Lightweight LLMs for Vascular Surgery Education

Faith Choe
MIT
faithfc@mit.edu

Hope Dargan
MIT CSAIL
hoped@mit.edu

Vivian Ding
MIT
vyd@mit.edu

Sasha Jovanovic-Hacon
MIT
ajovanov@mit.edu

Daniel Kassavin, MD, FSVS
Cambridge Health Alliance

Abstract

Existing research on medical large language models (LLMs) has primarily focused on large, proprietary models and general-domain benchmarks, with limited attention to vascular surgery or lightweight models suitable for clinical education. In this work, we curate a benchmark of 451 vascular surgery multiple-choice questions and systematically evaluate the performance of small open-source LLMs across three intervention levels: base model performance, retrieval-augmented generation (RAG), and retrieval-augmented fine-tuning (RAFT). Our goal is to assess whether lightweight models, when properly tuned, can provide accurate and explainable answers in a domain where safety and interpretability are paramount.

We find that while state-of-the-art models, like GPT-4o and gemini-2.0, achieve 50-52% accuracy, small models lag behind, with our best performing fine-tuned models reaching 29%. RAG performance was highly sensitive to retrieval quality, and RAFT improved performance only for smaller models when paired with short, focused context and explanation-based training. Medically fine-tuned models underperformed on multi-answer questions, suggesting overfitting to single-answer benchmarks. These findings highlight the potential and limitations of lightweight models for vascular education and motivate future work on high-quality domain data and instruction-following capabilities.

1 Introduction

Vascular surgery students are likely to encounter new patient scenarios as part of their clinical training. While students are encouraged to consult medical mentors or textbooks and clinical guidelines to answer their questions, mentors may have limited availability and textbooks may be difficult to search and apply to real-world scenarios. Large Language Models (LLMs) offer a potential pathway towards allowing students to quickly receive answers and thus learn more efficiently, but pose risks in terms of providing insufficient answers or hallucinating and misinforming a student in ways that could potentially cause harm to patients. Thus, evaluating the accuracy of current LLMs and finding the best models and methods to improve the quality and accuracy of their answers is vital.

Although a variety of medical question-answer benchmarks have been developed, there is no specific benchmark for vascular surgery that has been approved for use with LLMs and there has been limited research on tuning LLMs for the domain of vascular surgery. Previous LLM studies related to vascular surgery have focused on evaluating large LLMs. In contrast, we are interested in models that are capable of running locally offline: Large cloud-based LLMs raise concerns about HIPAA compliance should they be inadvertently used with patient data. They also require both internet connectivity and dramatically more processing power, which may be unavailable in some settings.

In this paper, we systematically evaluate the feasibility of using a lightweight, locally-run AI chatbot for vascular surgery education, targeting medical students, physician assistant students, and junior surgical residents. Such a tool ought to be capable of responding to a variety of patient cases with guidance on differential diagnosis or next steps that need to be taken, or walk through the steps of a specific procedure.

We curated a novel benchmark of 451 multiple choice questions (MCQs) on vascular surgery from a variety of sources and evaluated a variety of model sizes, sources, and tuning methods. Our study sought to answer the following research questions.

MLHC Final Project, May 9, 2025, Cambridge, MA, USA.

1. What is the baseline accuracy of the selected models on multiple choice questions? Are any question or model characteristics associated with higher or lower answer accuracy?
2. How do various methods including prompt engineering, Retrieval-Augmented Generation (RAG), and Retrieval-Augmented Fine Tuning (RAFT) affect model performance on multiple choice questions?

The contributions in this paper are a novel benchmark suite designed to mimic questions vascular surgery students might ask in real clinical scenarios, and a comparison of performance by various models and tuning methods on these benchmarks. Progress in this paper lays a foundation for future evaluation of new models and generation methods in this application domain. Source code can be found here.

2 Related Work

This section summarizes the results of previous work in a variety of subjects related to our investigation, including LLM performance in the domain of vascular surgery, methods such as prompt engineering, RAG, fine-tuning, and RAFT, and query characteristics that impact LLM performance.

2.1 Vascular Surgery LLM Studies

Several studies have explored using RAG to improve LLM performance on the Vascular Education and Self-Assessment Program (VESAP), a multiple choice vascular surgery question bank for trainees to use while studying for their licensing exams.[34] Feridooni et al. compared the performance of ChatGPT 3.5, 4, 4o, and VASC.AI (ChatGPT with a RAG containing over 200,000 clinical abstracts and surgery guidelines) on 244 text-based VESAP questions and found that accuracy varied from as low as 55% on ChatGPT3.5 to 94% on VASC.AI. Additionally, they found that across the models roughly three-fourths of incorrect answers were due to information errors, with the remaining classified as logic errors.[17] Vien et al. compared the performance of ChatGPT4o, Claude, LLAMA3 (70 billion parameters), with LLAMA3 models that had a variety of RAG setups on 680 VESAP questions and found that ChatGPT4o initially had the highest accuracy of 72%. While baseline LLAMA3 had only 61% accuracy, one LLAMA3 + RAG configuration based on AudibleBleeding podcast transcripts matched ChatGPT4o's performance, suggesting that smaller, customized, open-sourced LLMs can perform similarly to larger flagship models with properly curated source material.[30]

Other studies have explored large LLMs vascular surgery question answering abilities based on more open-ended prompts. Haidar et al. used ChatGPT 3.5 to generate information on three common procedures, and compared the results to information provided to patients in informational leaflets.[20] They concluded that AI-generated information about the procedures was currently poor in terms of both quantitative readability scores and quality of information compared to human-generated information. On the other hand, Javidan et al. found that ChatGPT 4 provided appropriate recommendations at a college-text level to 38/40 clinical scenario open-ended questions.[22] Le et al. found that in 25 fictional case studies ChatGPT, Falcon 40B, and Bard were far better at identifying whether the situation was an emergency than determining the appropriate next step in terms of care.[24]

