Small LMs Fine-Tuning and RAG for Reliable Medical Question-Answering

Hussein Abdallah Concordia University
hussein.abdallah@mail.concordia.ca,

Abstract

This report presents a methodology for enhancing medical question answering (QA) systems through the use of fine-tuned Small Language Models (SLMs). Focusing on various scales of the Qwen3 LLM family, the study demonstrates their effectiveness in delivering accurate and reliable responses in the medical domain. The proposed pipeline, MedQA, integrates BEFT-LoRA fine-tuning and document-based Retrieval-Augmented Generation (RAG) to improve performance on the MedQA dataset. While LoRA fine-tuning yields modest gains across different model scales, the incorporation of document-level RAG—providing on-demand access to relevant external information— enhanced the answer accuracy with lower cost. The combined approach achieves up to a 17% improvement in accuracy and consistently outperforms SOTA fine-tuned medical LLM baselines. the pipeline's code is available here.

1 Introduction

The integration of LLMs into healthcare has demonstrated strong potential to enhance the quality and efficiency of medical services. Among recent advances, medical question-answering (QA) systems have become essential for delivering accurate, timely responses to complex clinical queries (Anaissi et al., 2024). This report compares fine-tuned small Language Models (sLMs) with document-based Retrieval-Augmented Generation (RAG) techniques for medical QA. The goal is to improve access to medical knowledge, support clinical decisions, leverage multi-modal data (e.g., electronic health records), and ensure equitable access to reliable information.

LLMs Fine-tuning techniques have been introduced to adapt LLMs for domain-specific tasks while reducing the cost of training from scratch, especially in low-resource settings. BEFT-LoRA (Mangrulkar et al., 2022) enables efficient fine-tuning by injecting low-rank trainable adapters into selected model layers. Despite these advances, medical QA remains challenging. Current systems often struggle with complex terminology and contextually accurate responses, as shown in the Open Medical Leaderboard (Pal et al., 2024). Moreover, the rapid evolution of medical knowledge requires frequent, resource-intensive updates, risking catastrophic LLMs forgetting (Song et al., 2025). Generic models also risk propagating outdated or incorrect information, compromising their reliability in critical healthcare scenarios.

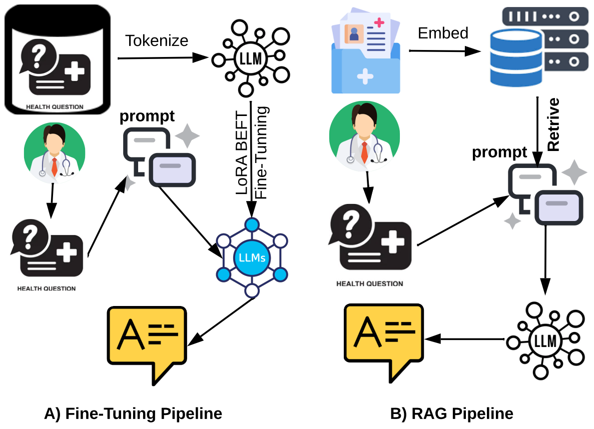

Figure 1: Medical QA Pipelines: (A) SLM Fine-Tuning – The SLM is first fine-tuned, then queried by the doctor with a medical question. (B) Document-based RAG – The doctor prompts a non-fine-tuned LLM, while relevant information is retrieved from a knowledge base and appended to the prompt.

Retrieval-Augmented Generation (RAG) has emerged as a prominent approach for enriching static LLMs with external knowledge by incorporating retrieved documents into the generation context (Gao et al., 2024). However, most existing RAG methods retrieve documents independently, limiting their ability to capture complex inter-document relationships. This constraint hinders performance in tasks requiring multi-hop reasoning, i.e., capturing the complex information exist in multi-modal EHR knowledge bases.

To address these challenges, this report demonstrates empirically how SLMs fine-tuning + RAG could improve the model's performance in medical QA datasets. Qwen3 LLM family is used at different scales, including 0.6B,4B,8B,14B parameters to benchmark these techniques. The MedQA (Jin et al., 2020) dataset is used to fine-tune the LLM and test its performance before and after the fine-tuning. The answer Accuracy (%) and Number of tokens are the metrics used to evaluate the models' performance.

2 The Proposed Pipelines

This report presents MedQA, a combined fine-tuning and RAG pipeline designed to enhance LLM performance on medical question-answering tasks. Figure 1 illustrates the two pipeline architectures and their respective steps.

