MedMobile: A mobile-sized language model with expert-level clinical capabilities

Krithik Vishwanath[1,3], Jaden Stryker[1], Anton Alaykin[1,4],
Daniel A. Alber[1], Eric K. Oermann[1,2,5]

[1]Department of Neurological Surgery, [2]Department of Radiology,
NYU Langone Medical Center, New York, New York, 10016

[3]Department of Aerospace Engineering and Engineering Mechanics,
The University of Texas at Austin, Austin, Texas, 78712

[4]Department of Neurosurgery,
Washington University School of Medicine in St. Louis, St. Louis, Missouri, 63110

[5]Center for Data Science,
New York University, New York, New York, 10016

Send correspondence to: eric.oermann@nyulangone.org, krithik.vish@utexas.edu

Abstract

Language models (LMs) have demonstrated expert-level reasoning and recall abilities in medicine. However, computational costs and privacy concerns are mounting barriers to wide-scale implementation. We introduce a parsimonious adaptation of phi-3-mini, MedMobile, a 3.8 billion parameter LM capable of running on a mobile device, for medical applications. We demonstrate that MedMobile scores 75.7% on the MedQA (USMLE), surpassing the passing mark for physicians (∼60%), and approaching the scores of models 100 times its size. We subsequently perform a careful set of ablations, and demonstrate that chain of thought, ensembling, and fine-tuning lead to the greatest performance gains, while unexpectedly retrieval augmented generation fails to demonstrate significant improvements.

Keywords: phi-3-mini, USMLE, MultiMedQA, medical Q&A, knowledge distillation, low-cost, open-source

GitHub: https://github.com/nyuolab/MedMobile

Preprint. Under review.

Main

In recent years, language models (LM) have shown notable promise in the medical domain for their quick decision-making and ability for reasoning and knowledge [1, 2, 3]. However, large-scale adaptation of LMs faces several barriers, including security risks and the significant computational costs of model serving [4, 5]. Furthermore, the most powerful large models are closed-source, hindering domain-specific adaptation [6]. To address these barriers, we fine-tune phi-3-mini, an open-sourced 3.8B parameter language model, on data from the medical domain. We name the resulting model MedMobile, as models of this size have been demonstrated to run on mobile devices and boast inexpensive inference costs [7]. MedMobile was fine-tuned with manual data (curated by human experts) and synthetic data (artificially generated from GPT-4 and textbooks), demonstrating the ability of smaller language models to mimic specific tasks using synthetic data generated from larger models with high levels of accuracy. We chose to use synthetically generated data in line with the original phi work, which demonstrated the ability of smaller language models to develop reasoning with less data and parameters [7]. To the best of our knowledge, MedMobile represents the smallest language model model to attain a passing score (∼60%) on the MedQA [8], a large collection of USMLE-style questions, achieving an accuracy of 75.7%.

Enabling smaller language models to achieve superior performance on the USMLE-style and other medical tasks is an active area of research [9, 10]. Due to advancements in language model architecture, higher quality training data, and novel prompt engineering techniques, recent open-source models in the ∼7-8B range, such as Meerkat [9] and UltraMedical Llama 3.1 [10], have achieved a passing score on USMLE-style Q&A (Fig 2A), even outperforming language models several times larger such as GPT-3.5 (175B) [11], the SOTA from two years ago. Meerkat [9], the first 7B parameter model to achieve such a distinction, focused on improving smaller models via synthetic textbook-based, USMLE-style questions generated by GPT-4.0. Another series of models, UltraMedical [10] expands this work, generating synthetic questions on a larger scale and across all question types in the MultiMedQA [12]. Generalist language models can be improved significantly with knowledge distillation via supervised fine-tuning (SFT) on synthetic data. In this regard, enhancing smaller language models with support from much larger models has emerged as a leading approach to achieving superior performance with low-compute requirements.

