npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01653-8

Small language models learn enhanced reasoning skills from medical textbooks

Hyunjae Kim[1], Hyeon Hwang[1], Jiwoo Lee[1], Sihyeon Park[1], Dain Kim[1], Taewhoo Lee[1], Chanwoong Yoon[1], Jiwoong Sohn[1], Jungwoo Park[1], Olga Reykhart[2], Thomas Fetherston[2], Donghee Choi[3], Soo Heon Kwak[4], Qingyu Chen[2] & Jaewoo Kang[1,5] ✉

Small language models (SLM) offer promise for medical applications by addressing the privacy and hardware constraints of large language models; however, their limited parameters (often fewer than ten billion) hinder multi-step reasoning for complex medical tasks. This study presents Meerkat, a new family of medical SLMs designed to be lightweight while enhancing reasoning capabilities. We begin by designing an effective and efficient training method. This involves extracting high-quality chain-of-thought reasoning paths from 18 medical textbooks, which are then combined with diverse instruction-following datasets within the medical domain, totaling 441K training examples. Fine-tuning was conducted on open-source SLMs using this curated dataset. Our Meerkat-7B and Meerkat-8B models outperformed their counterparts by 22.3% and 10.6% across six exam datasets, respectively. They also improved scores on the NEJM Case Challenge from 7 to 16 and from 13 to 20, surpassing the human score of 13.7. Additionally, they demonstrated superiority in expert evaluations, excelling in all metrics—completeness, factuality, clarity, and logical consistency—of reasoning abilities.

The healthcare and medical fields are highly complex, requiring in-depth expertise and knowledge. Despite the growing demand for healthcare services, a significant gap persists between this demand and the availability of skilled professionals[1–3]. While previous artificial intelligence (AI) technologies have partially alleviated this gap, they have either lacked linguistic capabilities or were limited to functioning within specialized, single tasks[4,5]. As a result, these technologies are not easily adaptable to the diverse and dynamic tasks encountered in healthcare settings, where language-based interaction between patients, medical staff, and healthcare professionals are essential.

Recent advancements in large language models (LLM) suggest a promising future for the application of AI in the field of healthcare and medicine, serving as efficient and rapid decision-making assistants for professionals[6,7]. These foundation models are capable of handling a wide range of language tasks in a generalizable manner. Several models have demonstrated their capability by surpassing the 60% passing threshold on the United States Medical Licensing Examination (USMLE) questions[8–12], recently reaching a remarkable accuracy rate of 95.0%[13]. Furthermore, their effectiveness has been highlighted in addressing real-world clinical case challenges, including responding to clinical inquiries related to daily practices, engaging in conversational history-taking, and diagnosing complex clinical cases[14–16].

Despite these advancements, the deployment of LLMs in the medical domain faces significant challenges in two key aspects. First, the use of commercial LLMs in healthcare, such as OpenAI's GPT-3.5[17] and GPT-4[18], is constrained by privacy and security concerns due to their closed-source nature[6,19–21]. These models rely on web-based APIs for data transmission, making the management of sensitive patient data particularly challenging in the absence of well-defined regulatory frameworks. Second, the high computational demands of LLMs pose a significant barrier to local deployment. Recent releases of open-source LLMs[22–24] have made on-premises deployment on in-house servers possible, offering a privacy-preserving alternative by eliminating the need to transmit sensitive data to external platforms. However, the substantial hardware requirements far exceed the capacity of standard desktop systems (requiring multiple 80GB A100/H100 GPUs), rendering on-premises deployment impractical for many hospitals and clinical research laboratories.

A practical solution likely involves deploying open-weight small language models (SLMs) with fewer than ten billion parameters locally [25–31]. These models can typically be run on high-end PCs (e.g., RTX 3090), making them accessible to a wider range of users and environments. However, a key challenge remains that these models lack the necessary multi-step reasoning capabilities to solve complex problems. In medicine,

[1]Korea University, Seoul, Republic of Korea. [2]Yale University, New Haven, CT, USA. [3]Imperial College London, London, UK. [4]Seoul National University Hospital, Seoul, Republic of Korea. [5]AIGEN Sciences, Seoul, Republic of Korea. ✉e-mail: kangj@korea.ac.kr

strong reasoning skills are particularly crucial for analyzing problems systematically, constructing logical paths, and accurately predicting answers. Thanks to their vast amount of parameters, often exceeding several hundreds of billion, LLMs naturally exhibit this "chain-of-thought" (CoT) reasoning ability[32], enabling them to provide step-by-step explanations to arrive at a conclusion for complex problems. In contrast, SLMs do not consistently acquire these abilities during pre-training[33–35].

Unfortunately, effective and efficient training methods for improving medical reasoning remain understudied. Existing medical SLMs are commonly initialized from general-domain SLMs and further trained on millions of domain-specific documents using a basic continuous pre-training method[28–31]. This approach not only demands substantial computational resources but also yields only limited effectiveness for training medical reasoning. For instance, PMC-Llama-7B required 32 A100 GPUs around 7 days to complete training, totaling around 5376 GPU hours[28]. With new and improved general-domain models being released on a monthly basis, this approach of continuously training models on medical corpora is becoming increasingly unsustainable. Furthermore, the performance improvements achieved often fail to justify the resource demands. For example, PMC-Llama-7B demonstrated only a 1.02% improvement over its counterpart, Llama-2-7B, on the MedQA benchmark.

We seek to address the following research questions: (1) What training method can be implemented to effectively enhance the limited reasoning capabilities of SLMs? (2) In light of the rapid release of new SLMs by both industry and research institutions, can this method be efficiently adapted to evolving models?

In this study, we introduce Meerkat, a new family of on-premises medical AI systems with enhanced reasoning skills acquired from textbooks. Our model is built upon the current state-of-the-art LMs, such as Mistral-7B[26] and Llama-3-8B[26], and fine-tuned using a diverse set of carefully crafted data. Specifically, we employed an LLM to extract CoT reasoning paths for 9.3K USMLE-style questions from the MedQA dataset[36]. To enhance diversity in reasoning, we further synthesized 78K additional questions along with their CoT reasoning paths using authoritative resources. This effort involved leveraging 18 textbooks that comprehensively span 16 medical disciplines. Furthermore, we aggregated existing instruction-following and chat datasets, authorized for research use, to address a broad range of applications in this domain. In total, the model was fine-tuned on 460K examples. The fine-tuning process takes only around 1 day on eight A100 GPUs, making it significantly more efficient compared to traditional continuous pre-training approaches.

Meerkat-7B and Meerkat-8B achieved an average accuracy of 64.5% and 66.7% across six benchmarks, surpassing the previous leading general- and medical-domain models, including Mistral-7B (41.2%), Llama-3-8B (56.1%), MediTron-7B (51.0%)[29], BioMistral-7B (55.4%)[30], and GPT-3.5 (54.8%)[17]. Notably, Meerkat-7B achieved scores of 77.1 on the MedQA[36], marking the first instance where a 7B model surpassed the USMLE's passing threshold of 60% accuracy, and also exceeding the previous best open-weight model performance of 70.2% set by MediTron-70B. In a test of NEJM Case Challenges, Meerkat-8B accurately diagnosed 20 cases, surpassing the human average of 13.8 and nearly matching GPT-4's score of 21.8. In human evaluations of the rationale generated by the models, Meerkat-8B outperformed its counterpart, Llama-3-8B, across all four metrics: completeness, factuality, clarity, and logical consistency. We underscore that our Meerkat models, along with the CoT fine-tuning approach, substantially narrowed the performance gap with commercial LMs, enabling smaller models to tackle challenging reasoning tasks. Our contributions are summarized as follows:

