Beyond Fine-tuning: Unleashing the Potential of Continuous Pretraining for Clinical LLMs.

Anonymous EMNLP submission

Abstract

Large Language Models (LLMs) have demonstrated significant potential in transforming clinical applications. In this study, we investigate the efficacy of four techniques in adapting LLMs for clinical use-cases: continuous pretraining, instruct fine-tuning, NEFTune (Jain et al., 2023), and prompt engineering. We employ these methods on Mistral 7B and Mixtral 8x7B models, leveraging a large-scale clinical pretraining dataset of 50 billion tokens and an instruct fine-tuning dataset of 500 million tokens. Our evaluation across various clinical tasks reveals the impact of each technique. While continuous pretraining beyond 250 billion tokens yields marginal improvements on its own, it establishes a strong foundation for instruct fine-tuning. Notably, NEFTune, designed primarily to enhance generation quality, surprisingly demonstrates additional gains on our benchmark. Complex prompt engineering methods further enhance performance. These findings show the importance of tailoring fine-tuning strategies and exploring innovative techniques to optimize LLM performance in the clinical domain.

1 Introduction

The advent of large language models (LLMs) has spurred a wave of innovation across various domains, with healthcare being a particularly promising area for their application. LLMs have the potential to transform clinical workflows, aid in diagnosis, and enhance patient care. However, effectively adapting these models to the nuances and complexities of the clinical domain remains a significant challenge.

Current approaches in the literature predominantly focus on either developing specialized clinical LLMs from scratch or fine-tuning existing models on large-scale clinical datasets. While these methods have shown promise, they often overlook the potential benefits of continuous pretraining on domain-specific data as a means to further enhance model performance. This is due in part to the complexities and potential instabilities associated with continued training of large models.

In this study, we take a comprehensive approach to optimizing clinical LLMs by systematically investigating the impact of continuous pretraining on in-domain data, in conjunction with instruct fine-tuning and advanced prompting strategies. We focus on the Mistral-7B (Jiang et al., 2023) and Mixtral-8x7B (Jiang et al., 2024) models, demonstrating that continuous pretraining, while yielding modest gains compared to fine-tuning and prompting, plays a crucial role in establishing a solid foundation for further specialization. By carefully balancing in-domain clinical data with general language data, we successfully mitigate instability issues and unlock the full potential of continuous pretraining for clinical LLMs.

Our work highlights the importance of understanding of the relationship between pretraining, fine-tuning, and prompting in adapting LLMs for clinical applications. By demonstrating the effectiveness of continuous pretraining on domain-specific data, we open doors for future research to further explore this underutilized technique to develop more accurate, reliable, and ultimately impactful clinical LLMs.

2 Related Works

The landscape of Large Language Models (LLMs) for healthcare is evolving rapidly, with most approaches involving either domain-specific pretraining or instruction fine-tuning of general-purpose models. OpenAI's GPT-3.5 and GPT-4 (OpenAI, 2023), alongside Google's Med-PaLM (Singhal et al., 2023a) and Med-PaLM 2 (Singhal et al., 2023b) have demonstrated impressive performance on medical benchmarks, despite limited transparency regarding their training details. Other models, such as GatorTron (Yang et al., 2022), and

PMC-LLaMA (Wu et al., 2023), have shown the potential of pretraining on extensive biomedical corpora to add domain-specific knowledge for clinical applications.

Instruction fine-tuning and dialogue datasets have also been instrumental in enhancing the zero-shot and few-shot generalization capabilities of LLMs. ChatDoctor (Li et al., 2023b) and MedAlpaca (Han et al., 2023), for instance, utilize medical conversations and other NLP tasks to improve LLaMA's performance on clinical queries. Recent models like Clinical Camel (Toma et al., 2023), MediTron (Chen et al., 2023) and Med42 (Christophe et al., 2024), based on LLaMA-2 (Touvron et al., 2023), further demonstrate the efficacy of this approach.

Building on the observation that models can learn from prompting alone (Brown et al., 2020), recent research has explored techniques to enhance clinical capabilities without additional training. These methods often extend the well-known Chain-of-Thought prompting technique, originally introduced by (Wei et al., 2022b), to better suit clinical use-cases. Notably, Microsoft's MedPrompt (Nori et al., 2023b) demonstrates significant improvements in GPT-4's performance on clinical QA tasks, while (Garikipati et al., 2024) apply similar strategies to the Yi family of models (Young et al., 2024). Google has also showcased the potential of complex prompting to boost the clinical capabilities of their Gemini model (Saab et al., 2024). However, while such complex prompting techniques can improve performance on standard benchmarks, their practicality and scalability in real-world clinical applications remain to be seen

Recent studies like LIMA (Zhou et al., 2024), FineWeb (Guilherme Penedo, 2024) and Phi (Li et al., 2023a) have highlighted the pivotal role of data quality in LLM training, emphasizing that it can often be more influential than architectural choices in determining model performance. High-quality data has been shown to significantly impact the model's ability to learn meaningful representations and generalize to new tasks. This shows the importance of our approach to dataset curation, ensuring that our models are trained on a robust and representative collection of clinical data.

