

Medical Adaptation of Large Language and Vision-Language Models: Are We Making Progress?

Daniel P. Jeong¹, Saurabh Garg^{1,2}, Zachary C. Lipton^{1,4}, Michael Oberst^{3,4}

¹Machine Learning Department, Carnegie Mellon University

²Mistral AI

³Department of Computer Science, Johns Hopkins University

⁴Abridge AI

{danielje, sgarg2, zlipton}@cs.cmu.edu, moberst@jhu.edu

Correspondence: danielje@cs.cmu.edu

Abstract

Several recent works seek to develop foundation models specifically for medical applications, adapting general-purpose large language models (LLMs) and vision-language models (VLMs) via continued pretraining on publicly available biomedical corpora. These works typically claim that such domain-adaptive pretraining (DAPT) improves performance on downstream medical tasks, such as answering medical licensing exam questions. In this paper, we compare seven public “medical” LLMs and two VLMs against their corresponding base models, arriving at a different conclusion: all medical VLMs and nearly all medical LLMs fail to consistently improve over their base models in the zero-/few-shot prompting regime for medical question-answering (QA) tasks. For instance, across the tasks and model pairs we consider in the 3-shot setting, medical LLMs only outperform their base models in 12.1% of cases, reach a (statistical) tie in 49.8% of cases, and are significantly worse than their base models in the remaining 38.2% of cases. Our conclusions are based on (i) comparing each medical model head-to-head, directly against the corresponding base model; (ii) optimizing the prompts for each model separately; and (iii) accounting for statistical uncertainty in comparisons. While these basic practices are not consistently adopted in the literature, our ablations show that they substantially impact conclusions. Our findings suggest that state-of-the-art general-domain models may already exhibit strong medical knowledge and reasoning capabilities, and offer recommendations to strengthen the conclusions of future studies.

1 Introduction

Recent advances in autoregressive large language models (LLMs) and vision-language models (VLMs) have attracted interest from practitioners in medicine, where these models hold great potential to transform various aspects of clinical practice

(e.g., medical diagnosis, information retrieval from clinical documents, patient triaging) (Fries et al., 2022a; Moor et al., 2023a). State-of-the-art performance on various medical benchmarks is typically achieved by massive-scale closed-source models, such as GPT-4 (OpenAI, 2023a,b), MED-GEMINI (Saab et al., 2024; Yang et al., 2024), and MED-PALM (Singhal et al., 2023a,b; Tu et al., 2024), often performing on par with humans on medical licensing exams and open-ended consumer health question-answering (QA) tasks. However, the general lack of transparency in these models, high API usage costs, and patient data privacy concerns make their integration into routine clinical workflows challenging (Marks and Haupt, 2023).

To address such concerns, recent works have proposed cheaper, open-source alternatives through *domain-adaptive pretraining* (DAPT; Gururangan et al., 2020), where a pretrained open-source general-domain model—such as LLAMA (Touvron et al., 2023a,b; Meta, 2024) or MISTRAL (Jiang et al., 2023) in the language space, and LLAVA (Liu et al., 2023) or OPEN-FLAMINGO (Awadalla et al., 2023) in the vision-language space—is continually pretrained on biomedical (image-)text corpora from public sources such as PubMed and medical textbooks. While some prior works show that medical models pretrained from scratch only using domain-specific corpora can outperform those trained via DAPT, both in the context of BERT-style encoder-only models (Devlin et al., 2019; Gu et al., 2021; Yang et al., 2022) and decoder models (Taylor et al., 2022; Luo et al., 2022; Hernandez et al., 2023; Bolton et al., 2024), the DAPT approach has become common practice, resulting in a trend where the release of a more capable general-domain model is typically followed by the release of its medical counterpart.

Despite the widespread adoption of medical DAPT, the claimed improvements in performance are worth scrutinizing. While the story is intu-

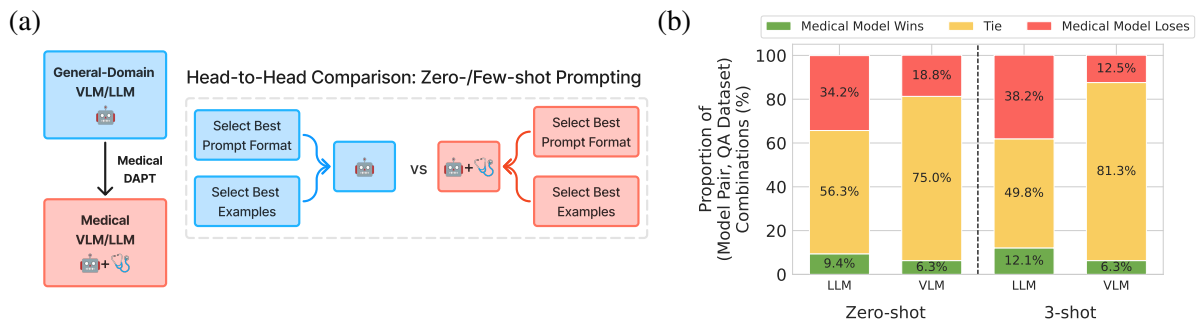


Figure 1: Medical LLMs and VLMs trained via domain-adaptive pretraining (DAPT) show limited improvement over their general-domain counterparts. (a) Overview of our head-to-head evaluation approach for each pair of general-domain (blue) and medically adapted LLM/VLM (red). (b) Win/tie/loss rate (%) of medical models vs. their corresponding base models across all (model pair, QA dataset) combinations. Win rate refers to the proportion of (model pair, QA dataset) combinations where a medical model shows a statistically significant improvement.

itive, more recent base models (e.g., LLAMA-3-8B (Meta, 2024)) already exhibit strong off-the-shelf performance on medical benchmarks without any adaptation (e.g., Open Medical LLM Leaderboard (Pal et al., 2024)), and given a lack of transparency about the pretraining corpora used to train the general-domain model in the first place, they may already be trained on relevant medical text.

Perhaps more concerning is the lack of apples-to-apples comparisons in the literature. First, medical models resulting from DAPT are often only compared against other baselines with different architectures (e.g., CLINICALCAMEL-70B (Toma et al., 2023) vs. GPT-4 (OpenAI, 2023a)) and under inconsistent evaluation setups (e.g., MEDITRON-70B (Chen et al., 2023) fine-tuned on MedQA (Jin et al., 2020) vs. non-fine-tuned MED42-V1-70B (Christophe et al., 2024)), which can confound the interpretation of results. Second, the common practice of using a single, fixed prompting setup (e.g., prompt format, choice of few-shot examples) for all models under evaluation also warrants concern, as LLM/VLM behavior is extremely sensitive to such design decisions (Jiang et al., 2020; Zhao et al., 2021; Ceballos-Arroyo et al., 2024), and the “optimal” choice of such details rarely correlates between different models (Sclar et al., 2024).

In this paper, we perform an apples-to-apples comparison that addresses these concerns, comparing seven medical LLMs and two medical VLMs against their general-domain base models. We find that, for all but one LLM pair—BIOMISTRAL-7B (Labrak et al., 2024) vs. MISTRAL-7B-INSTRUCT-V0.1 (Jiang et al., 2023), a pair of models that performs fairly poorly in absolute terms—the open-source medical LLMs and VLMs that we

evaluate do not consistently improve over their general-domain counterparts on various medical (visual) QA tasks (Figure 1). We compare several pairs of general-domain and medically adapted LLMs/VLMs (see Table 1), **whose only differences lie in medical DAPT** (i.e., one model is the base model, from which the other is derived via medical DAPT). For each pair, we compare their performances from zero-/few-shot prompting (Radford et al., 2019; Brown et al., 2020), after independently selecting the “best” prompt format and few-shot examples for each model based on the validation set and accounting for statistical uncertainty in model comparison.

Our findings (Section 4) suggest that state-of-the-art general-domain models may already exhibit strong medical knowledge and reasoning capabilities that can be leveraged effectively when prompted appropriately.

Our main contributions can be summarized as follows:

1. We provide a comprehensive head-to-head comparison between state-of-the-art general-domain LLMs/VLMs and their medical DAPT counterparts on various medical (visual) QA benchmarks, to investigate the effectiveness of DAPT for medical specialization.
2. We find that after optimizing the prompts for medical and general-domain models independently, all medical VLMs and nearly all medical LLMs that we evaluate fail to consistently improve over their corresponding general-domain base models.
3. We show that using a single, fixed prompt format and choice of few-shot examples for all models without testing for statistical signif-

Table 1: Summary of open-source autoregressive VLM and LLM pairs used for evaluation.

Model Class	General Domain	Medical Domain	Medical Adaptation Corpora
LLM	LLAMA-3-70B-INSTRUCT (Meta, 2024)	OPENBIOLLM-70B (Pal and Sankarasubbu, 2024)	Undisclosed
	LLAMA-2-70B (Touvron et al., 2023b)	MEDI TRON-70B (Chen et al., 2023)	Clinical Practice Guidelines (e.g., CDC, WHO) PubMed Articles (S2ORC; Lo et al., 2020)
	LLAMA-2-70B (Touvron et al., 2023b)	CLINICAL-CAMEL-70B (Toma et al., 2023)	ShareGPT 20k PubMed Articles Published Before 2021 Random 4k Subset of MedQA (Jin et al., 2020)
	LLAMA-3-8B (Meta, 2024)	OPENBIOLLM-8B (Pal and Sankarasubbu, 2024)	Undisclosed
	LLAMA-2-7B (Touvron et al., 2023b)	MEDI TRON-7B (Chen et al., 2023)	Clinical Practice Guidelines (e.g., CDC, WHO) PubMed Articles (S2ORC; Lo et al., 2020)
	MISTRAL-7B-INSTRUCT-V0.1 (Jiang et al., 2023)	BIOMISTRAL-7B (Labrak et al., 2024)	PubMed Articles (PMC Open Access Subset)
	LLAMA-2-7B-CHAT (Touvron et al., 2023b)	BIOMEDGPT-LM-7B (Luo et al., 2023)	PubMed Articles (S2ORC; Lo et al., 2020)
VLM	LLAVA-V0-7B (Liu et al., 2023)	LLAVA-MED-7B (Li et al., 2023)	PubMed Articles (PMC-15M; Zhang et al., 2023)
	OPEN-FLAMINGO-9B (Awadalla et al., 2023)	MED-FLAMINGO-9B (Moor et al., 2023b)	Medical Textbooks (MTB; Moor et al., 2023b) PubMed Articles (PMC-OA; Lin et al., 2023)

icance can lead to overly optimistic conclusions about the benefits from medical DAPT.

2 Related Work

DAPT (Gururangan et al., 2020) is a transfer learning approach, where a pretrained model is further pretrained on domain-specific data for better alignment to a target domain of interest (e.g., medicine, law). Several studies show that language models trained via DAPT often outperform their general-domain counterparts on domain-specific tasks, such as claim detection from blog posts (Chakrabarty et al., 2019), named entity recognition from German novels (Konle and Jannidis, 2020), and judgment prediction for legal cases (Xiao et al., 2021). In the medical domain, prior works based on BERT-style encoder-only language models (Devlin et al., 2019), such as BIOBERT (Lee et al., 2019) and CLINICALBERT (Alsentzer et al., 2019), show that medical DAPT improves fine-tuning performance on tasks such as medical concept extraction from patient reports (Uzuner et al., 2011), identification of gene-disease relations from PubMed abstracts (Doğan et al., 2014; Bravo et al., 2015; Krallinger et al., 2017), and natural language inference on clinical notes (Romanov and Shivade, 2018).

More recent works suggest that decoder-based autoregressive LLMs and VLMs trained via medical DAPT also show strong performance on various medical tasks. Medical LLMs such as MED-TRON (Chen et al., 2023), adapted from LLAMA-2 (Touvron et al., 2023b); and BIOMISTRAL (Labrak et al., 2024), adapted from MISTRAL-7B-INSTRUCT-V0.1 (Jiang et al., 2023); perform well on knowledge-intensive QA tasks based on medical licensing and academic exams (Jin et al., 2020; Pal et al., 2022; Hendrycks et al., 2021)

and PubMed abstracts (Jin et al., 2019). Medical VLMs such as LLAVA-MED (Li et al., 2023), adapted from LLAVA (Liu et al., 2023); and MED-FLAMINGO (Moor et al., 2023b), adapted from OPEN-FLAMINGO (Awadalla et al., 2023); also perform well on visual QA tasks based on radiology (Lau et al., 2018; Liu et al., 2021) and pathology images (He et al., 2020) and academic exams (Yue et al., 2024). These encouraging results have established DAPT as a go-to approach for training a medically specialized model, a conclusion that we re-examine in this work.

3 Experimental Setup

To investigate the effectiveness of medical DAPT in improving zero-/few-shot performance, we compare 7 medical LLMs and 2 medical VLMs against their general-domain counterparts in *pairs* (Figure 1(a)), on 13 textual QA datasets and 8 visual QA datasets, respectively. The models in each pair are *exactly identical* in model architecture and scale, and their only difference lies in whether they were additionally pretrained on medical data. We also note that while some of datasets used for evaluation contain both closed-ended (i.e., has clear ground-truth answers) and open-ended questions, we focus our evaluations on the former, where an objective, quantitative assessment of medical knowledge and reasoning capabilities is possible. For reproducibility of our results, we open-source the source code used for all of our evaluations described below via our GitHub repository¹.

Models. In Table 1, we provide a summary of all of the LLM and VLM pairs that we use for evaluation, along with details about the pretraining

¹<https://github.com/taekb/eval-medical-dapt>

corpora used for adaptation to the medical domain. For LLaVA (Liu et al., 2023), we use the very first version (v0) that uses VICUNA-v0 (Chiang et al., 2023) as the LLM backbone, as LLaVA-MED (Li et al., 2023) was adapted from that particular version. For all models, we use the checkpoints made available via HuggingFace. In all experiments, we generate predictions from each model via (i) greedy decoding (i.e., sampling with temperature $T = 0$) and (ii) constrained decoding. For constrained decoding, we constrain the token vocabulary to be one of the answer choice letters (e.g., one of [“A”, “B”, “C”, “D”] for a four-choice QA dataset) and treat the answer choice with the highest token probability as a given model’s prediction.