- We introduce Meerkat, a cutting-edge series of on-premises medical models with high-level reasoning capabilities. Meerkat represents the first instance of training a medical AI system using CoT data synthesized from raw textbooks and showing its effectiveness. Our fine-tuning approach is significantly more efficient than continuous pre-training, requiring approximately 28 times less GPU time for a 7B model. Moreover, it consistently demonstrates effectiveness regardless of the initial LM, meaning that our method can enhance performance even for newly released models through fine-tuning.
- Our models surpassed general and domain-specific open-weight models on six medical benchmarks. Meerkat-7B is the first 7B model to exceed the USMLE passing threshold, setting the standard as the leading open-weight model in its class. Additionally, Meerkat-8B surpassed the human benchmark score by 6.3 on the NEJM Case Challenges. In expert evaluations, Meerkat-8B significantly outperformed Llama-3-8B in four fine-grained metrics.
- We plan to release all associated artifacts, including our model weights and training data. The released data includes the new *MedQA-CoT* and *MedBooks-18-CoT* datasets, which comprises synthetic question-answer pairs with CoT reasoning paths extracted from a USMLE-style dataset and 18 textbooks. This can serve as a valuable resource for fine-tuning new models in the medical domain.

Results

Table 1 shows that Meerkat-7B significantly outperformed its counterpart, Mistral-7B, by an average of 22.3%, and the previous best 7B models, MediTron-7B and BioMistral-7B, by an average of 13.5% and 9.1%, respectively. Meerkat-8B also surpassed Llama-3-8B, MediTron-7B, and BioMistral-7B and with an average improvement of 10.6%, 15.7%, and 11.3%. Remarkably, our models outperformed GPT-3.5 (175B) by 9.7% and 11.9%, even with fewer parameters. To demonstrate the statistical significance, we performed bootstrapping to compare our Meerkat models with their counterparts (i.e., Mistral-7B and Llama-3-8B). Specifically, we generated 100 resampled datasets, each matching the original dataset in size, and conducted a paired t-test. Across all comparisons, the p value was less than 4.1993e−38, indicating an extremely low likelihood of the observed differences occurring by chance.

Figure 1a presents a more detailed depiction of the performance on MedQA. Meerkat-7B achieved scores of 71.2% and 77.1% in single-model and ensemble evaluation settings, respectively, significantly exceeding the 60% passing threshold. Meerkat-8B attained 74.2% and 77.3% in the same evaluation settings. Both models surpassed the previous state-of-the-art performance achieved by MediTron-70B (ensemble), exceeding its benchmark of 70.2% and setting a new standard among open-weight models.

Figure 1b highlights the performance of our models on the NEJM Case Challenges. The human score was determined through the majority vote of NEJM journal readers, following the methodology outlined in Eriksen et al.[16]. Our models consistently outperformed their counterparts, Mistral-7B and Llama-3-8B, as well as the human score of 13.7, demonstrating the potential effectiveness of our CoT fine-tuning approach in addressing real-world challenging problems. Notably, our best model, Meerkat-8B, achieved a score of 20 in the ensemble setting, performing comparably to GPT-4, which scored 21.8. While we also evaluated medical-specialized models, these models were excluded from detailed analysis due to their poor performance. This suggests that domain-specific continuous pre-training may struggle to generalize effectively when addressing real-world problems requiring complex medical reasoning.

Additionally, we conducted a human analysis of CoT reasoning for Meerkat-8B and Llama-3-8B models on a sample of 50 examples, selected from 101 cases where both models provided correct answers in the Medbullets-5 dataset. Ten experts evaluated anonymized model responses and choose the better option between the two. Additionally, we performed an model-based evaluation using GPT-4o, leveraging all 101 examples. Figure 2 demonstrates that Meerkat-8B consistently outperformed across four evaluation metrics in both human and GPT-4o assessments. Notably, Meerkat-8B showed significant advantages in completeness and logical consistency, reflecting its ability to perform detailed, step-by-step reasoning. In the factuality evaluation, GPT-4o frequently declared a draw, while human evaluators consistently identified Meerkat-8B as the superior model. This discrepancy underscores the challenges of model-based evaluation. Both Meerkat-8B and Llama-3-8B generally produced factually accurate outputs. However, responses occasionally included minor logical errors or

Table 1 | Performance of models for the six QA benchmarks: MedQA, USMLE sample test (USMLE), Medbullets-4 (MB-4), Medbullets-5 (MB-5), MedMCQA, and MMLU-Medical (MMLU)

Model	MedQA	USMLE	MB-4	MB-5	MedMCQA	MMLU	Avg.
Commercial LLMs							
o1	**95.7**	**95.7**	**87.9**	**86.0**	**82.9**	**95.2**	**90.6**
o3-mini	91.8	93.5	85.7	81.2	76.3	93.0	86.9
o1-mini	90.0	91.1	80.8	79.2	71.0	91.2	83.9
GPT-4o	83.6	89.2	76.3	66.5	63.1	86.2	77.5
GPT-4	81.4	86.6	68.8	63.3	72.4	87.1	76.6
GPT-3.5 (175B)	53.6	58.5	51.0	47.4	51.0	67.3	54.8
Open-source & Domain-specific SLMs							
Mistral-7B	43.2	40.5	38.8	32.8	40.7	51.0	41.2
Llama-3-8B	57.5	58.8	49.0	48.7	54.7	68.0	56.1
MediTron-7B	50.2	44.6	51.5	45.5	57.9	56.7	51.0
BioMistral-7B	54.3	51.4	52.3	48.7	61.1	64.6	55.4
Meerkat-7B (**Ours**)	71.2	70.1	**60.5**	52.8	61.5	70.7	64.5
Meerkat-8B (**Ours**)	**74.2**	**73.8**	59.7	**55.2**	**62.7**	**74.3**	**66.7**

Our Meerkat models generally performed better than existing 7B and 8B models and GPT-3.5 across all datasets. The scores in MMLU-Medical were calculated based on the average accuracies across the six medical-related subjects. Detailed results for the six subjects can be found in Supplementary Table 1. The scores of GPT-3.5 and GPT-4 are obtained from the papers of Nori et al.[11], Toma et al.[53], and Chen et al.[54].

The best performance for each category—Commercial LLMs and Open-source & Domain-specific SLMs—is highlighted in bold.

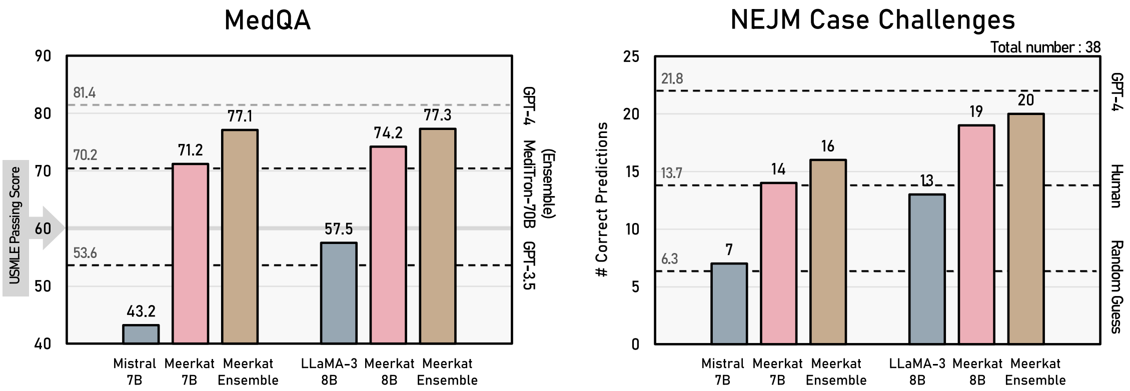


Fig. 1 | Meerkat-8B (ensemble) surpasses the previous best open-weight model, MediTron-70B (ensemble), on MedQA despite its smaller parameter size. Additionally, it significantly outperforms humans on NEJM Case Challenges and achieves performance comparable to GPT-4.

