LoRA vs Prompt Engineering: Trade-offs in Biomedical Question Answering with PubMedQA

ksheka2s pheddu2s

September 16, 2025

1. Introduction

Large Language Models (LLMs) such as LLaMA [4] have shown strong general-purpose reasoning ability, but their performance in specialized domains (e.g., biomedical literature) remains limited. Adapting LLMs to these domains is critical for applications such as medical question answering and literature review support.

Two promising adaptation strategies are:

(a) **Parameter-efficient fine-tuning with LoRA [3] (Low-Rank Adaptation):** Adds small trainable adapters to frozen LLM layers, enabling efficient domain adaptation with reduced computational cost.

(b) **Prompt engineering: [2]** Leverages pretrained models directly by carefully designing prompts (zero-shot or few-shot) without retraining.

This project systematically compares these approaches on the **PubMedQA** dataset, a benchmark for biomedical question answering. Our goal is to provide empirical evidence about the trade-offs between prompt engineering and LoRA fine-tuning, particularly in low-resource settings.

2. Research Hypotheses

- **H1 (Performance):** LoRA fine-tuning yields higher accuracy, F1 score, and BERTScore than prompt engineering as the number of training examples increases.
- **H2 (Stability):** Prompting exhibits higher variance, while LoRA produces more consistent results.
- **H3 (Efficiency):** Prompting is competitive in very low-resource regimes, but LoRA becomes more cost-effective.

3. Dataset

We selected PubMedQA [1] dataset :

- Task: Biomedical question answering with labels $\{$Yes, No, Maybe$\}$.
- Size:1,000 labeled examples.
- Subsampling: Create splits of 128, 512, and all available examples to simulate different data size.

4. Methodology

Our methodology involves preparing the dataset, applying two adaptation approaches (LoRA fine-tuning and prompt engineering).

- **Data Preparation:** We preprocess the PubMedQA dataset and create subsets containing 128, 512, and 1,000 examples for training, ensuring consistent evaluation across methods.
- **LoRA Fine-Tuning:**
  - Utilize LLaMA-2 (7B or smaller variant depending on compute resources) as the base model.
  - Insert low-rank adapters into the attention layers and train only these adapter parameters while freezing the backbone.
  - Experiment with two ranks, $r = 8$ and $r = 16$, across all subsets.
  - Evaluate the model's ability to answer biomedical questions based on the subset size.
- **Prompt Engineering:**
  - Design zero-shot and few-shot prompt templates tailored to the biomedical domain.
  - Evaluate LLaMA-2's performance on the same subsets without any parameter updates.

5. Evaluation

- Compare LoRA and prompt-based methods using metrics such as accuracy, F1 score, BERTScore, and confusion matrices.
- Perform error analysis to identify common failure modes and strengths of each method.

6. Contribution Plan

- **Member 1:** Responsible for LoRA fine-tuning, hyperparameter tuning, and running experiments.
- **Member 2:** Responsible for prompt engineering design, baseline implementation, and evaluation analysis.
- Both: Documentation, error analysis, and preparation for demo day.

References

[1] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, 2019.

[2] Shuo Li and Ning Ma. A study of large language model q&a based on lora fine-tuning and prompt engineering. In *Proceedings of the 2025 5th International Conference on Applied Mathematics, Modelling and Intelligent Computing*, pages 312–317, 2025.

[3] Martin Wistuba, Prabhu Teja Sivaprasad, Lukas Balles, and Giovanni Zappella. Choice of peft technique in continual learning: Prompt tuning is not all you need. *arXiv preprint arXiv:2406.03216*, 2024.

[4] X Zhang, N Talukdar, S Vemulapalli, S Ahn, J Wang, H Meng, et al. Comparison of prompt engineering and fine-tuning strategies in large language models in the classification of clinical notes [internet]. 2024.