Textual QA Datasets. For textual QA, we use MedQA (Jin et al., 2020), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), and MMLU-Medical (Hendrycks et al., 2021) for evaluation. MMLU-Medical refers to a subset of MMLU corresponding to 9 subjects related to medicine: anatomy, clinical knowledge, college biology, college medicine, high school biology, medical genetics, nutrition, professional medicine, and virology. For MedQA, we use the official train-validation-test splits as provided through BigBio (Fries et al., 2022b). We note that MedQA has two versions, one with four answer choices per question and the other with five, and we use both for evaluation. For MedMCQA, which does not have a public test set, we follow the approach taken by Wu et al. (2024) and Labrak et al. (2024), taking a random 80–20 train-validation split of the official training set and using the official validation set for testing. For PubMedQA, we follow Singhal et al. (2023a), using the 211k artificially generated QA samples for training, and taking a 50–50 split on the 1k expert-labeled examples. For MMLU-Medical, we use the official split as provided. We provide the remaining dataset details in Appendix A.

Visual QA Datasets. For visual QA, we use VQA-RAD (Lau et al., 2018), PathVQA (He et al., 2020), SLAKE (Liu et al., 2021), and MMMU-Medical (Yue et al., 2024) for evaluation. MMMU-Medical refers to a subset of MMMU corresponding to 5 subjects relevant to medicine: basic medical science, clinical medicine, diagnostics and laboratory medicine, pharmacy, and public health. For VQA-RAD, we address the train-test leakage and duplication issues in the official train-test splits, previously noted by Moor et al. (2023b), by remov-

ing the training examples repeated in the test set and removing all duplicates in both sets. We then take a random 80–20 split on the training set to create a new train-validation split, as the official split does not include a validation set. For MMMU-Medical, which does not have a public test set, we randomly select 5 examples from the official validation set for validation, and reserve the remaining 25 examples for testing. For all other datasets, we use the official split as provided. We provide the remaining dataset details in Appendix A.

Evaluation Metric. Since we focus on closed-ended QA tasks, we use exact-match accuracy as our main evaluation metric. Following the Holistic Evaluation of Language Models (HELM) benchmark (Liang et al., 2023), when we consider greedy decoding, we treat the text generated by a model (without any constraints on the vocabulary) to be its prediction, and check for an exact match between the prediction and the correct answer up to primitive string operations (e.g., lower-casing, removing white space/punctuation). To handle cases where the model simply repeats the list of answer choices or produces an ambiguous answer (e.g., selecting multiple answer choices), we take a conservative approach and treat the prediction to be incorrect, even if there is a match. Meanwhile, to quantify the extent of *improvement* from medical DAPT, we also consider the *relative* accuracy of the medical model with respect to the general-domain model. Formally, we define relative exact-match accuracy as $\mathbb{E}[\mathbb{1}[f_{\text{medical}}(x) = y] - \mathbb{1}[f_{\text{general}}(x) = y]] \in [-1, 1]$, where f_{medical} and f_{general} denote the medical and general-domain models, x and y denote the input prompt and answer in a QA pair from the test set, and $\mathbb{1}[\cdot]$ denotes the indicator function. This metric quantifies the difference in accuracy between the medical model and the general-domain model. To distinguish the two metrics, we refer to the former as the *absolute* exact-match accuracy in subsequent discussions.

Assessing Statistical Significance. Given the relatively small size of test datasets in medical QA benchmarks, it is important to assess whether the perceived improvements in performance from medical DAPT are attributable to chance. To account for statistical uncertainty, we use the percentile bootstrap, re-sampling (with replacement) questions from the test set to get a sample of the same size as the original test set. Within each resample, we compute the difference in accuracy for the

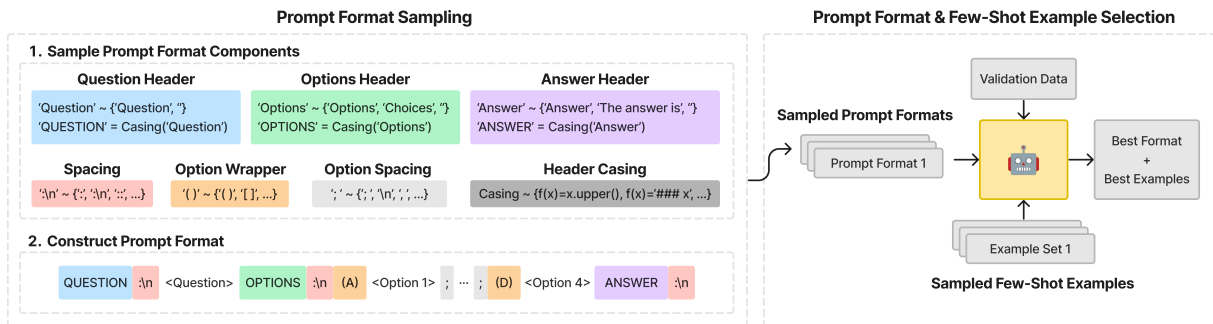


Figure 2: Overview of the prompt format sampling (left) and prompting strategy selection (right) process.

paired models, and repeat this process for 10,000 iterations. The resulting distribution of relative accuracy is used to derive a 95% confidence interval, and we judge a difference to be statistically significant if this interval does not cross zero. We do not perform any type of multiple-testing correction, which would have the effect of lowering the number of comparisons deemed to be significant.

3.1 Zero-/Few-shot Prompting with Model-Specific Prompt Selection

In this section, we provide an overview of our approach to assess whether medical DAPT leads to statistically significant improvements in zero-/few-shot medical QA performance. For few-shot prompting, we consider the 3-shot setting to ensure that the input prompt is shorter than the context window sizes for all models evaluated. For evaluation, we pay special attention to two aspects. First, language models are highly sensitive to the choice of prompting strategy (e.g., prompt format, choice of few-shot examples), where seemingly insignificant changes to the prompt can lead to idiosyncratic model behavior (Jiang et al., 2020; Zhao et al., 2021). Second, prior works show that the “optimal” choice of prompt format rarely correlates between different models (Sclar et al., 2024), suggesting that using a single, fixed prompt for all models for comparison can result in misleading conclusions.

To ensure a fair comparison that isolates the impact of medical DAPT, we treat the choice of prompt format and few-shot examples as additional hyperparameters when generating predictions, and tailor them to each model *independently* (Figure 2). We first randomly sample 10 plausible prompt formats from a predefined search space and 10 different sets of few-shot examples from the training set of each dataset. We then search over all pairs of prompt formats (plus one additional manually designed default format) and few-shot examples,

and select the best pair out of $(10 + 1) \times 10 = 110$ that results in the highest validation exact-match accuracy. Given that a grid search at this scale can be computationally expensive, especially for datasets like MedMCQA that contain 37k validation QA pairs (see Table A1), we randomly subsample 500 validation QA pairs for datasets that have more than 500. Using the vLLM framework (Kwon et al., 2023) for sampling model outputs, this leads to a runtime of around 5–15 minutes per trial, on 4 NVIDIA A6000 GPUs for the 70B models and 2 GPUs for the others. We then generate predictions on the test set using the selected prompt format and few-shot samples. In the zero-shot setting, we only search over the prompt formats.

To define the prompt format search space, we follow the approach by Sclar et al. (2024) and construct a context-free grammar of semantically equivalent yet syntactically distinct prompt formats (Figure 2, left). For the medical models that have a specific prompt format designed and recommended for closed-ended QA tasks (e.g., BIOMISTRAL (Labrak et al., 2024)), we fix the prompt format to what is provided and only search over the choice of few-shot examples. In the case when such information is missing or only partially available (see Table C1), we search over both the prompt formats and few-shot examples. For instruction-tuned models, which typically have a structured conversational format (e.g., “### User: . . . ### Assistant: . . .”) that is expected, we use the sampled question and answer templates to format each “user” query and “assistant” response. We provide the remaining details in Appendix B–C.

4 Results

Here, we summarize the main findings from the zero-/few-shot prompting experiments outlined in Section 3. Unless specified otherwise, we focus

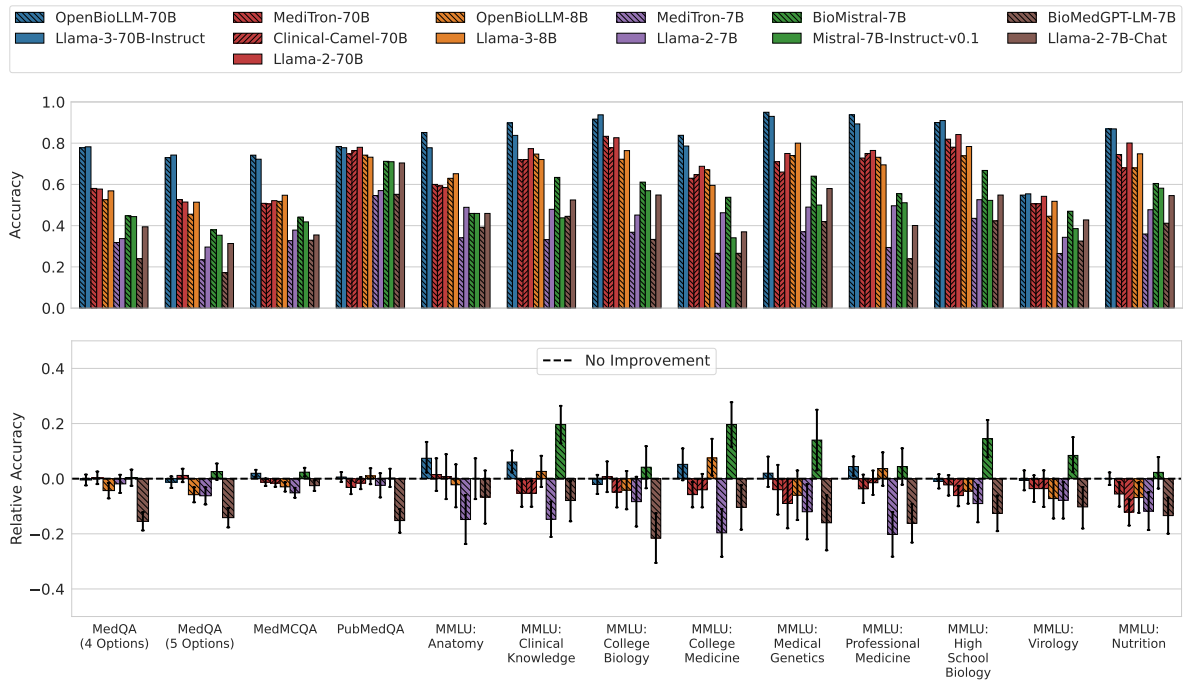


Figure 3: Medical LLMs do not consistently show a statistically significant improvement over their general-domain counterparts in the 3-shot setting, after independently selecting the best prompt format and examples for each model. Top row shows the absolute exact-match accuracies on the test set, and bottom row shows the relative exact-match accuracies along with 95% confidence intervals derived via bootstrapping on the test set (see Section 3). Here, we show the results for greedy decoding. The 3-shot results for constrained decoding are similar (see Figure E1(b)).

on the greedy decoding results in subsequent discussions and include the results for constrained decoding in Appendix E. Overall, we find that all medical VLMs and nearly all medical LLMs fail to consistently improve over their general-domain counterparts in the zero-shot and few-shot prompting regimes. Moreover, we demonstrate the importance of rigorous experimental design in surfacing this finding—performing pairwise model comparison with a single, fixed prompt optimized only for the medical model, while ignoring statistical uncertainty, paints a misleadingly optimistic picture of medical DAPT performance.

Finding 1: After model-specific prompt selection, the vast majority of medical models fail to consistently show a statistically significant improvement over the general-domain models. In Figures 3–4, we show the absolute and relative exact-match accuracies achieved by the medical and general-domain LLMs and VLMs in the zero-/few-shot prompting regime. For LLMs, we only show the 3-shot prompting results in the main text (see Appendix D for results in the zero-shot setting, which are similar). We exclude the results for CLINICAL-CAMEL-70B on both versions of

MedQA, as the model has already been trained on a subset of the official training split (see Table 1 in Toma et al. (2023)). For VLMs, we show both zero-shot and 3-shot results, as LLAVA-v0-7B and LLAVA-MED-7B were not pretrained to handle inputs with multiple images. We calculate the confidence intervals via bootstrapping on the test set, as described in Section 3.

The top row of Figure 3 shows that the absolute exact-match accuracies are mostly similar between each model pair across all datasets and model scales, with marginal performance improvements. In fact, the bottom row of Figure 3 shows that **only 2 out of 7 medical LLMs**—OPENBIOLLM-70B and BIOMISTRAL-7B—show statistically significant improvements in performance, with the 95% confidence intervals crossing zero relative accuracy in most cases for the other models. When compared against their corresponding base models, OPENBIOLLM-70B achieves a win rate of 30.8%, tie rate of 69.2%, and loss rate of 0%, while BIOMISTRAL-7B achieves a win rate of 46.2%, tie rate of 53.8%, and loss rate of 0% (Table D2). Notably, MEDI TRON-7B and BIOMEDGPT-LM-7B actually show significantly worse performance than their base models, with loss rates of 76.9%

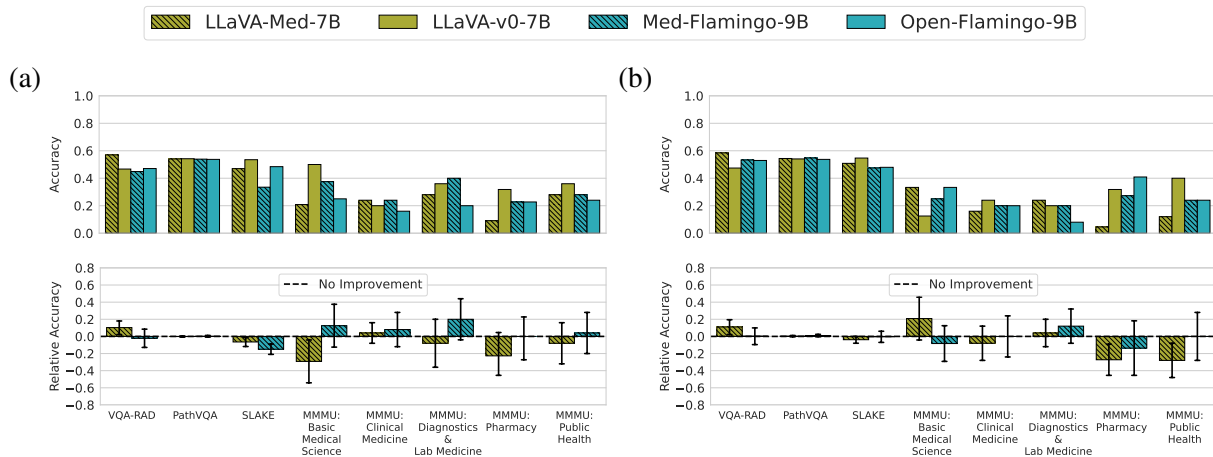


Figure 4: Medical VLMs do not show a statistically significant improvement over their general-domain counterparts in the (a) zero-shot and (b) 3-shot settings, after independently selecting the best prompt format and examples for each model. Top row shows the absolute exact-match accuracies on the test set, and bottom row shows the relative exact-match accuracies along with 95% confidence intervals derived via bootstrapping on the test set (see Section 3). Here, we show the results for greedy decoding. The results for constrained decoding are similar (see Figure E2).

and 92.3%, respectively. Similar trends hold for the zero-shot setting (Figure D1 and Table D1), where only CLINICAL-CAMEL-70B and BIOMISTRAL-7B show statistically significant improvements.