In Pipeline A (Fine-Tuning), an off-the-shelf small LLM is fine-tuned on a medical QA dataset such as

MedQA (Jin et al., 2020) to better understand medical terminology and concepts. The dataset is tokenized using the model's tokenizer, and a parameter-efficient fine-tuning method like PEFT-LoRA is applied to update only selective model layers, reducing the computational cost of full-model training. The fine-tuned model is then used to answer medical questions via a predefined prompt template. While effective, this approach is resource-demanding and requires repetitive fine-tuning to accommodate new concepts and data modalities, which often leads to catastrophic forgetting after repetitive fine-tuning (Song et al., 2025).

In Pipeline B (Document-based RAG), an off-the-shelf LLM is used alongside a retrieval step to enrich its limited domain-specific knowledge. Instead of answering the medical question directly, the model is provided with relevant context retrieved from an external medical knowledge base—comprising textbooks, research papers, medical records, etc. This knowledge base is chunked, embedded, and indexed in a vector database for efficient dense retrieval. At inference time, relevant chunks are retrieved and combined with the doctor's question into a single prompt, which is then passed to either a fine-tuned or non-fine-tuned LLM. The RAG pipeline incurs a one-time cost for embedding and indexing, allowing new or updated documents to be integrated incrementally. This approach offers a scalable and cost-efficient alternative to repeated fine-tuning while maintaining high performance with newly reasoning-led LLMs.

3 Experimental Results

3.1 system setup

- **Dataset**: The MedQA dataset (Jin et al., 2020), a multiple-choice medical question-answering benchmark, is employed in this study. It comprises questions derived from medical licensing exams, designed to assess physicians' clinical reasoning and domain knowledge. The dataset includes 10.2K English questions for fine-tuning and 1.2K for testing, covering a broad range of medical topics that require deep conceptual understanding. Additionally, a collection of reference books is compiled to construct the knowledge base for the RAG system. Due to time and resource constraints, evaluation is conducted on the first 100 questions from the test set.

- **Baselines**: The **Qwen3** family of LLMs is fine-tuned at multiple scales—0.6B, 4B, 8B, and 14B parameters—for 500 steps using a cosine learning rate scheduler to facilitate efficient convergence, particularly in text completion tasks. Qwen3 (Yang et al., 2025) achieves state-of-the-art performance in general QA and reasoning through Mixture-of-Experts (MoE) training. Additionally, **Medllama3-v20** (Probe Medical, 2025), a top-ranked medical LLM on the Open Medical Leaderboard, and **MedGemma** (Sellergren et al., 2025), Google's leading open-source model for medical text and image understanding, built on Gemma 3, are used for comparative evaluation.

- **Metric**: Accuracy is used to evaluate LLM performance on the MedQA multiple-choice questions, each containing one correct answer among four options. Additionally, the token usage is measured to assess computational efficiency and answer latency.

- **Hardware**: Fine-tuning was conducted on Google Colab using T4, L4, and A100 GPUs, while inference of the fine-tuned model was performed on a V100-16C virtual GPU hosted on Compute Canada resources.

- **Software**: The Unsloth fine-tuning library (Daniel Han and team, 2023) is employed for efficient and cost-effective model adaptation. For lightweight inference, the quantized GGUF Q4_K_M version of the fine-tuned models is hosted using the LLaMA-c++ server (Gerganov and community, 2025). In the RAG pipeline, Faiss vector databases (Johnson et al., 2019) are utilized for embedding indexing and retrieval, with embeddings generated by the Qwen3-0.6B model.

3.2 Results analysis

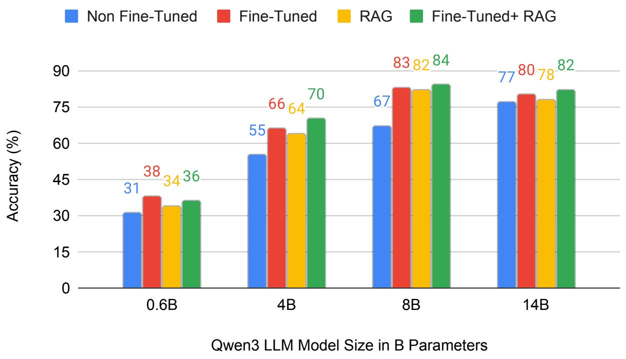

Figure 2: Accuracy results for the Qwen3 family show that non-fine-tuned models yield the lowest performance. Fine-tuning improves accuracy by up to 16%, while the RAG pipeline achieves comparable gains of up to 15% with lower embedding costs. Combining fine-tuning and RAG yields the highest improvement of up to 17%, albeit with increased computational expense. Among the models, the 8B variant offers the best balance between accuracy and cost.