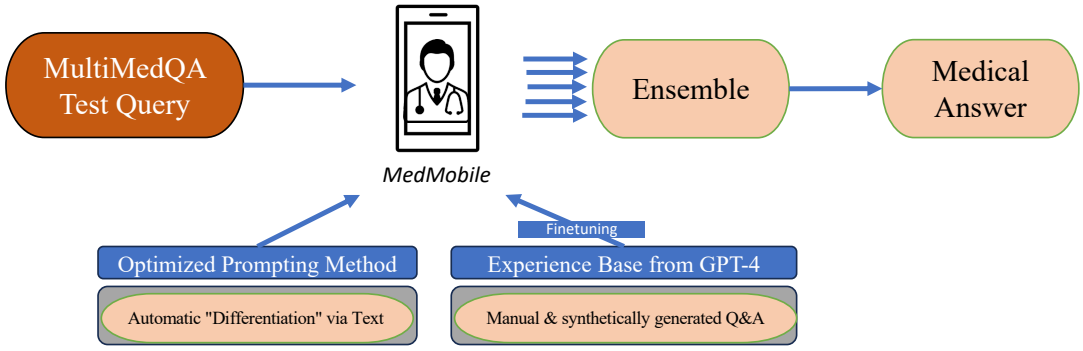
Since there is a significant loss of token generation speed and increase in power consumption on mobile devices after surpassing a model size of 5B parameters [13], we use the terminology "mobile-size" to refer to LMs smaller than 5B parameters. In this context, MedMobile is the first mobile-sized model to pass the USMLE. To achieve this with a low parameter count, we choose phi-3-mini as a backbone for our model as it exhibits enhanced reasoning capabilities relative to other models of the size [7]. Models that exhibit superior reasoning often utilize Chain-of-Thought (CoT), a technique that simulates human-like reasoning by using a sequence of logical steps to achieve an accurate final conclusion. Forefront language models, such as GPT-4, outputs a clear step-by-step process for arriving at its answer. By fine-tuning phi-3-mini using the CoT of GPT-4 (i.e., the logical process it uses to achieve it's final conclusion), we retain generalist reasoning capabilities while gaining medical domain knowledge, partially distilling the advanced problem-solving process and knowledge of GPT-4 to phi-3-mini [14]. Although MedMobile's performance doesn't exceed that of GPT-4, it marks a clinically significant advancement by enabling individuals to carry a board-certified clinical assistant in their pocket.

A)

Table 1. Multiple-choice question (MCQ) evaluation datasets part of the MultiMedQA.

Task	# of MCQs	Task Overview
MedQA (USMLE)	1273	MCQs based on the US medical licensing exam
MedMCQA	4183	MCQs based on Indian medical entrance exams (NEET, AIIMS)
MMLU (Clinical knowledge)	265	Clinical knowledge MCQs
MMLU (Medical genetics)	100	Medical genetics MCQs
MMLU (Anatomy)	135	Anatomy MCQs
MMLU (Professional medicine)	272	Professional medicine MCQs (based on USMLE)
MMLU (College biology)	144	College biology MCQs
MMLU (College medicine)	173	College medicine MCQs
PubMedQA	500	Research question (yes/no/maybe) with a corresponding PubMed abstract

B)



C)

Example Output from *MedMobile*

Question: A 55-year-old male visited his primary care physician complaining of chest pain and progressive shortness of breath worsened by exertion. His vital signs were within normal limits. He reports that he worked in a naval shipyard for 10 years during his twenties. A CT scan of his chest shows diffuse thickening of the pleura and a diagnosis of mesothelioma is made. Assuming this diagnosis is correct, which of the following is most likely to also be present in this patient?

A) Pneumothorax B) Pleural effusion C) Systemic inflammatory response syndrome D) Bronchioalveolar carcinoma

The patient's history of working in a naval shipyard and his current symptoms of chest pain and shortness of breath, along with the CT findings of diffuse thickening of the pleura and a diagnosis of mesothelioma, point towards . . . asbestos exposure. Mesothelioma . . . is strongly associated with asbestos exposure. One of the common complications of mesothelioma is pleural effusion, which . . . can contribute to the patient's symptoms of shortness of breath . . . Given the information provided and the typical complications associated with mesothelioma, pleural effusion is the most likely additional finding in this patient.

Therefore, the answer is [B].

MedMobile Output

Figure 1. Overview of MedMobile. Panel A) shows components of MultiMedQA [12], the evaluation tasks descriptions, and the number of questions in each data test set. MultiMedQA is a collection of 8 different datasets encompassing the medical domain. In tasks such as the MMLU, we test the model on its ability to perform complex reasoning tasks across medical and medical-adjacent domains. PubMedQA tests a model's ability to perform reasoned conclusions based on research-grade text. Finally, MedQA (USMLE) and MedMCQA evaluates the model on its ability to answer standardized medical questions necessary to be a licensed physician. In Panel B), we present a framework of medical Q&A evaluation and model building. MultiMedQA is used to evaluate a fine-tuned phi-3-mini model, MedMobile, and is optimized in its prompting using automatic differentiation with GPT-4 as described in TextGrad [15]. Responses are then filtered via an ensemble approach, where the most common answer is selected as the model's final answer. We also fine-tune our model on synthetic and manually medical questions, annotated with CoT by GPT-4. Panel C) exhibits a sample MedQA

(USMLE) question and MedMobile's response. Note that this is one of five responses generated before ensembling. MedMobile displays an ability to contextualize complex medical scenarios and develop expert-level conclusions. MedMobile's output is shortened in parts for visual purposes.

