deepseek实验二：Unsloth微调DeepSeek-R1蒸馏模型 - 构建医疗专家模型

准备环境

下载unsloth

```
pip install unsloth
```

下载vllm

```
pip install vllm
```

下载pillow

```
pip install --upgrade pillow
```

下载trl

```
pip install trl
```

导入和初始配置

```
from unsloth import FastLanguageModel
import torch
max_seq_length = 2048 # Choose any! We auto support RoPE Scaling internally!
dtype = None # None for auto detection. Float16 for Tesla T4, V100, Bfloat16 for
load_in_4bit = True # Use 4bit quantization to reduce memory usage. Can be False

model, tokenizer = FastLanguageModel.from_pretrained(
    model_name = "./DeepSeek-R1-Distill-Qwen-1.5B",
    max_seq_length = max_seq_length,
    dtype = dtype,
    load_in_4bit = load_in_4bit,
```

```
    # token = "hf_...", # use one if using gated models like meta-llama/Llama-2-
)

🦥 Unsloth: Will patch your computer to enable 2x faster free finetuning.
🦥 Unsloth Zoo will now patch everything to make training faster!
INFO 02-13 09:39:40 __init__.py:190] Automatically detected platform cuda.
==((====))==  Unsloth 2025.2.5: Fast Qwen2 patching. Transformers: 4.48.3.
   \\   /|    GPU: NVIDIA GeForce RTX 4090. Max memory: 23.546 GB. Platform: Lin
O^O/ \_/ \    Torch: 2.5.1+cu124. CUDA: 8.9. CUDA Toolkit: 12.4. Triton: 3.1.0
\        /    Bfloat16 = TRUE. FA [Xformers = 0.0.28.post3. FA2 = False]
 "-____-"     Free Apache license: http://github.com/unslothai/unsloth
Unsloth: Fast downloading is enabled - ignore downloading bars which are red col
./DeepSeek-R1-Distill-Qwen-1.5B does not have a padding token! Will use pad_toke
```

提示词模板

```
prompt_style = """Below is an instruction that describes a task, paired with an
Write a response that appropriately completes the request.
Before answering, think carefully about the question and create a step-by-step c

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

### Question:
{}

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

微调前的推理

```
question = "一个患有急性阑尾炎的病人已经发病5天，腹痛稍有减轻但仍然发热，在体检时发现右下腹


FastLanguageModel.for_inference(model)
inputs = tokenizer([prompt_style.format(question, "")], return_tensors="pt").to(
```

```
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>
好，我现在要处理一个关于急性阑尾炎患者的临床问题。首先，病人已经发病5天，腹痛有轻微减轻，但发

首先，急性阑尾炎通常由感染性或非感染性的因素引起，比如乙型脑炎。患者发热可能与感染有关，但也可

接下来，我需要评估患者的体征状况。腹痛和发热，但没有明显的低烧，所以不能确定是否感染。包块可能

首先，考虑抗生素治疗。急性阑尾炎可能有细菌感染，抗生素治疗可能有助于缓解症状，但可能引起炎症反

同时，需要考虑包块处理。包块可能影响排液，需要药物治疗，如β受体阻滞剂，或者手术，如开腹或镜下

此外，患者是否需要进一步的治疗，比如抗生素联合β受体阻滞剂，或者手术。如果包块较大或形状不规则

最后，评估患者的预后。药物治疗可能缓解症状，但可能引起炎症反应。手术可能需要 longer recover

总结一下，处理步骤可能包括抗生素治疗，同时考虑包块处理，可能需要药物或手术。
</think>

对于一个患有急性阑尾炎的病人，患者已经发病5天，腹痛轻微减轻但仍有发热，体检时右下腹有压痛包块

准备数据集

医疗数据集https://huggingface.co/datasets/FreedomIntelligence/medical-o1-reasoning-SFT/将用于训练所选模型。

```
train_prompt_style = """Below is an instruction that describes a task, paired wi
Write a response that appropriately completes the request.
Before answering, think carefully about the question and create a step-by-step c

### Instruction:
You are a medical expert with advanced knowledge in clinical reasoning, diagnost
Please answer the following medical question.
```

```
### Question:
{}

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

在每个训练数据集条目末尾添加 EOS（序列结束）标记至关重要

```
EOS_TOKEN = tokenizer.eos_token  # Must add 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("./medical-o1-reasoning-SFT", 'zh', split = "train[0:500]
print(dataset.column_names)
```

```
Generating train split: 0 examples [00:00, ? examples/s]


['Question', 'Complex_CoT', 'Response']
```

为了使Ollama和llama.cpp像自定义ChatGPT聊天机器人一样运行，我们必须只有 2 列 - 一个instruction和 一个output列。我们需要将数据集转换为适当的结构。

```
dataset = dataset.map(formatting_prompts_func, batched = True)
dataset["text"][0]
```

