Prescription for Efficiency: Making Medical LLMs Less Resource-Intensive

William Bush, Kendrick Hsu, Danny Stonberg,
Sahith Jagarlamudi, Vikramaditya Singh
University of Pennsylvania

Abstract

Large language models (LLMs) have shown remarkable capabilities in medical question-answering tasks, but their extensive computational resources and reliance on cloud-based servers present privacy and accessibility concerns. In this work, we investigate techniques for compressing medical LLMs through model pruning, quantization, and fine-tuning to run efficiently and privately on personal devices. We apply these techniques to benchmarks such as PubMedQA, MedQA, and MedMCQA, aiming to preserve accuracy while reducing model size. Our results demonstrate that careful optimization can maintain performance close to large, resource-intensive models, thus paving the way for more private, accessible, and resource-efficient medical NLP systems.

1 Introduction

Modern natural language processing (NLP) techniques have enabled significant advancements in medical question-answering (QA). Medical QA involves interpreting patient symptoms, summarizing clinical findings, and retrieving relevant medical literature to answer queries posed by both clinicians and patients. This task relates closely to computational linguistics: it requires understanding domain-specific vocabulary, complex medical terminologies, and subtle linguistic cues found in clinical notes and biomedical research articles. Unlike generic QA tasks, medical QA demands high reliability and accuracy due to the potential impact on healthcare decisions.

Illustrative Example: Consider a patient, Sarah, who experiences frequent, burning urination and suspects a urinary tract infection (UTI). She inputs her symptoms into an AI assistant: *"Based on these symptoms, could this be a UTI, how urgent is*

it to see a doctor, and what questions should I prepare for my appointment?" The system accesses a small medical LLM, running locally on her device, and provides a medically informed, accurate answer.

Formal Problem Definition: Formally, let C represent the medical context (patient symptoms, medical literature), and Q represent a medical question. The task is to produce an answer A that is both medically accurate and useful. The model: $f(C, Q) \rightarrow A$ should be efficient enough to run on limited hardware while maintaining strong accuracy.

Why This Task? State-of-the-art medical LLMs, while highly accurate, often have billions of parameters. This raises two concerns: (1) Resource-intensiveness limits their deployment on personal devices, and (2) Reliance on remote servers poses privacy risks when patients share sensitive health information. By exploring methods to compress and optimize large medical models, we can create systems that are both accessible and privacy-preserving. This project aims to produce smaller, efficient medical LLMs that approach the performance of their large-scale counterparts.

2 Literature Review

Several research efforts have sought to improve the main bottlenecks of large QA models - model size and data privacy. [Hinton et al.2015] introduced knowledge distillation, the practice of training a smaller model to mimic the original larger "teacher" model, thus effectively compressing the original model. Although not originally focused on the medical domain, this technique has informed subsequent attempts to build resource-efficient models for domain-specific tasks. To address privacy concerns, [McMahan et al.2017] proposed federated learning to train models across decentralized data sources without sharing raw information. Although this doesn't solve the model size or inference privacy issues, it was a step towards privacy-awareness when constructing models.

[Han et al.2015] explored more direct model compression techniques (e.g., pruning, quantization, Huffman coding) to reduce computational costs. While Han's work did not target medical QA specifically, it provided general methods that can be applied to large, more domain-specific models.

In the medical domain specifically, research involves adapting large pretrained language models, such as BioBERT or ClinicalBERT, to medical QA tasks [Lee et al.2019], achieving improved results over generic LLMs.

However, these domain-tuned models remain large, and few studies have explicitly addressed making them lightweight and privacy-preserving. Fortunately, Medical QA research has been accelerated by the release of benchmark datasets and shared tasks. For instance, **PubMedQA** [Jin et al.2019] focuses on yes/no/maybe reasoning from biomedical abstracts. **MedQA** [Zhang et al.2020] and **MedMCQA** [Pal et al.2022] provide multiple-choice clinical questions, simulating professional medical exam conditions. These datasets allow for standardized evaluation of model performance and drive progress in understanding domain-specific linguistic challenges.