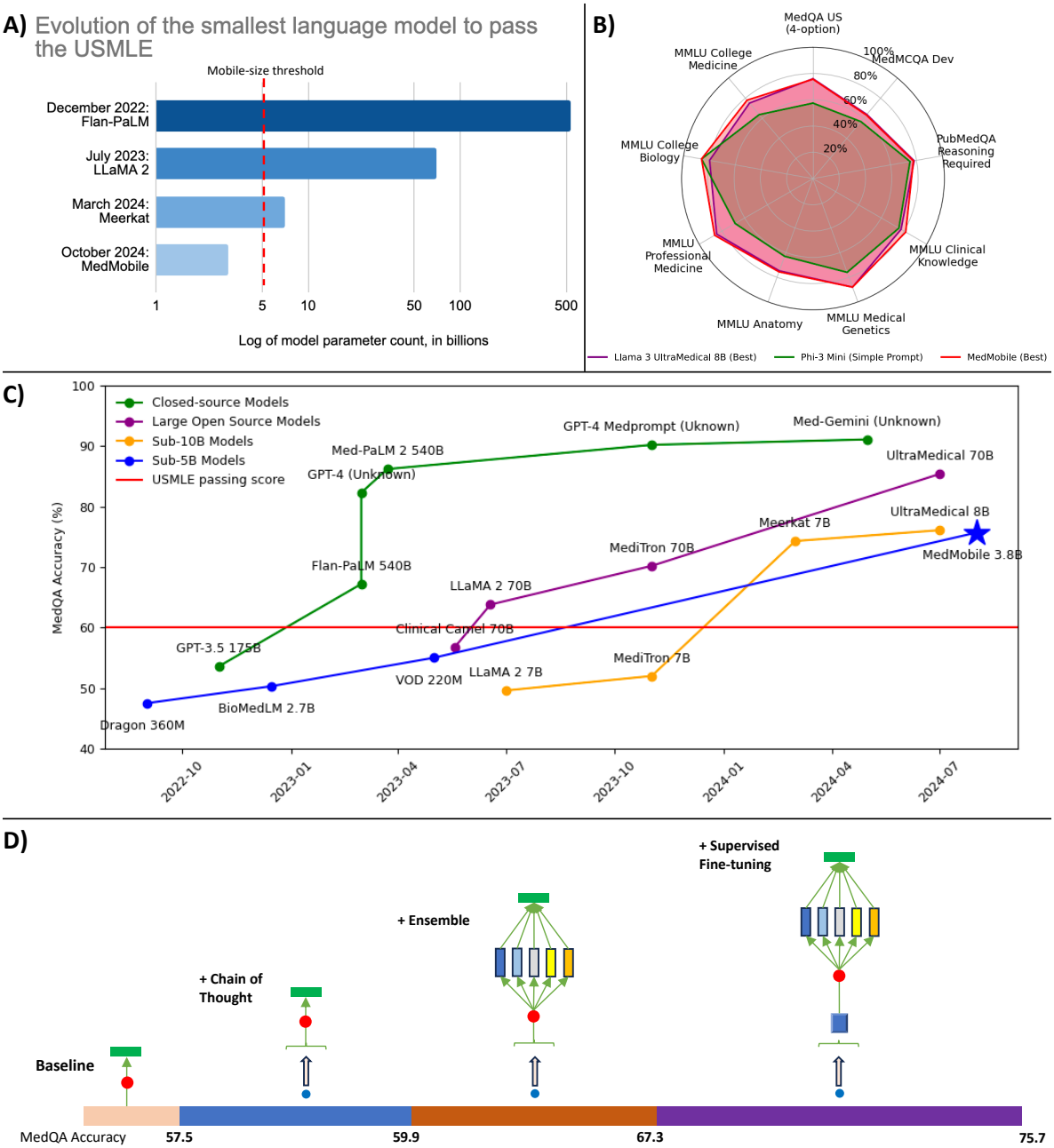


Figure 2. Overview of language models and their performance on USMLE-style questions, contextualized over time. Panel A) shows the progession of smallest language model that is able to pass the USMLE, based on the MedQA. Panel B) displays MedMobile (red) compared to Llama-3 UltraMedical 8B (purple), and a baseline phi-3-mini (green) model on the entire MultiMedQA. MedMobile achieves almost identical or superior performance across the entirety of the MultiMedQA compared to the SOTA of <10B parameter language models (UltraMedical 8B), with a fraction of the parameters. Panel C) presents the relative accuracy of MedMobile to other language models on the MedQA. Current models range vastly in parameter size; with closed-source models such as Med-Palm 2 requiring

540B parameters [16]. Open-source models for the medical domain are led by Llama 3.1 UltraMedical 70B, which achieves an accuracy of 85.4%. In the sub-10B parameter range, Llama 3.1 UltraMedical 8B leads SOTA with an accuracy of 76.1%. In "mobile-sized" models, which we define to have <5B parameters due to the significantly higher quantization, slower token generation, and compute requirements after surpassing the 5B parameter threshold [13], MedMobile beats previous SOTA by over 20% accuracy points. We note that there has not been significant development in the sub-5B parameter space for some time, and MedMobile is the first to surpass the passing score in this category. Panel D) shows a stepwise ablation study of components. We add individual components of the pipeline shown in Panel B and evaluate their impact on model accuracy before continuing. Through this method, we improve our accuracy from a baseline of 57.5% to a final accuracy of 75.7% on the MedQA test set.