```
Map:   0%|          | 0/500 [00:00<?, ? examples/s]
```

```
'Below is an instruction that describes a task, paired with an input that provid
```

训练模型

现在让我们使用 Huggingface TRL SFTTrainer。

```
model = FastLanguageModel.get_peft_model(
    model,
    r = 16, # Choose any number > 0 ! Suggested 8, 16, 32, 64, 128
    target_modules = ["q_proj", "k_proj", "v_proj", "o_proj",
                      "gate_proj", "up_proj", "down_proj",],
    lora_alpha = 16,
    lora_dropout = 0, # Supports any, but = 0 is optimized
    bias = "none",    # Supports any, but = "none" is optimized
    # [NEW] "unsloth" uses 30% less VRAM, fits 2x larger batch sizes!
    use_gradient_checkpointing = "unsloth", # True or "unsloth" for very long co
    random_state = 3407,
    use_rslora = False,  # We support rank stabilized LoRA
    loftq_config = None, # And LoftQ
)
```

```
Unsloth 2025.2.5 patched 28 layers with 28 QKV layers, 28 O layers and 28 MLP la
```

```
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,
    packing = False, # Can make training 5x faster for short sequences.
```

```
    args = TrainingArguments(
        per_device_train_batch_size = 2,
        gradient_accumulation_steps = 4,
        warmup_steps = 5,
        max_steps = 60,
        # num_train_epochs = 1, # For longer training runs!
        learning_rate = 2e-4,
        fp16 = not is_bfloat16_supported(),
        bf16 = is_bfloat16_supported(),
        logging_steps = 1,
        optim = "adamw_8bit",
        weight_decay = 0.01,
        lr_scheduler_type = "linear",
        seed = 3407,
        output_dir = "outputs",
        report_to = "none", # Use this for WandB etc
    ),
)
```

```
Map (num_proc=2):   0%|          | 0/500 [00:00<?, ? examples/s]
```

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

```
==((====))==  Unsloth - 2x faster free finetuning | Num GPUs = 1
   \\   /|    Num examples = 500 | Num Epochs = 1
O^O/ \_/ \    Batch size per device = 2 | Gradient Accumulation steps = 4
\        /    Total batch size = 8 | Total steps = 60
 "-____-"     Number of trainable parameters = 18,464,768
```

```
<div>

  <progress value='60' max='60' style='width:300px; height:20px; vertical-align:
  [60/60 01:21, Epoch 0/1]
</div>
<table border="1" class="dataframe">
```

Step Training Loss 1 2.913300 2 3.278500 3 3.003800 4 3.035400 5 2.532500 6 2.769800 7 3.066000 8 2.933600 9 2.928000 10 2.911500 11 2.601100 12 2.530900 13 2.529700 14 2.736300 15 2.349600 16 2.301700 17 2.714600 18 2.685500 19 2.321100 20 2.286600 21 2.194900 22 2.199000 23 2.242500 24 2.198800 25 2.464600 26 2.457300 27

2.336800 28 2.093200 29 2.422000 30 2.627400 31 2.384100 32 2.333400 33 2.288300 34 2.200900 35 2.377700 36 2.332500 37 2.373500 38 2.001200 39 2.431100 40 2.125000 41 2.425000 42 2.040500 43 2.297000 44 1.995700 45 2.391200 46 2.282700 47 2.379800 48 2.170800 49 2.170700 50 2.190800 51 2.019700 52 2.179700 53 2.073900 54 2.376400 55 2.308800 56 2.244400 57 2.010900 58 2.342400 59 2.070800 60 2.403900

微调后推理

我们再用同样的问题进行推理，看看有什么不同

```
print(question)
```

```
一个患有急性阑尾炎的病人已经发病5天，腹痛稍有减轻但仍然发热，在体检时发现右下腹有压痛的包块，
```

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

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>
嗯，病人已经发病5天了，腹痛稍微减轻了一点，但还是有一点发热。现在，我们看到右下腹有一个压痛包

首先，压痛包块可能是因为下腹部的炎症或者其他原因引起的。病人已经得了急性阑尾炎，所以下腹部的炎

接下来，我想了解一下，这种情况下，我们通常会考虑哪些处理方法呢？比如，先尝试一下冷敷，这样可以

不过，如果冷敷和止痛药都没能解决，那我们就得考虑更紧急的措施了。比如，如果病人还存在感染性包块

另外，病人有急性阑尾炎，所以下腹部的炎症可能会更严重。所以，我们可能需要考虑使用抗生素来帮助控

还有一点，病人有发热，这种情况通常需要更多的液体帮助，比如抗休克液，来帮助维持正常代谢。如果病
```

综合来看，我们首先应该尝试冷敷和止痛药，看看有没有缓解症状的效果。如果效果还不明显，再考虑手术

总之，根据这些考虑，我们可能会先试试冷敷，然后考虑止痛药和抗生素，最后再考虑手术。

</think>

根据病人的情况，右下腹压痛包块已经存在，结合急性阑尾炎的特征，建议进行冷敷处理。首先，建议使用