Model Accuracy: The accuracy of fine-tuned and RAG pipelines across Qwen3 models of varying scales on 100 MedQA questions (Figure 2). Non-fine-tuned models perform worst overall but remain competitive at the smallest (0.6B) and largest (14B) scales. Fine-tuning with BEFT-LoRA requires about 6 GPU hours and yields up to 16% accuracy improvement. RAG achieves similar gains of

Table 1: The Average Tokens Number per Prompt.

Model	Non-FT	FT	RAG	FT+RAG
Medgemma	**277.97**	N/A	580.69	N/A
Medllama3	**320.8**	N/A	631.93	N/A
Qwen3-0.6B	696.62	**429.26**	812.6	609.5
Qwen3-4B	934.13	**345.51**	968.15	2945.83
Qwen3-8B	**888.18**	1310.03	954.39	1594.16
Qwen3-14B	377.23	**290.38**	381.9	591.92

up to 17% with a one-time cost of 2 hours for embedding and indexing, adding roughly 500 tokens per prompt. The combined Fine-Tuned+RAG pipeline shows the highest accuracy by leveraging both approaches but incurs the full computational cost. RAG offers the best balance between accuracy and cost, particularly with the 8B model, while enabling flexible knowledge updates without the risk of catastrophic forgetting inherent in fine-tuning.

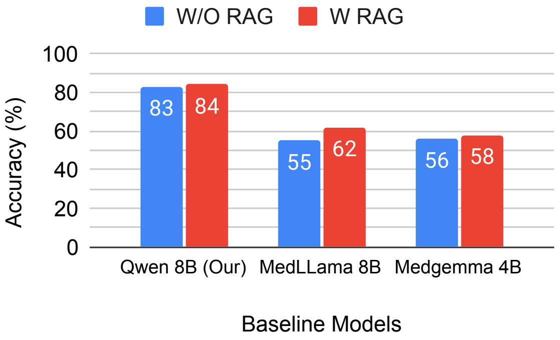

Figure 3: The fine-tuned Qwen3-8B LLM outperforms both Medllama-v3 and MedGemma baselines, with and without RAG integration. Qwen3 originally surpasses these models on reasoning benchmarks.

Comparative baselines: The MedQA fine-tuned Qwen3-8B model significantly outperforms baseline models, as shown in Figure 3. This advantage is attributed to the Qwen3 family's Mixture-of-Experts (MoE) training, which enhances its effectiveness over the baselines.

Number of Tokens: Fine-tuned models in Table 1 use fewer tokens on average, reducing cost and response time. Although RAG prompts require more tokens, this overhead is minimal when models are hosted locally with larger context windows.

3.3 Limitations & Opportunities

- **Independent chunks retrieval** may introduce noise and indirectly relevant contexts that may confuse the LLM reasoning. Supporting dense multi-modal data retrieval in a Graph-RAG architecture will improve the quality, reduce the noise, and reduce the overall response cost.

- **Large-scale** multi-modal medical knowledge base requires scalable/dynamic data processing pipelines and domain-specific/noise-less retrieval techniques for effective RAG pipelines.

- **LLMs hallucination** may introduce a critical non-factual medical content that is not easy to verify. Designing and implementing a medical trustworthy metric to evaluate the response credibility and trustworthiness is a demanding challenge.

- **Incomplete knowledge** usually hinder efficient retrieval. Representing medical knowledge bases as knowledge graphs and applying graph machine learning to predict missing nodes or edges can enhance RAG accuracy, particularly for unseen questions or cases.

References

Ali Anaissi, Ali Braytee, and Junaid Akram. 2024. Fine-tuning llms for reliable medical question-answering services. In *IEEE International Conference on Data Mining, ICDM 2024 - Workshops, Abu Dhabi, United Arab Emirates, December 9, 2024*, pages 146–153. IEEE.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Georgi Gerganov and community. 2025. Llama-c++.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*.

Jeff Johnson, Douze, and et.al. 2019. Billion-scale similarity search with GPUs.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

Ankit Pal, Pasquale Minervini, Andreas Geert Motzfeldt, and Beatrice Alex. 2024. open medical llm leaderboard. https://huggingface.co/spaces/openlifescienceai/open_medical_llm_leaderboard.

