**Fine-Tuning LLMs with Hugging Face**

**Step 1: Installing and Importing the Libraries**

Install the necessary libraries:

bash

Copy code

!pip install -q accelerate==0.21.0 peft==0.4.0 bitsandbytes==0.40.2 transformers==4.31.0 trl==0.4.7

!pip install huggingface\_hub

Import the libraries:

python

Copy code

import torch

from trl import SFTTrainer

from peft import LoraConfig

from datasets import load\_dataset

from transformers import (AutoModelForCausalLM, AutoTokenizer, BitsAndBytesConfig, TrainingArguments, pipeline)

**Step 2: Loading the Model**

Load the pre-trained model with quantization configuration:

python

Copy code

llama\_model = AutoModelForCausalLM.from\_pretrained(

pretrained\_model\_name\_or\_path = "aboonaji/llama2finetune-v2",

quantization\_config = BitsAndBytesConfig(

load\_in\_4bit = True,

bnb\_4bit\_compute\_dtype = getattr(torch, "float16"),

bnb\_4bit\_quant\_type = "nf4"

)

)

llama\_model.config.use\_cache = False

llama\_model.config.pretraining\_tp = 1

**Note**: Authentication may be required for Hugging Face Hub.

**Step 3: Loading the Tokenizer**

Load the tokenizer:

python

Copy code

llama\_tokenizer = AutoTokenizer.from\_pretrained(

pretrained\_model\_name\_or\_path = "aboonaji/llama2finetune-v2",

trust\_remote\_code = True

)

llama\_tokenizer.pad\_token = llama\_tokenizer.eos\_token

llama\_tokenizer.padding\_side = "right"

**Step 4: Setting the Training Arguments**

Define the training arguments:

python

Copy code

training\_arguments = TrainingArguments(

output\_dir = "./results",

per\_device\_train\_batch\_size = 4,

max\_steps = 100

)

**Step 5: Creating the Supervised Fine-Tuning Trainer**

Create the trainer:

python

Copy code

llama\_sft\_trainer = SFTTrainer(

model = llama\_model,

args = training\_arguments,

train\_dataset = load\_dataset(path = "aboonaji/wiki\_medical\_terms\_llam2\_format", split = "train"),

tokenizer = llama\_tokenizer,

peft\_config = LoraConfig(

task\_type = "CAUSAL\_LM",

r = 64,

lora\_alpha = 16,

lora\_dropout = 0.1

),

dataset\_text\_field = "text"

)

**Step 6: Training the Model**

Start training:

python

Copy code

llama\_sft\_trainer.train()

**Step 7: Chatting with the Model**

Use the model for generating text:

python

Copy code

user\_prompt = "Please tell me about Bursitis"

text\_generation\_pipeline = pipeline(

task = "text-generation",

model = llama\_model,

tokenizer = llama\_tokenizer,

max\_length = 300

)

model\_answer = text\_generation\_pipeline(f"[INST] {user\_prompt} [/INST]")

print(model\_answer[0]['generated\_text'])

**Note**: For memory-efficient attention, consider installing Xformers:

bash

Copy code

pip install xformers

**Example Output**

markdown

Copy code

<s>[INST] Please tell me about Bursitis [/INST] Bursitis is a condition where the bursae, small fluid-filled sacs that cushion and reduce friction between tendons, ligaments, and bones, become inflamed or irritated. Unterscheidung between bursitis and tendinitis is often difficult because the symptoms are similar.

Bursitis can occur in any bursa in the body, but it is most common in the shoulder, elbow, hip, and knee. The symptoms of bursitis can vary depending on the location of the affected bursa, but they typically include:

\* Pain or tenderness in the affected area

\* Swelling or redness in the affected area

\* Limited mobility or stiffness in the affected joint

\* Warmth or heat in the affected area

\* In severe cases, fever or chills

Bursitis can be caused by a variety of factors, including:

\* Overuse or repetitive motion of a joint

\* Trauma or injury to the affected area

\* Poor posture or body mechanics

\* Infection or infection of the bursa

\* Crystal-induced bursitis (e.g., gout)

\* Rheumatoid arthritis or other inflammatory conditions

\* Injection or need