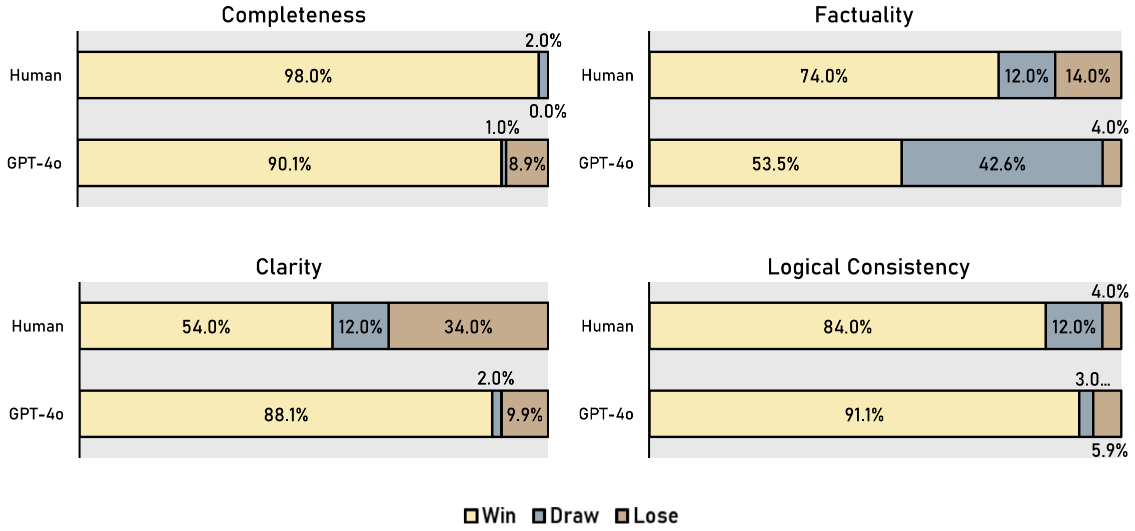


Fig. 2 | The Medbullets-5 dataset was used. “Win” refers to instances where Meerkat-8B received a higher score, “Lose” denotes cases where Llama-3-8B outperformed Meerkat-8B, and “Draw” indicates a tie between the two models.

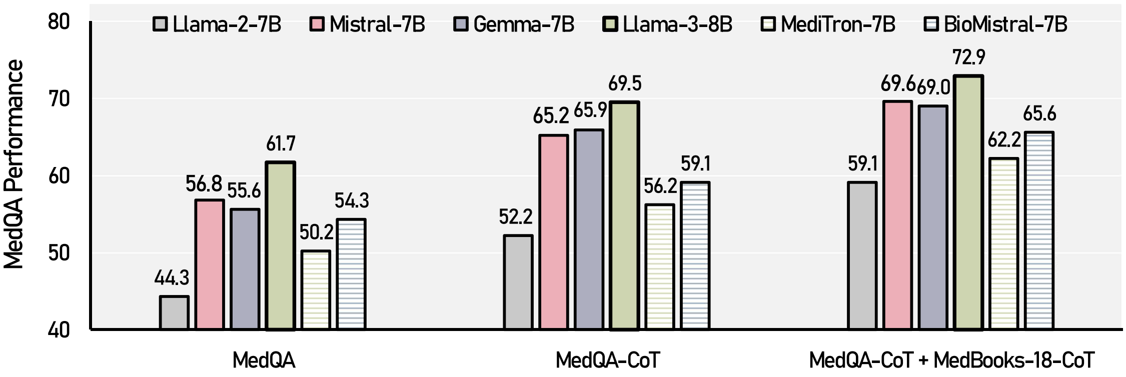


Fig. 3 | MediTron-7B and BioMistral-7B are domain-specific models. Mistral-7B, Gemma-7B, and Llama-3-8B performed better than MediTron-7B and BioMistral-7B, despite not being specialized models for biomedicine. “MedQA": training the model only using question-answer pairs in the MedQA training set. “MedQA-CoT": training the model using MedQA question-answer pairs and corresponding CoT reasoning data. “MedQA-CoT + MedBooks-18-CoT": training the model using the MedQA-CoT data and additional CoT data generated from textbooks.

potentially contentious content, while these issues did not significantly affect the overall conclusions. GPT-4o often overlooked them, leading to its higher rate of draw evaluations. In the clarity evaluation, human ratings for Meerkat-8B were the lowest among the criteria. This reflects Meerkat-8B’s tendency to provide more detailed and comprehensive responses than Llama-3-8B, which were perceived as overly verbose and less easy to follow. In contrast, GPT-4o, less affected by the verbosity of Meerkat-8B’s responses compared to humans, assigned higher clarity scores to Meerkat-8B. This underscores how metrics like clarity can reveal misalignments between human evaluations and model assessments. Also, while Meerkat’s responses are indeed clear, they could benefit from being more concise to improve readability for humans. Addressing this in future work could be an interesting direction.

Figure 3 presents ablation studies on CoT fine-tuning, textbook augmentation, and the backbone language model. First, we compared the performance of models trained solely on question-answer pairs from the MedQA training set (referred to as “MedQA”) with those trained on both MedQA data and CoT reasoning (referred to as “MedQA-CoT”). The results demonstrate that CoT fine-tuning dramatically improved MedQA performance by an average accuracy of 7.5% across the six models. These results emphasize the importance of training on CoT reasoning rather than solely relying on QA pair training. Second, augmenting the training data with additional QA pairs and CoT reasoning paths obtained from textbooks (referred to as “MedQA-CoT + MedBooks-18-CoT”) led to a further improvement in performance, with an average accuracy increase of 5.1% across the six models compared to those trained using MedQA-CoT alone. Lastly, we assessed six open-source LMs with 7B or 8B parameters released between July 2023 and April 2024 using the MedQA dataset. As a result, general-purpose models like Mistral-7B, Gemma-7B[27], and Llama-3-8B outperformed biomedical-specific models such as MediTron-7B and BioMistral-7B. Although the details of their pre-training corpus remain unknown, we hypothesize that these models achieved high MedQA performance due to extensive training on diverse corpora, likely including a large amount of biomedical literature. While Mistral and Gemma exhibited similar performance, we selected Mistral-7B as our backbone model because of its faster inference speed[26]. Following the release of Llama-3 in April 2024, we proceeded with training a new model, resulting in Meerkat-8B.

Discussion

Our Meerkat models, with 7B and 8B parameters, are compact yet powerful language models, designed to address the gap between performance and security. Designed for on-premises deployment, they ensure secure operation while maintaining compatibility with lower-spec GPUs, such as a single 24GB NVIDIA GeForce RTX 3090. This accessibility makes them suitable for a wide range of research institutions and healthcare facilities. Despite their smaller size, the Meerkat models demonstrate robust reasoning capabilities, as evidenced by improved performance on complex benchmark datasets and positive assessments from experts.

One of our key contributions is the introduction of an effective method to enhance smaller models with reasoning capabilities. In general domains, several studies have shown the promise of generating CoT reasoning using LLMs and fine-tuning SLMs with the generated data[33–35]. However, this approach had remained underexplored in the medical domain, with previous efforts predominantly relying on continuous pre-training[28–31]. A prior study [28] generated rationales for QA datasets like MedQA and MedMCQA to train their models; however, their approach relied on GPT-3.5 and simplistic prompts, leading to low-quality reasoning outputs. In contrast, we leveraged GPT-4 with highly refined prompts to produce substantially higher-quality reasoning data. Most notably, our study is the first to propose the augmentation of CoT data using raw medical textbooks. Additionally, as shown in Fig. 3, we demonstrated its effectiveness across several state-of-the-art language models, highlighting its robustness. In future research, it would be interesting to apply our method to newly released language models, as well as other corpora, such as clinical guidelines and PubMed articles.