3 Experiments

In this section, we present the four steps of our experimental framework: (1) continuous pretraining, (2) instruct fine-tuning, (3) NEFTune, and (4) complex prompt engineering.

3.1 Continuous Pretraining

Continuous pretraining involves extending the pretraining phase of a large language model (LLM) by exposing it to additional text data. This can be particularly beneficial in domain-specific applications, like healthcare, where models can be further trained on vast amounts of clinical literature. The goal is to refine the model's understanding of domain-specific terminology, relationships, and nuances, potentially leading to improved performance on relevant tasks. In our experiments, we investigate the impact of continuous pretraining on both Mistral 7B and Mixtral 8x7B models, utilizing a 50-billion-token clinical dataset.

Continuous pretraining of large language models, however, is not without its challenges. Typically, only the weights of the LLM are openly accessible, while the optimizer state remains unavailable. This lack of access can disrupt the training process, leading to instabilities and hindering the model's ability to effectively learn from the new data. Additionally, the potential distribution shift between the original pretraining data and the new clinical data can result in catastrophic forgetting, where the model loses proficiency on previously learned knowledge and tasks (Li and Lee, 2024).

Following the work presented in (Gupta et al., 2023), we implement a learning rate warm-up strategy, gradually increasing the learning rate over 1% of the total training steps. Specifically, we employ a linear warm-up, starting from 1/10th of our maximum learning rate and gradually ramping up to the full value. This gradual increase helps stabilize the training process and prevents drastic updates to the model's weights early on. Second, we address the potential distribution shift by blending our specialized clinical data with general language data from SlimPajama (Soboleva et al., 2023). This curated blend results in a 65-billion-token dataset, comprising 50 billion tokens of specialized clinical data and 15 billion tokens of general language data. We then perform continuous pretraining on this dataset for a total of 4 epochs, processing 260 billion tokens and allowing the model to acquire domain-specific knowledge while retaining its proficiency in general language understanding. In Figure 5, we illustrate the training loss curves over both the general and clinical data subsets. As depicted, our warmup strategy and data mixture effectively mitigate instabilities, demonstrating smooth convergence and a steady decrease in loss throughout the training process. This approach ensures the model's overall capabilities remain robust and facilitates the acquisition of specialized clinical knowledge.

3.2 Instruct Fine-Tuning

Instruct fine-tuning is a technique that aims to align large language models (LLMs) with human intentions and preferences by training them on a dataset of instructions and their corresponding desired outputs. This approach enables LLMs to better understand and respond to user prompts, improving their ability to generate relevant and useful responses in a variety of tasks.

To facilitate effective learning from instructions, we adopt a structured format incorporating the keywords <|system|>, <|prompter|>, and <|assistant|>. This format explicitly delineates the roles of the system, the user providing the prompt, and the assistant generating the response. By clearly defining these relationships, we guide the model to better understand the intent behind instructions and generate appropriate, medically relevant outputs

Each sample in our instruction-tuning dataset is composed of three elements: a system prompt, a user prompt, and the corresponding model response. To maximize the utilization of the model's available context length during training, we concatenate these samples across the entire dataset. The training process is auto-regressive and the loss is solely focused on the tokens comprising the responses. This targeted training strategy prioritizes the model's ability to generate accurate and relevant answers, rather than focusing on replicating prompts.

We train our models for 3 epochs using a cosine learning rate scheduler, which gradually decreases the learning rate over the course of training

3.3 NEFTune

NEFTune, a novel instruction fine-tuning technique introduced in (Jain et al., 2023), offers an alternative approach to our traditional pipeline. This method involves injecting noise into the embedding layer during training, a process that has shown improvements in the quality of the model's output generation. Furthermore, the introduced noise during training could act as a regularization method to stabilize the learning process. The noise vector is created by independently sampling each entry from a uniform distribution within the interval $[-1, 1]$. This vector is then scaled by a factor determined by the tunable parameter α, the sequence length L, and the embedding dimension d:

$$X'_{\text{emb}} \leftarrow X_{\text{emb}} + \left(\sqrt{\frac{\alpha}{Ld}}\right)\epsilon$$

During our experiments, we explored various values for α and discovered that the setting of $\alpha = 5$ yielded superior results. In our study, we explore NEFTune as a potential replacement for our standard instruct fine-tuning pipeline, investigating its impact on overall performance on clinical tasks.