2.2 LLM Tuning Methods

2.2.1 Prompt Engineering

Maharjan et al. found that applying several proven prompt engineering techniques such as kNN few-shot COT (finding the 5 nearest training questions and adding them along with chain of thought reasoning explanations before the actual query) and ensemble/self-consistency (prompting the model several times and taking the majority vote) on a Yi 34B base model led it to have state of the art performance on the MedQA question bank without additional fine tuning. Combining these "OpenMedLLM" prompt engineering techniques with open source models such as Yi 34B and Meditron resulted in improvements between 5-12% on question benchmarks.[28] Singhal et al. found that self-consistency and few-shot prompting improved MedPalm performance, but that chain of thought (COT) did not.[32]

2.2.2 Retrieval-Augmented Generation

Retrieval-Augmented Generation is a technique for improving LLM performance by providing relevant documents along with the initial query. It is particularly useful in knowledge-intensive scenarios, or when the knowledge required is domain-specific, as supporting information is directly brought into context. Both Feridooni et al. and Vien et al. found that RAG significantly improved multiple choice question answering performance on ChatGPT and other models.[17, 30] However, another recent RAG study by Chen et al. found that a RAG only had a limited impact on

Japanese medical question answering using small-scale open source models, and that the effectiveness of RAG is sensitive to the quality and relevance of the external retrieved content.[16]

2.2.3 Fine Tuning

Similarly, instruction fine-tuning LLMs for domain specific tasks has had mixed results. Singhal et al. reported that both MedPalm and MedPalm2 had improved performance due to careful instruction fine-tuning.[32, 33] However, studies have found that fine-tuning base models increases safety risks.[37] Additionally, instruction tuning is computationally expensive and can lead to catastrophic forgetting / knowledge forgetting.[27] Jeong et al. recently compared ten medical (pretrained on public biomedical corpora) LLMs against their corresponding base models, and found that nearly all medical LLMs failed to consistently improve over their base models.[23]

2.2.4 Retrieval-Augmented Fine Tuning

Retrieval-Augmented Fine Tuning (RAFT) is a recent method designed to enhance the reasoning capabilities of langauge models in domain-specific RAG settings.[38] Unlike traditional fine tuning or RAG methods, RAFT explicitly trains the model to reason over retrieved documents. During training, the model sees a mixture of queries with either relevant ("golden") or irrelevant ("distractor") pieces of the retrieved context. It is taught to generate chain-of-thought (CoT) style answers grounded in the golden sources while ignoring distractors. This improves not only factual accuracy, but also robustness to imperfect retrieval at test time. The RAFT approach has shown strong improvements across domains like domain-specific QA and API understanding compared to supervised fine-tuning or RAG alone.

2.3 Query Characteristics and LLM Accuracy

In addition to model architecture and tuning strategies, query characteristics have been shown to influence LLM performance. Prompt format, such as JSON, plain text, Markdown, or YAML, can significantly impact the performance of GPT models across tasks like multiple-choice questions, code generation, and translation.[21]. Larger models like GPT-4 were more robust to prompt format changes, but no single format worked best across all models and tasks.

Prompt length has also emerged as a key factor. Following an investaigation of how prompt length affects the performance of LLMs on domain-specific tasts, researchers found that longer prompts with more background information generally improve precision and recall scores.[26]. However, others caution that overly long prompts (i.e. padding prompts beyond 500 tokens) may overwhelm small-scale models, especially if additional tokens introduce irrelevant or distracting information.[25]. Notably, Levy et al.'s study involved answering questions solely from the input text, so adding extra text naturally increased distraction. In contrast, our longer prompts aim to enhance instructions rather than embed competing content.

LLMs have also been shown to struggle with negation, for example completing both sentences like "Ibuprofen is a kind of ____." and "Ibuprofen is not a kind of ____." with similar kinds of words (e.g., medicine, drug, painkiller).[36] We expect that certain kinds of negated question statements like one that asks "Which of the following is not true?" might suffer in terms of performance.

We apply these insights to evaluate lightweight LLMs for vascular surgery education, seeking prompt designs that maximize accuracy while preserving explainability and efficiency.

3 Methods

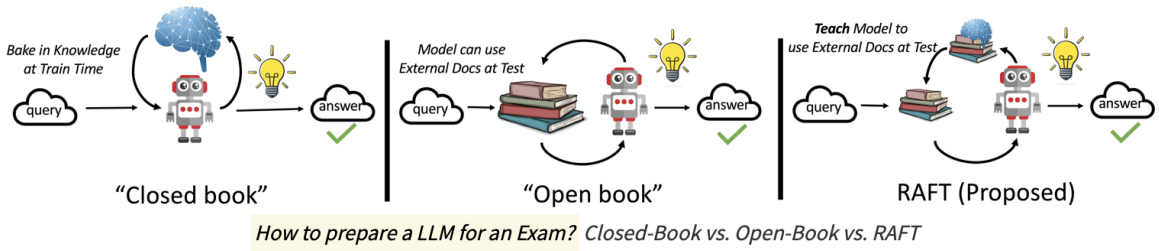

Figure 1: Comparison of base, RAG, and RAFT-style LLMs for question answering (adapted from[16]).

To assess how different intervention methods affect model performance on vascular surgery question answering, we evaluated each model under one or more of three setups: **Base** (no augmentation or tuning), **RAG** (retrieval-augmented generation), and **RAFT** (retrieval-augmented fine-tuning). These interventions represent a progression from no customization to lightweight adaptation using retrieval, and finally to full task-specific fine-tuning grounded in retrieved evidence.

3.1 Model Selection

3.1.1 Small Model Selection for Fine-Tuning

We selected nine lightweight models from the Hugging Face Open Medical LLM Leaderboard and evaluated their baseline performance on vascular surgery questions.[29] The selection prioritized models small enough to theoretically run on consumer-grade hardware, with parameter counts ranging from 124 million to 2.78 billion.

Each model was tested on 13 multiple-choice questions from the APDVS Medical Student Curriculum.[15] To assess robustness and generalizability, we created three distinct prompt variations for each question, yielding a total of 39 prompt-model evaluations. Our goal was to explore how different prompt structures might affect response quality across models of varying architecture and size.

