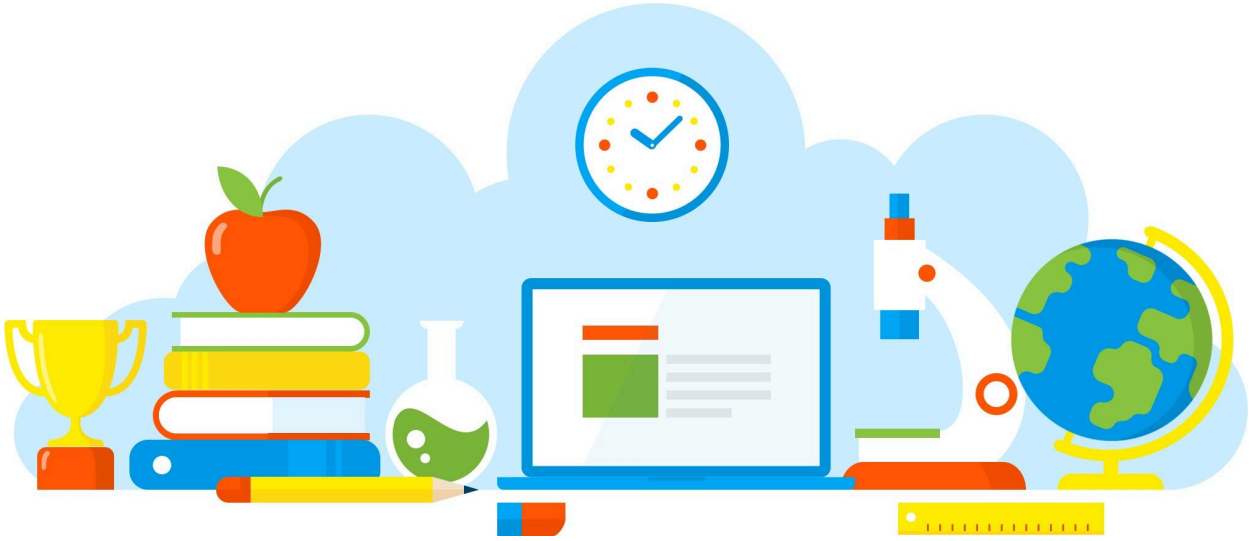

Fine-Tuning DeepSeek-R1 on Medical CoT Dataset

AI Doctor - Medical Reasoning with DeepSeek-R1

Intern Name: Muhammad Anis Faseel
Student ID: Arch-2507-0331
Contact: +923166461053

1. Objective

The objective of this project was to fine-tune a large language model on a medical dataset containing clinical reasoning examples. The aim was to enhance the model's ability to handle complex diagnostic questions using chain-of-thought (CoT) reasoning techniques.

---

2. Model & Dataset

- **Base Model:** `DeepSeek-R1-Distill-Llama-8B` from Hugging Face
- **Dataset:** `FreedomIntelligence/medical-o1-reasoning-SFT`
- **Samples Used:** 500 examples (training split)
- **Prompt Style:** Custom prompt with `<think>` tags for step-by-step reasoning

---

3. Tools & Libraries

- pip install torch==2.5.1 torchvision==0.20.1 --index-url https://download.pytorch.org/whl/cu121
- pip install triton==2.0.0 --force-reinstall
- pip install "unsloth[colab-new] @ git+https://github.com/unslothai/unsloth.git"
- pip install peft==0.10.0 trl==0.7.9 xformers==0.0.28.post3 accelerate bitsandbytes
- Unsloth (QLoRA + Flash Attention)
- Transformers (Hugging Face)

- Datasets (Hugging Face)
- PEFT & TRL for LoRA fine-tuning
- WANDB for experiment tracking
- PyTorch with CUDA 12.1

---

4. Fine-Tuning Flow

- **Model Loading:** 4-bit QLoRA model loaded using Unsloth API
- **Prompt Engineering:** Medical CoT examples framed using a templated prompt with `<think>` and `</think>` tags
- **LoRA Applied To:** `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`
- **Trainer Used:** `SFTTrainer` from `trl`
- **Training Arguments:**
  - Per Device Batch Size: 2
  - Accumulation Steps: 4
  - Max Steps: 60
  - Epochs: 1
  - Learning Rate: 2e-4
  - Mixed Precision: FP16/BF16 (auto-detected)

---

5. Inference Example

Sample inference after fine-tuning:

question = """A 59-year-old man presents with a fever..."""

FastLanguageModel.for_inference(model)

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)

print(tokenizer.decode(outputs[0]))
```

---

6. Results

- Fine-tuned model responded with accurate, coherent answers to complex medical queries.
- Efficient GPU usage due to 4-bit quantization
- WANDB tracking enabled for training visualization

---

7. Challenges

- **Triton Errors:** Resolved by pinning to version 2.0.0
- **Torch Compatibility:** Solved via correct installation using CUDA index URL
- **Runtime Instability:** Resolved by restarting Colab after installations

---

8. Conclusion

This project successfully showcases the power of LoRA fine-tuning using Unsloth for medical NLP tasks. The result is a competent medical reasoning AI agent capable of structured and logical diagnoses.

Credits

- Unsloth
- DeepSeek
- Medical CoT Dataset