3.4 Prompt Engineering

In-Context Learning refers to a model's ability to understand and generate relevant responses based on the context provided within a prompt. This capability allows the model to leverage previous examples or instructions given in the prompt to perform tasks more effectively without explicit training on new samples. Chain-of-Thought Reasoning (Wei et al., 2023) is a technique where the model is guided to generate a step-by-step explanation of its thought process before arriving at an answer. This approach encourages the model to articulate its reasoning, leading to more transparent and accurate outcomes. In our work, we harness these capabilities by implementing the 'Medprompt' prompting strategy as introduced by (Nori et al., 2023b). To thoroughly evaluate our models, we generate chain-of-thought explanations using four distinct prompt engineering methods:

- **Chain of thought (CoT):** Similar to (Kojima et al., 2023), we generate chain-of-thought on the evaluation dataset by appending "Let's think step-by step" to every sample. This method encourages the model to systematically break down its thought process, leading to more structured and transparent reasoning.

- **Few shot chain of thought:** In this approach, we improve the model's performance by providing context through static examples. Before generating the chain-of-thought explanation, we prepend the samples with five predefined few-shot examples. These examples serve as a guide, helping the model to understand and apply a consistent reasoning pattern.

- **Dynamic few shot chain of thought:** This advanced method combines dynamic retrieval

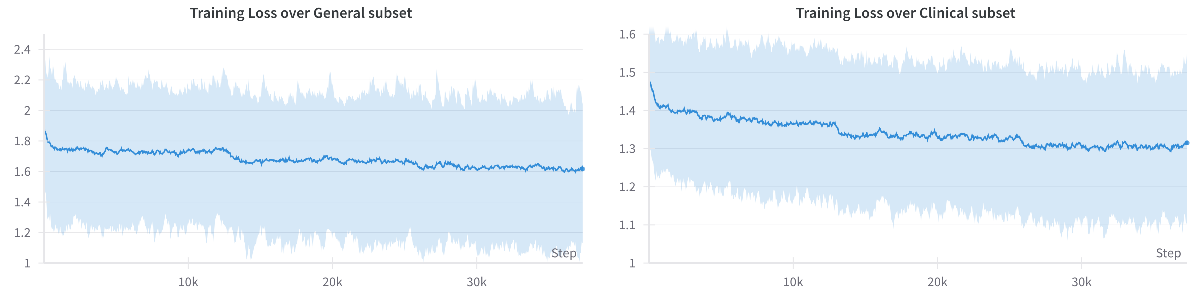

Figure 1: Training loss for Mixtral during continuous pretraining on the general (left) and clinical (right) subsets.

Hyperparameter	Mistral 7B			Mixtral 8x7B		
	Pretraining	Fine-tuning	NEFTune	Pretraining	Fine-tuning	NEFTune
Learning Rate Scheduler	Linear Warmup - Cosine			Linear Warmup - Cosine		
Max Learning Rate	7×10^{-6}	5×10^{-6}	5×10^{-6}	7×10^{-6}	1×10^{-6}	1×10^{-6}
Beta	(0.9, 0.95)			(0.9, 0.95)		
Alpha	-	-	5	-	-	5
Weight Decay	0.1			0.1		
Number of Steps	79,259	6,089	6,089	37,344	2,589	2,589

Table 1: Hyperparameters for Pretraining, Fine-tuning, and NEFTune on Mistral 7B and Mixtral 8x7B Models

and chain-of-thought generation. Initially, we create chain-of-thought reasoning for multiple-choice question-answering datasets and store these in a Milvus vector database (Wang et al., 2021). We then embed the training questions using gte-small embedding model (Li et al., 2023c). During evaluation, we retrieve the five most semantically similar training examples based on cosine similarity in the embedding space. These retrieved examples are used as few-shot examples, providing relevant context to the model for generating more accurate explanations.

- **Dynamic few shot chain of thought ensemble (CoT-En):** Building on the dynamic few-shot approach, this method introduces variability and robustness. Here, we shuffle the few-shot examples and the multiple-choice options, generating the chain-of-thought reasoning five times with a temperature setting of 0.2. This ensemble technique aims to produce a diverse set of reasonings.

3.5 Hardware infrastructure.

Our experiments were conducted on a high-performance computing cluster, utilizing a maximum of 10 nodes, each equipped with 8 NVIDIA H100 GPUs, for the continuous pretraining phase. For the subsequent fine-tuning stages, we employed 4 nodes of the same configuration. To efficiently train our large-scale models, we leveraged PyTorch's Fully Sharded Data Parallel (FSDP) (Zhao et al., 2023) framework, which enables distributed training across multiple GPUs while minimizing memory footprint. Additionally, we employed bfloat16 precision throughout our training pipeline.

