Technical Report: LLaMA-2-7B-HealthGPT

🧠 Project Overview

LLaMA-2-7B-HealthGPT is a prototype fine-tuned version of Meta's LLaMA-2-7B model tailored for health-related question answering and reasoning, with a focus on **Sri Lanka's healthcare system**. The model was fine-tuned using **LoRA adapters** and **4-bit quantization (NF4)** for memory-efficient training and inference.

🧪 Objectives

- Build a domain-specific LLM capable of reasoning over Sri Lanka's health system policies and practices.
- Prototype efficient finetuning using limited synthetic data.
- Evaluate the potential for expansion into multilingual and multimodal (audio-based) learning.

🧾 Dataset

- **Type**: Synthetic, GPT-4-generated
- **Samples**: 10 Q&A pairs
- **Format**: Prompt-response with instruction formatting
- **System Prompt**:
  `Given a puzzle-like, reasoning-heavy question about Sri Lanka's healthcare system.`
- **Sample Question**:
  "How does Sri Lanka manage to provide free healthcare despite limited funding?"
- **Sample Answer**:
  "Sri Lanka leverages community health programs, prioritizes PHC, and relies on government subsidies..."

🔧 Model & Training Details

- **Base Model**: `NousResearch/llama-2-7b-chat-hf`
- **Adapter**: LoRA (Low-Rank Adaptation)
- **Quantization**: 4-bit (NF4 via `bitsandbytes`)
- **Batch Size**: 1 (gradient accumulation steps = 4)

- **Max Grad Norm**: 0.3
- **Learning Rate**: 2e-4
- **Gradient Checkpointing**: Enabled
- **Tokenizer**: Right-padding; EOS as pad token
- **Training Framework**: Hugging Face Transformers + TRL

⚠️ Issues Faced

- **GPU Memory Overflow**
  Training the model on Google Colab Pro with 15–16 GB GPU failed intermittently due to high memory usage.
  🔄 **Solution**: Restarted runtime, adjusted batch size and used 4-bit quantization with `bnb_4bit_use_double_quant=True`.

📈 Results

- The model successfully learned the format and produced domain-aware completions.
- Output reflected structured reasoning, adherence to prompt format, and fluent English.

✅ Sample Inference

Prompt:
“Explain how to prevent the spread of infectious diseases in a rural village.”

Response:
"To prevent infectious diseases, focus on hygiene education, clean water access, vaccinations, and early health worker intervention..."

🔮 Future Enhancements

🌐 Multilingual Support

- Expand training data to support **Sinhala**, **Tamil**, and **English** for inclusive access across Sri Lanka.

🔊 Audio-Based Training

- Introduce **speech-to-text** prompts and audio completion using pretrained **Whisper**, **Wav2Vec 2.0**, or **AudioLM** models for health assistant speech agents.

📚 Dataset Expansion

- Use real data from:
    - Sri Lankan Ministry of Health public datasets
    - WHO health reports
    - Transcribed health interviews and media in all three languages
- Scale to 10k+ examples using GPT-4-driven synthetic augmentation.

🤖 Deployment Use Cases

- Chatbots for remote health advice
- Patient triage and education systems
- Public health awareness tools

🗂️ Files

File	Description
`train.jsonl`	9 examples used for fine-tuning
`test.jsonl`	1 example used for evaluation
`llama-2-7b-custom/`	Fine-tuned model adapter
`llama-2-7b-custom-merged/`	Final merged model with base weights