The success of our models can be attributed to the reasoning capabilities developed through CoT reasoning, primarily trained on USMLE-style questions. This methodology has significantly improved performance on USMLE-style benchmarks. Notably, the benefits extend beyond exam-focused tasks, as we observed a marked performance improvement on the NEJM Case Challenges, which are based on real patient cases. These results suggest that the advancements made by Meerkat are not confined to exam contexts but reflect a broader enhancement of reasoning abilities, as validated by expert evaluation (Fig. 2). We anticipate that these reasoning skills could be transferable to more complex, real-world clinical challenges in medicine. For example, our model could suggest additional treatment options and differential diagnoses based on a given patient’s presentation. This would provide candidates for treatment or diagnosis that the physician might have overlooked, encouraging a broader consideration of possibilities and helping to mitigate human biases.

Our fine-tuning approach is most effective with smaller models rather than larger ones. While we focused on evaluating models with fewer than ten billion parameters, which are well-suited for on-premises implementation, we also conducted a follow-up experiment on its effectiveness with larger models. When training Llama-3-70B on our data, the Meerkat-70B model significantly outperformed Llama-3-70B by an average score of 2.9% across six benchmarks, with the difference achieving high statistical significance ($p < 1.3978e^{5}$) (see Supplementary Table 2). Additionally, our model surpassed GPT-4 and GPT-4o by 1.4% and 0.5%, respectively. On the NEJM Case Challenges, the number of correct answers increased from 20 with

Llama-3-70B to 22. However, compared to the improvements observed with smaller models, the increase was relatively modest, likely due to the fact that larger models possess greater reasoning capabilities, leaving less room for learning from our CoT data. Therefore, we recommend testing with SLMs, such as those in the 7B to 8B range, or even smaller models, in future applications.

While our training approach significantly enhances the performance of SLM models, there is still considerable room for improvement to surpass the performance of commercial LLMs. Recent reasoning-centric models, such as O1 and DeepSeek-R1[24], have demonstrated remarkable improvements through reinforcement learning. Applying similar techniques to our model would be an intriguing direction for future research, which we leave as a topic for further exploration.

We highlight the need for further development towards more reliable AI systems in medical applications. We conducted a case study using several real-world clinical inquiries[37], comparing our model with the leading medical chat model, ChatDoctor-7B[38], GPT-4, and human responses (see Supplementary Tables 3–5). Overall, our model provided more detailed responses, outperforming ChatDoctor-7B. However, when compared to the larger model, GPT-4, our model occasionally produced inaccurate information, particularly regarding dosage. We attribute this not to limitations in reasoning, but to the inherent constraints of smaller models in terms of their medical knowledge and memorization capabilities. Approaches like retrieval-augmented generation (RAG) could potentially provide a promising solution[39]. Furthermore, given that our models were not fine-tuned using preference alignment techniques like reinforcement learning from human feedback (RLHF)[40], there is a possibility it could offer unsupported, unsafe, or biased responses. Hence, it is crucial to exercise caution and obtain expert validation before deploying the models in real-world scenarios to guarantee their reliability.

Table 2 | Statistics of our instruction-tuning datasets

Target application	Dataset	# Examples
Multiple-choice QA	MedQA-CoT[a36]	9308
	MedBooks-18-CoT[a]	77,660
	MedMCQA[41]	182,822
Free-form/Single-turn QA	LiveQA[42]	633
	MedicationQA[43]	689
	ChatDoctor-cleaned[a38]	111,902
Multi-turn QA	MedQA-dialog[a36]	4818
Clinical Note Generation	MTS-dialog[60]	1200
Miscellaneous	MedInstruct-52K[44]	52,002

"# Examples" denotes the number of training examples for each dataset.
[a]indicates that the dataset is newly constructed in this study. The total number of training examples is 441,034.

Methods

Meerkat models are based on state-of-the-art open-source models such as Mistral-7B and LLaMA-3-8B and are specifically instruct-tuned for the medical domain. The key highlight of the model training lies in constructing a high-quality instruction-tuning dataset. The training data preparation involves three primary steps: generating high-quality CoT data from a question-answering (QA) dataset, augmenting this CoT data using medical textbooks, and reusing/repurposing pre-existing instruction-following datasets covering various medical use cases with suitable modifications. Table 2 lists the datasets used for training the Meerkat models. Figure 4 depicts the overall process for generating and augmenting CoT data.

We distilled reasoning capabilities from a larger model to a smaller model. Specifically, we generated reasoning data from the larger model and fine-tuned the smaller model using this reasoning process. To achieve this, we prompted GPT-4 (the larger model) to analyze each given MedQA question and its options step-by-step, deriving the final conclusion through systematic reasoning (see Supplementary Table 6 for the input prompt). The MedQA dataset comprises USMLE-style questions specifically designed to assess human multi-step medical reasoning skills, making them suitable sources for obtaining CoT reasoning data. The reasoning generated by GPT-4 using this approach is highly detailed and more comprehensive than human responses (see Fig. 5). To ensure data quality, we kept questions only if the responses followed the specified output format and the answer predictions were correct; otherwise, we filtered them out, resulting in 9.3K out of 10K questions remaining. We call this CoT data *MedQA-CoT*.

Collecting CoT paths solely from a single MedQA dataset does not provide a wide enough variety of training examples to maximize the reasoning abilities of small LMs. To overcome this obstacle, we constructed *MedBooks-CoT-18*, a dataset containing an additional 78K question-answer pairs along with corresponding CoT paths, which are automatically generated from 18 English medical textbooks, spanning various medical disciplines–anatomy, biochemistry, cell biology, first aid, gynecology, histology, immunology, internal medicine, neurology, obstetrics, pathology, pediatrics, pharmacology, physiology, psychiatry, and surgery. They are provided by the study of Jin et al.[36] and released under license for research use. We initially generated question-answer pairs to construct the CoT from this synthetic QA data. We segmented textbooks into chunks with an average character-level length of up to 4K, 8K, and 12K, while allowing for overlap between chunks. Each text chunk was fed into GPT-4 with three USMLE-style questions sampled from MedQA for reference. GPT-4 was