We note that, while OPENBIO-LLM-70B shows improvement in the 3-shot setting, it does not show improvement in the zero-shot setting (winning on 7.7% and losing on 23.1% of tasks, see Table D1), and vice versa for CLINICAL-CAMEL-70B (winning on 0% of tasks and losing on 36.4% of tasks in the 3-shot setting, see Table D2), **leaving BIOMISTRAL-7B as the only medical LLM that wins more than it loses** against its base model (MISTRAL-7B-INSTRUCT-V0.1) in both settings, albeit with relatively low absolute performance.

In Figure 4, we make similar observations for medical VLMs in both zero-shot and 3-shot settings, where both LLaVA-MED-7B and MED-FLAMINGO-9B are **virtually indistinguishable from their base models in terms of performance, showing no statistically significant improvements**. Tables D1–D2 show that LLaVA-MED-7B achieves win/tie/loss rates of 12.5%/62.5%/25.0% in both zero-shot and 3-shot settings, while MED-FLAMINGO-9B achieves win/tie/loss rates of 0%/87.5%/12.5% in the zero-shot setting and 0%/100%/0% in the 3-shot setting. Meanwhile, we note that the confidence intervals for the MMMU-Medical datasets tend to be much wider than for the other visual QA datasets, as the test sets only include 25 QA examples for each subject (Table A1).

We similarly observe limited improvements overall with constrained decoding (see Appendix E.1). As shown in Figure E5(a), when we aggregate the results over all (model pair, QA dataset) combinations, medical LLMs achieve win/tie/loss rates of 16.9%/68.6%/14.5% in the zero-shot setting and 11.2%/74.1%/14.7% in the 3-shot setting, while medical VLMs achieve win/tie/loss rates of 6.3%/87.5%/6.3% in the zero-shot setting and 0%/93.8%/6.3% in the 3-shot setting. In fact, no medical VLM shows improvement over their base models regardless of the decoding strategy. Meanwhile, as shown in Tables E1–E2, we find that some medical LLMs show larger improvements with constrained decoding (notably, MEDITRON-70B and MEDITRON-7B), although the results are mixed (e.g., CLINICAL-CAMEL-70B performs worse in the zero-shot setting with constrained decoding).

In summary, these results suggest that when prompted with the “right” set of examples in an appropriate format, general-domain models may already exhibit the capacity to achieve performance competitive with medically adapted models, on various medical QA tasks.

Finding 2: Using a single, fixed prompt for all models and overlooking statistical uncertainty may overestimate the performance benefits of medical DAPT. Based on Finding 1, we further investigate whether the conclusions differ if the same prompt is used for each pair of medical and general-domain models. In particular, we consider

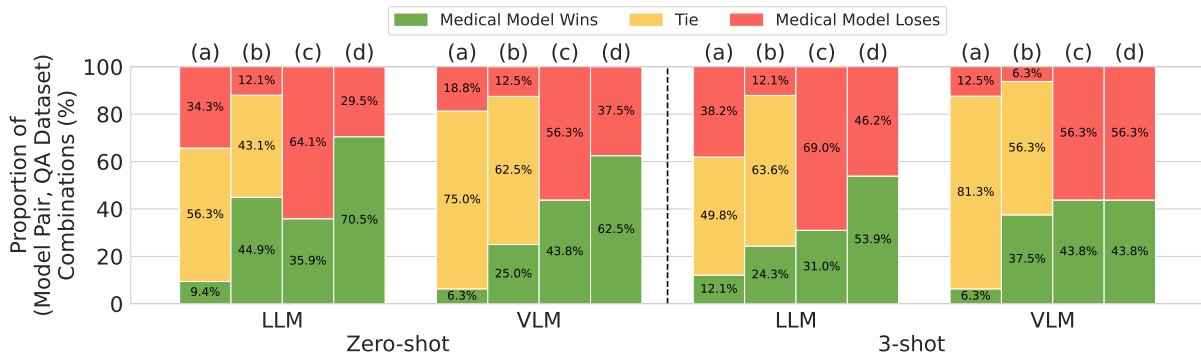


Figure 5: Optimizing the prompt for only the medical model and comparing models without accounting for statistical uncertainty can overestimate the performance improvements from medical DAPT. We show the win/tie/loss rate (%) of medical models vs. their base models across all (model pair, QA dataset) combinations, when (a) independently optimizing the prompt for each model and performing statistical testing, (b) optimizing the prompt only for the medical model and performing statistical testing, (c) independently optimizing the prompt for each model without statistical testing, and (d) optimizing the prompt only for the medical model without statistical testing. Here, we show the results for greedy decoding. The results for constrained decoding are similar (see Figure E5).

whether selecting a prompt only for the medical model, following Section 3.1, and using it for the corresponding general-domain model can widen the performance gap between each pair. We also assess whether this gap becomes amplified when models are compared without accounting for statistical uncertainty, which is often done in practice.

In Figure 5, we show how the win/tie/loss rates of the medical models, computed over all (model pair, QA dataset) combinations, change as we vary the following aspects of the experimental setup:

1. select prompts for each model independently vs. only based on the medical model;
2. determine a win for the medical model based on confidence intervals in relative accuracy vs. raw absolute accuracy.

We note that when comparing each model pair based on absolute accuracy, there are no ties, as the real-valued absolute accuracies are rarely identical. In Appendix D, we include Figures D2–D3 to show how the absolute and relative exact-match accuracies change when the prompt is only optimized for the medical model. We also include Tables D3–D4 to show changes in win/tie/loss rates. We show the same set of results for constrained decoding in Figures E3–E4 and Tables E3–E4 in Appendix E.

Overall, we find that for both LLMs and VLMs, the performance improvement from using a medically adapted model instead of its general-domain counterpart can be substantially overestimated when (i) the prompt is only tailored to the medical model; and (ii) the models are compared only based

on their absolute accuracies. Notably, in the zero-shot setting, the win rate increases from 9.4% to 70.5% for medical LLMs and from 6.3% to 62.5% for medical VLMs, when only performing prompt selection for the medical model and comparing based on absolute accuracy. Figure E5 in Appendix E.2 shows a similar trend in the win/tie/loss rates, when the model predictions are generated via constrained decoding. *These results highlight the importance of accounting for LLM/VLM sensitivity to the prompting details, as suggested by Sclar et al. (2024), and the statistical uncertainty in model comparison, in order to draw reliable conclusions about the effectiveness of medical DAPT.*

5 Discussion and Conclusion

In this work, we investigated the effectiveness of DAPT for training medically specialized LLMs and autoregressive VLMs suitable for knowledge-intensive medical (visual) QA tasks. To that end, we compared several pairs of state-of-the-art medical LLMs/VLMs to their general-domain counterparts, whose only differences lie in medical DAPT and are exactly identical in model architecture and scale. Our work diverges from prior works by providing a direct apples-to-apples comparison of medical and general-domain models while accounting for LLM/VLM sensitivity to prompting details and assessing the statistical significance of the results.

Across both model classes and all model scales, we found that the performance benefits from medical DAPT largely disappear when we (i) tailor the prompt format and choice of few-shot examples

to each medical and general-domain model separately; and (ii) account for statistical uncertainty in model comparison. In particular, we found that when we optimize the prompt only for the medical model and compare each model pair based on their absolute accuracies without accounting for uncertainty, the performance improvements from medical DAPT can be overestimated, potentially leading to unreliable conclusions about the benefits of medical DAPT. For example, in the zero-shot setting, evaluation under this setup leads to the conclusion that medical LLMs and VLMs, on average, outperform the corresponding general-domain models in 70.5% and 62.5% of all QA tasks, while the improvements are in reality statistically significant in only 9.4% and 6.3% of tasks after optimizing the prompt for each model to ensure a fair comparison.

Our findings suggest that for state-of-the-art general-domain LLMs and VLMs, the performance benefits from additionally pretraining on medical data from public sources such as PubMed may be limited. Notably, almost all of the medical models used in our evaluation use PubMed as the primary source of pretraining data for medical adaptation (Table 1), while open-source datasets commonly used for pretraining the general-domain base models in the first place (e.g., the Pile (Gao et al., 2020), S2ORC (Lo et al., 2020)) often already include PubMed data. Prior works also suggest that the intrinsic *capacity* of LLMs to solve a downstream task is largely obtained during the *initial* pretraining phase, and that post-training adjustments and prompt engineering efforts may only help elicit the existing capabilities (Reynolds and McDonell, 2021; Min et al., 2022). Thus, we argue that any claims about improvement from a proposed medical DAPT procedure should be evidenced by rigorous head-to-head comparisons against the corresponding general-domain model, in order to draw reliable conclusions about its effectiveness.

6 Limitations

We discuss our findings with the following caveats.

First, there is a vast and growing set of papers on applying medical DAPT to various general-domain base models, and we could not hope to compare all publicly available models here. While we selected the models to cover a wide range of general-domain base models and model scales (7B–70B) (Table 1) and included some of the latest models (e.g., OPEN-BIOLLM and LLAMA-3), it is always possible that

some newly released models do in fact yield better zero- or few-shot performance on medical QA.

Second, we focus in this paper on the narrower task of closed-ended medical QA. In part, this choice reflects the fact that such benchmarks are well-standardized and highly publicized. However, they do not reflect the breadth of possible applications of LLMs and VLMs in medical domains. For instance, Singhal et al. (2023b) show that medical LLMs such as MED-PALM-2 can produce physician-level answers to open-ended consumer health queries, and Agrawal et al. (2022) demonstrate the potential of using LLMs for extracting information from structured clinical notes. Some would argue that such tasks are a more realistic application of such models in practice, and it is certainly possible that an analysis like ours would find improved performance on such tasks, though we do not investigate these tasks in the present work.

Third, we do not consider downstream fine-tuning of models subject to medical DAPT. In part, this reflects issues of computational cost (e.g., to fine-tune 70B-parameter models) and the added complexity of reproducing a fine-tuning procedure, versus using publicly available model checkpoints. However, we acknowledge that zero- and few-shot performance are only part of a broader narrative around the claimed benefits of medical DAPT, which generally includes the additional claim that it provides a better initialization for downstream fine-tuning (Chen et al., 2023; Li et al., 2023).

While we acknowledge the limitations above, we do not believe they detract from the value of this work. We hope that our results call attention to a need for rigorous head-to-head evaluations when making similar claims of improved performance via medical DAPT, whether with other models, on other clinical tasks, or with respect to fine-tuning versus zero-/few-shot performance.

Acknowledgments

We gratefully acknowledge DARPA (FA8750-23-2-1015), ONR (N00014-23-1-2368), NSF (IIS2211955), UPMC, Highmark Health, Abridge, Ford Research, Mozilla, the PwC Center, Amazon AI, JP Morgan Chase, the Block Center, the Center for Machine Learning and Health, and the CMU Software Engineering Institute (SEI) via Department of Defense contract FA8702-15-D-0002, for their generous support of our research.