Among the nine models, we included two pairs of base and domain-fine-tuned variants to explore the impact of domain adaptation. For example, **GPT2 (124M)** and **GPT2-PMC** were fine-tuned on roughly 8000 question-answer pairs from PubMed.[8, 7] **TinyLlama-1.1B** and **MeLT-TinyLlama-1.1B** were fine-tuned on medical data.[11, 5] We also evaluated two models of different sizes from Apollo's Freedom Intelligence family; FreedomIntelligence/Apollo-0.5B (**Apollo-0.5B**- 464M parameters) and FreedomIntelligence/Apollo-2B (**Apollo-2B**- 2.51B parameters).[12, 13] This was to examine performance across scaled versions of the same architecture.

Models were evaluated both quantitatively and qualitatively. Quantitative evaluation was based on the number of correct answers out of 39. Qualitative feedback was provided by Dr. Daniel Kassavin, a board-certified vascular surgeon, who assessed response relevance, clinical accuracy, and clarity.

Some models demonstrated strong performance as lightweight baselines, while others failed to generate coherent or clinically relevant answers. These findings informed the selection of two small models for further fine-tuning: **Apollo-0.5B** and johnsnowlabs/JSL-MedPhi2-2.7B (**JSL-2.7B**).[12, 6]

Model	# Correct	Qualitative Summary
JSL-Medphi (2.78B) [6]	23	Most coherent and clinically relevant responses.
Apollo (2.51B) [13]	15	Similar to smaller Apollo; solid clinical reasoning, slightly verbose.
Apollo (464M) [12]	15	Matched Apollo-2B despite smaller size; efficient with MCQs.
MeLT-TinyLlama (1.1B) [5]	9	Often generated follow-up or hypothetical questions.
TinyLlama (1.1B) [11]	0	Failed to produce meaningful answers.
GPT2 (124M) [8]	0	Gave off-topic narrative-style outputs.
GPT2-PMC (124M) [7]	0	Failed to follow MCQ format.
Phi-1.5 (1.418B) [14]	0	Hallucinated random answers without explanation.
MEGA (126M) [10]	0	Unable to respond in MCQ format.

Table 1: Initial model evaluation results on 39 questions. Parameter counts are in billions (B) or millions (M) as noted.

3.1.2 Broader Model Set for Comparative Evaluation

We extended our initial small-model selection to a broader set of six models. The selected models were chosen to represent a diverse range of architectures, access methods, and instruction-following capabilities, and to benchmark under three intervention levels: Base (no retrieval or additional fine tuning), RAG, and RAFT.

The selected models are ChatGPT-4o (**GPT-4o**), gemini-2.0-flash-001 (**gemini-2.0**), microsoft/Phi-4-mini-instruct (**Phi-4B**), **Apollo-0.5B**, **JSL-2.7B**, and skumar9/Llama-medx_v3.2 (**LLAMA-Med-7B**). [2, 3, 14, 12, 6, 9] **GPT-4o** and **gemini-2.0** represent the performance of state-of-the-art (SOTA) LLMs in this task. Both were accessed via API and evaluated in their base forms to establish a strong upper-bound baseline. **LLAMA-Med-7B** was included as a

strong domain-specific baseline for base and RAG evaluation. **Phi-4B** is an open source instruction-tuned model from Microsoft. Because it has already seen general instruction fine-tuning, combinbing this model with RAG positions it as a candidate for out-of-the-box reasoning with minimal adaptation. **Apollo-0.5B** and **JSL-2.7B** were selected for full fine-tuning (including RAFT) based on their performance in the smaller-scale evaluation described in Section 3.1.

Model	Base	RAG	RAFT
GPT-4o	✓		
gemini-2.0	✓	✓	
LLAMA-Med-7B	✓	✓	
Phi-4B	✓	✓	
Apollo-0.5B	✓	✓	✓
JSL-2.7B	✓	✓	✓

Table 2: Model evaluation by intervention type. Each model was tested under one or more of three configurations: Base (no retrieval or fine-tuning), RAG (retrieval-augmented generation), and RAFT (retrieval-augmented fine-tuning).

3.2 Prompt Tuning

3.2.1 Prompt Format Evaluation

Before selecting a standardized prompt format for training and evaluation, we conducted a series of prompt engineering experiments on three of our smallest candidate models: Apollo-0.5B, JSL-2.7B, and LLAMA-Med-7B. Although LLAMA-Med-7B was ultimately not selected for fine-tuning due to resource constraints, its strong baseline performance made it a useful point of comparison.

Each model was evaluated on a shared set of 167 multiple-choice questions using four different prompt types:

- **Original**: Direct copy of the question and answer choices, formatted as in source material.
- **Medical Context**: The same as "Original" but with a short description of the domain (e.g., "This is a vascular surgery question").
- **Instructions**: The "Medical Context" format plus an explicit instruction (e.g., "Select the best answer choice. There may be more than one correct option.").
- **Few-Shot**: Multiple example Q&A pairs shown before the target question.

Model	Prompt	Num Correct	% Accuracy
Apollo-0.5B	original	33	20
Apollo-0.5B	medical context	32	19
Apollo-0.5B	**instructions**	**35**	**21**
Apollo-0.5B	few shot	1	1
JSL-2.7B	original	54	32
JSL-2.7B	medical context	56	34
JSL-2.7B	**instructions**	**59**	**35**
JSL-2.7B	few shot	56	34
LLAMA-Med-7B	original	61	37
LLAMA-Med-7B	**medical context**	**73**	**44**
LLAMA-Med-7B	instructions	67	40
LLAMA-Med-7B	few shot	59	35

Table 3: Prompt format evaluation across 167 multiple-choice questions. Bolded values indicate best-performing prompt for each model.

These results reveal several key trends. First, prompt format has a substantial impact on model performance, especially for smaller models. Apollo-0.5B showed the most sensitivity, performing best with instruction-style prompts and collapsing under the few-shot condition, often hallucinating new, unrelated questions. This behavior suggests that few-shot prompting may overload or confuse small models with limited context windows.

