MedReason-Dx: Benchmarking Step-by-Step Reasoning of Language Models in Medical Diagnosis

MedReason-Dx Team

Abstract

In high-stakes domains like medicine, **how** an AI arrives at an answer can be as critical as the answer itself. However, existing medical question answering benchmarks largely ignore the reasoning process, evaluating models only on final answer accuracy. This paper addresses the overlooked importance of reasoning path evaluation in medical AI. We introduce **MedReason-Dx**, a novel benchmark that assesses not just answers but the step-by-step reasoning behind them. MedReason-Dx provides expert-annotated step-by-step solutions for both multiple-choice and open-ended questions, spanning 24 medical specialties. By requiring models to produce and be evaluated on intermediate reasoning steps, our benchmark enables rigorous testing of interpretability and logical consistency in medical QA. We present the design of MedReason-Dx and outline diverse evaluation metrics that reward faithful reasoning. We hope this resource will advance the development of robust, interpretable medical decision support systems and foster research into large language models that can reason as well as they respond.

1 Introduction

Artificial intelligence systems for healthcare must not only deliver correct answers but also provide justifiable reasoning. In clinical decision support and medical question answering (QA), the reasoning path that leads to an answer is critical for trust and safety. A model that arrives at a diagnosis by flawed logic or guesswork poses significant risks, even if the final answer is correct. Conversely, a model that explains its reasoning can enable practitioners to verify each step, ensuring the conclusion is sound. Despite this, most existing benchmarks in medical AI evaluate models solely on whether the final answer is right, with little or no assessment of the reasoning process. This gap is problematic in high-stakes domains: evaluating only end answers may overlook dangerous reasoning errors and fails to encourage the development of models that "think" in a human-like, transparent manner. Recent advances in large language models (LLMs) and prompting techniques have brought reasoning to the forefront of AI research. In particular, *chain-of-thought* (CoT) prompting has demonstrated that LLMs can generate step-by-step solutions to complex problems, from math and logic puzzles to medical questions. By prompting models to articulate intermediate steps, researchers have achieved improved performance on challenging tasks and gained insight into model decision-making. For example, state-of-the-art medical LLMs can now produce explanations or rationales alongside their answers, showcasing the potential of AI to handle intricate clinical reasoning. These developments underscore an urgent need for benchmarks that can evaluate not just final accuracy but the quality of reasoning LLMs employ. If a model is prompted to reason but we lack ground truth reasoning paths for comparison, we cannot rigorously assess whether the model's reasoning is correct, complete, or clinically valid. Several medical QA datasets and benchmarks have emerged, yet they predominantly focus on answer correctness. Standard benchmarks drawn from medical exams (e.g., USMLE-style question banks, MedQA and MedMCQA) and research datasets like PubMedQA have driven progress in factual recall and question answering. Some of these resources include a short explanation or reference for the answer, but they do not provide a detailed, stepwise reasoning chain that could be used to evaluate a model's thought process. In other words, existing benchmarks treat reasoning as an

Preprint. Under review.

implicit skill, not an explicit target of evaluation. A model might earn full marks by selecting the correct option in a multiple-choice question, while in reality it could have arrived at that answer via incorrect assumptions or lucky guesswork. Conversely, a model might demonstrate mostly correct reasoning and make a minor error at the final step, but current benchmarks would simply mark the entire answer as wrong, offering no credit for nor analysis of the model's reasoning ability. This limitation hampers the development of robust medical AI: it is difficult to discern whether improvements in accuracy are due to better reasoning or just better pattern matching, and it provides no incentive for models to output interpretable solutions. To address these challenges, we propose **MedReason-Dx**, a new benchmark expressly designed to evaluate chain-of-thought reasoning in medical question answering. MedReason-Dx (where "CoT" denotes Chain-of-Thought) introduces several key innovations to the evaluation of medical AI:

- **Expert-annotated reasoning chains:** Each question in MedReason-Dx is accompanied by a step-by-step solution path crafted by medical experts. These reasoning chains detail the logical steps required to arrive at the correct answer, including relevant clinical facts, intermediate inferences, and elimination of distractors in the case of multiple-choice items. This provides a gold-standard trace of correct reasoning against which model-generated solutions can be compared.

