```
!pip install unsloth
!pip install --force-reinstall --no-cache-dir --no-deps git+https://github.com/unslothai/unsloth.git
```

```
from google.colab import userdata
hf_token = userdata.get('Hugging_face_new')
```

```
from unsloth import FastLanguageModel

max_seq_length = 2048
dtype = None
load_in_4bit = True
```

```
model, tokenizer = FastLanguageModel.from_pretrained(
    model_name = "unsloth/DeepSeek-R1-Distill-Llama-8B",
    max_seq_length = max_seq_length,
    dtype = dtype,
    load_in_4bit = load_in_4bit,
    token = hf_token,
)
```

```
<ipython-input-6-326b3facf0a4>:1: UserWarning: WARNING: Unsloth should be imported before trl, transformers to ensure all optimizations

Please restructure your imports with 'import unsloth' at the top of your file.
  from unsloth import FastLanguageModel
Unsloth: Failed to patch Gemma3ForConditionalGeneration.
🦥 Unsloth Zoo will now patch everything to make training faster!
==((====))==  Unsloth 2025.3.19: Fast Llama patching. Transformers: 4.51.1.
   \\   /|    NVIDIA A100-SXM4-40GB. Num GPUs = 1. Max memory: 39.557 GB. Platform: Linux.
O^O/ \_/ \    Torch: 2.6.0+cu124. CUDA: 8.0. CUDA Toolkit: 12.4. Triton: 3.2.0
\        /    Bfloat16 = TRUE. FA [Xformers = 0.0.29.post3. FA2 = False]
 "-____-"     Free license: http://github.com/unslothai/unsloth
Unsloth: Fast downloading is enabled - ignore downloading bars which are red colored!
model.safetensors: 100%  5.96G/5.96G [00:13<00:00, 607MB/s]
generation_config.json: 100%  236/236 [00:00<00:00, 29.5kB/s]
tokenizer_config.json: 100%  53.0k/53.0k [00:00<00:00, 3.13MB/s]
tokenizer.json: 100%  17.2M/17.2M [00:00<00:00, 169MB/s]
special_tokens_map.json: 100%  483/483 [00:00<00:00, 58.4kB/s]
```

```
train_prompt_style = """Below is an instruction that describes a task, paired with an input that provides further context.
Write a response that appropriately completes the request.
Before answering, think carefully about the question and create a step-by-step chain of thoughts to ensure a logical and accurate response.

### Instruction:
You are a medical expert with advanced knowledge in clinical reasoning, diagnostics, and treatment planning.
Please answer the following medical question.

### Question:
{}

### Response:
<think>
{}
</think>
{}"""
```

```
EOS_TOKEN = tokenizer.eos_token
```

```
def formatting_prompts_func(examples):
    inputs = examples["Question"]
    cots = examples["Complex_CoT"]
    outputs = examples["Response"]
    texts = []
    for input, cot, output in zip(inputs, cots, outputs):
        text = train_prompt_style.format(input, cot, output) + EOS_TOKEN
        texts.append(text)
    return {
```

```
        "text": texts,
    }
```

```
from datasets import load_dataset
dataset = load_dataset("FreedomIntelligence/medical-o1-reasoning-SFT","en", split = "train[0:1000]",trust_remote_code=True)
dataset = dataset.map(formatting_prompts_func, batched = True,)
dataset["text"][0]
```

```
README.md: 100%                                    1.65k/1.65k [00:00<00:00, 205kB/s]
medical_o1_sft.json: 100%                          74.1M/74.1M [00:00<00:00, 51.6MB/s]
Generating train split: 100%                       25371/25371 [00:01<00:00, 16331.30 examples/s]
Map: 100%                                          1000/1000 [00:00<00:00, 21831.46 examples/s]
'Below is an instruction that describes a task, paired with an input that provides further context. \nWrite a response that appropriate
ly completes the request. \nBefore answering, think carefully about the question and create a step-by-step chain of thoughts to ensure
a logical and accurate response.\n\n### Instruction:\nYou are a medical expert with advanced knowledge in clinical reasoning, diagnosti
cs, and treatment planning. \nPlease answer the following medical question. \n\n### Question:\nA 61-year-old woman with a long history
of involuntary urine loss during activities like coughing or sneezing but no leakage at night undergoes a gynecological exam and Q-tip
test. Based on these findings, what would cystometry most likely reveal about her residual volume and detrusor contractions?\n\n### Res
```

```
model = FastLanguageModel.get_peft_model(
    model,
    r=16,
    target_modules=[
        "q_proj",
        "k_proj",
        "v_proj",
        "o_proj",
        "gate_proj",
        "up_proj",
        "down_proj",
    ],
    lora_alpha=16,
    lora_dropout=0,
    bias="none",
    use_gradient_checkpointing="unsloth",  # True or "unsloth" for very long context
    random_state=3407,
    use_rslora=False,
    loftq_config=None,
)
```

```
Unsloth 2025.3.19 patched 32 layers with 32 QKV layers, 32 O layers and 32 MLP layers.
```

```
from trl import SFTTrainer
from transformers import TrainingArguments
from unsloth import is_bfloat16_supported

trainer = SFTTrainer(
    model=model,
    tokenizer=tokenizer,
    train_dataset=dataset,
    dataset_text_field="text",
    max_seq_length=max_seq_length,
    dataset_num_proc=2,
    args=TrainingArguments(
        per_device_train_batch_size=2,
        gradient_accumulation_steps=4,
        num_train_epochs = 1,
        warmup_steps=5,
        max_steps=60,
        learning_rate=2e-4,
        fp16=not is_bfloat16_supported(),
        bf16=is_bfloat16_supported(),
        logging_steps=10,
        optim="adamw_8bit",
        weight_decay=0.01,
        lr_scheduler_type="linear",
        seed=3407,
        output_dir = "/content/logs",
         report_to="none"
    ),
)
```

```
trainer_stats = trainer.train()
```

```
trainer_stats = trainer.train()
```

Show hidden output

```
question = "A 61-year-old woman with a long history of involuntary urine loss during activities like coughing or sneezing but no leakage at
```

```
FastLanguageModel.for_inference(model)  # Unsloth has 2x faster inference!
inputs = tokenizer([prompt_style.format(question, "")], return_tensors="pt").to("cuda")

outputs = model.generate(
    input_ids=inputs.input_ids,
    attention_mask=inputs.attention_mask,
    max_new_tokens=1200,
    use_cache=True,
)
response = tokenizer.batch_decode(outputs)
print(response[0].split("### Response:")[1])
```

```
<think>
Alright, let's think about this. We have a 61-year-old woman who's been dealing with involuntary urine loss whenever she coughs or sneez

But there's more to it. She's not having any leakage at night, which is interesting because urgency incontinence usually affects people

Okay, so when she's coughing or sneezing, her bladder is reacting quickly, which is typical of urgency incontinence. Now, let's consider

If we think about her symptoms and the findings from the Q-tip test, it's possible that there's an obstruction somewhere in her lower ur

Now, thinking about what cystometry would reveal, it's often used to assess bladder capacity and how it contracts. With urgency incontin

So, putting it all together, given her history and the Q-tip findings, it's likely that her cystometry would show she has a normal bladd
</think>
Based on the information provided, the cystometry would most likely reveal that this woman has a normal bladder capacity but exhibits sc
```

Start coding or generate with AI.