Second, simply appending a medical context line to the original format provided a minor benefit for larger models like LLAMA-Med-7B, but was not sufficient for improving the performance of Apollo-0.5B or JSL-2.7B. The best results

overall were achieved using instruction-based prompts that combined domain context with explicit task guidance (e.g., "Select the best option. There may be more than one correct answer."). This format was especially helpful in eliciting structured answers from smaller models without requiring training.

Based on these results, we standardized our prompt format for all subsequent experiments using an instructional format with embedded domain context. This format provided consistent gains in performance across models and offered a more robust foundation for training and evaluation.

3.2.2 Final Prompt Template

You are a vascular surgery medical expert. Answer the following multiple choice question based on the instructions below:

```
### Question: {question}
### Choices:
A. ...
B. ...
C. ...

### Answer: [model writes selected answer here]
### Reason: [model writes explanation here]
```

3.3 Retrieval-Based Augmentation Strategies

3.3.1 RAG

We constructed a custom corpus by collecting documents from vascular surgery training materials, including the APDVS curriculum and the most recent clinical guidelines from the Society for Vascular Surgery.[15, 35]To simulate retrieval, we used a dense retrieval system built with FAISS, embedding all documents and retrieving the top-k results for a given question. The documents were processed into chunk sizes of 100 and 500 tokens, with 20 and 100 tokens of overlap, respectively.

To ensure that models utilized the retrieved material effectively, retrieved chunks were always inserted before the question and answer prompt. This structure was critical as models were more likely to ground their answers in provided documents when context was placed upfront, and were less likely to hallucinate new context or additional questions, especially in small or instruction-tuned models.

3.3.2 RAFT

To fine-tune a model for vascular surgery question answering with robust retrieval-based reasoning, we implemented a domain-specific version of the RAFT methodology described by Zhang et al.[38] Our implementation is designed to help an LLM learn to answer questions using only relevant document excerpts generated with the RAG, while learning to ignore irrelevant content.

Retrieval Our corpus for RAFT is as described in section 3.3.1. For each question, we employ two retrieval strategies, one for each training objective. For golden examples, we retrieve the top-k relevant chunks for golden examples. For contrastive examples, we retrieve k distractor chunks randomly sampled from the corpus (excluding relevant ones).

Prompt Format Each training example consists of a prompt and a target. The prompt includes retrieved documents (formatted as `Document 1: ...`, etc.), followed by the question, multiple labeled answer choices (e.g., `A.`, `B.`, etc.), and ends with `Answer:` to signal where the model should begin generating. The target string follows a structured format: `##Answer: {letter} ##Reason: {explanation}`

Training Setup For each question, we create two training examples: one using only relevant context retrieved by our RAG system (golden), and one using only distractor documents randomly sampled from the rest of the corpus. All examples follow the prompt and target format described above. Distractor-only examples always used the fixed string `"##Reason: No sufficient information available."`

To populate the `##Reason:` field for golden examples, we used a two-phase approach. In the first iteration, we extracted source explanations for 52 of the 451 training questions. For golden examples without explanations, we used the placeholder `[Retrieved context can be used to explain the answer]` to encourage the model to ground its reasoning in context. In the second iteration, we generated explanations for all golden examples using gemini-2.0, conditioning each generation on the correct answer and the retrieved context. These synthetic rationales helped the model more directly learn to cite and reason from domain-specific sources.

This contrastive setup teaches the model to rely on relevant evidence when it exists and to abstain when no sufficient information is provided. While we do not yet mix golden and distractor chunks in a single prompt, this training scheme sets a foundation for future mixed-context fine-tuning.

RA-Style Answer Generation At inference time, we expect the model to replicate the behavior learned during training: identify the correct answer from retrieved context and justify it using supporting evidence. The structured format (`##Answer`, `##Reason`) encourages generation of interpretable explanations aligned with the provided documents. By exposing the model during training to examples where explanations are grounded in relevant retrieval, we aim to improve both answer accuracy and transparency at test time.

4 Data and Experiment setup

Instead of using VESAP, we opted to curate our own question bank.[1] Our question bank was designed to align with the types of queries that prospective users might pose to the model. We prioritized sourcing questions that reflect real-world clinical reasoning, including case-based scenarios and applied clinical knowledge, to simulate the kinds of patient-centered or diagnostic questions a user might ask in practice.

To achieve this, we systematically searched publicly available resources including vascular surgery textbooks, sample exam questions, professional guidelines, and academic publications. We specifically sought materials that contained multiple-choice questions, and clinical case studies. Under the guidance of vascular surgeon Dr. Daniel Kassavin, we verified the credibility and educational value of each source, selecting those with clinically accurate, well-structured, and practically relevant content.

Source	# Questions
American Board of Vascular Medicine [1]	4
APDVS Medical Student Curriculum [15]	61
Vascular Surgery: Cases, Questions and Commentaries (4th Ed., 2018) [19]	317
PrepLadder High Yield Cardiothoracic and Vascular Surgery Questions [4]	13
Royal Australian College of Surgeons Vascular Surgery Sample Exam [31]	56

Table 4: Summary of Resources Utilized for Question Bank Development

The majority of questions were manually selected through close reading and extraction of relevant entries. In some cases, basic parsing scripts were used to assist in extracting structured question-answer pairs from formatted documents. All entries underwent a final manual review to ensure consistency, accuracy, and relevance to clinical care.

To verify that our RAG implementation works as expected, we extracted and tested models on questions from the APDVS curriculum, which is included in the RAG corpus. These questions were used only for this sanity check; the remainder of our analyses exclude them.

Train/Test Split To create balanced training and test data sets, we stratified our 451-question dataset by both question source and the presence or absence of an explanation. Additionally, a subset of questions consisted of reworded pairs that test the same concept. To avoid data leakage, these linked pairs were always assigned to the same split. We ensured that each stratum was proportionally represented in both sets, while preserving pair integrity. The final split consists of 365 training questions and 86 test questions. Because each training question is used to create two training examples (one with relevant documents and one with distractors), the RAFT training set ultimately contains 730 examples.

Model	Num Correct / 451	% Accuracy
GPT-4o	224	50
gemini-2.0	**235**	**52**
Phi-4B	195	43
LLAMA-Med-7B	162	36
JSL-2.7B	152	34
Apollo-0.5B	109	24

Table 5: Base model evaluation results on 451 multiple choice vascular surgery questions.