References

- Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large Language Models are Few-Shot Clinical Information Extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*.
- Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. *arXiv:2308.01390*.
- Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjoui, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D. Manning. 2024. BioMedLM: A 2.7B Parameter Language Model Trained On Biomedical Text. *arXiv:2403.18421*.
- À Bravo, J Piñero, N Queralt-Rosinach, M Rautschka, and LI Furlong. 2015. Extraction of Relations Between Genes and Diseases from Text and Large-scale Data Analysis: Implications for Translational Research. *BMC Bioinformatics*, 16(55).
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Alberto Mario Ceballos-Arroyo, Monica Munnangi, Jinding Sun, Karen Zhang, Jered McInerney, Byron C. Wallace, and Silvio Amir. 2024. Open (Clinical) LLMs are Sensitive to Instruction Phrasings. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing (BioNLP)*.
- Tuhin Chakrabarty, Christopher Hidey, and Kathy MCKeown. 2019. IMHO Fine-Tuning Improves Claim Detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.
- Zeming Chen, Alejandro Hernández-Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. MediTron-70B: Scaling Medical Pretraining for Large Language Models. *arXiv:2311.16079*.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.
- Clément Christophe, Praveen K Kanithi, Prateek Mungal, Tathagata Raha, Nasir Hayat, Ronnie Rajan, Ahmed Al-Mahrooqi, Avani Gupta, Muhammad Umar Salman, Gurpreet Gosal, Bhargav Kanakiya, Charles Chen, Natalia Vassilieva, Boulbaba Ben Amor, Marco AF Pimentel, and Shadab Khan. 2024. Med42 – Evaluating Fine-Tuning Strategies for Medical LLMs: Full-Parameter vs. Parameter-Efficient Approaches. *arXiv:2404.14779*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.
- Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI Disease Corpus: A Resource for Disease Name Recognition and Concept Normalization. *Journal of Biomedical Informatics*, 47:1–10.
- Jason Fries, Ethan Steinberg, Scott Fleming, Michael Wornow, Yizhe Xu, Keith Morse, Dev Dash, and Nigam Shah. 2022a. How Foundation Models Can Advance AI in Healthcare.
- Jason Fries, Leon Weber, Natasha Seelam, Gabriel Al-tay, Debajyoti Datta, Samuele Garda, Sunny Kang, Rosaline Su, Wojciech Kusa, Samuel Cahyawijaya, Fabio Barth, Simon Ott, Matthias Samwald, Stephen Bach, Stella Biderman, Mario Sängler, Bo Wang, Alison Callahan, Daniel León Periñán, Théo Gigant, Patrick Haller, Jenny Chim, Jose Posada, John Giorgi, Karthik Rangasai Sivaraman, Marc Pàmies, Marianna Nezhurina, Robert Martin, Michael Cullan, Moritz Freidank, Nathan Dahlberg, Shubhan-shu Mishra, Shamik Bose, Nicholas Broad, Yanis Labrak, Shlok Deshmukh, Sid Kiblawi, Ayush Singh, Minh Chien Vu, Trishala Neeraj, Jonas Golde, Albert Villanova del Moral, and Benjamin Beilharz. 2022b. BigBio: A Framework for Data-Centric Biomedical Natural Language Processing. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn

- Presser, and Connor Leahy. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv:2101.00027*.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *Association for Computing Machinery (ACM) Transactions on Computing for Healthcare*, 3(1).
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. PathVQA: 30000+ Questions for Medical Visual Question Answering. *arXiv preprint arXiv:2003.10286*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations (ICLR)*.
- Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. 2023. Do We Still Need Clinical Language Models? In *Proceedings of the Conference on Health, Inference, and Learning (CHIL)*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth e Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv:2310.06825*.
- Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics*, 8:423–438.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What Disease Does This Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. *arXiv:2009.13081*.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.
- Leonard Konle and Fotis Jannidis. 2020. Domain and Task Adaptive Pretraining for Language Models. In *Workshop on Computational Humanities Research (CHR)*.
- Martin Krallinger, Obdulia Rabal, Saber Ahmad Akhondi, Mart n P rez P rez, Jesus Santamar a, Gael P rez Rodr guez, Georgios Tsatsaronis, Ander Intxaurreondo, Jos  Antonio Baso L pez, Umesh K. Nandal, Erin M. van Buel, Ambika Chandrasekhar, Marleen Rodenburg, Astrid L greid, Marius A. Doornenbal, Julen Oyarz bal, An lia Louren o, and Alfonso Valencia. 2017. Overview of the BioCreative VI Chemical-Protein Interaction Track. In *Proceedings of the BioCreative VI Workshop*.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *ACM Symposium on Operating Systems Principles*.
- Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. *arXiv:2402.10373*.
- Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A Dataset of Clinically Generated Visual Questions and Answers about Radiology Images. *Nature Scientific Data*, 5(180251).
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics*, 36(4):1234–1240.
- Chunyuhan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yan Zhang, Deepak Narayanan, Yuhuai Wu, Anyanya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher R , Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic Evaluation of Language Models. *arXiv:2211.09110*.

- Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. PMC-CLIP: Contrastive Language-Image Pre-training using Biomedical Documents. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*.
- Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. SLAKE: A Semantically-Labeled Knowledge-Enhanced Dataset for Medical Visual Question Answering. *arXiv:2102.09542*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
- Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining. In *Briefings in Bioinformatics*.
- Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. 2023. BioMedGPT: Open Multimodal Generative Pre-trained Transformer for Biomedicine. *arXiv:2308.09442*.
- Mason Marks and Claudia E. Haupt. 2023. AI Chatbots, Health Privacy, and Challenges to HIPAA Compliance. *Journal of American Medical Association (JAMA)*, 330(4):309–310.
- Meta. 2024. [Introducing Meta Llama 3: The Most Capable Openly Available LLM to Date](#).
- Sewon Min, Xinxu Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Michael Moor, Oishi Banerjee, Zahra Shakeri, Harlan Krumholz, Jure Leskovec, Eric Topol, and Pranav Rajpurkar. 2023a. Foundation Models for Generalist Medical Artificial Intelligence. *Nature*, 616:259–265.
- Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. 2023b. Med-Flamingo: A Multimodal Medical Few-shot Learner. *arXiv:2307.15189*.
- OpenAI. 2023a. GPT-4 Technical Report. *arXiv:2303.08774*.
- OpenAI. 2023b. [GPT-4V\(ision\) System Card](#).
- Ankit Pal, Pasquale Minervini, Andreas Geert Motzfeldt, Aryo Pradipta Gema, and Beatrice Alex. 2024. [Open Medical LLM Leaderboard](#).
- Ankit Pal and Malaikannan Sankarasubbu. 2024. [Open-BioLLMs: Advancing Open-Source Large Language Models for Healthcare and Life Sciences](#).
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. In *Proceedings of the Conference on Health, Inference, and Learning (CHIL)*.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. In *OpenAI Blog*.
- Laria Reynolds and Kyle McDonell. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*.
- Alexey Romanov and Chaitanya Shivade. 2018. Lessons from Natural Language Inference in the Clinical Domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambano Chaves, Szu-Yeu Hu, Mike Schaeckermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, Si-Wai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Ben Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale Webster, Joelle Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottwais, Alan Karthikesalingam, and Vivek Natarajan. 2024. Capabilities of Gemini Models in Medicine. *arXiv:2404.18416*.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying Language Models’ Sensitivity to Spurious Features in Prompt Design or: How I Learned to Start Worrying About Prompt Formatting. In *International Conference on Learning Representations (ICLR)*.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Mahdavi, Jason Wei, Hyung Chung, Nathan Scales,

- Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and Vivek Natarajan. 2023a. Large Language Models Encode Clinical Knowledge. *Nature*, 620:1–9.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023b. Towards Expert-Level Medical Question Answering with Large Language Models. *arXiv:2305.09617*.
- Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A Large Language Model for Science. *arXiv:2211.09085*.
- Augustin Toma, Patrick R. Lawler, Jimmy Ba, Rahul G. Krishnan, Barry B. Rubin, and Bo Wang. 2023. Clinical Camel: An Open Expert-Level Medical Language Model with Dialogue-Based Knowledge Encoding. *arXiv:2305.12031*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and Efficient Foundation Language Models. *arXiv:2302.13971*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shrutu Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv:2307.09288*.
- Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, Anil Palepu, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, Christopher Semturs, S. Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam, and Vivek Natarajan. 2024. Towards Generalist Biomedical AI. *New England Journal of Medicine (NEJM) AI*, 1(3).
- Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA Challenge on Concepts, Assertions, and Relations in Clinical Text. *Journal of American Medical Informatics Association (JAMIA)*, 18(5):552–556.
- Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. PMC-LLaMA: Toward Building Open-Source Language Models for Medicine. *Journal of the American Medical Informatics Association (JAMIA)*, page ocae045.
- Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A Pre-trained Language Model for Chinese Legal Long Documents. *AI Open*, 2:79–84.
- Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, Eric Wang, Ellery Wulczyn, Fayaz Jamil, Theo Guidroz, Chuck Lau, Siyuan Qiao, Yun Liu, Akshay Goel, Kendall Park, Arnav Agharwal, Nick George, Yang Wang, Ryutaro Tanno, David G. T. Barrett, Wei-Hung Weng, S. Sara Mahdavi, Khaled Saab, Tao Tu, Sreenivasa Raju Kalidindi, Mozziyar Etemadi, Jorge Cuadros, Gregory Sorensen, Yossi Matias, Katherine Chou, Greg Corrado, Joelle Barral, Shravya Shetty, David Fleet, S. M. Ali Eslami, Daniel Tse, Shruthi Prabhakara, Cory McLean, Dave Steiner, Rory Pilgrim, Christopher Kelly, Shekoofeh Azizi, and Daniel Golden. 2024. Advancing Multimodal Medical Capabilities of Gemini. *arXiv:2405.03162*.
- Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, AB Costa, Mona G Flores, Ying Zhang, Tanja Magoc, Christopher A Harle, Gloria Lipori, Duane A Mitchell, William R Hogan, Elizabeth A Shenkman, Jiang Bian, and Yonghui Wu. 2022. A Large Language Model for Electronic Health Records. *npj Digital Medicine*, 5(194).
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhao Chen. 2024. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In

Conference on Computer Vision and Pattern Recognition (CVPR).

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. 2023. BiomedCLIP: A Multimodal Biomedical Foundation Model Pretrained from Fifteen Million Scientific Image-Text Pairs. *arXiv:2303.00915*.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate Before Use: Improving Few-Shot Performance of Language Models. In *International Conference on Machine Learning (ICML)*.

A Additional Details on Datasets

Table A1: Summary of the number of examples in the train, validation, and test sets of all textual and visual QA datasets used for evaluation, in the top and bottom sections, respectively.

Dataset	Train	Validation	Test
MedQA (4 & 5 Options)	10178	1272	1273
MedMCQA	146257	36565	4183
PubMedQA	211269	500	500
MMLU: Anatomy	5	14	135
MMLU: Clinical Knowledge	5	29	265
MMLU: College Biology	5	16	144
MMLU: College Medicine	5	22	173
MMLU: High School Biology	5	32	310
MMLU: Medical Genetics	5	11	100
MMLU: Nutrition	5	33	306
MMLU: Professional Medicine	5	31	272
MMLU: Virology	5	18	166
VQA-RAD	820	205	272
PathVQA	9806	3135	3391
SLAKE	1943	422	415
MMMU: Basic Medical Science	5	5	25
MMMU: Clinical Medicine	5	5	25
MMMU: Diag. & Lab Medicine	5	5	25
MMMU: Pharmacy	5	5	25
MMMU: Public Health	5	5	25

In this section, we provide additional details on the textual and visual QA datasets introduced in Section 3. In Table A1, we summarize the number of QA examples included in the train, validation, and test sets of each dataset, after following the preprocessing steps detailed in Section 3. For VQA-RAD (Lau et al., 2018), PathVQA (He et al., 2020), and SLAKE (Liu et al., 2021), we only show the number of *closed-ended* visual QA examples, since our evaluations focus on closed-ended visual QA. For the datasets that required additional splits from the official train-validation-test split (e.g., due to the lack of a public test set), we include all of the fixed random seeds in our repository for reproducibility.

B Additional Details on Model-Specific Prompt Selection

In this section, we provide additional details on how we define the prompt format search space discussed in Section 3.1. We construct a context-free grammar of plausible prompt formats following the approach by Sclar et al. (2024) (see Section 3.1 and Appendix A of the paper for reference). Using the Backus-Naur notation, we first define the basic fields H_q for the question header (e.g., “### Question:”), H_c for the answer choice header

(e.g., “### Options:”), and H_a for the answer header (e.g., “### Answer:”) as

$$\begin{aligned} H_q(f_{\text{case}}, d_q, s_1) &::= f_{\text{case}}(d_q) s_1 \langle \text{text} \rangle, \\ H_c(f_{\text{case}}, d_c, s_1) &::= f_{\text{case}}(d_c) s_1, \\ H_a(f_{\text{case}}, d_a, s_1) &::= f_{\text{case}}(d_a) s_1 \langle \text{text} \rangle, \end{aligned}$$

where $f_{\text{case}} \in \mathcal{F}_{\text{case}}$ denotes the casing function (e.g., $x \mapsto \text{“###”} + x$, $x \mapsto x.\text{upper}()$), $d_q \in D_q$ denotes the question descriptor (e.g., “Question”), $d_c \in D_c$ denotes the answer choice descriptor (e.g., “Options”), $d_a \in D_a$ denotes the answer descriptor (e.g., “Answer”), $s_1 \in S_1$ denotes the header separator (e.g., ‘:’), and $\langle \text{text} \rangle$ denotes a text placeholder. For formatting the list of answer choices, we also define the basic fields C for formatting each answer choice (e.g., “(A) yes”) and L for the concatenation of all answer choices as follows:

$$\begin{aligned} C(f_{\text{wrap}}, f_{\text{index}}, i) &::= f_{\text{wrap}}(f_{\text{index}}(i)) \langle \text{text} \rangle, \\ L(f_{\text{wrap}}, f_{\text{index}}, n, s_2) &::= C(f_{\text{wrap}}, f_{\text{index}}, 0) s_2 \dots \\ &\quad s_2 C(f_{\text{wrap}}, f_{\text{index}}, n-1), \end{aligned}$$

where $f_{\text{wrap}} \in \mathcal{F}_{\text{wrap}}$ denotes the wrapper function for the answer choice letter (e.g., $x \mapsto \text{“}(\text{“} + x + \text{“})\text{”}$), $f_{\text{index}} \in \mathcal{F}_{\text{index}}$ denotes the numbering function that converts an integer index into a number format (e.g., $0 \rightarrow \text{“A”}$), $i \in \mathbb{Z}^+$ denotes the index of a particular answer choice from the list, $s_2 \in S_2$ denotes the answer choice separator, n denotes the number of answer choices, and $\langle \text{text} \rangle$ denotes a text placeholder. The full prompt format $P(f_{\text{case}}, f_{\text{wrap}}, f_{\text{index}}, d_q, d_c, d_a, s_1, s_2, n)$ is then constructed by concatenating all of the headers and the answer choices, while adding space $t \in T$ (e.g., “\n”) in-between:

$$P ::= H_q t H_c t L t H_a, \quad (1)$$

where we have left out the notations for the arguments for notational simplicity.

To define the prompt format search space, we instantiate the grammar above with the descriptors, separators, spaces, and functions shown below.