In the past few years, several techniques have demonstrated improvement in LMs' Q&A performance on various benchmarks [14, 17]. However, we find a lack of technique validation for our context, given that a technique, such as k-shot prompting, may only be valid for a specific size model, domain, or cocktail of techniques. To determine the positively contributing components of our pipeline, we add components one by one and evaluate after each addition (Fig. 2D). After component testing, we develop our final pipeline (Fig. 1B) built on SFT, CoT, response ensembling, and prompt optimization. While a baseline phi-3-mini at baseline achieves a score of 57.5% on the MedQA, adding CoT (+2.4%), ensembling responses (+7.4%), and conducting SFT (+8.4%) allows MedMobile to achieve an accuracy of 75.7%. In doing so, we noted several promising potential pipeline components did not favorably impact inference in medical Q&A, such as k-shot prompting with examples (−9.4%) and retrieval-augmented generation (RAG) (−12.6%) from high-quality sources (i.e., textbooks), perhaps due to an increased input token length. This improvement represents a substantial increase from the next best sub-5B parameter language model, VOD [18], at an accuracy of 55.0% on the MedQA. MedMobile's accuracy on the entirety of the MultiMedQA is comparable to the SOTA models in the medical domain with over double the number of parameters. In fact, MedMobile beats or matches UltraMedical 8B [10], the current model with the highest accuracy in the sub-10B parameter space, in 6 out of 9 evaluation tasks in the MultiMedQA (Fig. 2B). To the best of our knowledge, MedMobile is also the smallest model to achieve the distinction of passing USMLE-like questions on the MedQA.

MedMobile displays an ability to develop explainable responses to complex medical scenarios, carefully accounting for a complex combination of patient symptoms that may or may not influence treatment (Fig. 1C). There is a clear delineation of step-by-step logical CoT responses, evidence of medical knowledge distillation from GPT-4's and retention of reasoning capabilities. However, we note a decrease in performance based on token output length (see Supplemental Figure 1B). This can potentially be attributed to a loss of model CoT when crafting longer responses as fine-tuned smaller models tend to have reduced reasoning capabilities and weaker CoT. This can be compared to phi-3-mini baseline, which has a more consistent accuracy across the different token outputs it contains. However, across almost all bins, irrespective of output length, MedMobile outperforms phi-3-mini, highlighting the gain of domain-specific knowledge and the resultant gain in MedMobile's performance on medical tasks.

Contrary to popular literature, we do not utilize many of the prompt engineering approaches that are common for large language models including retrieval augmented generation (RAG) and k-shot prompting. We implemented these techniques (see Supplemental Figure 3), but they did not lead to any significant degree of improvement. We hypothesize that this is mainly driven by the context-window limitations small language models have, and note these are interesting barriers to tackle for

future research [7].

There are several limitations that apply to our work. While we demonstrate a significant improvement over previous open-source models of MedMobile's size, any-size models still demonstrate superior performance on medical tasks. Thus, ignoring the barriers of subscription fees and issues with uploading classified patient health information, GPT-4 can be used for quick and reliable online inference. We also note that real-world clinical and patient-facing deployment of MedMobile is yet to be evaluated, and is left for future works. Finally, MedMobile, in this work, is trained only on language and cannot receive image input.

This work can be easily expanded to vision-language models (VLMs) by building upon Phi-3-vision using this pipeline. VLMs have shown promise for superhuman predictive power and novel pattern recognition but notoriously require extensive training and inference costs due to the larger data sizes associated with high-resolution imaging [19, 20]. Using a smaller, domain-specific model, such as MedMobile, serves to combat these rising computational costs. Alongside this rise, we also note the increase in novel imaging methods that provide new dimensional data that machine learning models can leverage, such as photoacoustic imaging providing spectral information to individual voxels or shear wave elasticity imaging providing information about tissue stiffness [21, 22]. In light of the increase in imaging data from new dimensions (e.g., spectral data from photoacoustic imaging or tissue stiffness data from shearwave imaging) in these modalities, smaller language models may serve to foster new, cutting-edge insights and patterns that otherwise are hidden from humans, while bolstering quick compute times. In tandem with improvements of imaging modalities, VLM pattern recognition, and the increase in mobile-based ML platforms, such as Apple's new Apple Intelligence [23], we envision a method of use for mobile-sized VLMs centered around accessibility, where doctors and patients can take images with their iPhone and receive insights from a expert-level, fine-tuned LLM, without comprising personal security or requiring extensive computing power.

Recent studies in other domains have also demonstrated effective improvements to benchmark accuracy when using multi-LM agent-based systems [24]. A promising avenue for future research could be utilizing MedMobile as part of a multi-LM system, where problem-solving is divided into multiple iterations of MedMobile. Further distillation of GPT-4 on each agent in such an ensemble may allow for significant improvements to accuracy.

Highly accurate, expert-level mobile-sized language models, such as MedMobile, hold promise in low and middle-resource settings due to their reduced compute requirements and quicker inference times [25, 26, 27] and also serve to democratize access to the technical capabilities of LLMs beyond the domain of large technology companies and groups with substantial computing budgets. While we develop this work primarily for its impact in the medical domain, mobile-size language models and the related techniques in this work can be applied to any domain to train expert-level mobile assistants. We hope this work, and our open-source codebase, will contribute to the clinically meaningful development of mobile-sized language models that benefit physicians and patients.

