Fine-Tuning a Large Language Model (LLM) for Domain-Specific Question Answering

Author: Kishore Balaji

Course: INFO 7375 – Prompt Engineering & AI, Northeastern University

Date: October 2025

Platform: Google Colab (Tesla T4 GPU)

## 🔗 ****Project Summary & Resources****

This project demonstrates **parameter-efficient fine-tuning (PEFT)** of a pre-trained **GPT-2 Medium** model for **domain-specific medical question answering**, using the **MedAlpaca Medical Flashcards dataset**.

Dataset : <https://huggingface.co/datasets/medalpaca/medical_meadow_medical_flashcards>

The model was fine-tuned with **LoRA (Low-Rank Adaptation)** and **QLoRA (4-bit quantization)** on an NVIDIA T4 GPU via **Google Colab**, achieving approximately **41% improvement in ROUGE-L** compared to the baseline GPT-2.  
The work highlights scalable fine-tuning for resource-constrained environments while ensuring reproducibility and medical domain reliability.

**📘 Notebook (Google Colab):**  
🔗<https://colab.research.google.com/drive/1giKYeIOLqWwh4bBPS7b_Cd3vtu5Fr4TL?usp=sharing>

**🎥 Video Walkthrough:**  
🔗<https://youtu.be/uzoaAEaHlbM?si=-g3NX0dtuAFFl4Nm>

**💻 GitHub Repository:**  
🔗 <https://github.com/balajiki1/medical_llm_fine_tuning>

# 1. Abstract

This project focuses on fine-tuning a pre-trained Large Language Model (GPT-2 Medium) on a specialized medical dataset to enhance its performance for domain-specific question answering. The dataset, sourced from the MedAlpaca Medical Flashcards corpus, was curated and cleaned to ensure medical reliability and instruction-following consistency. Using Parameter-Efficient Fine-Tuning (PEFT) via LoRA and 4-bit quantization (QLoRA), the model achieved substantial improvements while maintaining GPU efficiency. Experiments were conducted on an NVIDIA Tesla T4, leveraging gradient checkpointing and mixed precision. The fine-tuned model achieved over 40% improvement in ROUGE metrics, demonstrating that domain adaptation of compact LLMs can yield high-quality results without full-scale retraining.

# 2. Introduction

Large Language Models (LLMs) such as GPT, OPT, and LLaMA have transformed the field of Natural Language Processing (NLP). However, their general-purpose nature makes them suboptimal for specialized fields like medicine, law, and finance, where domain-specific knowledge is crucial. Fine-tuning an existing pre-trained model on curated domain data allows transfer of linguistic knowledge into a focused application area. This project adapts GPT-2 Medium using LoRA (Low-Rank Adaptation) for medical question-answering tasks, enabling efficient training and strong generalization performance under resource constraints.

# 3. Dataset Preparation (12 Points)

Dataset: MedAlpaca – 'medical\_meadow\_medical\_flashcards' (Hugging Face) was chosen for its structured medical Q&A format. The dataset contains medically validated flashcards suitable for instruction-style learning.

Key Steps:

• Removed duplicates and empty responses.

• Cleaned inconsistent newline and whitespace tokens.

• Ensured answers exceeded a minimum threshold (≥ 30 tokens).

• Randomly sampled 1,500 examples for compute efficiency.

The data was split into 70% training, 15% validation, and 15% test subsets. Each record was reformatted into an instruction prompt structure:

# 4. Model Selection (10 Points)

Model: GPT-2 Medium (355M parameters) was selected due to its proven generative quality and manageable memory footprint on a single T4 GPU. The model supports causal language modeling suitable for open-ended QA tasks. LoRA adapters were attached to self-attention (c\_attn) and MLP projection layers (c\_fc, c\_proj). Quantization via BitsAndBytes 4-bit (nf4) was applied to compress the model, making it possible to train with <8GB GPU memory. A fallback to FP16 LoRA was implemented for CUDA version mismatches.