Descriptors:

$$\begin{aligned} D_q &= \{\text{“Question”, “”}\}; \\ D_c &= \{\text{“Options”, “Choices”, “”}\}; \\ D_a &= \{\text{“Answer”, “The answer is”}\}. \end{aligned}$$

Separators:

$$S_1 = \{“: ”, “ : ”, “ :: ”, “:\n”, “= ”, “ = ”, “ == ”, “=\n”, “ - ”, “ - ”, “_”, “\n”, “\n\n”\};$$
$$S_2 = \{“\n”, “\ ”, “; ”, “ || ”, “ ”, “;\n”, “;\n\n”, “, ”\}.$$

Spaces:

$$T = \{“\n”, “\n\n”, “ || ”, “ ”\}.$$

Casing, Wrapper, and Numbering Functions:

$$\mathcal{F}_{\text{case}} = \{x \mapsto x, x \mapsto x.\text{title}(), x \mapsto x.\text{upper}(), x \mapsto x.\text{lower}(), x \mapsto “### ” + x, x \mapsto “**” + x + “**”\};$$
$$\mathcal{F}_{\text{wrap}} = \{x \mapsto “(” + x + “)”, x \mapsto x + “.”, x \mapsto x + “)”, x \mapsto “[” + x + “]”, x \mapsto x + “)”, x \mapsto “<” + x + “>”\};$$
$$\mathcal{F}_{\text{index}} = \{x \mapsto \text{chr}(\text{ord}(\text{“A”}) + x)\}.$$

To randomly sample a prompt format accepted by the grammar, we randomly sample each of these components and construct the full prompt format, following Equation (1). Below, we show an example QA pair from the MedQA dataset (four answer choices), formatted according to the formats sampled from the prompt format space defined by the above context-free grammar.

Example 1:

A key factor facilitating the application of nested case-control studies from the MACS was:
OPTIONS – A) Data collection
B) Establishment of a repository of biologic specimens
C) Participant interest
D) Administration of the questionnaire by staff
THE ANSWER IS – B) Establishment of a repository of biologic specimens

Example 2:

QUESTION – A key factor facilitating the application of nested case-control studies from the MACS was:
CHOICES – [A] Data collection; [B] Establishment of a repository of biologic specimens;

[C] Participant interest; [D] Administration of the questionnaire by staff
ANSWER – [B] Establishment of a repository of biologic specimens

C Additional Details on Zero-/Few-shot Prompting

In this section, we summarize the prompting details made available for the medical LLMs and VLMs used in our evaluation (Appendix C.1), and the default prompt formats used for each LLM (Appendix C.2) and VLM (Appendix C.3), which have been reproduced based on the former.

C.1 Reproducibility of Prompting Details

In Table C1, we provide a summary of all of the prompting details available (in the context of closed-ended medical QA) for all medical LLMs and VLMs used in our evaluation. We share these details to demonstrate our best efforts with reproducing the original prompting setups considered for performing our evaluations. In particular, we focus on whether the following four components are explicitly made available, either in the original publications or the publicly released code repository: (i) system prompt; (ii) zero-/few-shot prompt format (used for closed-ended QA tasks); (iii) the choice of few-shot examples; and (iv) details on how the text generations are sampled (e.g., softmax temperature, top- p , beam size, random seeds used for sampling). Below, we provide detailed clarifications for each model.

OPENBIOLLM (Pal and Sankarasubbu, 2024).

For the OPENBIOLLM models, we follow the instructions provided in the model cards posted by the authors on HuggingFace, for the **70B-parameter** and **8B-parameter** models. We use the recommended system prompt and the LLAMA-3-based conversational prompt format. Meanwhile, in Table C1, we treat the prompt format as partially missing, as the exact format that was used to format each question (“user” query) and answer (“assistant” response) for evaluation on closed-ended multiple-choice questions is not provided. At the time of writing, there are no additional details about the models that have been publicly released, beyond what is provided in the model cards. We include the default prompt format used for OPENBIOLLM in Appendix C.2.1.

Table C1: Summary of all of the prompting details made available for each medical LLM and VLM used for evaluation. For each column, a checkmark (✓) indicates that the information was fully provided, a triangle (▲) indicates that the information was partially provided (e.g., random sampling without information about the seeds), and a cross (✗) indicates that the information was not provided at all. “N/A” indicates that the corresponding information is not available due to its irrelevance to the evaluation setup considered in the paper (e.g., lack of few-shot example details because the model was only originally evaluated in zero-shot or fine-tuning regimes).

Model	System Prompt	Zero-/Few-Shot Prompt Format	Few-Shot Examples	Sampling Details
OPENBIOLLM (Pal and Sankarasubbu, 2024)	✓	▲	✗	✗
CLINICAL-CAMEL (Toma et al., 2023)	✗	▲	✗	▲
BIOMISTRAL (Labrak et al., 2024)	✓	✓	✗	▲
MEDI TRON (Chen et al., 2023)	✓	▲	✓	✓
BIOMEDGPT-LM (Luo et al., 2023)	✗	✗	N/A	✗
LLAVA-MED (Li et al., 2023)	▲	▲	N/A	▲
MED-FLAMINGO (Moor et al., 2023b)	✓	▲	✗	✗

CLINICAL-CAMEL (Toma et al., 2023). For CLINICAL-CAMEL, we use the conversational prompt format used in the official GitHub repository, which corresponds to the official chat format for LLAMA-2 (Touvron et al., 2023b). As the system prompts and few-shot examples used for the main evaluations in the paper are not provided, we use our own manually designed default system prompt and search over different choices of few-shot examples. For sampling, the [evaluation code](#) uses default temperature setting of 0.7 (albeit without the random seeds), which differs from our evaluation setup. We include the default prompt format used for CLINICAL-CAMEL in Appendix C.2.2.

BIOMISTRAL (Labrak et al., 2024). For BIOMISTRAL, we use the system prompt and zero-/few-shot prompt format provided in Appendix F of the paper. At the time of writing, the code repository is not publicly available, and the paper does not provide details on what few-shot examples were used for evaluation. In Section 4.3 of the paper, Labrak et al. (2024) mention that the output vocabulary is constrained to be one of the answer choices in lettered format (e.g., one of [A,B,C,D]) to force the model to avoid generating irrelevant tokens in its output. Meanwhile, it is not explicitly clear whether (i) the filtered token with the highest probability was treated as the model’s prediction or (ii) a token was randomly sampled based on the renormalized token probabilities. We also note that the vocabulary filtering procedure makes their evaluation setup different from ours, as we use greedy decoding to sample the model outputs without any constraints on the vocabulary (see Sec-

tion 3). We include the default prompt format used for BIOMISTRAL in Appendix C.2.3.

MEDI TRON (Chen et al., 2023). For the MEDI TRON models, we use the system prompts—tailored specifically to MedQA, MedMCQA, PubMedQA, and the MMLU datasets—provided in Table 2 of the paper. For the prompt formats, we use the ones provided in the official GitHub repository, as the prompt formats (those with special ‘<|im_start|>’ and ‘<|im_end|>’ tokens, following the [ChatML format](#)) shown in the paper are only applicable to the fine-tuned models (see [this discussion](#) from the official GitHub repository). In particular, we refer to the prompt formats provided in the [dataset preprocessing](#) code and used for [evaluation](#) to determine the default prompt format for both the 70B- and 7B-parameter models. However, we were unable to reliably reproduce the zero-/few-shot prompting performance using this prompt format, and therefore perform a grid search over the prompt formats as well for model-specific prompt selection. In the [evaluation code](#), Chen et al. (2023) provide the random seeds used for sampling the few-shot examples; however, we also search over the set of few-shot examples to consider a larger number of few-shot example choices. For sampling, we use the same greedy decoding approach as considered in the paper (referred to “Top Token Selection” in Section 4.3 of the paper). We include the default prompt format used for MEDI TRON in Appendix C.2.4.

BIOMEDGPT-LM (Luo et al., 2023). While BIOMEDGPT-LM was evaluated on textual med-

ical QA datasets such as MedMCQA and PubMedQA, the evaluation was performed only in the supervised fine-tuning regime, and the prompt formats used for these datasets are not available, to the best of our knowledge. Meanwhile, the official GitHub repository provides [Jupyter notebook examples](#) containing a conversational prompt format used in the context of other QA tasks. We therefore use this format by default but search over the prompt formats for model-specific prompt selection, since it is not specifically designed for closed-ended multiple-choice QA tasks. Moreover, as the system prompt provided is not semantically applicable to the QA tasks that we consider (e.g., “You are working as an excellent assistant in chemistry and molecule discovery.”), we use our own manually designed default system prompt. We include the default prompt format used for BIOMEDGPT-LM in Appendix C.2.5.

LLAVA-MED (Li et al., 2023). For LLAVA-MED, we use the system prompt and conversational prompt format included in the “[simple_conv_med](#)” template from the official GitHub repository (for LLAVA-v0 (Liu et al., 2023), we use the “[simple_conv](#)” template) by default. For formatting the visual questions, we also refer to [this file](#) containing the raw visual QA results on VQA-RAD (“Please choose from the following two options: [yes, no]”). Meanwhile, we make these choices with the following caveats, to the best of our knowledge. First, the exact choice of system prompt and conversational prompt format used for evaluation are not discussed in the paper or the code repository, and we choose the one that has a system prompt specific to LLAVA-MED (“You are LLaVA-Med, a large language and vision assistant trained by a group of researchers at Microsoft . . .”) and follows the conversational format used for VICUNA-v0 (Chiang et al., 2023), which forms its LLM backbone. Second, details on how the answer choices should be formatted in the context of closed-ended QA tasks is only shown in the VQA-RAD results file. Given the uncertainty in such details, we also search over the prompt formats for model-specific prompt selection. We note that LLAVA-MED was not pretrained on multi-image inputs or evaluated in few-shot setting, and therefore details on the choice of few-shot examples are irrelevant. For sampling, the [evaluation code](#) uses a default temperature setting of 0.7 (albeit without the random

seeds), which differs from our evaluation setup. We include the default prompt format used for LLAVA-MED in Appendix C.3.1 and that for LLAVA-v0 in Appendix C.3.2.

MED-FLAMINGO (Moor et al., 2023b). For MED-FLAMINGO, we use the system prompt and prompt format provided in the [demo code](#) from the official GitHub repository by default. However, we search over the prompt formats when performing model-specific prompt selection, as the example prompt in the demo does not show details for formatting answer choices in a closed-ended QA context. The choice of few-shot examples and the sampling details used for the original evaluations on VQA-RAD and PathVQA are not available. We include the default prompt format used for MED-FLAMINGO in Appendix C.3.3 and that for OPEN-FLAMINGO in Appendix C.3.4.

C.2 Default LLM Prompt Formats

In this section, we share the *default* prompt formats that we use for each LLM, using MMLU (Clinical Knowledge) (Hendrycks et al., 2021) as a running example. We denote the system prompt in **red**, any few-shot examples in **green**, and the question being asked of the model in **purple**.

For models that do not have a specific system prompt and prompt format designed for closed-ended medical QA (see Section C.1), we use a manually designed prompt format by default. This includes all of the general-domain LLMs. For example, in the 1-shot setting, the default prompt for non-instruction-tuned models is as follows:

The following is a multiple-choice question about medical knowledge. Answer the question by choosing one of the options from A to D.

Question: Glycolysis is the name given to the pathway involving the conversion of:

- (A) glycogen to glucose-1-phosphate.
- (B) glycogen or glucose to fructose.
- (C) glycogen or glucose to pyruvate or lactate.
- (D) glycogen or glucose to pyruvate or acetyl CoA.

Answer: (C) glycogen or glucose to pyruvate or lactate.

Question: What size of cannula would you use in a patient who needed a rapid blood transfusion (as of 2020 medical knowledge)?

- (A) 18 gauge.

- (B) 20 gauge.
 - (C) 22 gauge.
 - (D) 24 gauge.
- ### Answer:

For instruction-tuned models, which typically expect a specific *conversational* format, we apply the above format to each “user” query and “assistant” response and remove the ‘###’ and ‘Answer:’ tags. For example, the input prompt to LLAMA-3-70B-INSTRUCT is as follows:

```
<|begin_of_text|>
<|start_header_id|> system <|end_header_id|>
The following is a multiple-choice question about medical knowledge. Answer the question by choosing one of the options from A to D.<|eot_id|>
<|start_header_id|>user<|end_header_id|>
Question: Glycolysis is the name given to the pathway involving the conversion of:
(A) glycogen to glucose-1-phosphate.
(B) glycogen or glucose to fructose.
(C) glycogen or glucose to pyruvate or lactate.
(D) glycogen or glucose to pyruvate or acetyl CoA.<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
(C) glycogen or glucose to pyruvate or lactate.<|eot_id|>
<|start_header_id|>user<|end_header_id|>
Question: What size of cannula would you use in a patient who needed a rapid blood transfusion (as of 2020 medical knowledge)?
(A) 18 gauge.
(B) 20 gauge.
(C) 22 gauge.
(D) 24 gauge.<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
```

In the following subsections, we show the system prompt and prompt formats used in the 1-shot setting for models that have a dedicated format. We exclude the model-specific special tokens (e.g., ‘[INST]’) for ease of presentation, and add ‘[User]’ and ‘[Model]’ to demarcate each question and answer for the instruction-tuned models.

C.2.1 OPENBIOLLM (Pal and Sankarasubbu, 2024)

You are an expert and experienced from the healthcare and biomedical domain with extensive medical knowledge and practical experience. Your name is OpenBioLLM, and you

were developed by Saama AI Labs. who’s willing to help answer the user’s query with explanation. In your explanation, leverage your deep medical expertise such as relevant anatomical structures, physiological processes, diagnostic criteria, treatment guidelines, or other pertinent medical concepts. Use precise medical terminology while still aiming to make the explanation clear and accessible to a general audience.

[User] Question: Glycolysis is the name given to the pathway involving the conversion of:
 (A) glycogen to glucose-1-phosphate.
 (B) glycogen or glucose to fructose.
 (C) glycogen or glucose to pyruvate or lactate.
 (D) glycogen or glucose to pyruvate or acetyl CoA.

[Model] (C) glycogen or glucose to pyruvate or lactate.

[User] Question: What size of cannula would you use in a patient who needed a rapid blood transfusion (as of 2020 medical knowledge)?
 (A) 18 gauge.
 (B) 20 gauge.
 (C) 22 gauge.
 (D) 24 gauge.

C.2.2 CLINICAL-CAMEL (Toma et al., 2023)

```
<s>[INST] «SYS»
The following is a multiple-choice question about medical knowledge. Answer the question by choosing one of the options from A to D.
«/SYS»
```

Question: Glycolysis is the name given to the pathway involving the conversion of:
 (A) glycogen to glucose-1-phosphate.
 (B) glycogen or glucose to fructose.
 (C) glycogen or glucose to pyruvate or lactate.
 (D) glycogen or glucose to pyruvate or acetyl CoA. [/INST] (C) glycogen or glucose to pyruvate or lactate.</s><s>[INST] Question: What size of cannula would you use in a patient who needed a rapid blood transfusion (as of 2020 medical knowledge)?
 (A) 18 gauge.
 (B) 20 gauge.
 (C) 22 gauge.
 (D) 24 gauge. [/INST]

C.2.3 BIOMISTRAL (Labrak et al., 2024)

The following are multiple choice questions (with answers) about medical knowledge.