MAILAB from Yonsei University Probe Medical. 2025. medllama3.

Andrew Sellergren, Kazemzadeh, and et.al. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Shezheng Song, Hao Xu, Jun Ma, Shasha Li, Long Peng, Qian Wan, Xiaodong Liu, and Jie Yu. 2025. How to alleviate catastrophic forgetting in llms finetuning? hierarchical layer-wise and element-wise regularization. *Preprint*, arXiv:2501.13669.

An Yang, Anfeng Li, and et.al. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

A Appendix

A.1 RAG documents Embeddings

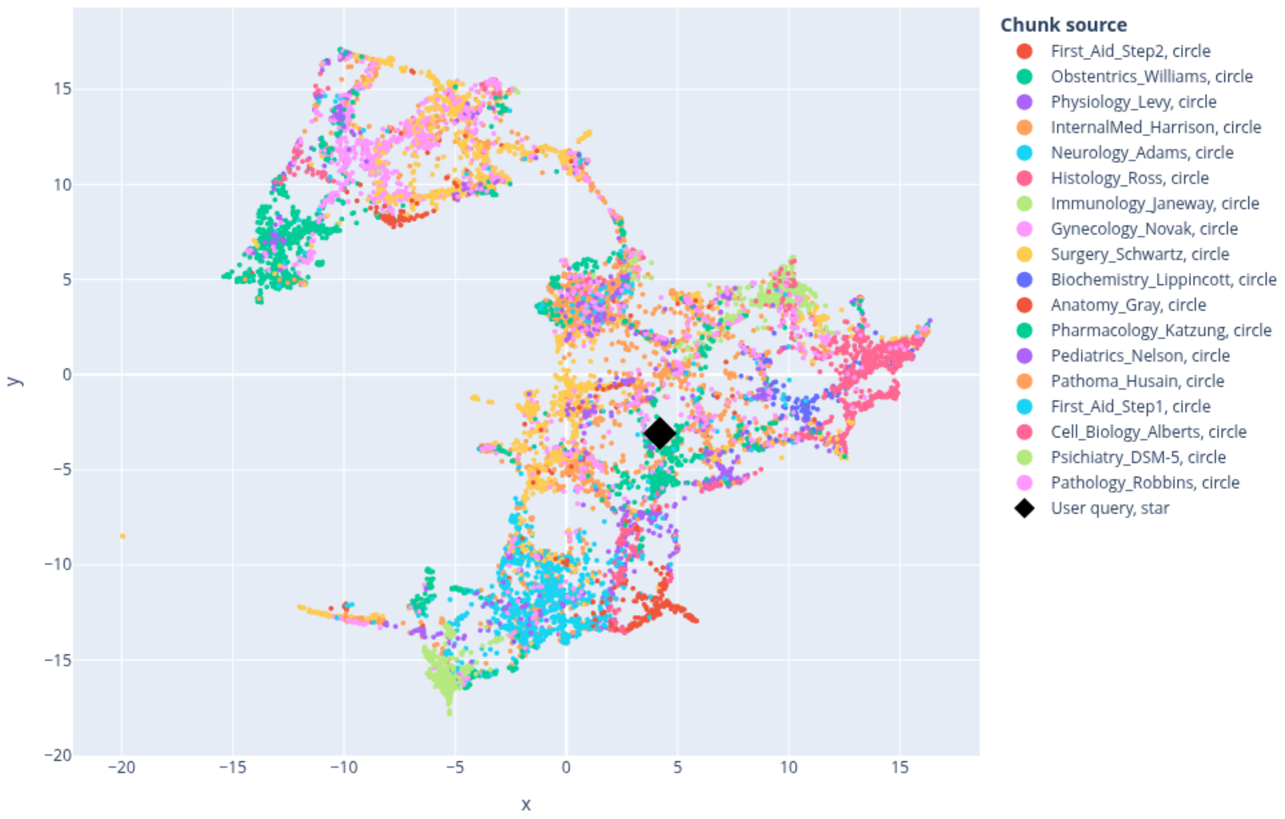

Figure 4: The MedQA external document Chunk's embedding projection in 2D. The document's knowledge base contains 18 book about different medical topics. The chunks are grouped by the source document. The chunks that contains similar content are clearly very close.

