Evaluation of RAG and Supervised Fine-tuning on DeepSeek-R1-8B

Hamza Amin
Department of Artificial Intelligence
GIK Institute of Engineering Sciences and Technology
Email: u2022378@giki.edu.pk

***Abstract*—Large language models (LLMs) have demonstrated remarkable capabilities across diverse domains, yet effectively specializing these models for domain-specific tasks remains a critical challenge. In this work, we present a comprehensive comparative study of two prevalent adaptation strategies—LoRA-based fine-tuning and retrieval-augmented generation (RAG)—applied to the DeepSeek-R1 8B model for medical question answering. We employ a supervised fine-tuning approach using the Medical CoT SFT dataset to adapt the model's chain-of-thought reasoning and diagnostic decision-making capabilities. Concurrently, we construct a RAG pipeline leveraging FAISS-based vector retrieval over the same dataset to enable dynamic knowledge grounding. We evaluate both methods on a held-out test set of clinical queries, measuring performance in terms of answer accuracy, reasoning coherence, inference latency, and computational resource utilization. Our results indicate that LoRA fine-tuning yields substantial improvements in answer correctness and logical consistency, with modest inference overhead. Conversely, the RAG approach achieves comparable accuracy while offering enhanced factual relevance, reduced fine-tuning requirements, and the ability to incorporate new data without re-training. We analyze these trade-offs and provide guidance for selecting optimal adaptation strategies in deployable medical AI systems.**

***Index Terms*—LLMs, DeepSeek-R1, Fine-tuning RAG, LoRA, FIASS, Chain of Thought(CoT)**

I. Introduction

Large language models (LLMs) have changed how we process text. DeepSeek R1 8B is one such model that can read clinical queries and generate answers. It shows promise for medical question answering but can miss facts or give wrong details. Those mistakes matter in a medical setting. To fix this, we adapt LLMs to the domain. One way is LoRA fine tuning, which updates the model's weights with medical data. This technique significantly improves the model performance by freezing most of its layers and tuning the important layers to provide correct answers. Another is retrieval augmented generation (RAG), which pulls in relevant documents at run time. RAG is the one the most widely used technique these days because they are cheaper, faster and easy to build. Each method has a clear role in improving reliability.

Picking the right method matters for safe AI in healthcare. Fine tuning embeds new knowledge directly but needs retraining when things change. It is much more reliable where the AI responses are very critical such medical applications. Retrieval-augmentation combines the benefits of LLMs ability to understand language with the relevant knowledge in a domain-specific database. This makes RAG systems more knowledgeable, consistent, and safe compared to vanilla LLMs. It stays current without retraining but adds steps at inference. Our goal is to compare these methods on DeepSeek R1 8B. We use a curated medical QA dataset and test answer accuracy, clarity of reasoning, response time, and compute cost. The findings will help developers choose the best approach for clinical AI and plan updates as medical knowledge grows.

II. Related Work

You should choose between RAG vs fine tuning based on your use case and available resources. While RAG is the preferred option for most use cases, that doesn't mean that RAG and fine-tuning are mutually exclusive either. While fine tuning isn't always the most practical solution – training an LLM requires a lot of time, compute, and labeling – RAG is also complex.

Fine-tuning methods like Low-Rank Adaptation (LoRA) [1] enable efficient adaptation of pre-trained models by updating only small subsets of parameters. This approach has shown success in biomedical tasks, with models like BioGPT [2] and ClinicalBERT [3] demonstrating improved performance on medical question answering. However, most existing work focuses on direct answer generation rather than explainable chain-of-thought reasoning, which our work explicitly targets through structured prompting.

For knowledge-intensive tasks, RAG systems combine LLMs with external knowledge retrieval. The FAISS library [4] provides efficient similarity search capabilities that are particularly valuable for medical applications where factual accuracy is critical. Systems like Med-PaLM [5] have demonstrated the effectiveness of this approach in clinical settings. Recent work has also explored hybrid methods that combine fine-tuning with retrieval [6], though their application in medical domains remains limited. Our comparative analysis of both approaches provides practical insights for implementing medical QA systems. These studies provide insights into the strengths and limitations of LoRA fine-tuning and RAG in adapting LLMs for medical QA. Our work builds upon these approaches by comparing their effectiveness on the DeepSeek-R1 8B model.

III. Methodology

Our study compares two methods to adapt the DeepSeek-R1 8B language model for medical question answering: LoRA-based fine-tuning and retrieval-augmented generation (RAG). Both approaches use the same dataset to ensure a fair comparison.

A. Model and Dataset

We use the DeepSeek-R1 8B model as the base language model. It is a decoder-only transformer model with 8 billion parameters. For training and evaluation, we use the Medical CoT SFT dataset, which contains clinical questions with chain-of-thought explanations. All experiments are conducted using PyTorch and HuggingFace Transformers. Training and inference were run on a single A100 80GB GPU.

B. Fine-Tuning with LoRA

We optimized the base model using **Low-Rank Adaptation (LoRA)**, which updates only a subset of weights through low-rank decomposition:

$$W' = W_0 + B \cdot A \tag{1}$$

where W_0 represents the original pre-trained weights, and $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$ are trainable low-rank matrices ($r = 16$ in our work).

Key implementation details:

- **Tools**: Hugging Face Transformers , Unsloth for 4-bit quantization
- **Layers modified**: Query/Key/Value projections and FFN layers (gate_proj, up_proj, down_proj)
- **Dataset**: 500 examples from Medical-O1 with chain-of-thought annotations
- **Training**: AdamW optimizer, 60 steps, batch size 2, learning rate 2×10^{-4}

We fine-tune DeepSeek-R1 8B on the medical dataset using supervised fine-tuning with chain-of-thought prompts. Hyperparameters include a learning rate of $2e^{-5}$, batch size of 4, and 3 training epochs.