Question: Glycolysis is the name given to the pathway involving the conversion of:

- (A) glycogen to glucose-1-phosphate.
- (B) glycogen or glucose to fructose.
- (C) glycogen or glucose to pyruvate or lactate.
- (D) glycogen or glucose to pyruvate or acetyl CoA.

Answer: (C)

Question: What size of cannula would you use in a patient who needed a rapid blood transfusion (as of 2020 medical knowledge)?

- (A) 18 gauge.
- (B) 20 gauge.
- (C) 22 gauge.
- (D) 24 gauge.

Answer: (

C.2.4 MEDITRON (Chen et al., 2023)

You are a medical doctor answering real-world medical entrance exam questions. Based on your understanding of basic and clinical science, medical knowledge, and mechanisms underlying health, disease, patient care, and modes of therapy, answer the following multiple-choice question. Select one correct answer from A to D. Base your answer on the current and standard practices referenced in medical guidelines.

Question: Glycolysis is the name given to the pathway involving the conversion of:

Options:

- A. glycogen to glucose-1-phosphate.
- B. glycogen or glucose to fructose.
- C. glycogen or glucose to pyruvate or lactate.
- D. glycogen or glucose to pyruvate or acetyl CoA.

The answer is: C

Question: What size of cannula would you use in a patient who needed a rapid blood transfusion (as of 2020 medical knowledge)?

Options:

- A. 18 gauge.
- B. 20 gauge.
- C. 22 gauge.
- D. 24 gauge.

The answer is:

C.2.5 BIOMEDGPT-LM (Luo et al., 2023)

The following is a multiple-choice question about medical knowledge. Answer the question by choosing one of the options from A to D.

Human: Glycolysis is the name given to the pathway involving the conversion of:

- (A) glycogen to glucose-1-phosphate.
- (B) glycogen or glucose to fructose.
- (C) glycogen or glucose to pyruvate or lactate.
- (D) glycogen or glucose to pyruvate or acetyl CoA.

Assistant: (C) glycogen or glucose to pyruvate or lactate.

Human: What size of cannula would you use in a patient who needed a rapid blood transfusion (as of 2020 medical knowledge)?

- (A) 18 gauge.
- (B) 20 gauge.
- (C) 22 gauge.
- (D) 24 gauge.

Assistant:

C.3 Default VLM Prompt Formats

In this section, we share the *default* prompt formats that we use for each general-domain/medical VLM, using VQA-RAD (Lau et al., 2018) as a running example. We denote the system prompt in red, any few-shot examples in green, and the question being asked of the model in purple. By default, we show the format used in the 1-shot setting.

C.3.1 LLaVA-MED (Li et al., 2023)

You are LLaVA-Med, a large language and vision assistant trained by a group of researchers at Microsoft, based on the general domain LLaVA architecture. You are able to understand the visual content that the user provides, and assist the user with a variety of medical and clinical tasks using natural language.

Follow the instructions carefully and explain your answers in detail.

Human: Does this patient have multiple lesions in their chest? Please choose from the following options: [yes, no]. <image>

Assistant: no

Human: Is there evidence of an aortic aneurysm? Please choose from the following options: [yes, no]. <image>

Assistant:

C.3.2 LLaVA-v0 (Liu et al., 2023)

A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human’s questions.

Human: Does this patient have multiple lesions in their chest? Please choose from the following options: [yes, no]. <image>

Assistant: no

Human: Is there evidence of an aortic aneurysm? Please choose from the following options: [yes, no]. <image>

Assistant:

C.3.3 MED-FLAMINGO (Moor et al., 2023b)

You are a helpful medical assistant. You are being provided with images, a question about the image and an answer. Follow the examples and answer the last question.

<image> Does this patient have multiple lesions in their chest?

(A) yes

(B) no

Answer: (B) no <lendofchunk>

<image> Is there evidence of an aortic aneurysm?

(A) yes

(B) no

Answer:

C.3.4 OPEN-FLAMINGO (Awadalla et al., 2023)

The following is a multiple-choice visual question requiring medical knowledge. Answer the question by choosing one of the provided answer options.

<image> Does this patient have multiple lesions in their chest?

(A) yes

(B) no

Answer: (B) no <lendofchunk> <image> Is there evidence of an aortic aneurysm?

(A) yes

(B) no

Answer:

D Additional Results for the Zero-/Few-Shot Prompting Evaluations with Greedy Decoding

In this section, we provide additional results for the main zero-/few-shot prompting experiments with greedy decoding, which are discussed in Section 4.

Table D1: The win, tie, and loss rates (%) of all medical LLMs (top) and VLMs (bottom) in the zero-shot setting, after independently optimizing the prompts for both medical and general-domain models. Model predictions are generated via greedy decoding.

Model	Win	Tie	Loss
OPENBIOLLM-70B (Pal and Sankarasubbu, 2024)	7.7	69.2	23.1
MEDI TRON-70B (Chen et al., 2023)	0	61.5	38.5
CLINICAL-CAMEL-70B (Toma et al., 2023)	27.3	63.6	9.1
OPENBIOLLM-8B (Pal and Sankarasubbu, 2024)	0	46.2	53.8
MEDI TRON-7B (Chen et al., 2023)	0	69.2	30.8
BIO MISTRAL-7B (Labrak et al., 2024)	30.8	69.2	0
BIO MED GPT-LM-7B (Luo et al., 2023)	0	15.4	84.6
LLAVA-MED-7B (Li et al., 2023)	12.5	62.5	25.0
MED-FLAMINGO-9B (Moor et al., 2023b)	0	87.5	12.5

Table D2: The win, tie, and loss rates (%) of all medical LLMs (top) and VLMs (bottom) in the 3-shot setting, after independently optimizing the prompts for both medical and general-domain models. Model predictions are generated via greedy decoding.

Model	Win	Tie	Loss
OPENBIOLLM-70B (Pal and Sankarasubbu, 2024)	30.8	69.2	0
MEDI TRON-70B (Chen et al., 2023)	0	69.2	30.8
CLINICAL-CAMEL-70B (Toma et al., 2023)	0	63.6	36.4
OPENBIOLLM-8B (Pal and Sankarasubbu, 2024)	7.7	61.5	30.8
MEDI TRON-7B (Chen et al., 2023)	0	23.1	76.9
BIO MISTRAL-7B (Labrak et al., 2024)	46.2	53.8	0
BIO MED GPT-LM-7B (Luo et al., 2023)	0	7.7	92.3
LLAVA-MED-7B (Li et al., 2023)	12.5	62.5	25.0
MED-FLAMINGO-9B (Moor et al., 2023b)	0	100.0	0

D.1 Finding 1 (Section 4)

Figure D1 shows the absolute and relative exact-match accuracies achieved by the medical and general-domain LLMs in the zero-shot prompting regime, after independently optimizing the prompt for each model. In Tables D1–D2, we also show the zero-shot and 3-shot win/tie/loss rates achieved by the medical LLMs and VLMs. For CLINICAL-CAMEL-70B, we compute the win/tie/loss rates while excluding the MedQA datasets, as discussed in Section 4. We boldface the medical models that win on more datasets than they lose when compared against their general-domain counterparts.

As discussed in Finding 1 of Section 4, we find that in both the zero-shot and 3-shot settings, only 2 out of 7 medical models show statistically significant improvements over their corresponding base models (CLINICAL-CAMEL-70B and BIO MISTRAL-7B for zero-shot; OPENBIOLLM-70B and BIO MISTRAL-7B for 3-shot), albeit by a limited margin in terms of absolute accuracy. For all other models, the win rates are less than or equal to the loss rates, and the majority of cases result

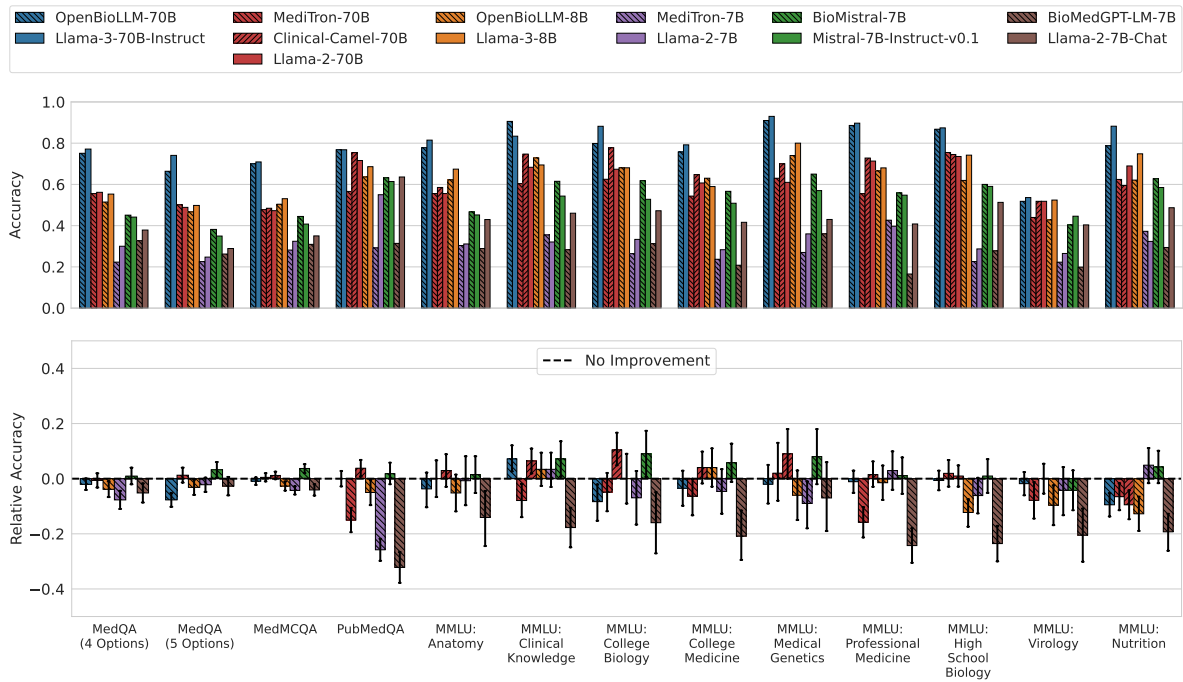


Figure D1: Medical LLMs do not show a statistically significant improvement over their general-domain counterparts in the zero-shot setting, after independently selecting the best prompt format and examples for each model. Top row shows the absolute exact-match accuracies on the test set, and bottom row shows the relative exact-match accuracies along with 95% confidence intervals derived via bootstrapping on the test set (see Section 3). We show the results for when model predictions are generated via greedy decoding.

Table D3: The win, tie, and loss rates (%) of all medical LLMs (top) and VLMs (bottom) in the zero-shot setting, when using a single, fixed prompt optimized only for the medical model. Model predictions are generated via greedy decoding.

Model	Win	Tie	Loss
OPENBIOLLM-70B (Pal and Sankarasubbu, 2024)	23.1	61.5	15.4
MEDI TRON-70B (Chen et al., 2023)	23.1	69.2	7.7
CLINICAL-CAMEL-70B (Toma et al., 2023)	90.9	9.1	0
OPENBIOLLM-8B (Pal and Sankarasubbu, 2024)	69.2	30.8	0
MEDI TRON-7B (Chen et al., 2023)	76.9	23.1	0
BIO MISTRAL-7B (Labrak et al., 2024)	30.8	69.2	0
BIO MEDGPT-LM-7B (Luo et al., 2023)	0	38.5	61.5
LLAVA-MED-7B (Li et al., 2023)	25.0	62.5	12.5
MED-FLAMINGO-9B (Moor et al., 2023b)	25.0	62.5	12.5

Table D4: The win, tie, and loss rates (%) of all medical LLMs (top) and VLMs (bottom) in the 3-shot setting, when using a single, fixed prompt optimized only for the medical model. Model predictions are generated via greedy decoding.

Model	Win	Tie	Loss
OPENBIOLLM-70B (Pal and Sankarasubbu, 2024)	38.5	61.5	0
MEDI TRON-70B (Chen et al., 2023)	0	100.0	0
CLINICAL-CAMEL-70B (Toma et al., 2023)	54.5	45.5	0
OPENBIOLLM-8B (Pal and Sankarasubbu, 2024)	30.8	69.2	0
MEDI TRON-7B (Chen et al., 2023)	38.5	23.1	38.5
BIO MISTRAL-7B (Labrak et al., 2024)	0	100.0	0
BIO MEDGPT-LM-7B (Luo et al., 2023)	7.7	46.2	46.2
LLAVA-MED-7B (Li et al., 2023)	50.0	37.5	12.5
MED-FLAMINGO-9B (Moor et al., 2023b)	25.0	75.0	0

in a tie (i.e., the confidence interval crosses zero relative accuracy).