Methods

Evaluation Data

To determine an LM's ability in the medical domain, we evaluate the model on the MultiMedQA, a multi-dataset of medical questions [12]. The MultiMedQA is composed of 8 individual datasets ranging

from USMLE-style questions (MedQA) to College Biology (MMLU College Biology) and is outlined in Figure 1A. We choose to evaluate on these datasets due to the expert level of medical reasoning and knowledge required for USMLE-style questions, and to test the model's ability against the range of medical tasks with the other datasets. Testing on the PubMedQA also demonstrates MedMobile's ability to perform on research-related medical inquiries. These results are displayed in Supplemental Table 1.

Supervised Fine-Tuning (SFT)

To train phi-3-mini's baseline parameters to the medical domain, we utilize the UltraMedical dataset, a collection of over 400K synthetic and manual-curated multiple-choice questions [10]. In particular, we instruction-fine-tune phi-3-mini using CoT responses from GPT-4 for each of these questions, allowing for knowledge distillation from GPT-4's much larger parameter set. To perform SFT, we train phi-3-mini for 3 epochs on 4 A100 nodes for 83 hours on the UltraMedical dataset. We also utilize a learning rate of 1×10^{-4} and an effective batch size of 32.

Prompt Optimization

To ensure streamlined and favorable prompting, we utilize TextGrad [15], a multi-LLM system for improving prompting verbiage using. TextGrad automatically develops improvements to smaller language models' prompts by utilizing a much stronger model (in this case, we use GPT-4). GPT-4, as an optimizer model, generates new prompting templates. Then, a loss is calculated based on the accuracy generated by MedMobile on the prompt. While TextGrad finds an improved prompt verbiage for phi-3-mini baseline, it also supports that MedMobile does best with no additional prompting instructions. Due to the limited context window capabilities of a model of this size, it is likely that the additional text only hinders the model from domain-specific tasks that it is already trained to reason within. By utilizing the CoT responses of GPT-4 to fine-tune MedMobile, MedMobile exhibits high levels of medical reasoning without the necessity of additional prompting.

Other Techniques

MedMobile was also assessed with other prompting methods such as k-shot prompting, retrieval augmented generation, BM-25 searching, and additional prompting techniques. To develop retrieval-based prompting methods, we feed in paragraphs from Harrison's Principles of Internal Medicine, 21e [28]. We attempt a variety of retrieval-based scenarios, such as the lucine implementation of BM-25, RAG based on cosine similarity with questions and paragraphs embedded with MedCPT [29], and combination methods that use both scores to select the best contextual paragraphs related to the question. However, we note that these additions did not broadly improve model performance.

Acknowledgements

EKO is supported by the National Cancer Institute's Early Surgeon Scientist Program (3P30CA016087-41S1) and the W.M. Keck Foundation. We would like to acknowledge Nader Mherabi and Dafna Bar-Sagi, Ph.D., for their continued support of medical AI research at NYU. We thank Michael Constantino, Kevin Yie, and the NYU Langone High-Performance Computing (HPC) Team for supporting computing resources fundamental to our work.

Author Contributions

EKO conceptualized and supervised the study. KV designed the MedMobile LLM pipeline. KV, JS, and AA implemented and trained the LLM. KV evaluated and tested the LLM. JS aided with LLM serving and deployment. KV wrote the initial draft of the manuscript. KV, AA, DAA, EKO edited the manuscript. All authors revised and approved the manuscript.

Competing Interests

Disclosures: EKO and DA report consulting income with Sofinnova Partners. EKO reports equity in Eikon Therapeutics, Artisight Incorporate. The other authors have no personal, financial, or institutional interest pertinent to this article.

Data availability

The datasets generated or analysed during the current study are available in the nyuolab/MedMobile repository, https://github.com/nyuolab/MedMobile. The model weights are available on https://huggingface.co/KrithikV/MedMobile.

Code availability

Our code is shared publicly on GitHub upon publication of this work and can be found at https://github.com/nyuolab/MedMobile.

References

[1] Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.

[2] Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

[3] Lavender Yao Jiang, Xujin Chris Liu, Nima Pour Nejatian, Mustafa Nasir-Moin, Duo Wang, Anas Abidin, Kevin Eaton, Howard Antony Riina, Ilya Laufer, Paawan Punjabi, et al. Health system-scale language models are all-purpose prediction engines. *Nature*, 619(7969):357–362, 2023.

[4] Ehsan Ullah, Anil Parwani, Mirza Mansoor Baig, and Rajendra Singh. Challenges and barriers of using large language models (llm) such as chatgpt for diagnostic medicine with a focus on digital pathology–a recent scoping review. *Diagnostic pathology*, 19(1):43, 2024.

[5] Xiuquan Li and Tao Zhang. An exploration on artificial intelligence application: From security, privacy and ethic perspective. In *2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pages 416–420, 2017.

