

## ✓ Install packages and libraries

```
!pip install trl
!pip install -U bitsandbytes
!pip install datasets
```

 [Show hidden output](#)

```
import warnings
warnings.filterwarnings("ignore")

import pandas as pd
import torch
from transformers import AutoTokenizer, AutoModelForCausalLM, BitsAndBytesConfig, pipeline, TrainingArguments, Trainer
from peft import AutoPeftModelForCausalLM, LoraConfig, prepare_model_for_kbit_training, get_peft_model
from trl import SFTTrainer
from google.colab import drive
from datasets import Dataset
```

+ Code

+ Text

## ✓ Use pre-trained GPT2 model from OpenAI

```
model_name = "openai-community/gpt2"

tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForCausalLM.from_pretrained(model_name)
```

## ✓ Load Kaggle dataset


Link to dataset: <https://www.kaggle.com/datasets/divyanshu2000/doctor-healthcare-100k>

```
import os

current_dir = os.getcwd()
file_dir = os.path.join(current_dir, "Doctor-HealthCare-100k.csv")
data = pd.read_csv(file_dir)
```


## ✓ Healthcare dataset has 112,156 rows and no null values

```
data.head()
```



	instruction	input	output
0	If you are a doctor, please answer the medical...	I woke up this morning feeling the whole room ...	Hi, Thank you for posting your query. The most...
1	If you are a doctor, please answer the medical...	My baby has been pooing 5-6 times a day for a ...	Hi... Thank you for consulting in Chat Doctor....
2	If you are a doctor, please answer the medical...	Hello, My husband is taking Oxycodone due to a...	Hello, and I hope I can help you today.First, ...
3	If you are a doctor, please answer the medical...	lump under left nipple and stomach pain (male)...	Hi. You have two different problems. The lump ...
4	If you are a doctor, please answer the medical...	I have a 5 month old baby who is very congeste...	Thank you for using Chat Doctor. I would sugge...

```
data.info()
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 112156 entries, 0 to 112155
Data columns (total 3 columns):
#   Column      Non-Null Count  Dtype
---  -
0   instruction  112156 non-null object
1   input        112156 non-null object
2   output       112156 non-null object
dtypes: object(3)
memory usage: 2.6+ MB
```

```
data.shape
```

```
(112156, 3)
```

```
data.isnull().sum()
```

```
0
instruction 0
input       0
output      0
```

```
dtype: int64
```

## ✓ Rename columns for clarity

```
data.rename(columns={"input":"question", "output":"response"}, inplace=True)
data.head()
```

```
0      If you are a doctor, please answer the medical...  I woke up this morning feeling the whole room ...  Hi, Thank you for posting your query. The most...
1      If you are a doctor, please answer the medical...  My baby has been pooing 5-6 times a day for a ...  Hi... Thank you for consulting in Chat Doctor....
2      If you are a doctor, please answer the medical...  Hello, My husband is taking Oxycodone due to a...  Hello, and I hope I can help you today.First, ...
3      If you are a doctor, please answer the medical...  lump under left nipple and stomach pain (male)...  HI. You have two different problems. The lump ...
4      If you are a doctor, please answer the medical...  I have a 5 month old baby who is very congeste...  Thank you for using Chat Doctor. I would sugge...
```

## ✓ instruction column has the same task for every entry

```
data["instruction"].nunique()
```

```
1
```

## ✓ A quick look at the first index's instruction, question, and response

```
data["instruction"].iloc[0]
```

```
'If you are a doctor, please answer the medical questions based on the patient's description.'
```

```
data["question"].iloc[0]
```

```
'I woke up this morning feeling the whole room is spinning when i was sitting down. I went to the bathroom walking unsteadily, as i tried to focus i feel nauseous. I try to vomit but it wont come out.. After taking panadol and sleep for few hours, i still feel the same.. By the way, if i lay down or sit down, my head do not spin, only when i want to move around then i feel the whole world is spinning... And it is normal stomach discomfort at the same time? Earlier after i relieved myself. the spinning lessen so i am not sure whether
```

```
data["response"].iloc[0]
```

```
'Hi, Thank you for posting your query. The most likely cause for your symptoms is benign paroxysmal positional vertigo (BPPV), a type of peripheral vertigo. In this condition, the most common symptom is dizziness or giddiness, which is made worse with movements. Accompanying nausea and vomiting are common. The condition is due to problem in the ear, and improves in a few days on own. Betahistine tablets would help relieve your symptoms. Doing vestibular rehabilitation or adaptation exercises would prevent the recurrence of these symptoms
```

