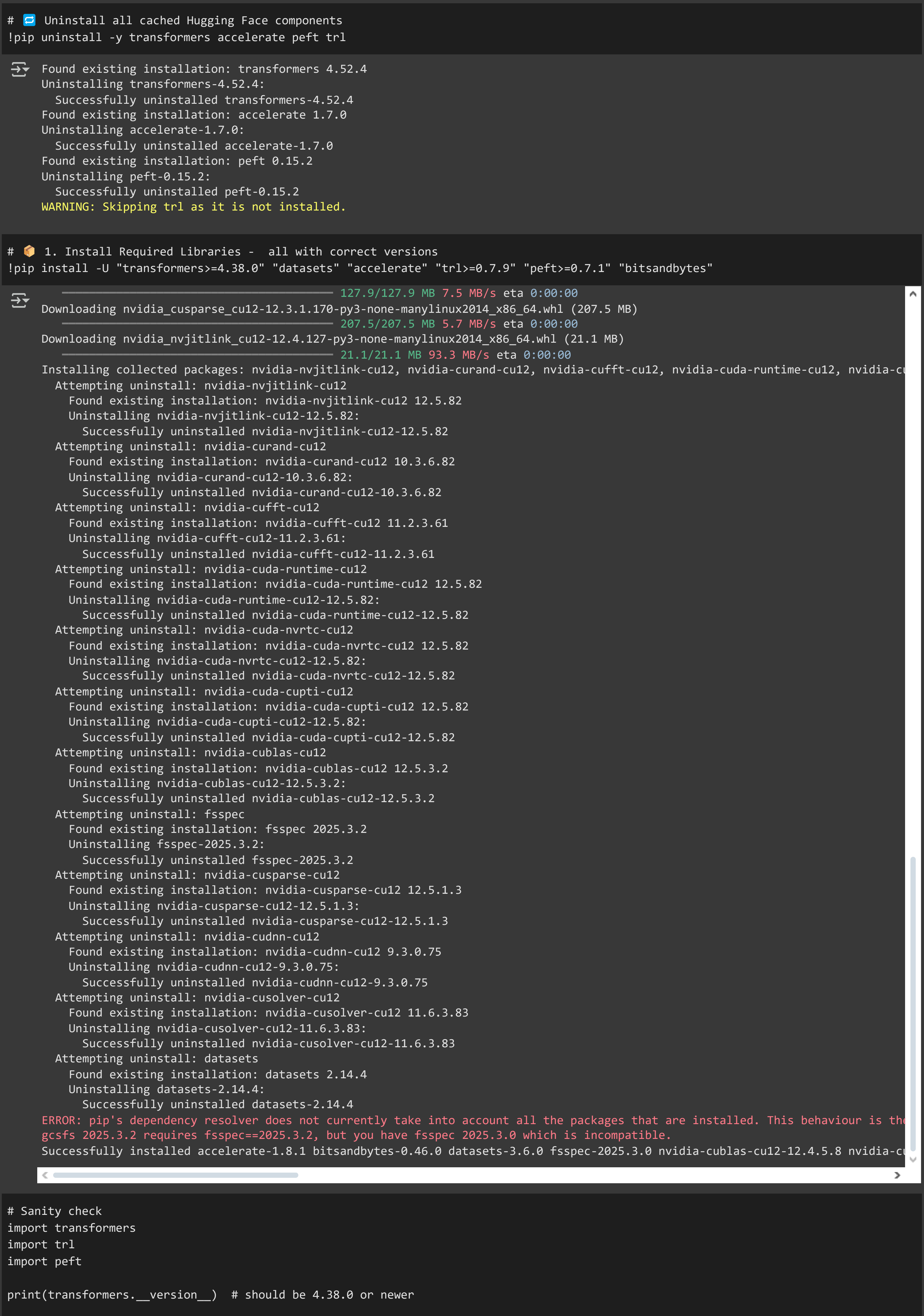


```
# 🔁 Uninstall all cached Hugging Face components
!pip uninstall -y transformers accelerate peft trl
```

```
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"
```

```
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 127.9/127.9 MB 7.5 MB/s eta 0:00:00
Downloading nvidia_cusparse_cu12-12.3.1.170-py3-none-manylinux2014_x86_64.whl (207.5 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 207.5/207.5 MB 5.7 MB/s eta 0:00:00
Downloading nvidia_nvjitlink_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl (21.1 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 21.1/21.1 MB 93.3 MB/s eta 0:00:00
Installing collected packages: nvidia-nvjitlink-cu12, nvidia-curand-cu12, nvidia-cufft-cu12, nvidia-cuda-runtime-cu12, nvidia-cu
  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: 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-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: datasets
    Found existing installation: datasets 2.14.4
    Uninstalling datasets-2.14.4:
      Successfully uninstalled datasets-2.14.4
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is the
gcsfs 2025.3.2 requires fsspec==2025.3.2, but you have fsspec 2025.3.0 which is incompatible.
Successfully installed accelerate-1.8.1 bitsandbytes-0.46.0 datasets-3.6.0 fsspec-2025.3.0 nvidia-cublas-cu12-12.4.5.8 nvidia-cu
```

```
# 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
```

```
4.52.4
0.19.0
0.15.2
```

```
# 🔐 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)
```

```
# 🧠 Step 2: Setup

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

# Tokenizer + 4-bit config
model_id = "microsoft/phi-2"
tokenizer = AutoTokenizer.from_pretrained(model_id)
tokenizer.pad_token = tokenizer.eos_token

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)
```