5 Results

Table 5 shows the results of running all 451 multiple choice questions on the six models we selected for final evaluation. We found that the state of the art gemini-2.0 performed the best with 52% accuracy, slightly outperforming GPT-4o. Surprisingly, Phi-4B outperformed the medically tuned models, including LLAMA-Med-7B, which had more parameters.

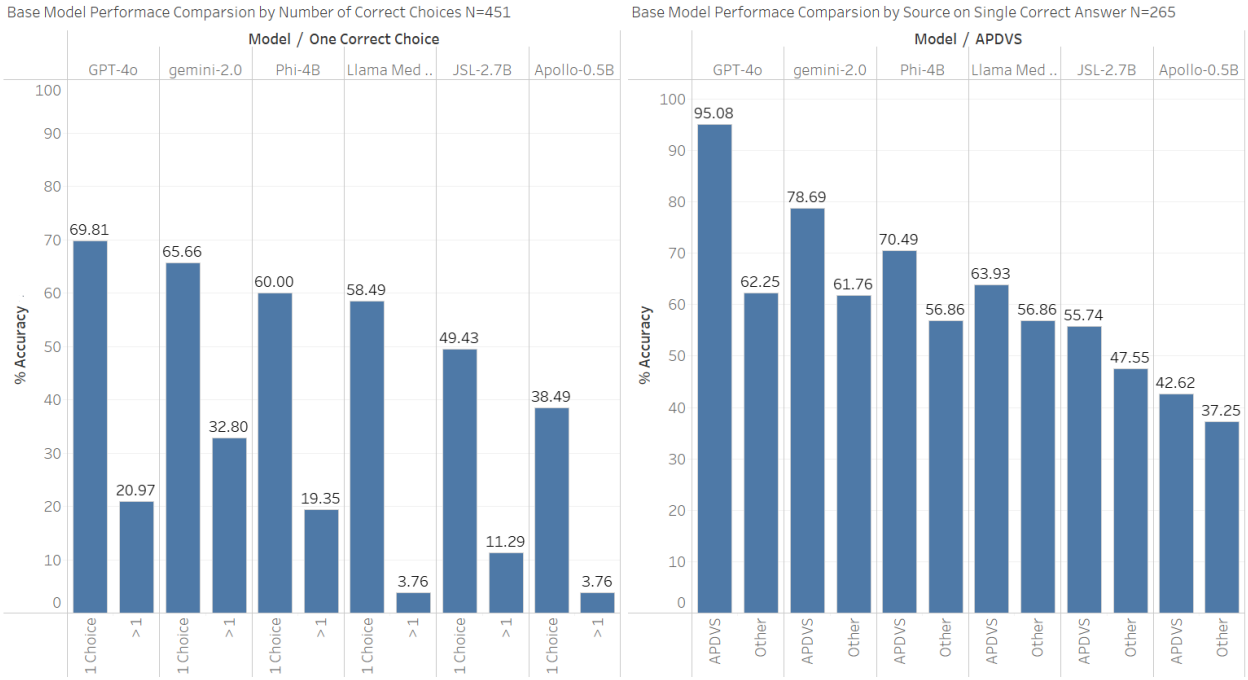

Figure 2: The left bar chart shows the base model performance on questions with exactly one correct answer choice versus questions with more than one correct answer choice. The right bar chart shows that for the 265 questions with one correct answer choice, the base models performed better on the APDVS questions than the questions from other sources.

We also investigated how base model performance varied across a number of question characteristics including source (see Table 4), number of correct answer choices (1 (265), 2 (78), 3 (57), 4 (36), 5 (12), 6 (1), and 7 (2)), question topic (lower extremity arterial disease (128), venuous disease (126), aortic disorders (107), carotid disorders (32), endovascular surgery (9), other (49)), and question type (general knowledge (235) or clinical case study (216)). Two relevant factors appear to be the number of correct answer choices and the question source, as shown in Figure 2. In particular, the medically tuned models had particularly low performance on questions with more than one correct choice, suggesting that they may be overfitting to medical QA benchmarks, which typically have a single correct answer choice.

[1]Note to instructors: When we asked the Society of Vascular Surgery whether we could use the VESAP question bank for our study, they said no and that the prior studies involving VESAP had used the question bank without authorization. Because of this and the $825 access fee, we opted to create our own question bank.

All model's better performance on APDVS questions may suggest that this textbook was included in training material or that the questions were easier or phrased in a way that better prompted the model.

5.1 RAG

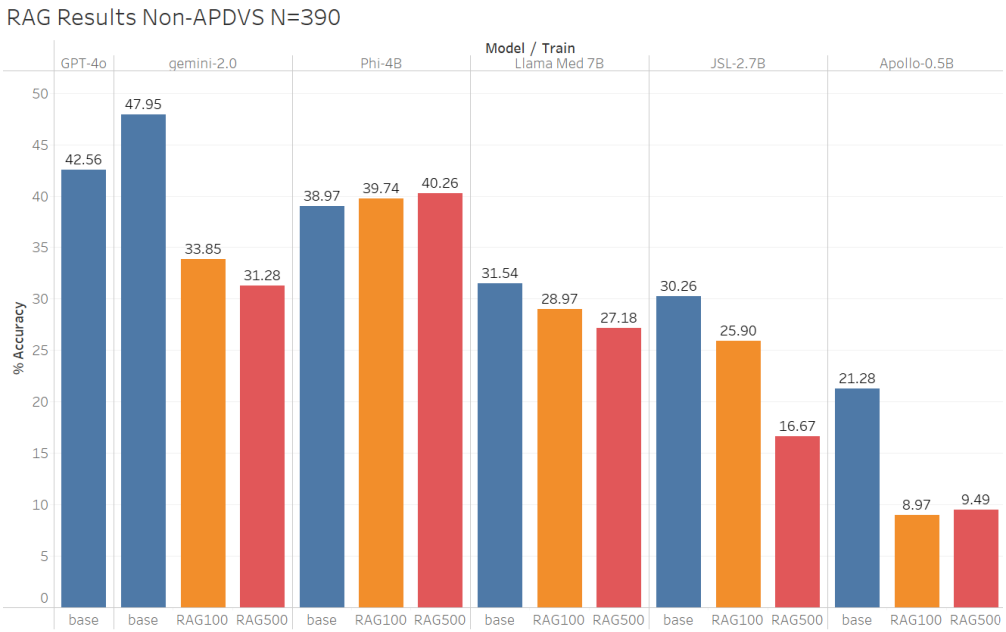

Figure 3: The bar chart shows the base model performance compared to base model provided with context from a 100 token RAG (orange bars) or a 500 token RAG (red bars). APDVS questions were excluded because the RAG contained this textbook.