- **Diverse question formats and topics:** Our benchmark covers a broad spectrum of medical knowledge. It includes both multiple-choice questions (with several answer options) and open-ended questions that require free-form answers, ensuring that models are tested on various response formats. The questions span 24 medical specialties, ranging from internal medicine and cardiology to pediatrics, surgery, and more. This diversity reflects the real-world breadth of medical practice and ensures that the benchmark evaluates reasoning across different sub-domains and problem types (diagnosis, treatment decisions, biomedical mechanism explanations, etc.).

- **Evaluation metrics for reasoning quality:** MedReason-Dx departs from traditional single-metric evaluation by introducing multiple criteria to assess model performance. In addition to standard answer accuracy, we define metrics that measure the fidelity of a model's reasoning to the expert-provided chain. For instance, we evaluate whether the model's reasoning covers the same key steps or medical facts as the reference solution, and whether the logical progression is sound. This could involve step-wise accuracy scoring, similarity measures between generated and reference reasoning, and expert review of reasoning coherence. By quantifying reasoning quality, the benchmark encourages models that are not only correct, but correct for the right reasons.

- **Interpretability and robustness focus:** By requiring and evaluating intermediate reasoning, MedReason-Dx places interpretability at the core of model assessment. This is especially crucial for medical AI systems that clinicians need to trust. A model that can articulate a valid reasoning chain is inherently more transparent and easier to debug than one that only outputs an answer. Furthermore, focusing on reasoning helps reveal when a model's knowledge is superficial. We anticipate that models performing well on MedReason-Dx will demonstrate greater robustness, as they must handle complex multi-step problems in a principled way rather than relying on shallow cues. Our benchmark thus serves as a stress test for genuine reasoning ability in medical contexts.

In summary, MedReason-Dx is the first benchmark to comprehensively target reasoning path evaluation in medical QA. It offers the community a testbed to develop and rigorously vet models that aim to be not just answer engines, but reliable reasoning assistants for healthcare. We describe the construction of MedReason-Dx, including the data collection and expert annotation process, and provide an analysis of its contents. We also outline an evaluation framework and baseline results using current LLMs (without revealing any performance outcomes here). By emphasizing how answers are derived, our work addresses a critical gap for high-stakes AI: the need for systems whose decisions can be inspected and trusted. We believe MedReason-Dx will facilitate research into interpretable and robust medical AI, ultimately contributing to safer and more effective clinical decision support tools.

2 Related Works

2.1 Medical LLMs

The evolution of medical large language models (Med-LLMs) has led to advancements in model architectures, training paradigms, and domain-specific adaptations, enabling applications in information extraction, clinical decision support, dialogue systems, and multimodal medical AI.

Early Med-LLMs, such as BioBERT (Lee et al., 2020) and PubMedBERT (Gu et al., 2021), were trained on extensive biomedical literature and PubMed abstracts, excelling in tasks like named entity recognition, relation extraction, and text classification. Models such as ClinicalT5 (Lu et al., 2022) and GatorTron (Yang et al., 2022) extend this capability to clinical text summarization and report generation, while Codex-Med (Liévin et al., 2024) specializes in structured medical documentation. Galactica (Taylor et al., 2022), designed for scientific and medical applications, enhances literature analysis and information retrieval.

Recent models incorporate instruction fine-tuning (IFT) and reinforcement learning from human feedback (RLHF) to improve the accuracy of medical text generation and knowledge extraction. Med-PaLM (Singhal et al., 2023) and Med-PaLM 2 (Singhal et al., 2025) exemplify this trend, refining medical question answering and clinical decision-making. Med-Alpaca (Han et al., 2023) further demonstrates the adaptability of fine-tuned language models for specialized healthcare applications. Meanwhile, GatorTronGPT (Peng et al., 2023) builds on the GatorTron architecture with targeted fine-tuning, enhancing its precision in medical report generation. Conversational AI models like ChatDoctor (Li et al., 2023b) is tailored for virtual medical consultations, offering patient triage assistance and personalized recommendations.