4 Datasets

In this section, we detail our approach to constructing both the pretraining and fine-tuning datasets. Our primary objective is to curate datasets that optimize model performance while maintaining the highest standards of quality and relevance to the clinical domain.

4.1 Pretraining Dataset

Our pretraining corpus comprises a mix of biology and healthcare data from publicly available sources, including full-text research articles, abstracts, open textbooks, and Wikipedia articles. We excluded data containing personally identifiable information as well as data without a permissive license for commercial use.

Pretraining data for large language models

(LLMs) typically requires several normalization and cleaning steps to make it suitable for training. However, since we have controlled the input sources and limited them to trusted sources, our pretraining pipeline primarily involves five major steps: 1) document parsing, 2) low-length filtering, 3) document-level deduplication, 4) exact deduplication, and 5) data chunking.

Document parsing involves either scraping webpages or extracting text from research articles. Once the text is extracted from all sources, we remove sources with insufficient information by applying a length threshold filter. As our data mix mainly consists of full-text research articles, there is a high likelihood of document-level duplication with different DOI IDs. To address this, we used the MinHash (Broder, 1997) deduplication technique with a similarity threshold of 0.85: for each document, we compute a sketch and measure its approximate similarity with other documents, removing pairs with high overlap. We perform MinHash deduplication using 9,000 hashes per document, calculated over 5-grams and divided into 15 buckets of 400 hashes each.

Document-level deduplication removes similar documents across different data sources, but there could still be some text duplication within the documents. Therefore, we additionally employed an exact deduplication step (Lee et al., 2021) to eliminate identical text segments from the dataset. As advised in the original literature, we ran the exact deduplication twice with length thresholds of 400 and 100 bytes, since duplicates may persist even after the first pass. Finally, the entire dataset is tokenized, concatenated, and split into chunks with a predefined context length for continuous pretraining.

4.2 Finetuning Dataset

Our instruction-tuning dataset is a curated blend of open-source medical question-answering data, sourced primarily from medical forums like Stack Exchange, rich in expert discussions and patient inquiries. We also integrate relevant medical segments extracted from general domain datasets, ensuring a diverse representation of medical subfields and contexts. This comprehensive dataset provides a solid foundation for training our model to accurately understand and generate medically relevant content.

To improve the chain-of-thought capabilities of the fine-tuned model, we generate chain-of-thought explanations for datasets that benefit from reasoning chains. After generating these reasoning chains, we discard those that do not correspond with the correct answers and use these samples as zero-shot examples. We employ the Mixtral-Instruct model for both generating and verifying the reasoning chains. For more details on the composition of the finetuning dataset, please refer to Table 3.

5 Evaluations

To rigorously assess the efficacy of our fine-tuning approaches, we focus on a comprehensive evaluation of the models' capabilities across a spectrum of clinical question-answering (QA) tasks. We employ a diverse suite of QA datasets, including MedQA (Jin et al., 2020), USMLE sample exam and self-assessment (Nori et al., 2023a; Han et al., 2023), MMLU (medical subset)(Hendrycks et al., 2021), and MedMCQA(Pal et al., 2022), to ensure a thorough and representative assessment of model performance in various clinical scenarios.

Our evaluation methodology uses the EleutherAI Harness framework (Gao et al., 2021), which focuses on the likelihood of a model generating each proposed answer rather than directly evaluating the generated text itself. To enhance the granularity and relevance of our analysis, we introduce modifications to the Harness codebase. Instead of computing the likelihood of generating only the answer choice labels (a, b, c, or d), we extend the computation to encompass the likelihood of generating the complete answer text. This modification provides a more detailed understanding of the model's performance, as it takes into account the entire answer generation process, including the ability to articulate reasoning and justify the selected answer choice.

To evaluate the efficacy of MedPrompt prompting strategies, we integrate these prompts into the Harness framework. This involves generating reasoning chains based on the prompts and then using Harness to assess the likelihood of the model producing the final answer derived from these chains. This approach allows us to evaluate the impact of specific prompting techniques on the model's ability to reason through complex clinical scenarios.

Throughout our evaluation, we report accuracy as the primary metric across all tables, providing a clear and interpretable measure of the models' proficiency in clinical QA tasks.