We then measured model performance with providing RAG context with 100 token chunk sizes and 500 token chunk sizes as shown in Figure 3. Because the RAG included the APDVS textbook, we expected and found that RAG improved the performance for all models on those questions, particularly the 500 token chunk size, which was generally long enough to include both the question and the answer from the text itself. However, on non-APDVS questions, all models except Phi-4B saw a drop in performance. However, Phi-4B's improvement was small—it went from 152 correct answers to 155 correct with RAG 100 and 157 with RAG 500.

5.2 RAFT Performance and Instruction Fine-Tuning

Our experiments with RAFT revealed several notable patterns regarding model size, context length, and the use of context-grounded explanations (see Figure 4).

Apollo-0.5B benefited significantly from RAFT compared to the untuned baseline. The most effective setup for Apollo-0.5B used 100-token context chunks with training explanations generated by GPT-4o, (`RAFT_100_gpt`), achieving a 10.5% absolute improvement over the base model (28.95% vs. 18.42%). In contrast, performance dropped when using 500-token chunks, suggesting that shorter, focused context is more beneficial for small models, which may become prone to hallucination when provided with excessive information.

JSL-2.7B showed negative gains from RAFT. All RAFT variants underperformed compared to the model's base accuracy of 22%, with most RAG and RAFT methods reaching only between 9% and 13%. One plausible explanation is that instruction fine-tuning was less effective for larger models in our low-data setting, as our dataset (365 training questions $\times 2$ document variants) may not have been sufficient for them to leverage their full capabilities. In some cases, fine-tuning may have restricted generalization by reinforcing brittle reasoning on a narrow domain distribution.

Finally, we observed that using GPT-4o to generate explanations grounded in retrieved content sometimes improved performance. For example, `RAFT_100_gpt` outperformed `RAFT_100` on Apollo-0.5B (28.95% vs. 22.37%) and `RAFT_500_gpt` outperformed `RAFT_500` on JSL-2.7B (13.16% vs 11.84%).

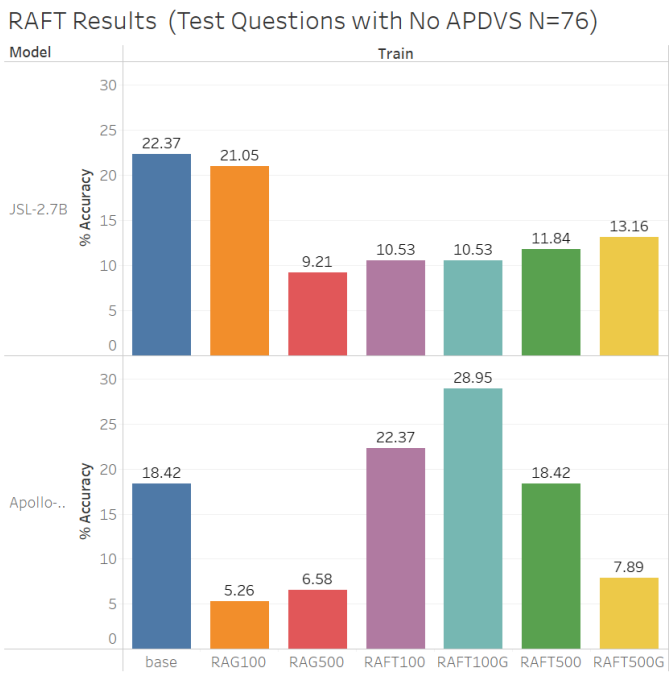

Figure 4: Base model performance compared to RAG performance and RAFT with various training configurations including varying token sizes (100 and 500) and GPT generated explanations (G). Questions that were used for training and APDVS questions were excluded.

Overall, our results may suggest that lightweight models benefit most from fine-tuning when context is tightly scoped and explanations are grounded, while larger models may require more data or different tuning strategies to fully realize their potential in domain-specific question answering.

6 Discussion

Overall, regardless of method, all models performed quite poorly on our curated multiple choice questions. Even large models such as gemini-2.0 and GPT-4o barely answered more than 50% of questions correctly. In particular, all models correctly answered less than a third of the 190 questions with more than one correct answer choice. We also observed that the medically tuned models were more prone to errors, for example outputting text that simply repeated the prompt, contained additional questions, a different language, random nonsense, or completely irrelevant context.

We experimented with a number of techniques to address these shortcomings, which included adjusting the prompt format and implementing various RAG and RAFT fine-tuning configurations. While Apollo-0.5B saw significant improvement with the addition of `RAFT_100_gpt`, JSL-2.7B had negative gains from RAFT. Most models experienced negative gains from both RAG implementations, but microsoft/Phi-4-mini-instruct, a model that has already been instruction fine-tuned, slightly improved with the additional context. Consistent with Jeong et al., we found that models that were fine-tuned for the medical domain did not perform better than general-purpose models.[23]

6.1 Limitations

We prioritized MCQs due to limited data availability during question bank development and because they allow for more straightforward evaluation. Unfortunately, MCQs are not fully representative of real-world clinical decision making, where physicians must independently generate differential diagnoses and management plans without relying on predefined choices. For example, some questions were phrased using negations (e.g. “Which of the following does not apply?”), which may be more difficult for an LLM to respond to correctly. Language models have been shown to struggle with negation in the past[18], and while the state of the art is improving, there may still be some performance degradation compared to questions phrased only in the positive. Additionally, it is unclear how providing LLMs with multiple choices from which they need to select changes their reasoning.

