via SFTTrainer
- Framework: 🤗 Transformers + TRL

## 🚀 Quickstart

```python
from transformers import pipeline
pipe = pipeline("text-generation", model="gauri-sharan/phi2-alpaca-lora")
print(pipe("### Instruction:\nExplain quantum tunneling.\n### Response:\n")[0]['generated_text'])
```

"""

with open("README.md", "w") as f:
    f.write(readme)

```

```

# 🚀 8. Train + Save mid-run + Push
trainer.train()
trainer.save_model("checkpoint_final")

```

```

🔄 [3750/3750 2:06:15, Epoch 3/3]

```

| Step | Training Loss | Validation Loss |
|------|---------------|-----------------|
| 500 | 1.008100 | 0.935242 |
| 1000 | 0.945200 | 0.917384 |
| 1500 | 0.965500 | 0.912245 |
| 2000 | 0.931600 | 0.908042 |
| 2500 | 0.972800 | 0.905064 |
| 3000 | 1.001000 | 0.903870 |
| 3500 | 0.951500 | 0.903118 |

```

No files have been modified since last commit. Skipping to prevent empty commit.
WARNING:huggingface_hub.hf_api.No files have been modified since last commit. Skipping to prevent empty commit.
-----
TypeError                                Traceback (most recent call last)
/tmp/ipython-input-12-4272791645.py in <cell line: 0>()
      2 trainer.train()
      3 trainer.save_model("checkpoint_final")
----> 4 trainer.push_to_hub(commit_message="QLoRA finetuning complete", readme=readme)

/usr/local/lib/python3.11/dist-packages/transformers/trainer.py in push_to_hub(self, commit_message, blocking, token, revision,
**kwargs)
    4855         kwargs["tags"].append(model_tag)
    4856
-> 4857         self.create_model_card(model_name=model_name, **kwargs)
    4858
    4859         # Wait for the current upload to be finished.

TypeError: SFTTrainer.create_model_card() got an unexpected keyword argument 'readme'

```

```

trainer.push_to_hub(commit_message="QLoRA finetuning complete")

```

```

🔄 No files have been modified since last commit. Skipping to prevent empty commit.
WARNING:huggingface_hub.hf_api.No files have been modified since last commit. Skipping to prevent empty commit.
CommitInfo(commit_url='https://huggingface.co/gauri-sharan/phi2-alpaca-lora-4bit/commit/d56bbf659a1ff950869fd706c8ecfe1a0f64f9e9', commit_message='QLoRA finetuning complete', commit_description='',
oid='d56bbf659a1ff950869fd706c8ecfe1a0f64f9e9', pr_url=None, repo_url=RepoUrl('https://huggingface.co/gauri-sharan/phi2-alpaca-lora-4bit'), endpoint='https://huggingface.co', repo_type='model', repo_id='gauri-sharan/phi2-alpaca-lora-4bit'),
or revision=None, or num=None)

```

```

import json
from huggingface_hub import upload_file

model_id = "gauri-sharan/phi2-alpaca-lora-4bit"
config = {
    "architectures": ["AutoModelForCausalLM"],
    "model_type": "phi",
    "transformers_version": "4.45.1",
    "torch_dtype": "float16"
}


```

```

}
with open("config.json", "w") as f:
    json.dump(config, f, indent=2)

upload_file(
    path_or_fileobj="config.json",
    path_in_repo="config.json",
    repo_id=model_id,
    repo_type="model"
)

```

 CommitInfo(commit_url='https://huggingface.co/gauri-sharan/phi2-[alpaca-lora-4bit/commit/1fc92ce4f74cf5a3b50601c95d9ccb2568d19c3b](#)', commit_message='Upload config.json with huggingface_hub', commit_description='', oid='1fc92ce4f74cf5a3b50601c95d9ccb2568d19c3b', pr_url=None, repo_url=RepoUrl('https://huggingface.co/gauri-sharan/phi2-[alpaca-lora-4bit](#)', endpoint='https://huggingface.co', repo_type='model', repo_id='gauri-sharan/phi2-[alpaca-lora-4bit](#)'), pr_revision=None, pr_num=None)

```

from transformers import AutoModelForCausalLM
model = AutoModelForCausalLM.from_pretrained("gauri-sharan/phi2-alpaca-lora-4bit")

```

 Fetching 2 files: 100% 2/2 [00:00<00:00, 110.81it/s]

Loading checkpoint shards: 100% 2/2 [00:33<00:00, 14.22s/it]

adapter_model.safetensors: 100% 10.5M/10.5M [00:00<00:00, 24.5MB/s]

```

from ipywidgets import IntProgress
from IPython.display import display

# dummy widget to force widget infra to register state
f = IntProgress(min=0, max=10)
display(f)
f.value = 7

```



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