	# of parameters	MedQA	USMLE	MMLU	MedMCQA
BioMistral (Labrak et al., 2024)	7B	45.09	46.67	63.63	44.58
Clinical Camel (Toma et al., 2023)	70B	53.42	54.35	69.75	47.01
MediTron (Chen et al., 2023)	70B	51.14	57.31	68.26	42.36
Med42 (Christophe et al., 2024)	70B	61.52	72.01	76.71	60.93
Mistral 7b Instruct	7B	42.89	48.18	62.75	43.32
Mistral 7b F (ours)	7B	54.28	62.63	68.30	58.11
Mistral 7b N (ours)	7B	60.72	61.97	70.35	58.57
Mistral 7b $P + F$ (ours)	7B	58.36	63.84	72.28	**60.84**
Mistral 7b $P + N$ (ours)	7B	**62.69**	**63.98**	**73.45**	59.79
Mixtral 8x7b Instruct	46.7B	52.55	65.99	75.78	53.74
Mixtral 8x7b F (ours)	46.7B	62.60	72.68	79.10	62.85
Mixtral 8x7b N (ours)	46.7B	66.93	70.05	79.57	64.64
Mixtral 8x7b $P + F$ (ours)	46.7B	67.09	**73.57**	**79.92**	65.29
Mixtral 8x7b $P + N$ (ours)	46.7B	**68.34**	72.82	79.84	**65.34**

Table 2: Accuracy over multiple clinical QA tasks. F stands for Instruct-Finetuning, P stands for Pretraining, and N stands for NEFTune. We show that our models improve on all tasks as we gradually add more training techniques.

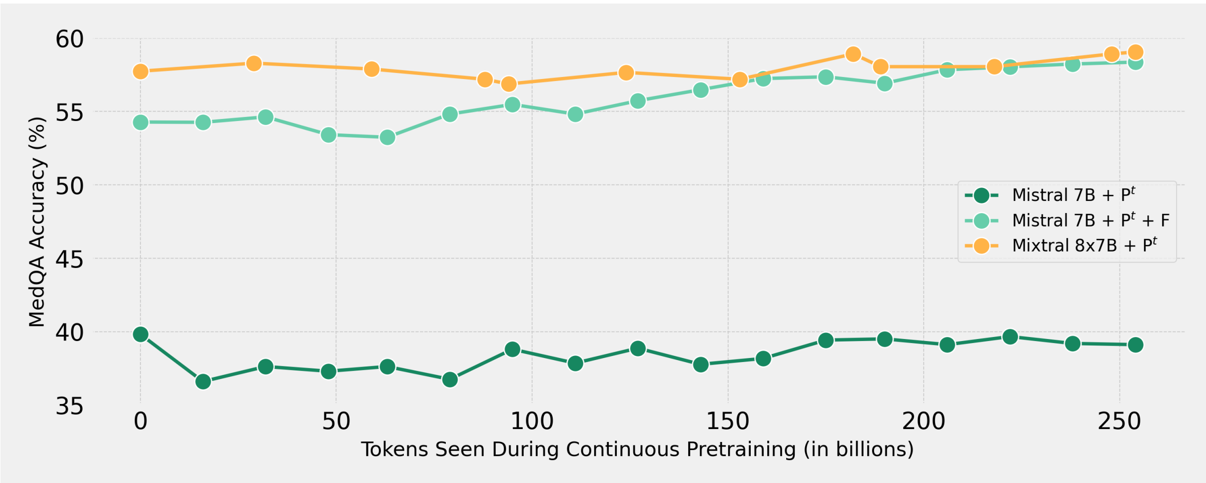

Figure 2: Evolution of MedQA accuracy for Mistral-7b and Mixtral 8x7b base models as well as our instructed versions of Mistral-7b during continuous pretraining. $P\hat{}t$: Continuous Pretrained with variable numbers of tokens t, F: Instruct Finetuned. We show that, while base model accuracy remains consistent, applying instruct-finetuning leads to notable improvements.

6 Results

In this section, we present the results of our experiments, revealing key insights into the effectiveness of different training approaches for clinical language models.

Non-instructed models can't be evaluated on QA tasks. Throughout our continuous pretraining process, we saved multiple checkpoints and assessed their performance on our clinical QA benchmarks. Given the absence of instruction fine-tuning, we opted for a no-prompt evaluation format. As illustrated in Figure 2, a slight performance decline is observed between the base model (with no pretraining) and the initial checkpoints. While subsequent checkpoints exhibit gradual improvement with increased exposure to clinical data, their performance consistently trails behind the original base model. This observation underscores the critical role of instruction fine-tuning in equipping LLMs with the necessary skills to effectively comprehend and respond to questions in the clinical domain.