## ✓ Create a prompt to structure the given text within a function

```
def create_prompt(data):
```

```
    prompt = f"""
```

```
    Below is the same instruction for every task. The patient will ask a question as the input, and a doctor is requested to provide a response
```

```
    """ Instruction:
```

```
{data["instruction"]}

### Question:
{data["question"]}

### Response:
{data["response"]}
"""

return prompt
```

## ✓ Generate text for the first few indices with the pre-trained model

```
pipe = pipeline(task='text-generation', model=model, tokenizer=tokenizer, truncation=True, max_length=1024)
result = pipe(create_prompt(data.iloc[0]))
print(result[0]['generated_text'])
```

🔄 Device set to use cuda:0  
Setting `pad\_token\_id` to `eos\_token\_id`:50256 for open-end generation.

Below is the same instruction for every task. The patient will ask a question as the input, and a doctor is requested to provide a response.

```
### Instruction:
If you are a doctor, please answer the medical questions based on the patient's description.
```

```
### Question:
I woke up this morning feeling the whole room is spinning when i was sitting down. I went to the bathroom walking unsteadily, as i tri
```

```
### Response:
Hi, Thank you for posting your query. The most likely cause for your symptoms is benign paroxysmal positional vertigo (BPPV), a type c
A Question that has been addressed:
Thanks. This article is an interview with Dr. Shirer which is part of a series. Clicking on the "Questions" button may reveal informat
```

E-mail: [info@medicalmagazine.org](mailto:info@medicalmagazine.org)...

A question that has been addressed:

Hi Dr. My name is Ashok, and i am a physician and a retired American Army med platoon commander. Your question about the presence of Par

Question: Did you know that there's an issue of paroxysmal paroxysmal vestibular syndrome (PPDS)?

Dr. Dr

```
pipe = pipeline(task='text-generation', model=model, tokenizer=tokenizer, truncation=True, max_length=1024)
result = pipe(create_prompt(data.iloc[1]))
print(result[0]['generated_text'])
```

🔄 Device set to use cuda:0  
Setting `pad\_token\_id` to `eos\_token\_id`:50256 for open-end generation.

Below is the same instruction for every task. The patient will ask a question as the input, and a doctor is requested to provide a response.

```
### Instruction:
If you are a doctor, please answer the medical questions based on the patient's description.
```

```
### Question:
My baby has been pooing 5-6 times a day for a week. In the last few days it has increased to 7 and they are very watery with green str
```

```
### Response:
Hi... Thank you for consulting in Chat Doctor. It seems your kid is having viral diarrhea. Once it starts it will take 5-7 days to com
### Response: Hi Dr. My baby has been pooing for most of our 3 years of life, only pooping often. It was my husband that started wit
```

Dr. Michael J. Johnson

I'm a New Orleans native whose job as a medical provider is caring for people with severe diarrhea.

Most of our clients have mild and moderate to severe dry and swollen or discolored stools (norto sintro cristacea, septica pustularis).

In addition, there have been 6 major bacterial infections, some of which I can attest to to 3 of which I have seen in this relationship

I have had a lot of people ask me why I don't vaccinate their children. I would like a vaccine to help to prevent infection. But I don't

I'm very aware of the need for a vaccine program, and I am extremely patient with anyone that is taking the virus to help protect against it. As of 1st April, 2012 I have had an outbreak of severe and long-lived respiratory illnesses that have affected about 5,000 people as I suspect I have no experience with a "water source" (such as the street, or even the hospital if the situation is like mine.)

I've had 5 people over the years that I've heard that their children have been poisoned. All I can do is just watch...

I've had 5 kids (3 years old) that were poisoned by using any means imaginable. I have seen no signs of this, nothing to worry to be safe. It's a serious but non-accidental disease. But it seems that not only has this happened, but in my own private life as well.

But still, I'm concerned for my child's well being, and for the well being of my child and staff all across the United States. And I have

```
pipe = pipeline(task='text-generation', model=model, tokenizer=tokenizer, truncation=True, max_length=1024)
result = pipe(create_prompt(data.iloc[2]))
print(result[0]['generated_text'])
```

Device set to use cuda:0  
Setting `pad\_token\_id` to `eos\_token\_id`:50256 for open-end generation.