D.2 Finding 2 (Section 4)

Figures D2–D3 show how the absolute and relative exact-match accuracies change for LLMs and VLMs in the zero-shot and 3-shot settings, when we use a single, fixed prompt that is only optimized for the medical model. In Tables D3–D4, we also show the zero-shot and 3-shot win/tie/loss rates in this scenario. We boldface the medical models that

win on more datasets than they lose when compared against their general-domain counterparts. Compared to when the prompt is independently optimized for each model, we see that a greater number of medical models show statistically significant improvements. Notably, all medical VLMs outperform their general-domain counterparts in both zero-shot and few-shot accuracy under this setup, and all but one medical LLMs outperform their general-domain counterparts in the zero-shot setting. These results suggest that using a single,

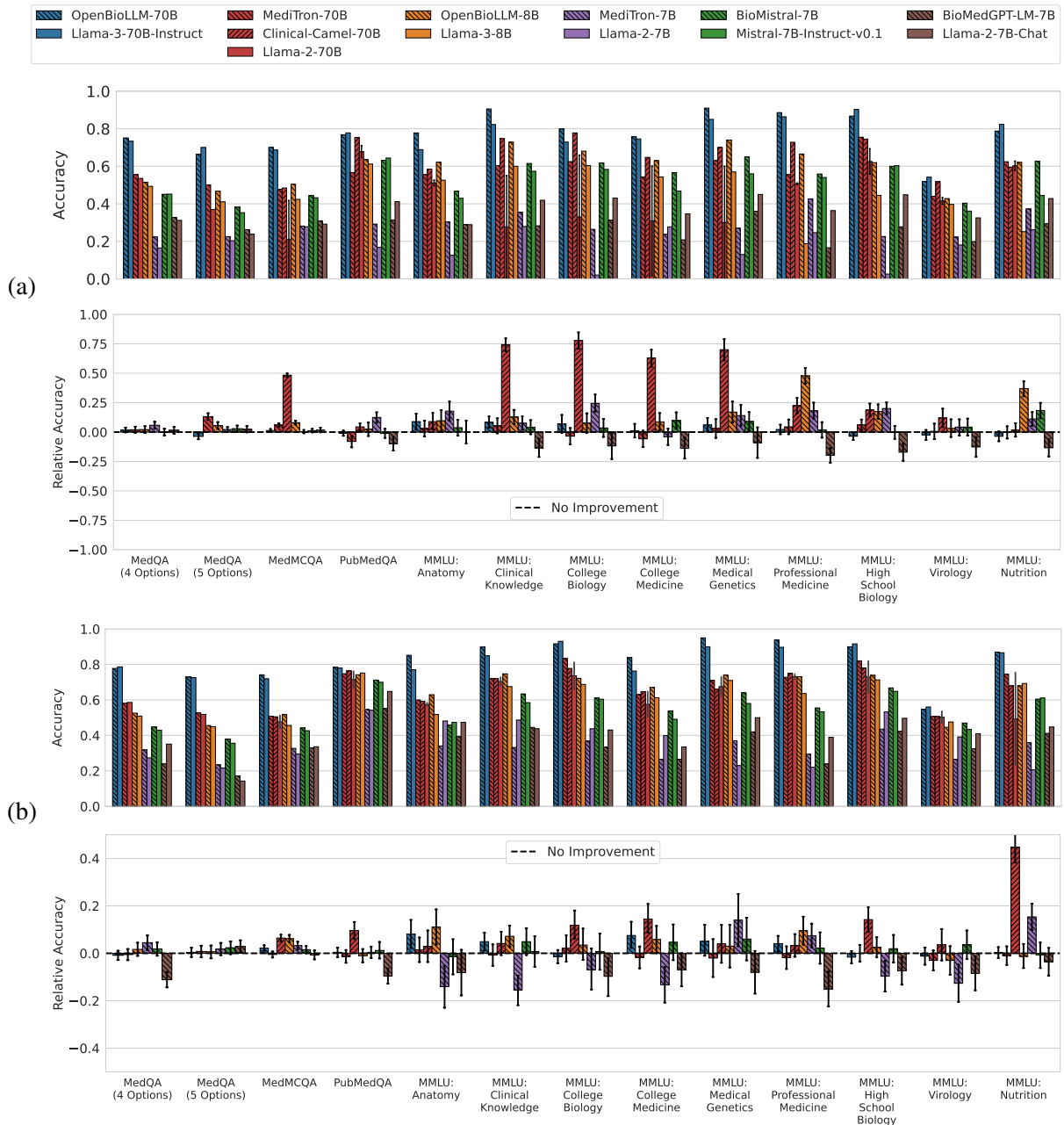


Figure D2: Using a single, fixed prompt format only optimized for the medical model can overestimate the performance improvements from medical DAPT, in both (a) zero-shot and (b) 3-shot settings. For each setting, top row shows the absolute exact-match accuracies on the test set, and bottom row shows the relative exact-match accuracies along with 95% confidence intervals derived via bootstrapping on the test set (see Section 3). For LLAMA-2-70B, which has multiple corresponding medical LLMs (MEDITRON-70B and CLINICAL-CAMEL-70B), we include a min-max error bar in the absolute accuracy plots to show how the absolute accuracy changes with respect to each prompt. We show the results for when model predictions are generated via greedy decoding.

fixed prompt that is only tailored to one model can result in an unfair comparison and can potentially lead to an overestimation of the performance benefits of medical DAPT.

E Results for the Zero-/Few-Shot Prompting Evaluations with Constrained Decoding

In this section, we provide all of the results for the zero-/few-shot prompting evaluations described in Section 3, where the model predictions are generated via *constrained* instead of greedy decoding.

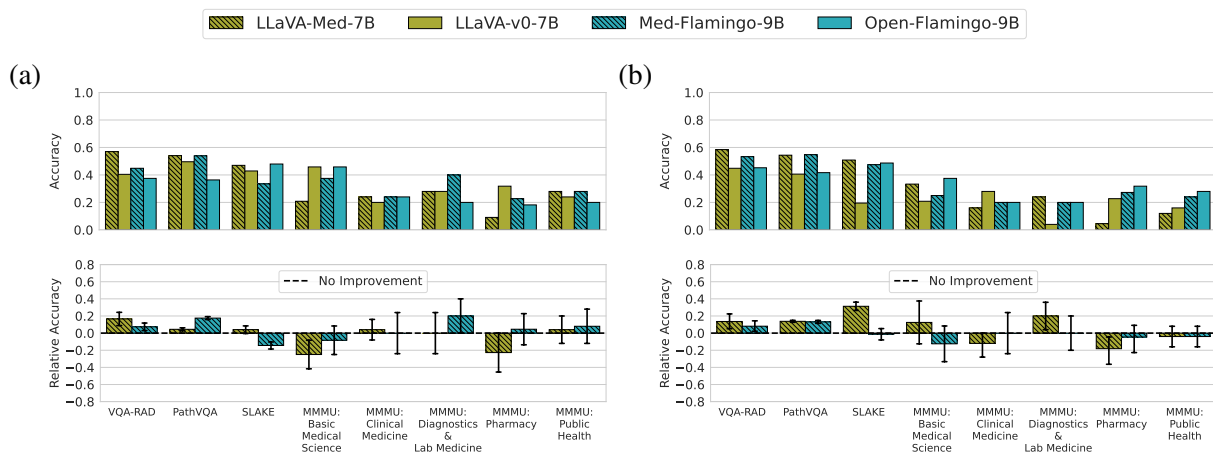


Figure D3: Using a single, fixed prompt format only optimized for the medical model can overestimate the performance improvements from medical DAPT, in both (a) zero-shot and (b) 3-shot settings. For each setting, top row shows the raw exact-match accuracies on the test set, and the bottom row shows the *relative* exact-match accuracies along with 95% confidence intervals derived via bootstrapping on the test set (see Section 3). We show the results for when model predictions are generated via greedy decoding.

E.1 Finding 1 (Section 4)

Here, we show the constrained decoding results for the medical LLMs and VLMs after independently optimizing the prompt for each model. In Figures E1–E2, we show the absolute and relative exact-match accuracies achieved by all LLMs and VLMs in the (a) zero-shot and (b) 3-shot prompting regimes. In Tables E1–E2, we show the zero-shot and 3-shot win/tie/loss rates achieved by each model. For CLINICAL-CAMEL-70B, we compute the win/tie/loss rates while excluding the MedQA datasets, as discussed in Section 4. We boldface the medical models that win on more datasets than they lose when compared against their general-domain counterparts.

Figure E1(a) and Table E1 show that 4 out of 7 medical LLMs show improvements over their general-domain counterparts in the zero-shot setting, albeit by a limited margin in absolute terms. In the 3-shot setting, Figure E1(b) and Table E2 show that only 2 out of 7 medical LLMs—MEDI TRON-7B and BIOMISTRAL-7B—show improvements over their general-domain counterpart, but with a tie on 92.3% of all datasets. For all other models, the win rates are less than or equal to the loss rates, and the majority of cases result in a tie. Meanwhile, Figure E2 and Tables E1–E2 show that no medical VLM shows a statistically significant improvement over its general-domain counterpart in either the zero-shot or 3-shot setting.

Table E1: The win, tie, and loss rates (%) of all medical LLMs (top) and VLMs (bottom) in the zero-shot setting, after independently optimizing the prompts for both medical and general-domain models. Model predictions are generated via constrained decoding.

Model	Win	Tie	Loss
OPENBIO LLM-70B (Pal and Sankarasubbu, 2024)	7.7	76.9	15.4
MEDI TRON-70B (Chen et al., 2023)	30.8	46.2	23.1
CLINICAL-CAMEL-70B (Toma et al., 2023)	18.2	72.7	9.1
OPENBIO LLM-8B (Pal and Sankarasubbu, 2024)	0	53.8	46.2
MEDI TRON-7B (Chen et al., 2023)	23.1	76.9	0
BIOMISTRAL-7B (Labrak et al., 2024)	30.8	69.2	0
BIO MED GPT-LM-7B (Luo et al., 2023)	7.7	84.6	7.7
LLAVA-MED-7B (Li et al., 2023)	0	100.0	0
MED-FLAMINGO-9B (Moor et al., 2023b)	12.5	75.0	12.5

Table E2: The win, tie, and loss rates (%) of all medical LLMs (top) and VLMs (bottom) in the 3-shot setting, after independently optimizing the prompts for both medical and general-domain models. Model predictions are generated via constrained decoding.

Model	Win	Tie	Loss
OPENBIO LLM-70B (Pal and Sankarasubbu, 2024)	23.1	53.8	23.1
MEDI TRON-70B (Chen et al., 2023)	15.4	69.2	15.4
CLINICAL-CAMEL-70B (Toma et al., 2023)	9.1	72.7	18.2
OPENBIO LLM-8B (Pal and Sankarasubbu, 2024)	7.7	69.2	23.1
MEDI TRON-7B (Chen et al., 2023)	7.7	92.3	0
BIOMISTRAL-7B (Labrak et al., 2024)	7.7	92.3	0
BIO MED GPT-LM-7B (Luo et al., 2023)	7.7	69.2	23.1
LLAVA-MED-7B (Li et al., 2023)	0	87.5	12.5
MED-FLAMINGO-9B (Moor et al., 2023b)	0	100.0	0

E.2 Finding 2 (Section 4)

Here, we present the constrained decoding results for medical LLMs and VLMs when using a single, fixed prompt format only optimized for the medical

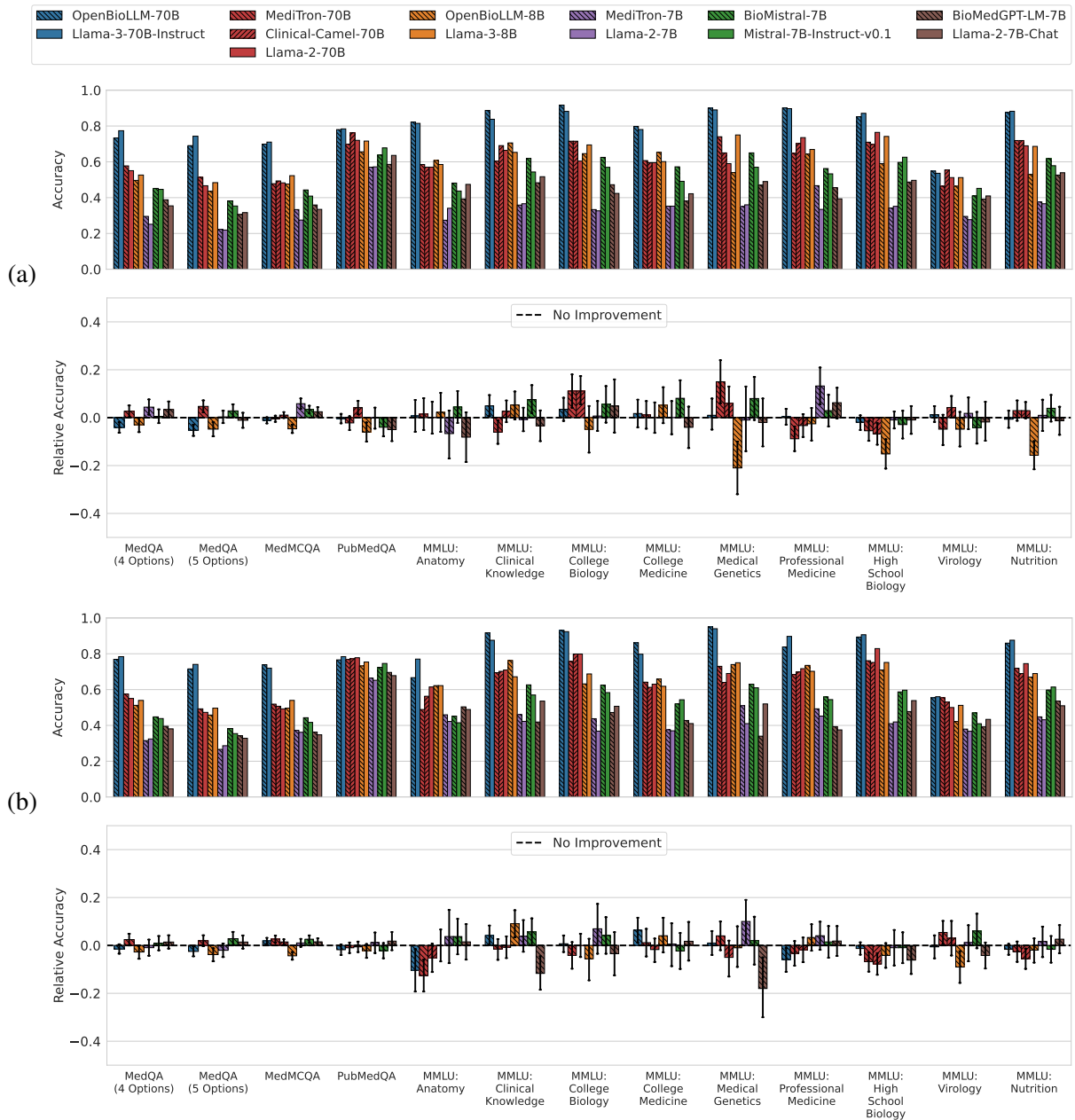


Figure E1: Medical LLMs do not show a statistically significant improvement over their general-domain counterparts in both (a) zero-shot and (b) 3-shot settings, after independently selecting the best prompt format and examples for each model. Top row shows the absolute exact-match accuracies on the test set, and bottom row shows the relative exact-match accuracies along with 95% confidence intervals derived via bootstrapping on the test set (see Section 3). We show the results for when model predictions are generated via constrained decoding.