Beyond these, several Med-LLMs focus on domain-specific adaptations. PMC-LLaMA (Wu et al., 2023) enhances biomedical literature processing, aiding both academic research and clinical applications. GPT-4-Med (Nori et al., 2023), a refined adaptation of GPT-4, excels in complex clinical text processing and high-quality medical content generation. In the field of Traditional Chinese Medicine (TCM), models like Taiyi-LLM (Luo et al., 2024), and Zhongjing (Yang et al., 2024) integrate classical TCM literature with modern medical insights, supporting diagnosis and treatment planning. Additionally, advancements in multilingual and multimodal medical models have broadened AI's applicability in global healthcare. HuatuoGPT (Zhang et al., 2023) and its successor HuatuoGPT-II (Chen et al., 2023) leverage expanded datasets and optimized architectures to improve clinical report generation and diagnostic decision support. Med-Flamingo (Moor et al., 2023) extends Med-LLM capabilities to multimodal medical tasks, integrating textual and visual information. Med-Gemini (Saab et al., 2024), a bilingual model, facilitates cross-lingual medical communication, promoting international healthcare collaboration.

These advancements underscore the ongoing evolution of Med-LLMs, enhancing their ability to process complex medical language, integrate multimodal data, and support diverse healthcare applications. As these models continue to evolve, they hold the potential to significantly improve clinical decision-making, personalized medicine, and cross-cultural medical communication.

2.2 Medical Benchmarks

The development of diverse and standardized datasets, along with robust evaluation platforms, is essential for advancing AI applications in the medical domain. Existing research in this area can be broadly categorized into two main directions: (1) datasets tailored for various medical AI tasks and (2) automated benchmarks designed to assess the clinical capabilities of large models.

The first category consists of datasets that support tasks such as information extraction, question answering, text generation, and natural language inference. For instance, datasets like GENIA (Kim et al., 2003), CADEC (Karimi et al., 2015), and BC5CDR (Li et al., 2016) are widely used for named entity recognition, relation extraction, and event detection across biomedical literature and clinical records. Meanwhile, MedQA (Jin et al., 2021), PubMedQA (Jin et al., 2019), CMCQA (Xia et al., 2022), and Huatuo-26M (Li et al., 2023a) have been developed to evaluate models' abilities in medical knowledge retrieval, clinical reasoning, and diagnostic decision-making. Additionally, datasets such as MIMIC-III (Johnson et al., 2016), MIMIC-CXR (Johnson et al., 2019), HealthSearchQA (Singhal et al., 2023), and CORD-19 (Wang et al., 2020) facilitate tasks like clinical report generation, summarization, and case-based discussions. In the natural language inference domain,

Table 1: Comparison with existing Medical QA benchmarks.

Benchmark	CoT Evaluation	No. Domains	reasoning intensive	MCQ	OEQ	Expert Annotation
MMedBench	✗	21	✓	✓	✗	✓
MedQA	✗	-	✗	✓	✗	✗
MedMCQA	✗	21	✗	✓	✗	✓
MMLU	✗	6	✗	✓	✗	✗
Medbullets	✗	-	✓	✓	✗	✓
JAMA Challenge	✗	13	✓	✓	✗	✓
LiveQA	✗	-	✗	✗	✓	✓
ClinicBench	✗	-	✗	✓	✓	✓
Ours	✓	24	✓	✓	✓	✓

Table 2: Benchmarking the accuracy (%) performane of the existing models.

	multiple-choice		open-ended	
	CoT	Direct	CoT	Direct
DeepSeeK R1	65.03	64.36	40.14	42.39
DeepSeeK V3	60.47	59.79	33.56	37.02
GPT-4o	58.28	59.12	37.72	47.70
o1-mini	60.47	62.67	37.37	41.18
Baichuan4-Turbo	46.62	42.74	27.16	28.37

MedNLI (Romanov and Shivade, 2018) provides a benchmark for understanding logical relationships in medical texts. Recently, MedReason (Wu et al., 2025) was proposed to address the scarcity of high-quality, step-by-step reasoning data in the medical domain. Unlike datasets distilled directly from general-purpose LLMs, MedReason constructs 32,682 question–answer pairs with detailed Chain-of-Thought explanations, leveraging a structured medical knowledge graph to extract and guide the reasoning paths. These reasoning chains are factually grounded and validated through both automated answer checking and expert review by medical professionals from diverse specialties.

The second category focuses on automated benchmarks for evaluating large medical models, reducing reliance on expert-driven manual assessments. MedBench (Cai et al., 2024) provides a broad evaluation platform with 40,041 questions covering various medical fields. AutoEval (Liao et al., 2023) reformats USMLE questions into multi-turn dialogues, assessing models based on information coverage and task accuracy. LLM-Mini-CEX (Shi et al., 2023) leverages patient simulators and ChatGPT to evaluate diagnostic dialogue quality. MedGPTEval (Xu et al., 2023) integrates Chinese medical datasets and public benchmarks, using 16 expert-refined indicators to measure professional competence. LLM-Human Evaluation (Chiang and Lee, 2023) examines automated assessment feasibility, showing alignment with human evaluators in adversarial and open-ended tasks. These frameworks systematically measure model performance, lower assessment costs, and support medical AI optimization.