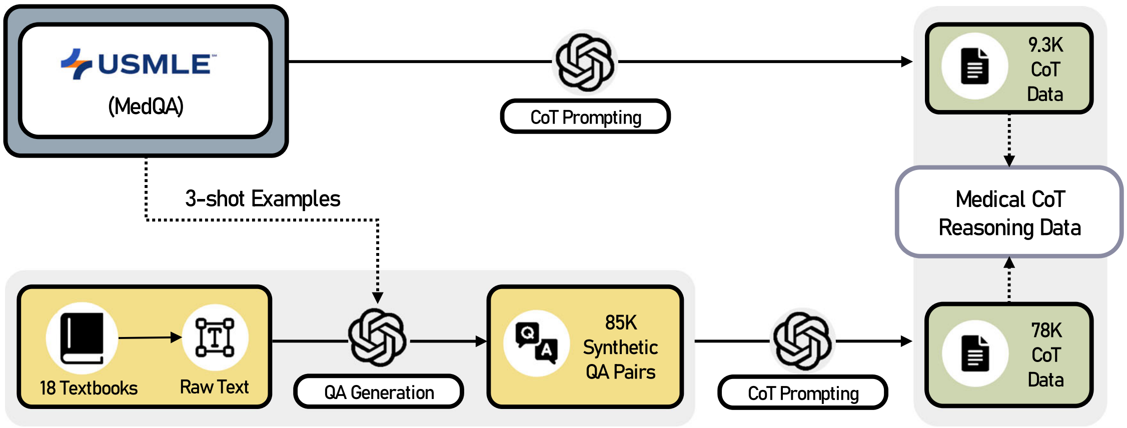
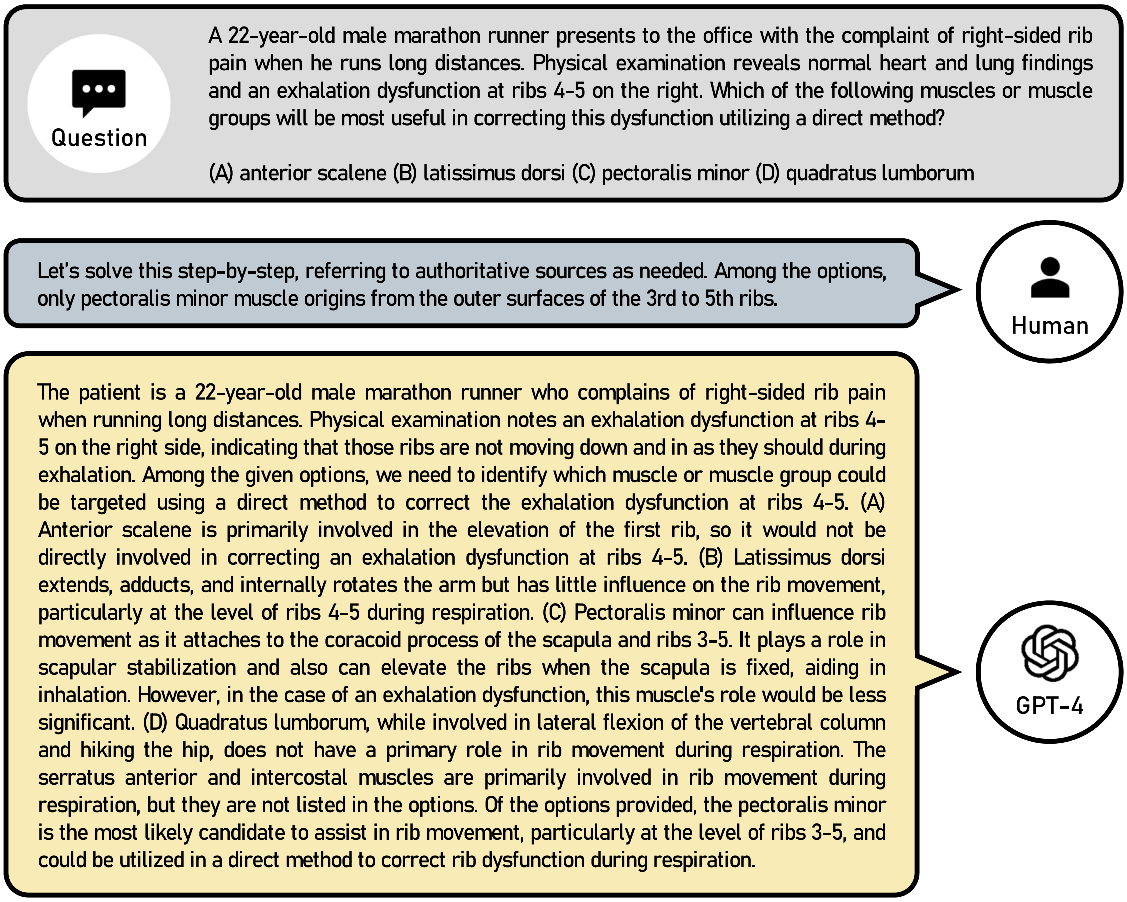


Fig. 4 | GPT-4 was prompted to provide answers, along with step-by-step explanations, for USMLE-style questions from MedQA, resulting in the creation of 9.3K CoT data. GPT-4 also received three randomly sampled questions from MedQA and text chunks from medical textbooks as input to produce synthetic question-answer pairs. GPT-4 was then prompted to generate step-by-step explanations for these generated questions, resulting in an additional 78K CoT data.



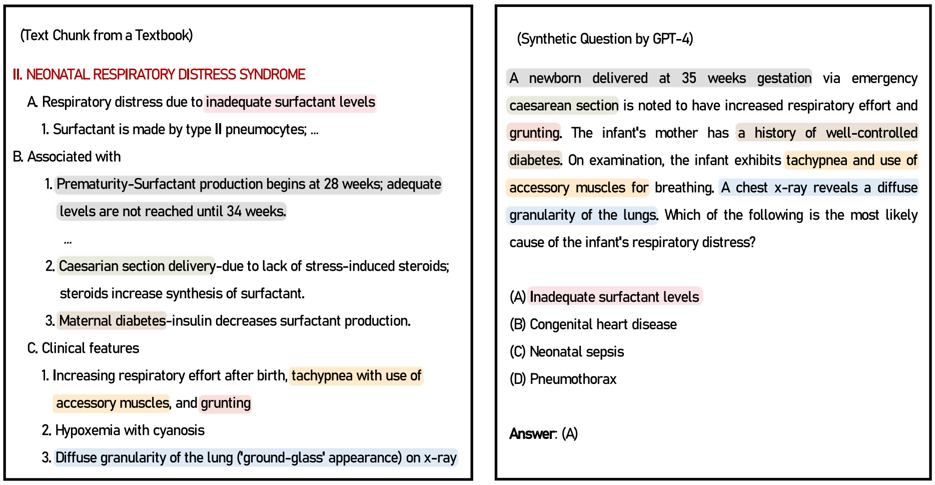
A 22-year-old male marathon runner presents to the office with the complaint of right-sided rib pain when he runs long distances. Physical examination reveals normal heart and lung findings and an exhalation dysfunction at ribs 4-5 on the right. Which of the following muscles or muscle groups will be most useful in correcting this dysfunction utilizing a direct method?

(A) anterior scalene (B) latissimus dorsi (C) pectoralis minor (D) quadratus lumborum

Let's solve this step-by-step, referring to authoritative sources as needed. Among the options, only pectoralis minor muscle origins from the outer surfaces of the 3rd to 5th ribs.

The patient is a 22-year-old male marathon runner who complains of right-sided rib pain when running long distances. Physical examination notes an exhalation dysfunction at ribs 4-5 on the right side, indicating that those ribs are not moving down and in as they should during exhalation. Among the given options, we need to identify which muscle or muscle group could be targeted using a direct method to correct the exhalation dysfunction at ribs 4-5. (A) Anterior scalene is primarily involved in the elevation of the first rib, so it would not be directly involved in correcting an exhalation dysfunction at ribs 4-5. (B) Latissimus dorsi extends, adducts, and internally rotates the arm but has little influence on the rib movement, particularly at the level of ribs 4-5 during respiration. (C) Pectoralis minor can influence rib movement as it attaches to the coracoid process of the scapula and ribs 3-5. It plays a role in scapular stabilization and also can elevate the ribs when the scapula is fixed, aiding in inhalation. However, in the case of an exhalation dysfunction, this muscle's role would be less significant. (D) Quadratus lumborum, while involved in lateral flexion of the vertebral column and hiking the hip, does not have a primary role in rib movement during respiration. The serratus anterior and intercostal muscles are primarily involved in rib movement during respiration, but they are not listed in the options. Of the options provided, the pectoralis minor is the most likely candidate to assist in rib movement, particularly at the level of ribs 3-5, and could be utilized in a direct method to correct rib dysfunction during respiration.