This study was also limited by the small number of custom questions that we curated. For example, the small training set may be one potential explanation for why RAFT led to improvements on Apollo-0.5B but not JSL-2.7B. Also, because we did not have access to the VESAP question bank, we cannot compare our results to prior work with state of the art models as described in 2.1. Our limited access to GPU also prevented us from trying RAFT on larger models such as LLAMA-Med-7B.

Our study was also likely limited by our small RAG corpus. We faced challenges sourcing high-quality data for RAG: Many useful sources such as textbooks require payment or other permissions to access. While medical research papers are high-quality and often publicly available, they are not always directly relevant to clinical scenarios, and we were concerned that small models could be easily distracted by irrelevant retrieved contexts.

6.2 Recommendations and Future work

Overall, we would not recommend relying on LLMs for vascular surgery education, at least in their current form. Although RAG and RAFT provided modest improvements in some models, none of the models approached an accuracy level that would be acceptable for educational or clinical decision support. Given the relatively low accuracy even after intervention, we believe current models are not sufficiently robust or reliable for independent use in this domain.

First, future work should prioritize evaluating open-ended, free-response questions, as this more closely reflects real-world clinical reasoning. Unlike MCQs, open-ended response require models to synthesize information, justify decisions, and generate differential diagnoses, skills that are essential in clinical practical but harder to evaluation automatically. Designing evaluation frameworks and labeled datasets for these types of responses is a key challenge.

Second, expanding high-quality, domain-specific training resources will be critical to improving model accuracy and generalizability. Our question bank, while curated from reputable sources, included only 451 questions, a relatively small dateset for fine-tuning large models. Access to more diverse, comprehensive, and expertly annotated data could enhance both retrieval-based methods and supervised training.

Finally, future research should continue to explore the tradeoffs between model size, accuracy, and deployment feasibility. While smaller models offer advantages in speed, cost, and local deployment, our findings show a substantial performance gap compared to state-of-the-art large models. Identifying strategies that maintain interpretability and safety while narrowing this gap will be crucial for practical adoption in clinical education.

7 Conclusion

In this study, we systematically evaluated the feasibility of using lightweight LLMs for vascular surgery education. We curated a benchmark of 451 vascular surgery multiple-choice questions and assessed a range of model and interventions, including prompt engineering, RAG, and RAFT.

Our results show that smaller models remain substantially less accurate than state-of-the-art models like GPT-4o and Gemini when used without additional customization, achieving accuracies between 24-36%, compared to 50% for GPT-4o and 52% for Gemini. Even with interventions, the lightweight models still fall short in terms of accuracy and reliability. RAG approaches were highly sensitive to retrieval quality, and improvements from RAFT depended on model size and training setup. Across models, challenges in maintaining accuracy, robustness, and interpretability remain.

These findings highlight the need for continued research into creating lightweight, explainable, and dependable AI tools for clinical education. Future work should prioritize expanding high-quality training and retrieval resources, developing evaluation methods for open-ended clinical reasoning, and exploring ways to balance model size, performance, and deployment feasibility. Despite current limitations, lightweight models remain a promising foundation for building accessible AI-powered educational tools in medicine.

8 Acknowledgments and Member Contributions

8.1 Acknowledgments

We thank Dr. Daniel Kassavin for originating the project idea and serving as our clinical mentor. As the only member of our team with a background in vascular surgery, Dr. Kassavin provided critical guidance throughout the project and met regularly with the team to ensure clinical accuracy and progress.

We also thank Professor David Sontag for his thoughtful feedback during our two progress meetings. His guidance helped us frame our evaluation in a more rigorous and comprehensive way.

8.2 Member Contributions

Sasha led model selection, implemented RAFT training and evaluation, and managed the GitHub repository and documentation.

Vivian helped with question bank development and curated and processed the RAFT document corpus.

Hope evaluated base and RAG model performance, maintained a dataset of nearly 8000 question-answer pairs across over 30 model-implementation configurations, and analyzed model outputs to identify trends.

Faith curated the question bank by searching publicly available sources and manually extracting and formatting questions from the case studies textbook. She evaluated thousands of Q&A model responses for factual accuracy, clarity, and explainability.

All members contributed to writing the paper.

References

[1] ABVM - Sample Questions. https://www.vascularboard.org/examinfo_questions.cfm.

[2] ChatGPT-4o. https://openai.com/index/hello-gpt-4o/.

[3] Gemini 2.0 Flash. https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash.

[4] High Yield Cardiothoracic and Vascular Surgery Questions. https://www.prepladder.com/neet-ss-surgery/cardiothoracic-and-vascular-surgery/high-yield-cardiothoracic-and-vascular-surgery-questions.

[5] IBI-CAAI/MELT-TinyLlama-1.1B-Chat-v1.0. https://huggingface.co/ibi-caai/melt-tinyllama-1.1b-chat-v1.0.

[6] johnsnowlabs/JSL-MedPhi2-2.7B. https://huggingface.co/johnsnowlabs/jsl-medphi2-2.7b.

[7] manupande21/GPT2_pmc. https://huggingface.co/manupande21/gpt2_pmc.

[8] openai-community/gpt2. https://huggingface.co/openai-community/gpt2.

[9] skumar9/Llama-medx_v3.2. https://huggingface.co/skumar9/llama-medx_v3.2.

[10] BEE-spoke-data/mega-ar-126m-4k. https://huggingface.co/bee-spoke-data/mega-ar-126m-4k, Oct. 2023.

[11] TinyLlama/TinyLlama_v1.1. https://huggingface.co/tinyllama/tinyllama_v1.1, July 2023.

[12] FreedomIntelligence/Apollo-0.5B. https://huggingface.co/freedomintelligence/apollo-0.5b, Dec. 2024.

[13] FreedomIntelligence/Apollo-2B. https://huggingface.co/freedomintelligence/apollo-2b, Dec. 2024.

[14] microsoft/Phi-4-mini-instruct. https://huggingface.co/microsoft/phi-4-mini-instruct, May 2025.