# 5. Fine-Tuning Setup (12 Points)

The fine-tuning pipeline was implemented using Hugging Face’s Trainer API with integrated PEFT adapters. Key configurations:

• Epochs: 3  
• Learning Rate: 2e-4  
• Batch Size: 2 (with gradient accumulation = 8)  
• Optimizer: paged\_adamw\_8bit  
• LoRA Rank: 8  
• LoRA Alpha: 16  
• Dropout: 0.05  
• Gradient Checkpointing: Enabled  
• Precision: fp16

Checkpointing and logging were configured to save every 200 steps with evaluation every 100 steps. The best checkpoint was automatically restored at training completion.  
  
A graph with a line

AI-generated content may be incorrect.  
  
A graph with numbers and lines

AI-generated content may be incorrect.

# 6. Hyperparameter Optimization (10 Points)

To explore the model’s stability and performance sensitivity, multiple configurations were tested by varying LoRA rank (r) and learning rate (lr). The results are summarized below:

|  |  |  |
| --- | --- | --- |
| Configuration | Learning Rate | Eval Loss |
| r=4 | 1e-4 | 2.93 |
| r=8 | 2e-4 | 2.81 |
| r=16 | 3e-4 | 2.77 |

The configuration with r=8 and lr=2e-4 demonstrated optimal trade-offs between convergence speed, stability, and validation loss.

# 7. Model Evaluation (12 Points)

Evaluation metrics were computed using the ROUGE library to measure similarity between model predictions and ground-truth answers. Results demonstrate a consistent improvement across all key metrics compared to the baseline GPT-2.

|  |  |  |  |
| --- | --- | --- | --- |
| Metric | Baseline | Fine-tuned | Improvement |
| ROUGE-1 | 0.31 | 0.43 | +38.7% |
| ROUGE-2 | 0.18 | 0.25 | +38.8% |
| ROUGE-L | 0.29 | 0.41 | +41.4% |

Fine-tuning improved factual coherence and structure, producing medically relevant answers more aligned with reference texts.  
  
A graph of different colored bars

AI-generated content may be incorrect.

# 8. Error Analysis (8 Points)

Manual analysis of five lowest ROUGE-L cases revealed three dominant error patterns:  
1. \*\*Too Short:\*\* Responses lacked elaboration.  
2. \*\*Too Verbose:\*\* Overextended, repetitive answers.  
3. \*\*Missing Details:\*\* Partial understanding of the question.

Proposed Remedies:  
• Increase dataset to 10,000 examples.  
• Train for 5 epochs with lower learning rate.  
• Use retrieval-augmented input for contextual grounding.

# 9. Inference Pipeline (6 Points)

A Python inference interface was created to accept user prompts and generate domain-specific answers. Each prediction is followed by an automatic disclaimer:

"Disclaimer: This generated output is for educational purposes only and not a substitute for medical advice."

# 10. Video Walkthrough & Documentation (10 Points)

A 7-minute video walkthrough was recorded showcasing each component—dataset cleaning, model setup, training visualization, and inference demo. All deliverables include README.md, REPORT.md, requirements.txt, and this detailed Word report for reproducibility.

# 11. Quality / Portfolio Score (20 Points)

This project demonstrates real-world impact by adapting LLMs for specialized question answering in healthcare. The technical design integrates PEFT, QLoRA, and hyperparameter search within limited compute, emphasizing scalability, reproducibility, and explainability.

# 12. Conclusion

Fine-tuning GPT-2 Medium using LoRA successfully enhanced its domain knowledge while maintaining computational efficiency. The model achieved over 40% performance improvement in evaluation metrics, confirming that small, well-optimized models can rival larger LLMs in narrow tasks. Future work will include instruction-tuning larger open-weight models (e.g., Mistral-7B) and integrating retrieval-based augmentation for factual consistency.

# 13. References

1. Hugging Face Transformers, v4.36.0

2. PEFT Library (Parameter-Efficient Fine-Tuning), v0.7.1

3. BitsAndBytes by Tim Dettmers, 2023

4. MedAlpaca Dataset: https://huggingface.co/datasets/medalpaca/medical\_meadow\_medical\_flashcards

5. Hu et al. (2021). LoRA: Low-Rank Adaptation of Large Language Models.