Instruct Fine-tuning Specializes the Model for QA Data The remarkable leap in performance observed in Table 2 after instruct fine-tuning highlights the efficacy of this approach in aligning LLMs with the specific demands of clinical question-answering tasks. While this outcome is not surprising or novel, it reaffirms the established effectiveness of fine-tuning methodologies in adapting models to specific domains. By exposing the models to a curated dataset of instructions and corresponding answers, we effectively specialize both Mistral and Mixtral to formulate answers in the clinical domain. This targeted training approach enhances the models' ability to understand the intent behind questions, use the knowledge acquired during pretraining, and generate accurate, relevant, and informative responses. The substantial gains observed across all benchmarks show the critical role of instruct fine-tuning in bridging the gap between general language understanding and specialized clinical expertise, ultimately empowering LLMs to excel in medical question answering.

Continuous Pretraining Shows Consistent Performance Gains To assess the impact of continuous pretraining on downstream performance, we conducted a comprehensive evaluation by instruct-fine-tuning various checkpoints saved during the pretraining process. Figure 2 illustrates the performance trajectory of these models as they are exposed to increasing amounts of pretraining data. Initially, the gains are relatively minor, particularly within the first 100 billion tokens. However, as the models continue to learn from the vast corpus of clinical text, we observe a gradual and steady improvement in their performance across a range of QA benchmarks.

This trend suggests that continuous pretraining serves as a valuable foundation, gradually enhancing the model's understanding of clinical concepts and terminology. As the models assimilate more domain-specific knowledge, they become better equipped to leverage the instruction data during fine-tuning, ultimately leading to superior performance on clinical QA tasks.

Table 2 provides a detailed breakdown of the performance gains achieved through continuous pretraining across various benchmarks. Notably, our continuously pretrained models consistently outperform state-of-the-art models, including the instruct-tuned versions of both Mistral and Mixtral. These results underscore the efficacy of continuous pretraining in equipping LLMs with the necessary domain knowledge to excel in clinical applications.

The magnitude of these gains, however, varies across model sizes. Mistral-7B demonstrates significant improvement, while the larger Mixtral 8x7B model exhibits more marginal, yet still consistent, benefits. This suggests that while continuous pretraining remains valuable for larger models, its impact may be less pronounced compared to smaller counterparts, potentially due to the diminishing returns of additional data for already extensive architectures. These findings demonstrate the importance of carefully weighing the computational costs and performance benefits of continuous pretraining, particularly for larger LLMs.

Adding noise helps finetuning. In our experiments, we observed an intriguing phenomenon with the NEFTune technique, originally proposed in (Jain et al., 2023). While the authors demonstrated that NEFTune applied to LLaMA-2 7B maintained Harness accuracy across several QA tasks, we show in Table 2, that for Mistral-7B, it not only preserved but, in most cases, even improved the model's performance. This performance increase was consistent across both the base model and the continuously pretrained model. This result is particularly surprising as NEFTune was primarily designed to enhance generation quality, not necessarily benchmark accuracy. We hypothesize that the injection of noise during training might act as a form of regularization, preventing overfitting and leading to better generalization on downstream tasks. However, the exact mechanisms behind this improvement warrant further investigation. This result suggests that the benefits of NEFTune extend beyond its intended purpose, potentially influencing the model's ability to reason and select the most likely answer.

Prompt Engineering makes the difference. Figure 3 showcases the potential of MedPrompt as a viable alternative to traditional fine-tuning and pretraining techniques. By incorporating Chain-of-Thought (CoT) prompting and KNN CoT ensembles, we achieve substantial performance gains for the Mixtral-Instruct model on various clinical QA tasks. The effectiveness of MedPrompt is consistently observed across different model configurations: fine-tuned models outperform their non-fine-tuned counterparts, and pretraining followed by fine-tuning further amplifies these improvements. Notably, by employing MedPrompt with CoT and KNN CoT ensembles, we elevate MedQA accuracy

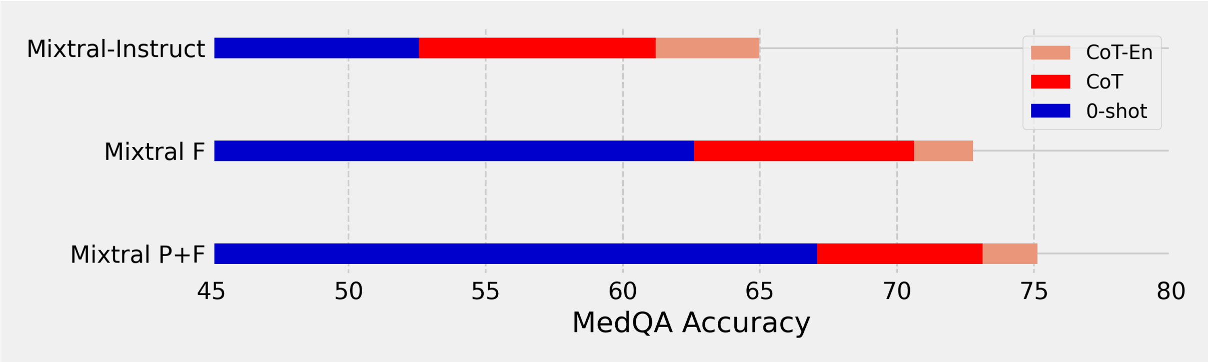

Figure 3: Evolution of MedQA accuracy using MedPrompt over different versions of Mixtral.