Copy a token from your Hugging Face tokens page and paste it below.

Immediately click login after copying your token or it might be stored in plain text in this notebook file.

Token:

□ Add token as git credential?

Login

Pro Tip: If you don't already have one, you can create a dedicated 'notebooks' token with 'write' access, that you can then easily reuse for all notebooks.

```
tokenizer_config.json: 100%  7.34k/7.34k [00:00<00:00, 160kB/s]
vocab.json: 100%  798k/798k [00:00<00:00, 5.39MB/s]
merges.txt: 100%  456k/456k [00:00<00:00, 5.64MB/s]
tokenizer.json: 100%  2.11M/2.11M [00:00<00:00, 13.3MB/s]
added_tokens.json: 100%  1.08k/1.08k [00:00<00:00, 28.1kB/s]
special_tokens_map.json: 100%  99.0/99.0 [00:00<00:00, 3.48kB/s]
config.json: 100%  735/735 [00:00<00:00, 21.5kB/s]
model.safetensors.index.json: 100%  35.7k/35.7k [00:00<00:00, 907kB/s]
Fetching 2 files: 100%  2/2 [00:56<00:00, 56.06s/it]
model-00001-of-00002.safetensors: 100%  5.00G/5.00G [00:55<00:00, 202MB/s]
model-00002-of-00002.safetensors: 100%  564M/564M [00:10<00:00, 37.5MB/s]
Loading checkpoint shards: 100%  2/2 [00:33<00:00, 14.34s/it]
generation_config.json: 100%  124/124 [00:00<00:00, 11.7kB/s]
```

```
# 📚 Step 3: 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]
```

```
# ⚙️ Step 4: 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 is1. Turn it to False automatically.
```