The PubMedQA dataset was curated from PubMed abstracts to test biomedical reading comprehension. MedQA, on the other hand, is constructed from medical licensing exam questions, testing the model's ability to choose the best answer from multiple choices. MedMCQA expands on this by introducing a large-scale dataset of multiple-choice clinical questions, enabling researchers to evaluate scalability and generalization.

Using these datasets and the learnings from previous research, we intend to expand the literature by integrating pruning and quantization with medical-domain fine-tuning to create an efficient model suitable for on-premise inference. By building upon the insights and targeted goals from previous approaches, we aim to produce a smaller, more private, and more efficient medical QA model that still performs competitively on established benchmarks like PubMedQA, MedQA, and MedMCQA.

3 Experimental Design

3.1 Data

We use three datasets: **PubMedQA**, **MedQA**, and **MedMCQA**. PubMedQA includes yes/no/maybe labels derived from biomedical abstracts [Jin et al.2019], MedQA consists of multiple choice questions from medical exams [Zhang et al.2020], and MedMCQA offers a large-scale multiple-choice QA setting [Pal et al.2022].

Table 1 shows the size of each split. For PubMedQA, we have 1,000 training, 800 test, and 200 validation samples. MedQA offers 10,178 training samples, 1,273 test samples, and 1,272 validation samples. MedMCQA is much larger, with 182,822 training samples, 6,150 tests, and 4,183 validation samples. The label distribution is somewhat balanced for multiple choice tasks, but the PubMedQA yes/no/maybe split is slightly skewed, favoring yes/no answers.

Dataset	Train	Test	Validation
PubMedQA	1,000	800	200
MedQA	10,178	1,273	1,272
MedMCQA	182,822	6,150	4,183

Table 1: Dataset sizes for PubMedQA, MedQA, and MedMCQA.

Data Characterization: In PubMedQA, the reasoning involves comprehending short biomedical texts and producing yes/no/maybe answers. MedQA and MedMCQA require selecting the best answer from multiple choices, often involving deep clinical reasoning. Input lengths vary, with some questions referencing specific medical conditions, treatments, or outcomes.

3.2 Evaluation Metric

We use standard classification metrics: Accuracy, Precision, Recall, and F1-score. Accuracy measures how often the predicted answer matches the correct one. Precision and Recall quantify correctness among predicted positives and total positives, respectively, and F1 provides a harmonic mean of Precision and Recall.
For binary or yes/no tasks:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \quad \text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Predictions}}$$

3.3 Simple Baseline

As a simple baseline, we use a majority-class predictor. For PubMedQA, it always selects the most frequent answer type. For multiple-choice tasks (MedQA, MedMCQA), it chooses the most common choice. This baseline characterizes dataset difficulty. Table 2 shows that while PubMedQA's majority baseline reaches about 0.61 accuracy, MedQA and MedMCQA baselines are lower, highlighting the wider range of likely answers and thus the increased complexity of the task.

Dataset	Acc	Prec	Rec	F1
PubMedQA Baseline	0.6100	0.3721	0.6100	0.4622
MedQA Baseline	0.2193	0.0481	0.2193	0.0789
MedMCQA Baseline	0.3223	0.1038	0.3223	0.1571

Table 2: Majority-class baseline performance on the test sets.

4 Experimental Results

4.1 Published Baseline

We implemented Llama-3.2-7b a large pretrained LLM known for strong general language understanding. This model serves as a published baseline. Although not originally fine-tuned for these medical QA tasks, it demonstrates zero-shot capabilities.

We evaluated Llama-3.2-7b on the test sets of PubMedQA, MedQA, and MedMCQA using our defined metrics. On PubMedQA, it achieved around 0.4850 accuracy, below the majority baseline. MedQA accuracy was about 0.2050, and MedMCQA accuracy was 0.2634. These results are closer, but still slightly lower than those reported for domain-specialized models in prior work, likely due to domain mismatch and lack of task-specific tuning. Additionally, results are not directly comparable to the original papers if they used different test splits or domain-specific pretraining.