Fig. 5 | GPT-4 offers answers that are notably more detailed, comprehensive, and accurate compared to those from humans. The human explanation is sourced from the study of Singhal et al.[9].



(Text Chunk from a Textbook)

II. NEONATAL RESPIRATORY DISTRESS SYNDROME

A. Respiratory distress due to inadequate surfactant levels

1. Surfactant is made by type II pneumocytes; ...

B. Associated with

1. Prematurity-Surfactant production begins at 28 weeks; adequate levels are not reached until 34 weeks.

...

2. Caesarian section delivery-due to lack of stress-induced steroids; steroids increase synthesis of surfactant.

3. Maternal diabetes-insulin decreases surfactant production.

C. Clinical features

1. Increasing respiratory effort after birth, tachypnea with use of accessory muscles, and grunting

2. Hypoxemia with cyanosis

3. Diffuse granularity of the lung ('ground-glass' appearance) on x-ray

(Synthetic Question by GPT-4)

A newborn delivered at 35 weeks gestation via emergency caesarean section is noted to have increased respiratory effort and grunting. The infant's mother has a history of well-controlled diabetes. On examination, the infant exhibits tachypnea and use of accessory muscles for breathing. A chest x-ray reveals a diffuse granularity of the lungs. Which of the following is the most likely cause of the infant's respiratory distress?

(A) Inadequate surfactant levels

(B) Congenital heart disease

(C) Neonatal sepsis

(D) Pneumothorax

Answer: (A)

Fig. 6 | The left panel represents the textbook chunk, while the right panel displays data generated based on that chunk. Spans highlighted in the same color indicate related content. It is evident that GPT-4 heavily relies on the textbook to generate questions, resulting in a significant reduction in hallucination and ensuring that the generated QA examples cover essential content from the textbook. This leads to an overall improvement in quality.

subsequently prompted to generate the correct answer, plausible options, and a corresponding question based on the provided textbook chunk (see Supplementary Table 7 for prompt details). The question included a case representation containing the patient's demographic information and symptoms. By instructing GPT-4 to generate questions based on textbooks, as depicted in Fig. 6, we can reduce potential hallucinations during the question generation process and ensure that the questions reflect essential medical knowledge covered in the textbooks, relevant to both medical exams and real-world clinical practice. Additionally, we instructed the model to refrain from generating questions if the provided text chunk contained significant noise or lacked adequate information to create QA examples. After question-answer pairs were created from textbooks, we followed a similar procedure to generate CoT reasoning paths using GPT-4 as we did with the MedQA dataset to ensure data quality.

In addition to the CoT datasets that we constructed, we incorporated existing instruction-following datasets into the model training to enhance the versatility of our model for various medical-domain applications (see Table 2 for the summary). We refined or repurposed several datasets to better suit the model training and align with the target applications. Below are detailed descriptions of each dataset:

- MedMCQA[41]: this large dataset comprises exam questions from the two Indian affiliations, AIIMS (All India Institute of Medical Sciences) and NEET PG (National Eligibility cum Entrance Test for Post Graduate courses). We leveraged this dataset because it spans a broad spectrum of medical knowledge across 21 subjects, which could complement the medical knowledge of small LMs. Although the dataset also includes human explanations for the questions, we did not utilize them because they were too brief and not sufficiently detailed.
- LiveQA[42]: this dataset contains healthcare-related questions received by the U.S. National Library of Medicine (NLM), accompanied by free-form responses from experts. The questions span various topics including diseases, drugs, and more, making it ideal for training our model on real-world queries.
- MedicationQA[43]: this dataset comprises consumer questions, particularly focusing on inquiries related to drugs, along with expert responses. Since these types of questions constitute a significant portion of healthcare inquiries, they serve as valuable resources for developing practical medical AI models.
- ChatDoctor-cleaned (ours): the data is derived from ChatDoctor[38], a collection of real patient inquiries and doctor responses obtained from an online medical consultation platform. While ChatDoctor provides rich and useful data examples, it also contains noise inherent to online Q&A platforms, such as greetings or closing remarks by the doctors (e.g., they often begin the response with “Welcome to Chat Doctor” or end the response with “Best wishes, Chat Doctor.”). To address this, we manually created three noisy inputs and corresponding corrected outputs, utilizing them as in-context examples. We then employed GPT-3.5 to remove noise from 112K original responses, resulting in our ChatDoctor-cleaned dataset. See Supplementary Table 8 for the input prompt.
- MedQA-dialog (ours): while engaging in multi-turn dialog with users is a crucial requirement for medical AI, there’s a lack of suitable training datasets for this purpose. To fill this gap, we instructed GPT-3.5 to generate conversations by role-playing as both patients and doctors based on MedQA questions and corresponding CoT reasoning. In the dialog, the patient should minimally communicate their symptoms and medical history, while the doctor should guide the conversation, asking follow-up questions to gather a thorough medical history and records. We generated 4.8K conversations corresponding to the intended output format from the entire 9.3K MedQA examples, creating the MedQA-dialog dataset. Details of the input prompt is provided in Supplementary Table 9.
- MedInstruct-52K[44]: this dataset is a collection of synthetic 52K medical-domain instructions generated by GPT-3.5 and GPT-4, similar to the self-instruct approach[45,46]. We included this dataset to improve the model’s generalizability to various user queries and use cases.

We initialized our models with the Mistral-7B-v0.1 and Meta-Llama-3-8B-Instruc weights, respectively. We selected these models through a validation process, in which the models were evaluated on MedQA after fine-tuning with MedQA-CoT and MedBooks-18-CoT (see Ablation Study for details). We fine-tuned the models on a combined dataset comprising the nine training datasets listed in Table 2, using a standard next-token prediction objective. The 7B model was trained for three epochs with a batch size of 128, a learning rate of 2e-6, a warm-up ratio of 0.04, and a maximum length of 2048 tokens, using eight 80G A100 GPUs, which took approximately 1 days to complete. The 8B model was trained on Google TPUs with batch sizes of 128, and learning rates of 7e-6. We employed FlashAttention[47] and the fully sharded data parallel approach (FSDP) for efficiency.

We evaluated our models against three categories of robust baseline models: (1) proprietary LLMs, including GPT-3.5[17], GPT-4[18], GPT-4o[48], o1-mini[49], o3-mini[50], and o1[51]; (2) open-source SLMs, such as Mistral-7B-Instruct-v0.1[26] and Llama-3-8B-Instruct[22]; and (3) domain-specific SLMs, including MediTron-7B[29] and BioMistral-7B[30]. We fine-tuned MediTron-7B and BioMistral-7B for three epochs using the MedQA and MedMCQA training sets, as these models were neither instruction-tuned nor aligned with the required answer formats, even when carefully prompted. In contrast, we did not further fine-tune the Mistral and Llama models, which are already instruction-tuned, following the approach of Chen et al.[29].

Table 3 | Statistics of benchmark datasets and the evaluation methods used

Type	Dataset	Metric	# Examples
Quantitative assessment			
USMLE Exams	MedQA[36]	Accuracy	1273
	USMLE sample test		325
	Medbullets-4[54]		308
	Medbullets-5[54]		308
Other Exams	MedMCQA[41]	Accuracy	4182
	MMLU-Medical[55]	Average Accuracy	1089
	- Clinical knowledge		265
	- Medical genetics		100
	- Anatomy		135
	- Professional medicine		272
	- College biology		144
	- College medicine		173
Case Challenges	NEJM Case Challenges[16]	Accuracy	38
Qualitative assessment			
Human A/B Test (Comparison of Models’ Reasoning)	Medbullets-5[54] (Subset)	Completeness Factuality Clarity Logical Consistency	50

“# Examples”: the number of test examples for each dataset. The seven datasets for quantitative assessment comprise multiple-choice question-answering tasks, where the correct answer must be selected from a set of provided options. The Human A/B test qualitatively evaluates the reasoning process.