[15] APDVS. *The APDVS Medical Student Curriculum*. Association of Program Directors in Vascular Surgery (APDVS), 2023.

[16] Y. Chen, F. Li, X. Song, T. Li, Z. Xu, X. Chen, I. Sukeda, and I. Li. Exploring the Role of Knowledge Graph-Based RAG in Japanese Medical Question Answering with Small-Scale LLMs, Apr. 2025. arXiv:2504.10982 [cs].

[17] T. Feridooni, A. P. Javidan, D. N. Mahmood, Z. Gomes, A. Dueck, M. Wheatcroft, and D. Szalay. Development of a vascular surgery-specific artificial intelligence chat interface using retrieval-augmented generation: VASC.AI, a specialized vascular surgery chatbot. *JVS-Vascular Insights*, 2:100137, Jan. 2024.

[18] I. García-Ferrero, B. Altuna, J. Alvez, I. Gonzalez-Dios, and G. Rigau. This is not a dataset: A large negation benchmark to challenge large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[19] G. Geroulakos and B. Sumpio, editors. *Vascular Surgery: Cases, Questions and Commentaries*. Springer International Publishing, Cham, 2018.

[20] O. Haidar, A. Jaques, P. W. McCaughran, and M. J. Metcalfe. AI-Generated Information for Vascular Patients: Assessing the Standard of Procedure-Specific Information Provided by the ChatGPT AI-Language Model. *Cureus*, 15(11):e49764, Nov. 2023.

[21] J. He, M. Rungta, D. Koleczek, A. Sekhon, F. X. Wang, and S. Hasan. Does Prompt Formatting Have Any Impact on LLM Performance?, Nov. 2024. arXiv:2411.10541 [cs].

[22] A. P. Javidan, T. Feridooni, L. Gordon, and S. A. Crawford. Evaluating the progression of artificial intelligence and large language models in medicine through comparative analysis of ChatGPT-3.5 and ChatGPT-4 in generating vascular surgery recommendations. *JVS-Vascular Insights*, 2:100049, Jan. 2024.

[23] D. P. Jeong, P. Mani, S. Garg, Z. C. Lipton, and M. Oberst. The Limited Impact of Medical Adaptation of Large Language and Vision-Language Models, Feb. 2025. arXiv:2411.08870 [cs].

[24] Q. Le, K. S. Lavingia, and M. Amendola. The performance of large language models on fictional consult queries indicates favorable potential for AI-assisted vascular surgery consult handling. *JVS-Vascular Insights*, 2:100052, Jan. 2024.

[25] M. Levy, A. Jacoby, and Y. Goldberg. Same Task, More Tokens: the Impact of Input Length on the Reasoning Performance of Large Language Models, July 2024. arXiv:2402.14848 [cs].

[26] Q. Liu, W. Wang, and J. Willard. Effects of Prompt Length on Domain-specific Tasks for Large Language Models, Feb. 2025. arXiv:2502.14255 [cs].

[27] Y. Luo, Z. Yang, F. Meng, Y. Li, J. Zhou, and Y. Zhang. An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-tuning, Jan. 2025. arXiv:2308.08747 [cs].

[28] J. Maharjan, A. Garikipati, N. P. Singh, L. Cyrus, M. Sharma, M. Ciobanu, G. Barnes, R. Thapa, Q. Mao, and R. Das. OpenMedLM: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. *Scientific Reports*, 14(1):14156, June 2024. Publisher: Nature Publishing Group.

[29] A. Pal, P. Minervini, A. G. Motzfeldt, and B. Alex. openlifescienceai/open_medical_llm_leaderboard. `https://huggingface.co/spaces/openlifescienceai/open_medical_llm_leaderboard`, 2024.

[30] Peter Vien, Elie Donath MD, Joshua Le, Aryan Naik, Sean Liebscher MD, Benjamin Carriveau, and Daniel Bertges MD. Exploring The Use Of Customized Large Language Models For Potential In Vascular Surgery: A Pilot Study Involving The VESAP Board Exam. Autsin, Texas, Mar. 2025.

[31] Royal Australasian College of Surgeons (RACS). Sample Written Exam – Vascular Surgery.

[32] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, A. Babiker, N. Schärli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, B. Agüera y Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam, and V. Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, Aug. 2023. Publisher: Nature Publishing Group.

[33] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, M. Amin, L. Hou, K. Clark, S. R. Pfohl, H. Cole-Lewis, D. Neal, Q. M. Rashid, M. Schaekermann, A. Wang, D. Dash, J. H. Chen, N. H. Shah, S. Lachgar, P. A. Mansfield, S. Prakash, B. Green, E. Dominowska, B. Agüera y Arcas, N. Tomašev, Y. Liu, R. Wong, C. Semturs, S. S. Mahdavi, J. K. Barral, D. R. Webster, G. S. Corrado, Y. Matias, S. Azizi, A. Karthikesalingam, and V. Natarajan. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, Mar. 2025. Publisher: Nature Publishing Group.

[34] Society for Vascular Surgery. VESAP® | Society for Vascular Surgery.

[35] Society for Vascular Surgery. Vascular Surgery Clinical Practice Guidelines, May 2025.

[36] T. H. Truong, T. Baldwin, K. Verspoor, and T. Cohn. Language models are not naysayers: An analysis of language models on negation benchmarks, June 2023. arXiv:2306.08189 [cs].

[37] N. Zhang, Y. Yao, B. Tian, P. Wang, S. Deng, M. Wang, Z. Xi, S. Mao, J. Zhang, Y. Ni, S. Cheng, Z. Xu, X. Xu, J.-C. Gu, Y. Jiang, P. Xie, F. Huang, L. Liang, Z. Zhang, X. Zhu, J. Zhou, and H. Chen. A Comprehensive Study of Knowledge Editing for Large Language Models, Nov. 2024. arXiv:2401.01286 [cs].

[38] T. Zhang, S. G. Patil, N. Jain, S. Shen, M. Zaharia, I. Stoica, and J. E. Gonzalez. RAFT: Adapting Language Model to Domain Specific RAG, June 2024. arXiv:2403.10131 [cs].