from 52.55% with the baseline Mixtral-Instruct model to a value exceeding 75%. These results not only show the potential of advanced prompting strategies like MedPrompt to significantly enhance LLM performance in clinical applications without requiring computationally expensive fine-tuning or pretraining procedures, but also highlight the crucial role of pretraining in establishing a strong foundation for further improvement.

7 Conclusion and Discussions

In this study, we have systematically investigated the impact of continuous pretraining on in-domain clinical data, in conjunction with instruct fine-tuning and advanced prompting strategies, on the performance of LLMs in clinical question-answering tasks. Our findings demonstrate that continuous pretraining, while yielding modest improvements compared to other techniques, remains a valuable tool for enhancing LLM performance in the clinical domain. While continuous pretraining can often be challenging due to instability issues, we have shown that by carefully balancing in-domain clinical data with general language data, we can effectively mitigate these challenges and achieve consistent performance gains.

Furthermore, we have demonstrated that the benefits of continuous pretraining extend beyond the initial training phase, as it lays a solid foundation for subsequent instruct fine-tuning and the application of complex prompting techniques like MedPrompt. The synergy between continuous pretraining and these additional methods results in state-of-the-art performance on a variety of clinical QA benchmarks, outperforming existing models like the instruct-tuned versions of Mistral and Mixtral.

Our research opens up several avenues for future exploration. Further ablation studies could examine the effects of different domain data sources, beyond the clinical realm, on LLM performance. Additionally, a more comprehensive analysis of the optimal data mix for continuous pretraining, including varying proportions of in-domain and general language data, could yield valuable insights for maximizing the benefits of this technique.

This study provides a comprehensive framework for optimizing clinical LLM performance. Our findings offer valuable insights for future research and development efforts aimed at leveraging LLMs to address challenges and opportunities presented by the healthcare domain.

8 Limitations

While our research offers valuable insights into optimizing clinical LLMs, it is not without limitations. Primarily, our study focused on a specific set of models (Mistral and Mixtral) and a limited number of clinical QA datasets. While we strive for diversity in our benchmark selection, the generalizability of our findings to other LLM architectures or clinical tasks remains an open question.

Additionally, the computational resources required for continuous pretraining, particularly for larger models, may pose a barrier for widespread adoption. Further investigation into more efficient pretraining methods could address this limitation.

Finally, while our evaluation framework provides a comprehensive assessment of model performance on QA tasks, it does not fully capture the nuances of real-world clinical applications, where factors like explainability, bias mitigation, and safety are paramount. Future research should explore these aspects in greater detail to ensure the responsible and effective deployment of LLMs in healthcare settings.

References

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinform.*, 20(1):511:1–511:23.

A.Z. Broder. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, pages 21–29.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Clément Christophe, Praveen K Kanithi, Prateek Munjal, Tathagata Raha, Nasir Hayat, Ronnie Rajan, Ahmed Al-Mahrooqi, Avani Gupta, Muhammad Umar Salman, Gurpreet Gosal, et al. 2024. Med42–evaluating fine-tuning strategies for medical llms: Full-parameter vs. parameter-efficient approaches. *arXiv preprint arXiv:2404.14779*.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Anurag Garikipati, Jenish Maharjan, Navan Preet Singh, Leo Cyrus, Mayank Sharma, Madalina Ciobanu, Gina Barnes, Qingqing Mao, and Ritankar Das. 2024. Openmedlm: Prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*.

Loubna Ben Allal Anton Lozhkov Colin Raffel Leandro Werra Thomas Wolf Guilherme Penedo, Hynek Kydlíček. 2024. Fineweb: decanting the web for the finest text data at scale. `https://huggingface.co/spaces/HuggingFaceFW/blogpost-fineweb-v1`.

Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats L Richter, Quentin Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. 2023. Continual pre-training of large language models: How to (re) warm your model? *arXiv preprint arXiv:2308.04014*.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Neel Jain, Ping-yeh Chiang, Yuxin Wen, John Kirchenbauer, Hong-Min Chu, Gowthami Somepalli, Brian R Bartoldson, Bhavya Kailkhura, Avi Schwarzschild, Aniruddha Saha, et al. 2023. Neftune: Noisy embeddings improve instruction finetuning. *arXiv preprint arXiv:2310.05914*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

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

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Matt Gardner Johannes Welbl, Nelson F. Liu. 2017. Crowdsourcing multiple choice science questions.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. *arXiv preprint arXiv:2010.09926*.