model. In Figures E3–E4, we show how the absolute and relative exact-match accuracies change for all LLMs and VLMs in the zero-shot and 3-shot settings. In Tables E3–E4, we also show the zero-shot and 3-shot win/tie/loss rates in this scenario. We boldface the medical models that win on more datasets than they lose when compared against their general-domain counterparts. Compared to when the prompt is independently optimized for each

model, we see that a greater number of medical models show statistically significant improvements. In Figure E5, we also show how the win/tie/loss rates of the medical models, computed over all (model pair, QA dataset) combinations, change as we vary the prompting setups as in Finding 2 of Section 4. As in the greedy decoding setup, we find that for both LLMs and VLMs, the performance improvements from medical DAPT can be substan-

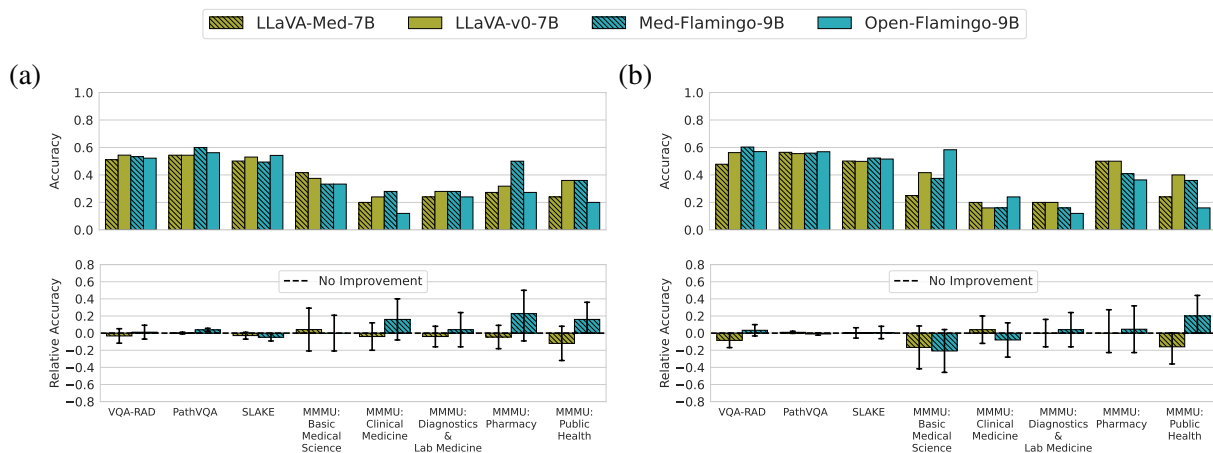


Figure E2: Using a single, fixed prompt format only optimized for the medical model can overestimate the performance improvements from medical DAPT, in both (a) zero-shot and (b) 3-shot settings. For each setting, top row shows the raw exact-match accuracies on the test set, and the bottom row shows the *relative* exact-match accuracies along with 95% confidence intervals derived via bootstrapping on the test set (see Section 3). We show the results for when model predictions are generated via constrained decoding.

Table E3: The win, tie, and loss rates (%) of all medical LLMs (top) and VLMs (bottom) in the zero-shot setting, when using a single, fixed prompt optimized only for the medical model. Model predictions are generated via constrained decoding.

Model	Win	Tie	Loss
OPENBIOLLM-70B (Pal and Sankarasubbu, 2024)	30.8	69.2	0
MEDI TRON-70B (Chen et al., 2023)	30.8	53.8	15.4
CLINICAL-CAMEL-70B (Toma et al., 2023)	63.6	36.4	0
OPENBIOLLM-8B (Pal and Sankarasubbu, 2024)	46.2	46.2	7.7
MEDI TRON-7B (Chen et al., 2023)	7.7	76.9	15.4
BIO MISTRAL-7B (Labrak et al., 2024)	23.1	76.9	0
BIO MED GPT-LM-7B (Luo et al., 2023)	30.8	69.2	0
LLAVA-MED-7B (Li et al., 2023)	0	100.0	0
MED-FLAMINGO-9B (Moor et al., 2023b)	25.0	75.0	0

Table E4: The win, tie, and loss rates (%) of all medical LLMs (top) and VLMs (bottom) in the 3-shot setting, when using a single, fixed prompt optimized only for the medical model. Model predictions are generated via constrained decoding.

Model	Win	Tie	Loss
OPENBIOLLM-70B (Pal and Sankarasubbu, 2024)	23.1	61.5	15.4
MEDI TRON-70B (Chen et al., 2023)	7.7	92.3	0
CLINICAL-CAMEL-70B (Toma et al., 2023)	36.4	63.6	0
OPENBIOLLM-8B (Pal and Sankarasubbu, 2024)	30.8	61.5	7.7
MEDI TRON-7B (Chen et al., 2023)	7.7	92.3	0
BIO MISTRAL-7B (Labrak et al., 2024)	7.7	16.7	0
BIO MED GPT-LM-7B (Luo et al., 2023)	30.8	69.2	0
LLAVA-MED-7B (Li et al., 2023)	25.0	75.0	0
MED-FLAMINGO-9B (Moor et al., 2023b)	0	100.0	0

tially overestimated when (i) the prompt is only tailored to the medical model; and (ii) the models are compared only based on their absolute accuracies. For example, in the zero-shot setting, the

win rate increases from 16.9% to 68.0% for medical LLMs and from 6.3% to 56.3% for medical VLMs, when only performing prompt selection for the medical model and comparing based on raw absolute accuracy.

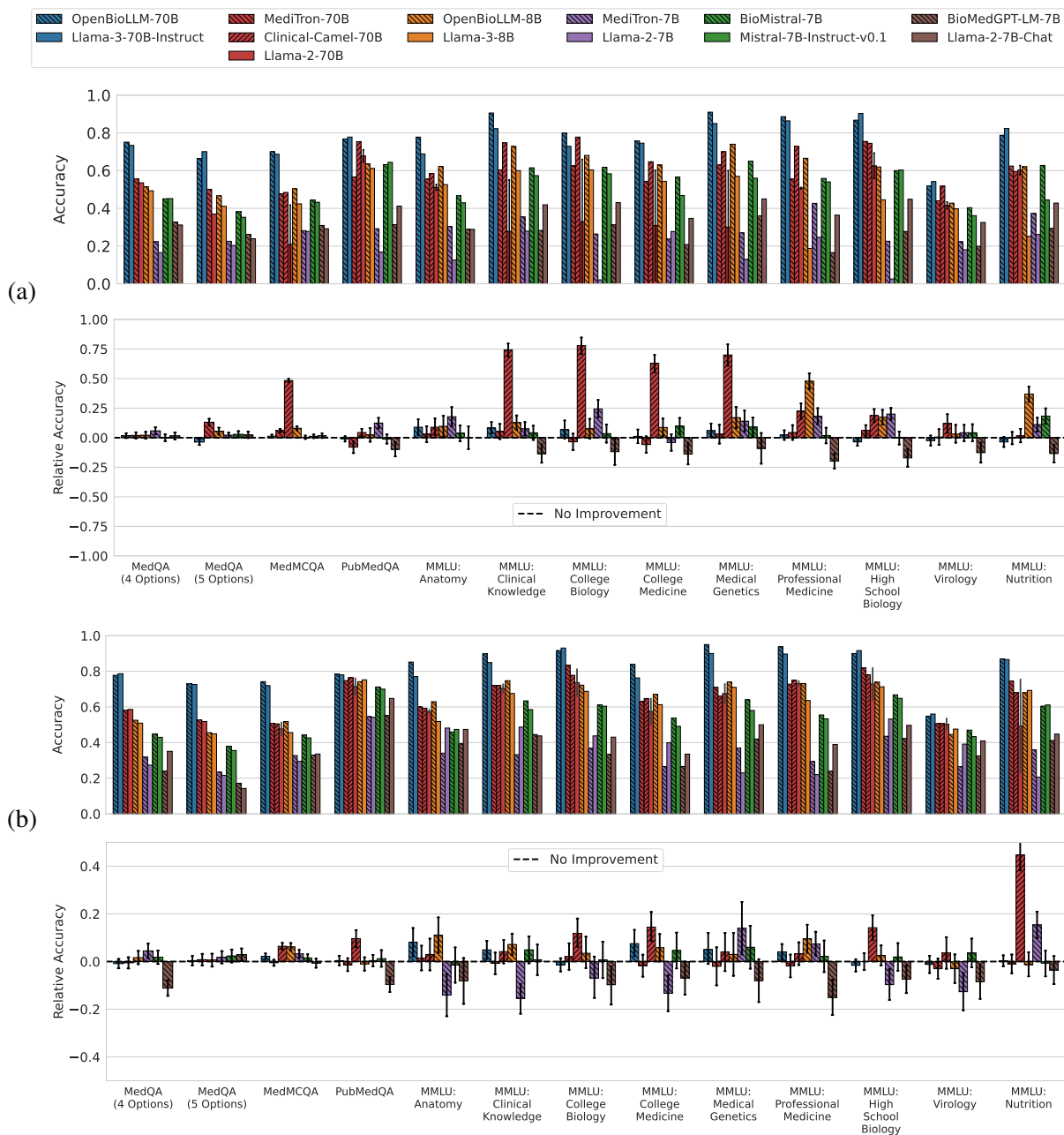


Figure E3: Using a single, fixed prompt format only optimized for the medical model can overestimate the performance improvements from medical DAPT, in both (a) zero-shot and (b) 3-shot settings. For each setting, top row shows the absolute exact-match accuracies on the test set, and bottom row shows the relative exact-match accuracies along with 95% confidence intervals derived via bootstrapping on the test set (see Section 3). For LLAMA-2-70B, which has multiple corresponding medical LLMs (MEDITRON-70B and CLINICAL-CAMEL-70B), we include a min-max error bar in the absolute accuracy plots to show how the absolute accuracy changes with respect to each prompt. We show the results for when model predictions are generated via constrained decoding.

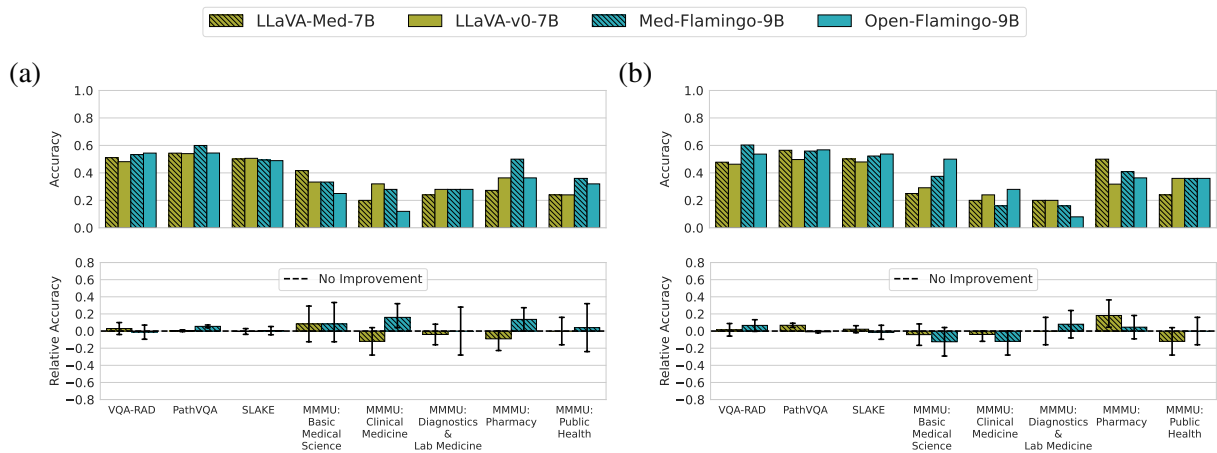


Figure E4: Using a single, fixed prompt format only optimized for the medical model can overestimate the performance improvements from medical DAPT, in both (a) zero-shot and (b) 3-shot settings. For each setting, top row shows the raw exact-match accuracies on the test set, and the bottom row shows the *relative* exact-match accuracies along with 95% confidence intervals derived via bootstrapping on the test set (see Section 3). Here, we show the results for when model predictions are generated via constrained decoding.

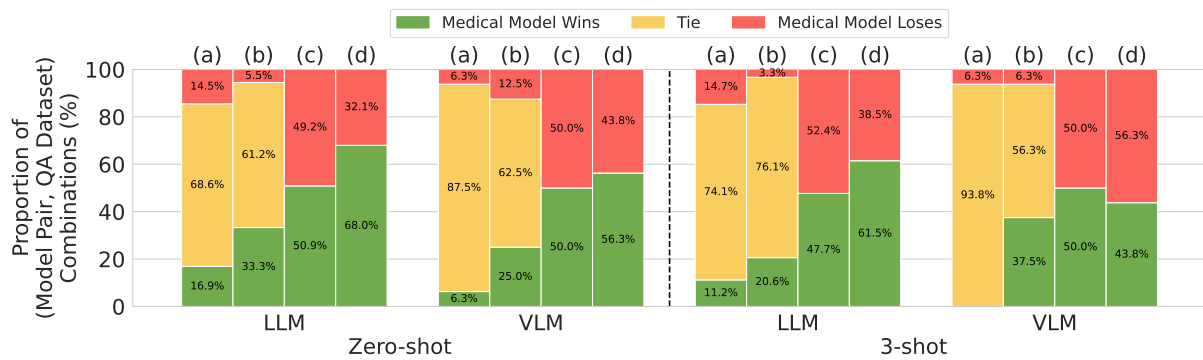


Figure E5: Optimizing the prompt for only the medical model and comparing models without accounting for statistical uncertainty can overestimate the performance improvements from medical DAPT. We show the win/tie/loss rate (%) of medical models vs. their base models across all (model pair, QA dataset) combinations, when (a) independently optimizing the prompt for each model and performing statistical testing, (b) optimizing the prompt only for the medical model and performing statistical testing, (c) independently optimizing the prompt for each model without statistical testing, and (d) optimizing the prompt only for the medical model without statistical testing. Here, we show the results when model predictions are generated via constrained decoding.