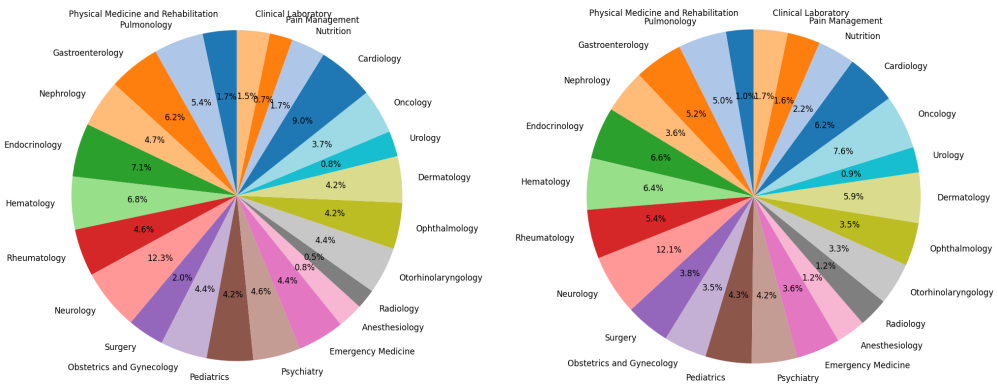
3 Medical Benchmark with Step-wise evaluation

3.1 Data curation

In this section, we detail our process for curating the data and the specifics of the dataset we collected. The dataset we curated consisted of two types: multiple-choice questions and open-ended questions. In the following, we develop a description of the process of curating each of these two types of data.

3.1.1 Data curation for multiple-choice question

The data collection for our MedReason-Dx benchmark is designed to create a challenging reasoning dataset that diverges from typical knowledge-based question-answer datasets. The objective is to curate a dataset where models are required to perform complex, multi-step reasoning to derive the



(a) (b)

Figure 1: t-SNE visualization of natural images and AI-generated images. (a): ● and ● are natural vs. AI-generated on the original model; (b): ● and ● on the pruned model.

correct answers, reflecting the intricate processes involved in real-world clinical diagnostics. To achieve this, we employed a rigorous data selection process that involved filtering questions from well-established medical datasets, which encompass real-world clinical cases across various medical disciplines. Each problem is associated with a detailed series of reasoning steps that mirror the diagnostic workflow, ensuring that the reasoning process is both comprehensive and contextually relevant. We define a set of 24 medical domains, derived from common hospital departments, including: "Cardiology", "Pulmonology", "Gastroenterology", "Nephrology", "Endocrinology", "Hematology", "Rheumatology", "Neurology", "Surgery", "Obstetrics and Gynecology", "Pediatrics", "Psychiatry", "Emergency Medicine", "Anesthesiology", "Radiology", "Otorhinolaryngology", "Ophthalmology", "Dermatology", "Urology", "Oncology", "Physical Medicine and Rehabilitation", "Nutrition", "Pain Management" and "Clinical Laboratory". The selection of questions prioritizes diversity in the types of clinical challenges and the reasoning methods required for problem-solving. This diversity encompasses a wide array of diagnostic tasks that span both common and rare clinical conditions. Questions are specifically chosen for their requirement of complex multi-step reasoning processes, including, but not limited to, physiological mechanism analysis, differential diagnosis, hypothesis testing, exclusionary reasoning, and the integration of cross-disciplinary knowledge. By focusing on the reasoning complexity and diversity, the dataset reflects the multifaceted nature of clinical decision-making and the diverse set of cognitive strategies employed by healthcare professionals in practice. The aim is to ensure that the dataset not only captures the breadth of medical knowledge but also challenges models to engage in higher-order reasoning reflective of real-world medical diagnostic scenarios

After finalizing the selection of challenging questions, we create the step-by-step answers and extract key points with the help of medical experts from diverse specialties. This approach allows for a comprehensive evaluation of the model's reasoning ability, focusing not only on the correctness of the final answer but also on the clarity and logic of the reasoning process itself. The step-by-step answers break down the reasoning into clear, logical steps, each representing a critical part of the decision-making process. The key points highlight the most important information, such as clinical findings or diagnostic considerations, necessary for reaching the correct diagnosis. These key points help ensure that the model's response covers all relevant aspects required for accurate medical decision-making. The goal of generating these steps and key points is to assess the reasoning process in detail, ensuring that the model's explanation is complete and logically sound. An example piece of data created is shown in Figure 2.