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.

Nathan Lambert, Lewis Tunstall, Nazneen Rajani, and Tristan Thrush. 2023. Huggingface h4 stack exchange preference dataset.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2021. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*.

Chen-An Li and Hung-Yi Lee. 2024. Examining forgetting in continual pre-training of aligned large language models. *arXiv preprint arXiv:2401.03129*.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023a. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023b. ChatDoctor: A medical chat model Fine-Tuned on a large language model Meta-AI (LLaMA) using medical domain knowledge. *Cureus*, 15(6):e40895.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023c. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Wing Lian, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023. Openorca: An open dataset of gpt augmented flan reasoning traces. `https://https://huggingface.co/Open-Orca/OpenOrca`.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023a. Capabilities of gpt-4 on medical challenge problems.

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

OpenAI. 2023. GPT-4 technical report.

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*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

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

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization.

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

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023b. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. 2023. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti 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. 2023. Llama 2: Open foundation and fine-tuned chat models.

David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of* the Association for Computational Linguistics, pages 960–966, Florence, Italy. Association for Computational Linguistics.

Jianguo Wang, Xiaomeng Yi, Rentong Guo, Hai Jin, Peng Xu, Shengjun Li, Xiangyu Wang, Xiangzhou Guo, Chengming Li, Xiaohai Xu, et al. 2021. Milvus: A purpose-built vector data management system. In *Proceedings of the 2021 International Conference on Management of Data*, pages 2614–2627.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. PMC-LLaMA: Towards building open-source language models for medicine. *arXiv [cs.CL]*.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B 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 Digit Med*, 5(1):194.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. 2023. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

A Appendix: Supplementary Materials

A.1 Finetuning Dataset Mix

Dataset	# Samples	Mixture ratio (%)
Medical domain		
MedMCQA (Pal et al., 2022)	180,462	23.49
Medical Flashcards (Han et al., 2023)	30,106	3.92
StackExchange (Lambert et al., 2023)	64,246	8.36
MedQA (USMLE) (Jin et al., 2020)	11,290	1.47
CORD-19 (Wang et al., 2020)	17,721	2.31
PubMedQA (Jin et al., 2019)	499	0.06
HeadQA (Vilares and Gómez-Rodríguez, 2019)	2,657	0.35
MediQA (Han et al., 2023)	1,950	0.25
SciQ (Johannes Welbl, 2017)	11,679	1.52
PubMed Causal (Han et al., 2023)	2,169	0.28
OpenGPT	66,026	8.59
MedQUAD (Ben Abacha and Demner-Fushman, 2019)	14,553	1.89
MMLU (Hendrycks et al., 2021)	244	0.03
Niv2* (Wang et al., 2022)	11,447	1.49
Pubhealth (Kotonya and Toni, 2020)	9,804	1.28
Total	424,853	55.29
General domain		
SlimOrca T0 (Lian et al., 2023; Sanh et al., 2022)	109,235	14.22
SlimOrca Flan (Lian et al., 2023; Longpre et al., 2023)	109,169	14.21
SlimOrca CoT (Lian et al., 2023; Wei et al., 2022a)	74,172	9.65
Ultrachat (Ding et al., 2023)	50,953	6.63
Total	343,529	44.71

[†] The following categories were included: “academia", “bioinformatics”, “biology", “cogsci", “fitness", “health".

[‡] Only samples in English were used.

[$] The following subjects were included: “anatomy", “clinical knowledge", “college medicine", “medical genetics", “professional medicine", “college biology", “high-school biology", “professional psychology", “high-school psychology", “human sexuality", “human aging", “nutrition", and “virology".

* Samples from 47 tasks (from over 1,000 tasks) related to science, healthcare and medicine were included.

Table 3: Summary of subsets of the data used for fine-tuning the models. Note that medical-domain data correspond to approximately 60% of the entire dataset.

A.2 Prompt formats

```
<|system|>: You are a biomedical expert. Select the correct
option for the following question:
<|prompter|>: {question}
{Option 1}
{Option 2}
{Option 3}
{Option 4}
{Option 5}
<|assistant|>: The correct answer is: {Correct option}
```

Figure 4: Zero-shot prompt format on a sample from MedQA

```
<|system|>: You are a biomedical expert. Select the correct
option for the following question:
<|prompter|>: {question}
{Option 1}
{Option 2}
{Option 3}
{Option 4}
{Option 5}
<|assistant|>: Let's think step-by-step. {COT reasoning} Th
erefore, the correct answer is: {Correct option}
```

Figure 5: Chain-of-thought prompt format on a sample from MedQA