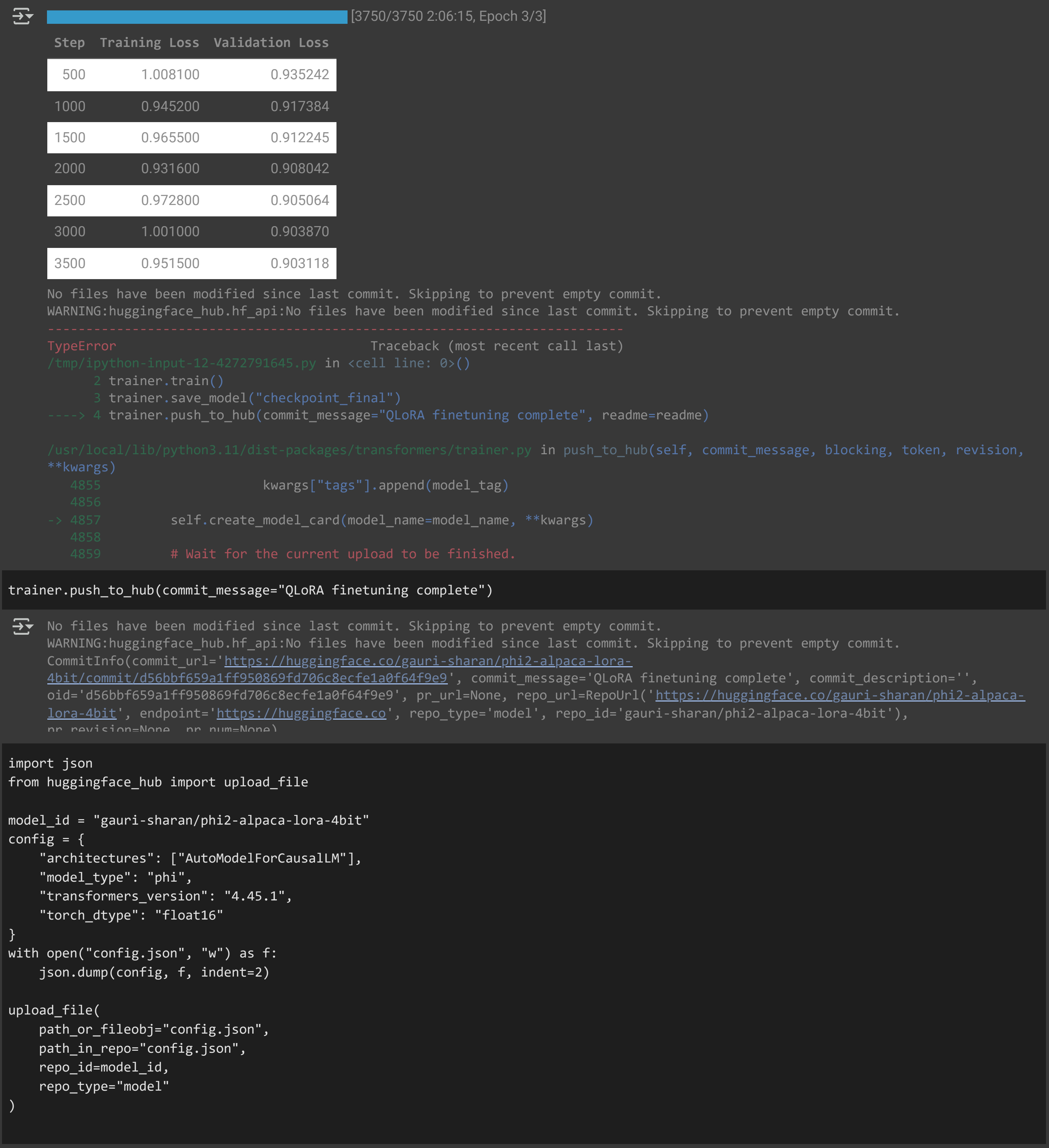
```
# 🏗️ Step 5: Initialize Trainer
trainer = SFTTrainer(
    model=model,
    args=sft_cfg,
    train_dataset=ds["train"],
    eval_dataset=ds["test"],
    processing_class=tokenizer,
    formatting_func=formatting_func
)
```

```
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](ht

## 🧠 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)
```

```
# 🚀 Step 6: 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.
```

```python
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'),
pr_revision=None, pr_num=None)
```

```python
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"
)
```