[6] Ruei-Shan Lu, Ching-Chang Lin, and Hsiu-Yuan Tsao. Empowering large language models to leverage domain-specific knowledge in e-learning. *Applied Sciences*, 14(12):5264, 2024.

[7] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

[8] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

[9] Hyunjae Kim, Hyeon Hwang, Jiwoo Lee, Sihyeon Park, Dain Kim, Taewhoo Lee, Chanwoong Yoon, Jiwoong Sohn, Donghee Choi, and Jaewoo Kang. Small language models learn enhanced reasoning skills from medical textbooks, 2024.

[10] Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, Xingtai Lv, Hu Jinfang, Zhiyuan Liu, and Bowen Zhou. Ultramedical: Building specialized generalists in biomedicine, 2024.

[11] Openai OpenAI. Openai: Introducing chatgpt. *URL https://openai. com/blog/chatgpt*, 2022.

[12] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[13] Stefanos Laskaridis, Kleomenis Kateveas, Lorenzo Minto, and Hamed Haddadi. Melting point: Mobile evaluation of language transformers. *arXiv preprint arXiv:2403.12844*, 2024.

[14] Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*, 2024.

[15] Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. Textgrad: Automatic "differentiation" via text, 2024.

[16] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[17] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 2024.

[18] Valentin Liévin, Andreas Geert Motzfeldt, Ida Riis Jensen, and Ole Winther. Variational open-domain question answering. In *International Conference on Machine Learning*, pages 20950–20977. PMLR, 2023.

[19] Chen Zhu, Wei Ping, Chaowei Xiao, Mohammad Shoeybi, Tom Goldstein, Anima Anandkumar, and Bryan Catanzaro. Long-short transformer: Efficient transformers for language and vision. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 17723–17736. Curran Associates, Inc., 2021.

[20] Yixing Jiang, Jesutofunmi A Omiye, Cyril Zakka, Michael Moor, Haiwen Gui, Shayan Alipour, Seyed Shahabeddin Mousavi, Jonathan H Chen, Pranav Rajpurkar, and Roxana Daneshjou. Evaluating general vision-language models for clinical medicine. *medRxiv*, pages 2024–04, 2024.

[21] Armen P Sarvazyan, Oleg V Rudenko, Scott D Swanson, J Brian Fowlkes, and Stanislav Y Emelianov. Shear wave elasticity imaging: a new ultrasonic technology of medical diagnostics. *Ultrasound in medicine & biology*, 24(9):1419–1435, 1998.

[22] Nilesh Mathuria, Krithik Vishwanath, Blake C. Fallon, Antonio Martino, Giorgio Brero, Richard R Willson, Miguel Valderrabano, Carly S. Filgueira, and Richard R. Bouchard. In vivo assessment of cardiac radiofrequency ablation in a large-animal model using photoacoustic-ultrasound imaging. *medRxiv*, 2024.

[23] Tom Gunter, Zirui Wang, Chong Wang, Ruoming Pang, Andy Narayanan, Aonan Zhang, Bowen Zhang, Chen Chen, Chung-Cheng Chiu, David Qiu, et al. Apple intelligence foundation language models. *arXiv preprint arXiv:2407.21075*, 2024.

[24] Weizhou Shen, Chenliang Li, Hongzhan Chen, Ming Yan, Xiaojun Quan, Hehong Chen, Ji Zhang, and Fei Huang. Small llms are weak tool learners: A multi-llm agent. *arXiv preprint arXiv:2401.07324*, 2024.

[25] Nethra Venkatayogi, Maanas Gupta, Alaukik Gupta, Shreya Nallaparaju, Nithya Cheemalamarri, Krithika Gilari, Shireen Pathak, Krithik Vishwanath, Carel Soney, Tanisha Bhattacharya, et al. From seeing to knowing with artificial intelligence: A scoping review of point-of-care ultrasound in low-resource settings. *Applied Sciences*, 13(14):8427, 2023.