4.2 Extensions

We explored three key extensions: Pruning, Quantization, and Fine-Tuning on LLama-3.2-7B, as well as Quantization and Fine-Tuning on Mistral-7B.

Pruning: We performed an intra-level weight pruning of the Llama-3.2-7b model to remove redundant or low-weight parameters. This means that for each level of the network, we removed the lowest performing 20% of nodes. While this significantly reduced model size and inference latency, it lowered accuracy. For example, PubMedQA accuracy dropped from 0.4850 to around 0.1100, illustrating the trade-off between size and performance. Because of these low results we then chose to focus on quantization instead of pruning. As for pruning however, our attempts were not exhaustive, and further research might explore structural tuning (removing entire node pathways) or more targeted pruning.

Quantization: Next, we used the bits and bytes packages to quantize the model's weights to 4-bits, reducing memory usage and allowing for faster, lower-precision calculations during inference. Quantization preserved accuracy more effectively than pruning alone, indicating that important information was stored in each node, but maybe the precision of accuracy on that node was less important. PubMedQA accuracy remained close to 0.48, nearly matching the uncompressed baseline. This suggests quantization is a viable strategy for efficiency without severe performance loss.

Fine-Tuning: The fine-tuning process aimed to adapt the pre-trained LLaMA 3.2-3B model for medical question-answering tasks using the MedQuad dataset. To achieve this, the dataset was reformatted into prompt-response pairs and tokenized for uniform input representation. The model leveraged Parameter-Efficient Fine-Tuning (PEFT) with LoRA (Low-Rank Adaptation) to train only a subset of parameters, significantly reducing memory and computational requirements. Additionally, quantization with BitsAndBytes further optimized the model's resource usage by enabling 8-bit precision. Through a carefully configured training pipeline, the model was fine-tuned using a limited dataset with gradient accumulation and mixed-precision training, providing a robust yet efficient approach to domain-specific language model adaptation. We see marked increases from the pruned model after fine-tuning as we see accuracy increasing to 0.2145 from 0.1200 for the MedQA dataset and 0.2606 from 0.1562 for the MedMCQA dataset - with both metrics beating the LLama-3.2-7b baseline. These improvements show the impact of the domain-specific fine-tuning on the light weight model. However, for the PubMedQA we see only a slight increase to 0.1500 from 0.1100 and our fine-tuned model underperforms the strong baseline. Recent research on lightweight models show the impressive results of chain of thought reasoning in smaller models resulting in performance above that of larger models. This allows us to use edge computational abilities along with finetuning as a direct replacement to the memory constraints.

Mistral-7B Fine-Tuning: As a further extension, we fine-tuned the 4-bit Mistral-7B model on the MedQuAD dataset using LoRA to enhance its domain-specific performance on PubMedQA, MedQA, and MedMCQA. This fine-tuning aimed to assess its impact on accuracy and explanation quality. On PubMedQA, fine-tuning significantly increased precision ($0.70 \rightarrow 0.79$) but reduced accuracy ($0.71 \rightarrow 0.64$), indicating a more cautious but less accurate response strategy. For MedQA and MedMCQA, fine-tuning failed to improve performance, with accuracy remaining at 0.19 and 0.25, respectively, highlighting challenges in adapting to complex

reasoning and multiple-choice formats.

Compared to the baseline LLama-3.2-7B, Mistral-7B achieved better PubMedQA accuracy (0.71 vs. 0.485 pre-finetuning) and precision after fine-tuning but struggled similarly on MedQA and MedMCQA. While fine-tuning slightly improved explanation quality on PubMedQA, the shallow, generic responses for MedQA and MedMCQA persisted. These findings emphasize the need for better-aligned training data and advanced techniques to address real-world medical QA complexities.

Extension Results: Table 3 shows performance comparisons. Each row represents a configuration, and each column corresponds to a test dataset. Despite not reaching state-of-the-art results, the extensions demonstrate the feasibility of creating smaller, resource-friendly models that approach baseline accuracy levels.