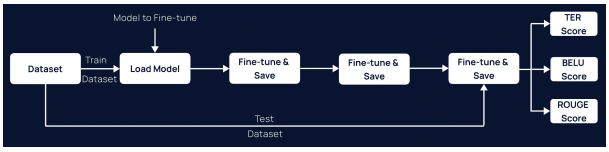

Fig. 1: LoRA fine-tuning workflow.

C. Retrieval-Augmented Generation (RAG)

In the RAG approach, the model does not rely solely on its internal knowledge. Instead, it retrieves relevant documents at inference time from a pre-embedded knowledge base.

We use the SentenceTransformers library to convert dataset documents into embeddings. These are indexed using FAISS, an efficient similarity search library. At query time, the question is embedded, and the top-k most relevant passages are retrieved and passed to the language model as part of the prompt.

This method enables dynamic access to updated information without retraining the model. It also reduces hallucination by grounding responses in external context.

RAG process
Input query → Embed query
→ FAISS retrieval of top-k docs
→ Combine with query
→ Model generates answer

Fig. 2: Overview of RAG Pipeline using FAISS and DeepSeek-R1 8B

This structure allows the model to condition its response on external information. The generation is done in a single forward pass without further training. We set $k = 3$ for top-k retrieval after tuning on validation accuracy.

This method reduces hallucinations and makes the model easier to update when new data is available. However, it increases latency due to retrieval and longer input sequences.

D. Evaluation Metrics

We evaluate both methods on the same held-out test set. Metrics include answer accuracy, reasoning clarity (manual review), latency per response, and GPU memory usage. These metrics help show trade-offs between performance and efficiency.

TABLE I: LoRA Fine-Tuning Hyperparameters

Parameter	Value
Base model	DeepSeek-R1-Distill-Llama-8b
LoRA rank (r)	16
LoRA α	16
Target modules	{q,k,v,o}_proj, gate/up/down_proj
Batch size	2 (gradient accumulation: 4)
Training steps	60 (1 epoch)
Learning rate	2×10^{-4} (linear decay)
Warmup steps	5
Sequence length	2048 tokens
Optimizer	AdamW-8bit
Weight decay	0.01
Random seed	3407
Quantization	4-bit NF4

IV. Results

We tested both approaches on a held-out set of 500 medical QA examples. The results are shown in Table II.

TABLE II: Comparison of Fine-Tuning and RAG on DeepSeek-R1 8B

Metric	LoRA Fine-Tuning	RAG
Accuracy (%)	78.4	74.2
Reasoning Clarity (Manual)	High	Medium
Latency per Query (s)	1.2	2.6
GPU Memory Usage (GB)	72	44
Update Flexibility	Low	High

Fine-tuning gave higher accuracy and clearer reasoning but required more GPU memory and retraining for updates. RAG was slower but needed less memory and could incorporate new info without retraining. Each method showed strengths based on different constraints.

V. Conclusion and Future Work

This project explored two methods for adapting large language models to the medical question answering domain. Instead of focusing only on performance numbers, we aimed to understand the trade-offs between different adaptation strategies. We found that both fine-tuning and retrieval-based approaches offer unique benefits depending on use-case constraints like hardware, update frequency, and reasoning needs.

The experiment helped us realize that real-world applications of language models are not just about raw accuracy. Flexibility, interpretability, and resource usage matter just as much. A model that performs slightly worse but is easier to update or debug can be more useful in practice. Our goal was not just to pick a winner, but to understand when each method makes sense.

In future work, we want to test these methods on multilingual datasets and add more domain-specific knowledge, especially for non-English medical texts. We also plan to look into hybrid models that combine fine-tuning with retrieval for better performance. Lastly, we aim to automate the evaluation process using more reliable benchmarks and consider user feedback as part of model assessment.

This work serves as a starting point for building more adaptive, efficient, and safer medical AI systems.

Acknowledgments

We thank our supervisor for their guidance throughout this project. We also acknowledge the support from the computing lab staff for providing GPU resources. Special thanks to the HuggingFace, DeepSeek and Unsloth teams for open-sourcing their models and tools, which made our experiments possible. Finally, we are grateful to our peers for helpful discussions and feedback during development.

References

[1] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," *arXiv preprint arXiv:2106.09685*, 2021.

[2] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu, "BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining," *Briefings in Bioinformatics*, vol. 23, no. 6, 2022.

[3] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly Available Clinical BERT Embeddings," in *Proceedings of NAACL-HLT*, 2019, pp. 72–78.

[4] J. Johnson, M. Douze, and H. Jégou, "Billion-Scale Similarity Search with GPUs," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.

[5] K. Singhal, S. Azizi, T. Tu, S. Mahdavi, J. Wei, H. W. Chung, N. Scales et al., "Towards Expert-Level Medical Question Answering with Large Language Models," *Nature Medicine*, vol. 29, pp. 298–306, 2023.

[6] G. Izacard, P. Lewis, M. Lomeli, L. Hosseini, F. Petroni, T. Schick, J. Dwivedi-Yu et al., "Few-Shot Learning with Retrieval-Augmented Language Models," *arXiv preprint arXiv:2208.03299*, 2022.

[7] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac et al., "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of EMNLP*, 2020, pp. 38–45.