In qualitative assessment, we evaluated the models using six medical exam datasets and one additional dataset that is more challenging and features realistic patient cases. Medical exams are widely used testbeds for evaluating the foundational medical knowledge and reasoning abilities of AI models[8–12,28–31,52]. Case challenges are used to assess advanced reasoning in real-world cases[16]. We provide an overview of the benchmark datasets in Table 3 with detailed descriptions below.

- MedQA[36]: MedQA is one of the most widely used benchmarks in the medical domain. The dataset consists of USMLE-style questions curated by medical examination experts from various medical question banks (an example of USMLE-style questions can be found in Fig. 5). These questions are structured in a multiple-choice format, with four options provided for each question.
- USMLE sample test: this resource is an official study material for students preparing for the USMLE, closely mirroring the style and difficulty level of the actual tests. Each question is accompanied by a varying number of options, ranging from four to nine. We utilized the preprocessed version of this data as provided by Toma et al.[53].
- Medbullets[54]: this dataset comprises USMLE-style questions sourced from tweets posted since April 2022. Compared to questions in MedQA or the USMLE sample test, these questions are less likely to have been encountered during pre-training, making them more challenging to solve. We utilized both Medbullets-4, which provides four options, and Medbullets-5, which offers five options.
- MedMCQA[41]: this benchmark corresponds to the test split of the MedMCQA dataset, which consists of medical exam questions with four options.
- MMLU-Medical[55]: MMLU was originally designed to assess the world knowledge of models across various subjects including mathematics,

physics, history, and law. Singhal et al.[9] created MMLU-Medical by extracting six subjects relevant to the medical field from MMLU, clinical knowledge, medical genetics, anatomy, professional medicine, college biology, and college medicine, aiming to evaluate medical-specialized systems.

- NEJM Case Challenges[16]: this dataset, sourced from the New England Journal of Medicine website (https://www.nejm.org/case-challenges) comprises long and complex real-world clinical cases, requiring deep clinical reasoning and a thorough understanding of complex health conditions. Most questions involve diagnosing or determining the next steps based on the provided case, which includes extensive text, laboratory results in tables, and imaging results as captions. The dataset adopts a multiple-choice QA format, offering six options for each question.

Evaluating models solely based on their accuracy on benchmarks is insufficient, as flaws often emerge in the rationale behind the model's answer, even when the answer itself is correct[56]. These flaws can undermine the reliability and robustness of the model's decision-making process. Therefore, a comprehensive qualitative analysis of the model's reasoning process is essential to fully understand its performance and to identify areas for improvement. We conducted a human A/B test to thoroughly evaluate the CoT reasoning of two models on the Medbullets-5 questions. A total of ten evaluators participated, including four medical students and six MD physicians who had completed their licensure process in South Korea. Among the evaluators, five had experience studying for or passing the USMLE step-1 exam. To account for differences in experience, we divided the evaluators into two groups of five based on their medical licensure status and USMLE experience. Evaluators within the same group assessed the same set of 25 questions and the final evaluation for each question was determined by a majority vote within the group. Evaluators selected the model they deemed superior, with the option to choose "Draw" only when it was difficult to determine a clear advantage between the models. We held a separate Q&A session to provide detailed instructions for the evaluation process, ensuring that all evaluators were familiar with the evaluation guidelines and key considerations. Additionally, we provided evaluators with the human gold-standard explanations, which were provided by the Medbullets-5 data, to enhance the objectivity of the evaluation. We employed the following four metrics in the evaluation process:

- Completeness: Does the explanation cover all necessary information to fully answer the question, without leaving any important details out?
- Factuality: Is the information provided accurate and reliable?
- Clarity: How clearly does the explanation communicate its points? Ensure the language is easy to understand and free from ambiguity.
- Logical Consistency: Does the explanation maintain internal consistency, with ideas presented in a coherent manner?

The model responses were anonymized before being presented to the evaluators. To minimize the influence of each model's response style, we applied preprocessing steps such as removing special characters favored by specific models. Additionally, the model responses were randomly shuffled. We selected 50 cases in which both models produced correct answers to eliminate any discrepancies in evaluation scores arising from answer accuracy. The models used for this evaluation were Meerkat-8B and its counterpart, Llama-3-8-Instruct.

During inference, we used the vLLM platform for fast inference[57]. We applied BFloat16 and greedy decoding for the single-model evaluation in the QA tasks. In the ensemble evaluation, we used a temperature of 0.7, and a repetition penalty of 1.0. We utilized a choice shuffling ensemble technique[58], which involves randomizing the given options before presenting them to the models and subsequently conducting a majority vote to determine the final predictions. This helps mitigate potential biases in the position of the correct answer[59].

Data availability

Training Data: https://huggingface.co/datasets/dmis-lab/meerkat-instructions, MedQA: https://github.com/jind11/MedQA, USMLE sample test: https://www.usmle.org/exam-resources, MedMCQA: https://medmcqa.github.io/, Medbullets: https://github.com/HanjieChen/ChallengeClinicalQA, NEJM Case Challenges: https://www.nejm.org/case-challenges, MMLU: https://huggingface.co/datasets/lukaemon/mmlu.

Code availability

Our models are publicly available at our official Hugging Face repository (https://huggingface.co/collections/dmis-lab/meerkat-6710b7ae0258fc540c475eec). Meerkat-7B and Meerkat-8B were trained based on the following GitHub repositories: FastChat (https://github.com/lm-sys/FastChat) and LLaMA-Factory (https://github.com/hiyouga/LLaMA-Factory/tree/main). The pre-trained weights of Mistral-7B-v0.1 and Meta-Llama-3-8B-Instruct are available at Hugging Face, huggingface.co/mistralai/Mistral-7B-v0.1 and huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct, respectively.

Received: 26 January 2025; Accepted: 19 April 2025;
Published online: 02 May 2025

References

1. Liu, J. X., Goryakin, Y., Maeda, A., Bruckner, T. & Scheffler, R. Global health workforce labor market projections for 2030. *Hum. Resour. Health* **15**, 1–12 (2017).
2. Scheffler, R. M. & Arnold, D. R. Projecting shortages and surpluses of doctors and nurses in the OECD: what looms ahead. *Health Econ. Policy Law* **14**, 274–290 (2019).
3. Tamata, A. T. & Mohammadnezhad, M. A systematic review study on the factors affecting shortage of nursing workforce in the hospitals. *Nurs. Open* **10**, 1247–1257 (2023).
4. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
5. Norgeot, B. et al. Protected health information filter (philter): accurately and securely de-identifying free-text clinical notes. *NPJ Digital Med.* **3**, 57 (2020).
6. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
7. Tian, S. et al. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Brief. Bioinforma.* **25**, bbad493 (2024).
8. Kung, T. H. et al. Performance of ChatGPT on USMLE: potential for ai-assisted medical education using large language models. *PLoS Digital health* **2**, e0000198 (2023).
9. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
10. Singhal, K. et al. Toward expert-level medical question answering with large language models. *Nat. Med.* **31**, 943–950 (2025).
11. Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of gpt-4 on medical challenge problems. *arXiv* https://doi.org/10.48550/arXiv.2303.13375 (2023).
12. Brin, D. et al. Comparing chatgpt and gpt-4 performance in usmle soft skill assessments. *Sci. Rep.* **13**, 16492 (2023).
13. Xie, Y. et al. A preliminary study of o1 in medicine: are we closer to an AI doctor? *arXiv* https://doi.org/10.48550/arXiv.2409.15277 (2024).
14. Zakka, C. et al. Almanac-retrieval-augmented language models for clinical medicine. *NEJM AI* **1**, AIoa2300068 (2024).
15. Tu, T. et al. Towards conversational diagnostic ai. *arXiv* https://doi.org/10.48550/arXiv.2401.05654 (2024).
16. Eriksen, A. V., Möller, S. & Ryg, J. Use of gpt-4 to diagnose complex clinical cases. (2023).
17. OpenAI. Introducing chatgpt. https://openai.com/blog/chatgpt (2022).