Model	**PubMedQA Acc**	**MedQA Acc**	**MedMCQA Acc**
Llama-3.2-7b Baseline -26GB	0.4850	0.2050	0.2634
Pruned Only -20.8GB	0.1100	0.1200	0.1562
Quantized Only (LLama 3.2 - 6.5GB)	0.4800	0.2020	0.2600
Quantized + Fine-Tuned (LLama 3.2 - 6.5GB)	0.1500	0.2145	0.2606
Quantized Only (Mistral 3.2 - 1.8GB)	0.7100	0.1900	0.2500
Quantized + Fine-Tuned (Mistral-7B -1.8GB)	0.6400	0.1900	0.2660

Table 3: Performance of various model configurations on test sets. Pruning reduces accuracy but improves efficiency, while quantization preserves accuracy better while still improving efficiency. Fine-tuning recovers performance lost due to pruning and may slightly improve upon quantized results.

In comparison to the strong Llama-3.2-7b baseline, the results show a mixed picture. Quantized models, particularly the Mistral-7B variant, demonstrated competitive accuracy on PubMedQA and MedMCQA, nearing or exceeding baseline performance in some cases (e.g., PubMedQA accuracy of 0.71 vs. baseline 0.4850). However, on MedQA, both quantized and fine-tuned models struggled to match baseline performance, highlighting the challenge of adapting to complex datasets.

One notable highlight is the precision achieved by the fine-tuned Mistral-7B model, particularly on PubMedQA, where precision rose significantly to 0.79 from 0.70. In the medical domain, precision is critical as it reflects the model's ability to avoid false positives, ensuring that responses are accurate and reliable. This is especially important in clinical applications where incorrect answers can lead to harmful outcomes or misinformed decisions.

Fine-tuning yielded modest improvements over quantized-only models, with MedQA accuracy rising to 0.2145 for LLama-3.2 and MedMCQA increasing to 0.2660 for Mistral-7B. However, on PubMedQA, fine-tuning slightly reduced Mistral-7B's accuracy (0.6400 vs. 0.7100).Mistral-7B's strong precision, particularly in PubMedQA (0.79), underscores its potential for high-reliability medical applications. While the extensions did not fully match the baseline in all cases, they highlight viable trade-offs between efficiency and performance for tasks demanding high precision.

4.3 Error Analysis

We examined errors made by our best performing systems: Quantized + Fine-Tuned Llama 3.2 and Quantized + Fine-Tuned Mistral-7B.

Quantized + Fine-Tuned Llama 3.2: Common errors involved subtle clinical reasoning or ambiguous medical terms. About 20% of errors were due to confusion in interpreting complex conditions, and around 15% involved misunderstanding medical abbreviations.

Compared to the published baseline, our fine-tuned quantized model corrected some straightforward factual mistakes. For instance, if the baseline frequently confused treatments for related conditions, the fine-tuned model got these correct more often. However, the model still struggled with multi-step reasoning questions that required combining multiple pieces of clinical knowledge. While the compression methods introduced some noise, careful fine-tuning mitigated many of these issues.

- Example:
  - Question: "A 7-year-old boy is brought to his pediatrician's office for a follow-up visit. He was diagnosed with asthma when he was 3 years old and has since been on treatment for the condition. He is currently on a β-agonist inhaler because of exacerbation of his symptoms. He has observed that his symptoms are more prominent in springtime, especially when the new flowers are blooming. His mother has a backyard garden and whenever he goes out to play there, he experiences chest tightness with associated shortness of breath. He has been advised to take more precaution during this seasonal change and to stay away from pollen. He is also being considered for an experimental therapy, which attenuates the activity of certain mediators which cause his asthmatic attack. The targeted mediator favors the class switching of antibodies. A reduction in this mechanism will eventually reduce the exaggerated

response observed during his asthmatic attacks, even when exposed to an allergen. Which of the following mediators is described in this experimental study?"