A.2 MedQA Prompt Examples

Fine-Tuned Models prompt:

<System Role>: You are a professional, highly experienced doctor professor. Please answer the patients' questions using only one of the options in the brackets.
<User Role>: Patient Case: A junior orthopaedic surgery resident is completing a carpal tunnel repair with the department chairman as the attending physician. During the case, the resident inadvertently cuts a flexor tendon. The tendon is repaired without complication. The attending tells the resident that the patient will do fine, and there is no need to report this minor complication that will not harm the patient, as he does not want to make the patient worry unnecessarily. He tells the resident to leave this complication out of the operative report. Which of the following is the correct next action for the resident to take?
Choices: A: Disclose the error to the patient and put it in the operative report.
B: Tell the attending that he cannot fail to disclose this mistake.
C: Report the physician to the ethics committee.
D: Refuse to dictate the operative report.
You can only output the predicted label in exact words. No other words should be included.
Answer:

RAG prompt:

<System Role>: You are a professional, highly experienced doctor professor. Please answer the patients' questions using only one of the options in the brackets.
<Assistant Role>: Use the given Supportive Information to refine and think about the answer.
<User Role>: Patient Case: A junior orthopaedic surgery resident is completing a carpal tunnel repair with the department chairman as the attending physician. During the case, the resident inadvertently cuts a flexor tendon. The tendon is repaired without complication. The attending tells the resident that the patient will do fine, and there is no need to report this minor complication that will not harm the patient, as he does not want to make the patient worry unnecessarily. He tells the resident to leave this complication out of the operative report. Which of the following is the correct next action for the resident to take?
Choices: A: Disclose the error to the patient and put it in the operative report.
B: Tell the attending that he cannot fail to disclose this mistake.
C: Report the physician to the ethics committee.
D: Refuse to dictate the operative report.
You can only output the predicted label in exact words. No other words should be included.
Answer:
Supportive Information:
Database of High-Yield Facts
The seventh edition of First Aid for the USMLE Step 2 CK contains a revised and expanded database of clinical material that student authors and faculty have identified as high yield for boards review. The facts are organized according to subject matter, whether medical specialty (e.g., Cardiovascular, Renal) or high-yield topic (e.g., Ethics) in medicine. Each subject is then divided into smaller subsections of related facts. Individual facts are generally presented in a logical approach, from basic definitions and epidemiology to History/Physical Exam, Diagnosis, and Treatment. Lists, mnemonics, and tables are used when helpful in forming key associations.
The content is mostly useful for reviewing material already learned. This section is not ideal for learning complex or highly conceptual material for the first time. Black-and-white images appear throughout the text. In some cases, reference is made to the "clinical image" section at the end of Section 2, which contains full-color glossy plates of histology and patient pathology by topic. At the end of Section 2, we also feature a Rapid Review chapter of key facts and classic associations to cram a day or two before the exam.
The Database of High-Yield Facts is not comprehensive. Use it to complement your core study material and not as your primary study source.
The facts and notes have been condensed and edited to emphasize the essential material. Work with the material, add your own notes You can only output the predicted label in exact words. No other words should be included.

A.3 Fine-tuning details

SFTConfig Parameters:

- per_device_train_batch_size: 4
- gradient_accumulation_steps: 16
- warmup_steps: 0
- max_steps: 500
- num_train_epochs: 3 For longer training runs!
- learning_rate: 2e-5
- logging_steps: 10
- optim : "adamw_8bit"
- weight_decay : 0.01
- lr_scheduler_type : "cosine"

Training loss:

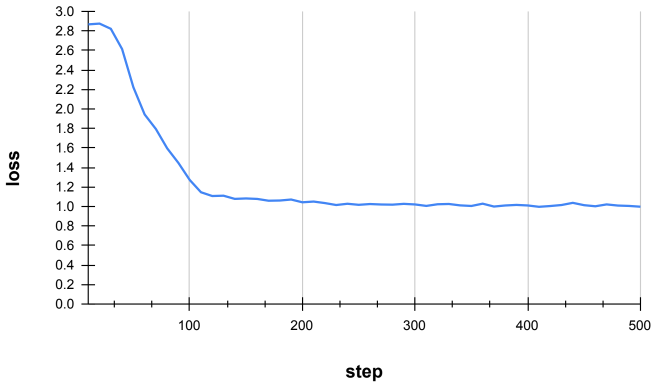

Figure 5: The Qwen3-14B fine-tuning loss.