18. Achiam, J. et al. Gpt-4 technical report. *arXiv* https://doi.org/10.48550/arXiv.2303.08774 (2023).
19. Li, X. & Zhang, T. An exploration on artificial intelligence application: from security, privacy and ethic perspective. In *2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, 416–420 (IEEE, 2017).
20. Bartoletti, I. AI in healthcare: ethical and privacy challenges. In *Artificial Intelligence in Medicine: 17th Conference on Artificial Intelligence in Medicine, AIME 2019, Poznan, Poland, June 26–29, 2019, Proceedings 17*, 7–10 (Springer, 2019).
21. Meskó, B. & Topol, E. J. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digital Med.* **6**, 120 (2023).
22. AI@Meta. Llama 3 model card. (2024).
23. xAI. Open release of grok-1. https://x.ai/blog/grok-os (2024).
24. Guo, D. et al. Deepseek-r1: incentivizing reasoning capability in llms via reinforcement learning. *arXiv* https://doi.org/10.48550/arXiv.2501.12948 (2025).
25. Touvron, H. et al. Llama: open and efficient foundation language models. *arXiv* https://doi.org/10.48550/arXiv.2302.13971 (2023).
26. Jiang, A. Q. et al. Mistral 7b. *arXiv* https://doi.org/10.48550/arXiv.2310.06825 (2023).
27. Google. Gemma: introducing new state-of-the-art open models. https://blog.google/technology/developers/gemma-open-models/ (2024).
28. Wu, C. et al. PMC-LLaMA: toward building open-source language models for medicine. *J. Am. Med. Inf. Assoc.* **31**, 1833–1843 (2024).
29. Chen, Z. et al. Meditron-70b: scaling medical pretraining for large language models. *arXiv* https://doi.org/10.48550/arXiv.2311.16079 (2023).
30. Labrak, Y. et al. Biomistral: a collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics ACL 2024*, 5848–5864 (Association for Computational Linguistics, 2024).
31. Xie, Q. et al. Medical foundation large language models for comprehensive text analysis and beyond. *npj Digit. Med.* **8**, 141 (2025).
32. Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **35**, 24824–24837 (2022).
33. Wei, J. et al. Emergent abilities of large language models. *Transactions on Machine Learning Research* https://openreview.net/forum?id=yzkSU5zdwD (2022).
34. Tay, Y. et al. Ul2: Unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations* (OpenReview.net, 2022).
35. Chung, H. W. et al. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.* **25**, 1–53 (2024).
36. Jin, D. et al. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Appl. Sci.* **11**, 6421 (2021).
37. Manes, I. et al. K-qa: a real-world medical q&a benchmark. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, 277–294 (Association for Computational Linguistics, 2024).
38. Li, Y. et al. Chatdoctor: a medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus* **15**, e40895 (2023).
39. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv. Neural Inf. Process. Syst.* **33**, 9459–9474 (2020).
40. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **35**, 27730–27744 (2022).
41. Pal, A., Umapathi, L. K. & Sankarasubbu, M. Medmcqa: a large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, 248–260 (PMLR, 2022).
42. Abacha, A. B., Agichtein, E., Pinter, Y. & Demner-Fushman, D. Overview of the medical question answering task at trec 2017 liveqa. In *TREC*, 1–12 (National Institute of Standards and Technology (NIST), 2017).
43. Abacha, A. B. et al. Bridging the gap between consumers' medication questions and trusted answers. In *MedInfo*, 25–29 (IOS Press, 2019).
44. Zhang, X. et al. Alpacare: Instruction-tuned large language models for medical application. *arXiv* https://doi.org/10.48550/arXiv.2310.14558 (2023).
45. Wang, Y. et al. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Vol 1: Long Papers (2023).
46. Taori, R. et al. Stanford alpaca: an instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca (2023).
47. Dao, T., Fu, D., Ermon, S., Rudra, A. & Ré, C. Flashattention: fast and memory-efficient exact attention with io-awareness. *Adv. Neural Inf. Process. Syst.* **35**, 16344–16359 (2022).
48. OpenAI. Gpt-4o system card. https://openai.com/index/gpt-4o-system-card/ (2024).
49. OpenAI. Openai o1-mini. https://openai.com/index/openai-o1-mini-advancing-cost-efficient-reasoning/ (2024).
50. OpenAI. Openai o3-mini. https://openai.com/index/openai-o3-mini/ (2025).
51. OpenAI. Introducing OpenAI o1. https://openai.com/o1/ (2024).
52. Saab, K. et al. Capabilities of gemini models in medicine. *arXiv* https://doi.org/10.48550/arXiv.2404.18416 (2024).
53. Toma, A. et al. Clinical camel: an open-source expert-level medical language model with dialogue-based knowledge encoding. *arXiv* https://doi.org/10.48550/arXiv.2305.12031 (2023).
54. Chen, H., Fang, Z., Singla, Y. & Dredze, M. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol 1: Long Papers, 3563–3599 (2025).
55. Hendrycks, D. et al. Measuring massive multitask language understanding. In *International Conference on Learning Representations* (OpenReview.net, 2020).
56. Jin, Q. et al. Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine. *npj Digital Med.* **7**, 190 (2024).
57. Kwon, W. et al. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles* (Association for Computing Machinery, 2023).
58. Nori, H. et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. *arXiv* https://doi.org/10.48550/arXiv.2311.16452 (2023).
59. Ko, M., Lee, J., Kim, H., Kim, G. & Kang, J. Look at the first sentence: position bias in question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1109–1121 (Association for Computational Linguistics, 2020).
60. Abacha, A. B., Yim, W.-W., Fan, Y. & Lin, T. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2283–2294 (Association for Computational Linguistics, 2023).

Acknowledgements

We thank Heeju Jin, Yena Joo, Seihee Park, Hyunsuk Lee, Sebin Lee, Ji Yeong Park, Jieun Kim, Yujin Park, Seoyoung Choi, and Soobin Cho for their valuable time and feedback. The authors affiliated with Korea University and S.H.K. were supported by the National Research Foundation of Korea (RS-2023-00262002). Additionally, the Korea University-affiliated authors received funding from the following sources: the National Research

Foundation of Korea (NRF-2023R1A2C3004176); the MSIT (Ministry of Science and ICT), Korea, under the ICT Creative Consilience program (IITP-2025-RS-2020-II201819) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation); a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HR20C00213); and the Artificial Intelligence Industrial Convergence Cluster Development Project, funded by MSIT and Gwangju Metropolitan City. H.K. was supported by a Korea University research grant.

Author contributions

H.K. and J.K. designed the study. H.K. wrote the main manuscript text. H.K. and J.P. conducted the experiments. S.P. drafted the figures. H.H., J.L., S.P., D.K., T.L., C.Y., and J.S., performed the data preprocessing and analysis. O.R., T.F., and S.H.K. contributed to the interpretation of the results. D.C., Q.C., and J.K. reviewed the manuscript and provided critical feedback. All authors have read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01653-8.

Correspondence and requests for materials should be addressed to Jaewoo Kang.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.