Below is the same instruction for every task. The patient will ask a question as the input, and a doctor is requested to provide a response.

```
### Instruction:
If you are a doctor, please answer the medical questions based on the patient's description.
```

```
### Question:
Hello, My husband is taking Oxycodone due to a broken leg/surgery. He has been taking this pain medication for one month. We are trying
```

```
### Response:
Hello, and I hope I can help you today. First, there is no medication that can be taken by the father that has any way to get into your system.
It feels okay if you don't believe it. I am still going to give you the opportunity to find out what you need. If you do now you can
```

I hope that I could answer your questions on this very important topic once I get more time and time and time I will. If it sounds cool to you, please let me know. Thank you in advance for stopping by my site and checking out my work. If you have any questions there, feel free to e-mail me at: [uk.st@protonmail.com](mailto:uk.st@protonmail.com)

Cheers.

-Stent

## ✓ Finetune the GPT2 model

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

model = AutoModelForCausalLM.from_pretrained(
    model_name,
    quantization_config=bnb_config,
    use_cache=False,
    use_flash_attention_2=False,
    device_map="auto",
    torch_dtype=torch.float16
)

model.config.pretraining_tp = 1

# Load tokenizer
tokenizer = AutoTokenizer.from_pretrained(model_name)
tokenizer.pad_token = tokenizer.eos_token
tokenizer.padding_side = "right"

peft_config = LoraConfig(
    lora_alpha=30,
    lora_dropout=0.2,
    r=16,
    bias="none",
```

```

        task_type="CAUSAL_LM",
    )

model = prepare_model_for_kbit_training(model)
model = get_peft_model(model, peft_config)

```

## ✓ Save model in Google Colab directory

```

drive.mount('/content/drive')

output_dir = '/content/drive/My Drive/Colab Notebooks/Projects/Finetuned_GPT_Text_Gen_Model_Healthcare'

🔗 Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).

args = TrainingArguments(
    output_dir=output_dir,
    per_device_train_batch_size=10,
    gradient_accumulation_steps=1,
    gradient_checkpointing=True,
    optim="adafactor",
    logging_steps=10,
    save_strategy="epoch",
    learning_rate=1e-5,
    num_train_epochs=3,
    fp16=True,
    max_grad_norm=0.3,
    warmup_ratio=0.03,
    lr_scheduler_type="constant",
    disable_tqdm=False,
    report_to="none"
)

```

## ✓ Reformat Kaggle dataset and prompt function to fit SFTTrainer formatting

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

## ✓ Return prompt for individual entries with brackets

```

def formatting_func(data):

    prompt = f"""
    Below is the same instruction for every task. The patient will ask a question as the input, and a doctor is requested to provide a response.

    ### Instruction:
    {data["instruction"]}

    ### Question:
    {data["question"]}

    ### Response:
    {data["response"]}
    """

    return [prompt]

trainer = SFTTrainer(
    model=model,
    train_dataset=train_dataset,
    peft_config=peft_config,
    tokenizer=tokenizer,
    formatting_func=formatting_func,
    args=args
)

```

🔗 Map: 100%

112156/112156 [01:59<00:00, 1066.61 examples/s]

```
os.environ["WANDB_DISABLED"] = "true"
```

## ✓ Train the finetuned model

Loss is decent for only 3 epochs and a limited batch size of 10.

```
trainer.train()
```

 [36/36 00:44, Epoch 3/3]

Step	Training Loss
10	0.318000
20	0.318500
30	0.312000

```
TrainOutput(global_step=36, training_loss=0.3145848744445377, metrics={'train_runtime': 46.019, 'train_samples_per_second': 7.367, 'train_steps_per_second': 0.782, 'total_flos': 178384492560384.0, 'train_loss': 0.3145848744445377, 'epoch': 3.0})
```

```
trainer.save_model(output_dir)
```

## ✓ Test new model and tokenizer with finetuned parameters

```
model = AutoModelForCausalLM.from_pretrained(
    args.output_dir,
    quantization_config=bnb_config,
    use_cache=False,
    use_flash_attention_2=False,
    device_map="auto",
    torch_dtype=torch.float16
)
```

```
model.config.pretraining_tp = 1
```

```
tokenizer = AutoTokenizer.from_pretrained(args.output_dir)
```

### ✓ Generated text for first entry mentions BPPV, and seems more related to the prompt than the pre-trained generated text.