  - Options: ['IL-5', 'IL-2', 'IL-10', 'IL-13', 'IL-4']
  - True Answer: IL-4
  - Model Prediction: IL-5

Quantized + Fine-Tuned Mistral-7B: When correct, the model often provides explanations referencing studies, patient populations, or interventions. However, these explanations can be factually dubious, and hence a large portion of errors arise when the model provides explanations which are superficially plausible but lack true domain-specific reasoning or evidence.

- Example:
  - Question: "Does self-efficacy mediate the relationship between transformational leadership behaviors and healthcare workers' sleep quality?"
  - True Answer: Maybe
  - Model Prediction: Yes
  - Explanation: "Sleep quality is a construct in itself, it is not an outcome of another variable."
- The explanation does not address the relationship between the variables and instead provides generic reasoning.

In addition, the model frequently selects incorrect answers for complex multiple-choice questions. This is often due to weak reasoning chains or the inability to differentiate between distractor options. An example of this is shown below from the MedQA dataset:

- Example:
  - Question: "A 3-month-old infant... has a high-pitched holosystolic murmur. The most likely cause of this patient's symptoms is associated with which of the following abnormalities?"
  - Options: ['22q11 deletion', 'Deletion of genes on chromosome 7', 'Lithium exposure in utero', 'Maternal alcohol consumption', 'Retinoic acid exposure in utero']
  - True Answer: 22q11 deletion

  - Model Prediction: Lithium exposure in utero

- ”Lithium exposure in utero” is selected, but the model fails to connect lithium exposure to the symptoms described. Instead, the correct answer involves 22q11 deletion, which is associated with congenital heart defects.

5 Conclusions

This project explored compression techniques—quantization and fine-tuning—on medical language models (LLama-3.2-7B and Mistral-7B) to balance accuracy and efficiency for tasks like PubMedQA, MedQA, and MedMCQA. Quantization preserved most baseline performance, with Mistral-7B achieving 0.71 accuracy on PubMedQA, outperforming LLama-3.2-7B in key metrics. Fine-tuning using MedQuAD yielded modest gains for MedMCQA (0.2660 for Mistral-7B) and MedQA (0.2145 for LLama-3.2), though it reduced PubMedQA accuracy (0.64 vs. 0.71 for Mistral-7B), highlighting dataset alignment challenges.

Precision improvements (e.g., 0.79 for Mistral-7B on PubMedQA) underscore the potential for high-reliability applications in medical contexts, where minimizing false positives is critical. While the compressed models did not surpass the baseline in all tasks, they achieved competitive performance with significantly reduced resource demands. This work underscores the feasibility of efficient, privacy-preserving medical NLP solutions for on-device deployment, with future directions including better-aligned fine-tuning datasets and advanced reasoning strategies for complex tasks.

Acknowledgements

We thank the course staff for their guidance, and the authors of the PubMedQA, MedQA, and MedMCQA datasets for making their resources publicly available.

A Appendix

Below is example code used for tokenizing input data:

```
from transformers import AutoTokenizer

tokenizer = AutoTokenizer.from_pretrained("meta-llama/Llama-2-7b-hf")
def preprocess(question, context):
    input_text = f"Context: {context}\nQuestion: {question}\nAnswer:"
```

```
    return tokenizer(input_text, truncation=True, padding=True,
                     max_length=512)
```

References

[Hinton et al.2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv:1503.02531*.

[Han et al.2015] Song Han, Jeff Pool, John Tran, and William Dally. 2015. Learning both weights and connections for efficient neural network. In *NIPS*.

[Jin et al.2019] Qiao Jin, Rodney Kinney, and Zhiyong Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *EMNLP Workshop BioASQ*.

[Lee et al.2019] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, et al. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36(4).

[McMahan et al.2017] H. Brendan McMahan, Eider Moore, Daniel Ramage, et al. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *AISTATS*.

[Pal et al.2022] Ankit Pal, et al. 2022. MedMCQA: A Large-scale Multiple-Choice Dataset for Medical Domain QA. In *Findings of ACL*.

[Zhang et al.2020] Yicheng Zhang, et al. 2020. MedQA: A Large-scale Medical Domain QA Dataset. In *Proceedings of ...*.