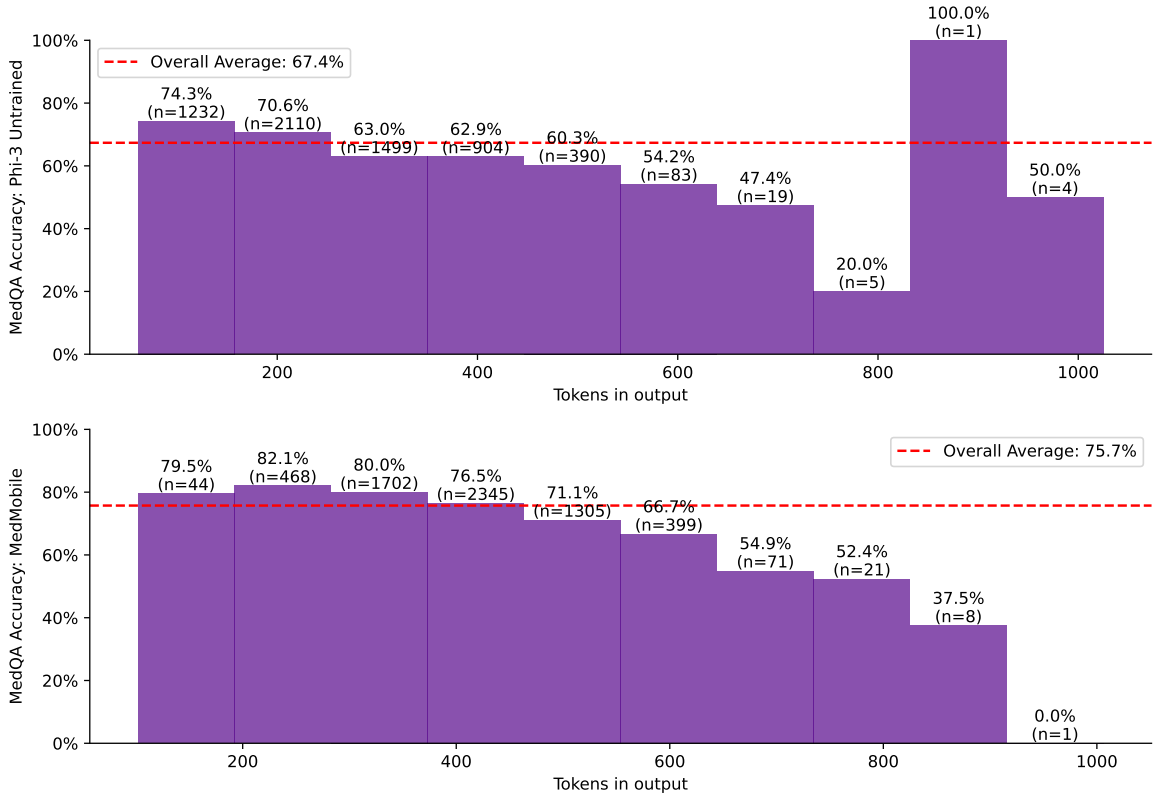
[26] Diego M López, Carolina Rico-Olarte, Bernd Blobel, and Carol Hullin. Challenges and solutions for transforming health ecosystems in low-and middle-income countries through artificial intelligence. *Frontiers in Medicine*, 9:958097, 2022.

[27] Tadeusz Ciecierski-Holmes, Ritvij Singh, Miriam Axt, Stephan Brenner, and Sandra Barteit. Artificial intelligence for strengthening healthcare systems in low-and middle-income countries: a systematic scoping review. *npj Digital Medicine*, 5(1):162, 2022.

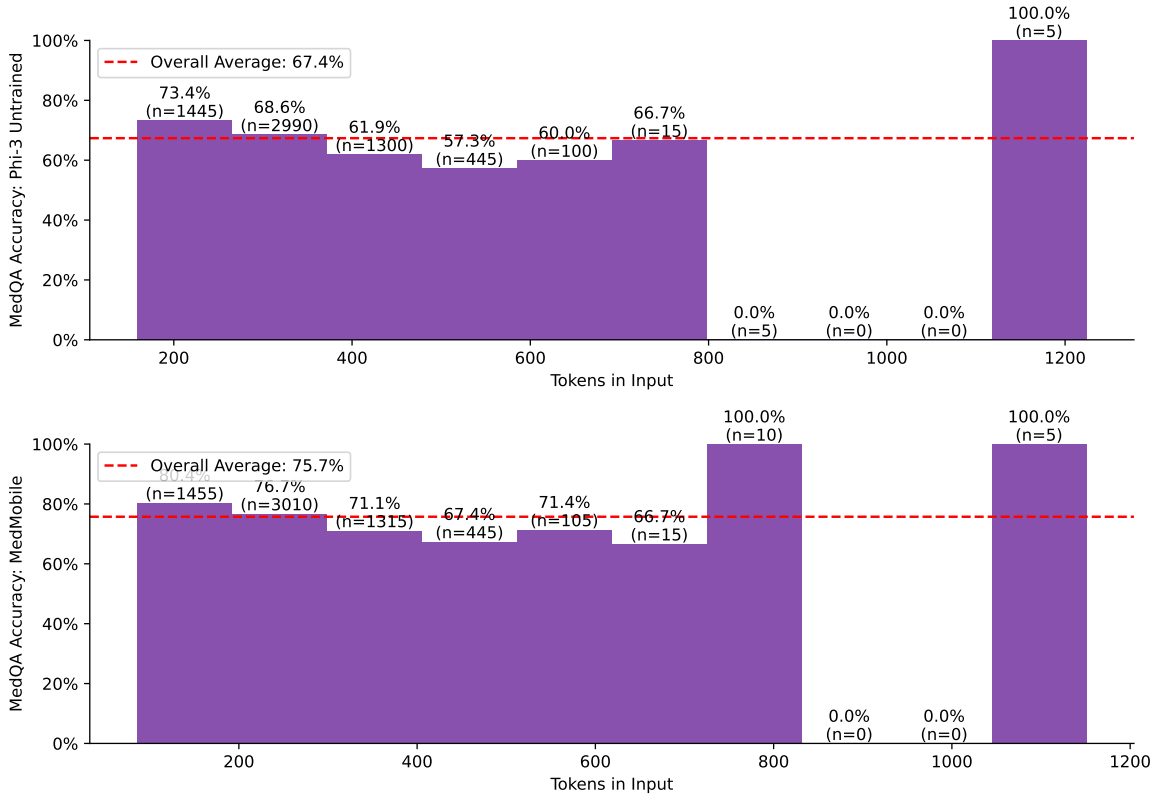
[28] E Silverman, J Crapo, B Make, J Jameson, A Fauci, D Kasper, S Hauser, D Longo, and J Loscalzo. Harrison's principles of internal medicine 21e, 2022.