```
pipe = pipeline(task='text-generation', model=model, tokenizer=tokenizer, truncation=True, max_length=1024)
result = pipe(create_prompt(data.iloc[0]))
print(result[0]['generated_text'])
```

 Device set to use cuda:0

Below is the same instruction for every task. The patient will ask a question as the input, and a doctor is requested to provide a response.

### Instruction:

If you are a doctor, please answer the medical questions based on the patient's description.

### Question:

I woke up this morning feeling the whole room is spinning when i was sitting down. I went to the bathroom walking unsteadily, as i tried to get up.


### Response:

Hi, Thank you for posting your query. The most likely cause for your symptoms is benign paroxysmal positional vertigo (BPPV), a type of inner ear disorder that causes brief episodes of intense dizziness or vertigo.

# Question: Hi Dr. Fenn: It is important for us to get rid of BPPV and the BSPV. What would you advise us to do? There is an old adage that says "BPPV is a curse, but it can be cured".

### ✓ The generated text mentions a baby is the patient along with urinary test results, which align with the prompt.

```
pipe = pipeline(task='text-generation', model=model, tokenizer=tokenizer, truncation=True, max_length=1024)
result = pipe(create_prompt(data.iloc[1]))
print(result[0]['generated_text'])
```

 Device set to use cuda:0

Below is the same instruction for every task. The patient will ask a question as the input, and a doctor is requested to provide a res

### Instruction:

If you are a doctor, please answer the medical questions based on the patient's description.

### Question:

My baby has been pooping 5-6 times a day for a week. In the last few days it has increased to 7 and they are very watery with green str

### Response:

Hi... Thank you for consulting in Chat Doctor. It seems your kid is having viral diarrhea. Once it starts it will take 5-7 days to com

► Reply All ► Thank you,

Am I able to work on you? ► No, the patient is quite weak. So if a medicine worked really well for you, then maybe use my advice. It w  
Cheryl will check your doctor's notes before doing further tests on you. ► The patient is currently under 18. ► It does not appear th

▫ I do not know whether you understand now or not. ► You would have a hard time convincing, if you didn't understand. ► Your doctor

Results mentions Oxycodone and concerns about its impact on the baby. However, it implies the mother is taking Oxycodone as opposed to the father from the prompt's information.

```
pipe = pipeline(task='text-generation', model=model, tokenizer=tokenizer, truncation=True, max_length=1024)
result = pipe(create_prompt(data.iloc[2]))
print(result[0]['generated_text'])
```

🔗 device set to use cuda:0

Below is the same instruction for every task. The patient will ask a question as the input, and a doctor is requested to provide a res

### Instruction:

If you are a doctor, please answer the medical questions based on the patient's description.

### Question:

Hello, My husband is taking Oxycodone due to a broken leg/surgery. He has been taking this pain medication for one month. We are tryin

### Response:

Hello, and I hope I can help you today. First, there is no medication that can be taken by the father that has any way to get into your  
.

f you do not know what Oxycodone does for the baby, try to remember its name and its role on the brain in pregnancy.

f you have any questions or comments or if you have taken some medications that have not yet been prescribed as a medication, please le  
hank you again for coming here.

s you can see in the pictures, here is an instruction to your doctor for each of the 10 tasks:

) Question: Hello, My husband is on this, since he is doing a lot of pain meds. I am wondering if it is legal to take a medication with  
his page has been updated to include more information.

or example, to give it that meaning, I have been taking a little more pain medication to try and decrease my pain to help my son recover  
would also ask my patients if they are making any use of the medication without it.

still don't know why I take so many medications. This is just me, and I will see about this soon.

lease help by donating!

elp by donating!

elp by donating!

elp by donating!

elp by donating!

his page has been updated to include more information.

or example, I have been taking a little more pain medication to try and decrease my pain to help my son recover from it. However, this  
would also ask my patients if they are making any use of the medication without it. If some of the children have pain or are sufferin

lease help by donating!

elp by donating!

## Conclusion

Unfortunately, the number of epochs and batch size is restricting due to a limited GPU capacity. However, the generated text does do slightly better with the finetuned model than compared to the pretrained model.

Ideally, more epochs, bigger batch sizes, playing around with an evaluation dataset, and more complex hyperparameter tuning and function could contribute to more accurately generated text.