[29] Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023.

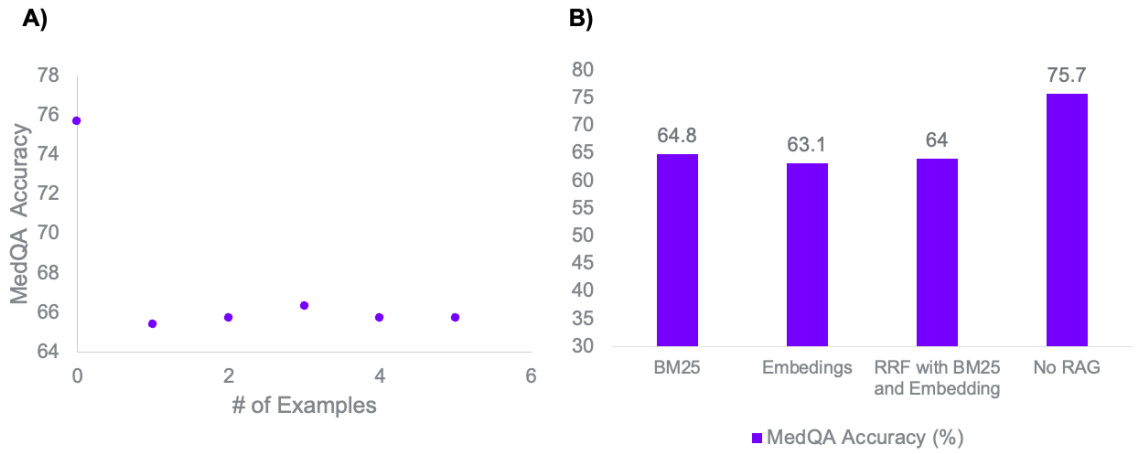
Supplementary Material



Supplemental Figure 1. Comparison of number of output tokens in a response and accuracy on MedQA questions. Each question of the MedQA test set is represented 5x in this figure due to the ensemble performed. Some questions are not included in the plots (< 20) as model response exceeded maximum generation output and an accuracy could not be evaluated. Top panel is a CoT enhanced baseline phi-3-mini model, whereas the bottom panel is our fine-tuned model, MedMobile.



Supplemental Figure 2. Comparison of number of input tokens in a response and accuracy on MedQA questions. Each question of the MedQA test set is represented 5x in this figure due to the ensemble performed. Some questions are not included in the plots (< 20) as model response exceeded maximum generation output and an accuracy could not be evaluated. Top panel is a CoT enhanced baseline phi-3-mini model, whereas the bottom panel is our trained model, MedMobile.



Supplemental Figure 3. Panel A) depicts the accuracy of MedMobile on the MedQA relative to the number of k-shot prompting (i.e., number of examples given to the model alongside the evaluation question). Panel B) shows different forms of retrieval for RAG and their resultant effects on the accuracy of MedMobile on the MedQA dataset. To conduct RAG based on vector embeddings, we compute cosine similarity based on MedCPT vectors generation between the question and paragraphs in the textbook. RAG built on BM-25 is developed through the lucine implementation, and selects the paragraph with the highest score for a particular question. While all forms of RAG achieve sub-optimal results, we note that BM25 seemed to affect the model least negatively with the addition of context. The source of information for these evaluations is from Harrison's Principles of Internal Medicine, 21e [28].

Supplemental Table 1. Evaluation results across the MultiMedQA, for phi-3-mini, MedMobile, UltraMedical 8B, and Flan-Palm. Scores for UltraMedical 8B and Flan-Palm are sourced from literature [1, 10].

	phi-3-mini Baseline (3.8B)	MedMobile (3.8B)	UltraMedical (8B)	Flan-PaLM (540B)
MedQA (USMLE)	57.5	75.7	76.1	67.6
MedMCQA Dev	56.7	63.2	63.8	57.6
PubMedQA Reasoning Required	75	77.6	78.2	79
MMLU (Clinical Knowledge)	75.5	81.5	77.4	80.4
MMLU (Medical Genetics)	76	88	88	75
MMLU (Anatomy)	63	75.6	74.8	63.7
MMLU (Professional Medicine)	68.4	86.4	84.6	83.8
MMLU (College Biology)	86.1	86.1	79.9	88.9
MMLU (College Medicine)	63.6	78	75.1	76.3