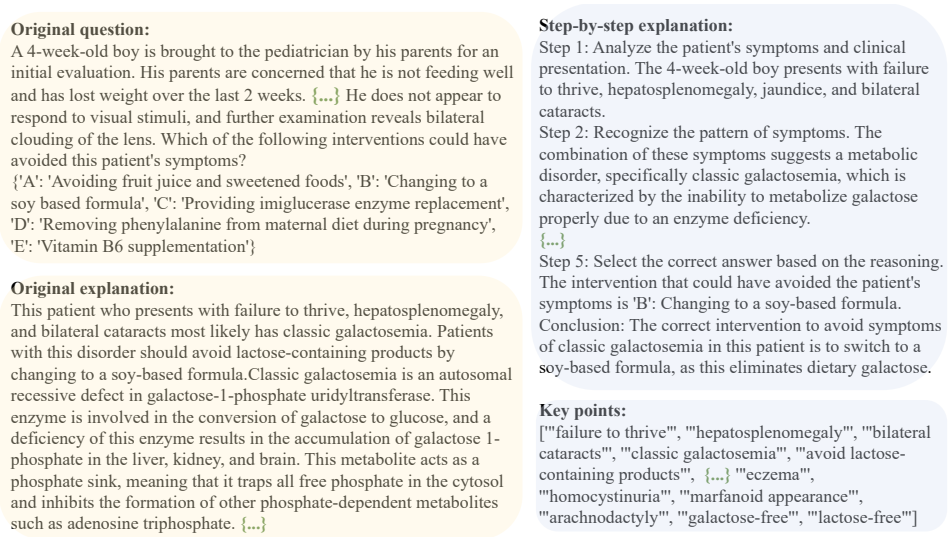


Figure 2: As the model unlearns, the model's ability to represent natural and generated images decays at different rates. Dots of different colors represent features of different classes of images.

3.1.2 Data curation for open-ended question

Multiple-choice questions often simplify the difficulty of the questions and fail to accurately reflect real-world scenarios, as human doctors can't make diagnoses based on predefined options. Consequently, we further construct open-ended reasoning questions.

"Please revise the following multiple-choice question to an open-ended question:"

"You are a professional doctor, please remove the information about the wrong options from the diagnostic results based on the questions and diagnostic results given and keep only the correct results."

When rewriting the problem, we only change the last question in the problem to ensure minimal changes to the original problem. Again, we invite human experts to monitor this change to ensure that it is done correctly. Upon obtaining the open-ended questions and answers, we reformulate the answers in a STEP-BY-STEP format, similar to the approach used for multiple-choice questions. Additionally, we extract the key points from the answers to facilitate subsequent assessments.

3.2 Evaluation

As previously mentioned, in addition to emphasizing the accuracy of the final answer, we also place significant attention on the comprehensiveness of the reasoning process and the thoroughness with which the relevant keywords are captured during this process. In the following, we provide a detailed description of the evaluation metrics employed to assess these three critical aspects of our analysis.

Correctness of the final answer. Firstly, in accordance with traditional evaluation methodologies, we assess the accuracy of the final answers provided by the model. For multiple-choice questions, we directly compare the option selected by the model with the correct one. For open-ended questions, we instruct the model to provide the answer in a specified format: "Please give your answer in the following format: "Therefore, the answer is \box{your answer}." Then we use LLM to determine whether the given answer is equivalent to the correct answer and calculate accuracy.

Completeness and necessity of reasoning steps. In addition to evaluating the correctness of the final answers, we also assess the completeness and necessity of the model's reasoning process. While current evaluations of large medical models typically emphasize the correctness of the results, it is crucial to recognize that, in high-stakes domains such as medicine, the transparency and rationality of the model's reasoning process are equally important. This is because physicians rely on the reasoning behind the model's conclusions to assess the reliability of the final results. Therefore, we also conduct a thorough evaluation of the model's reasoning process to ensure its robustness and interpretability.

Completeness of keywords. In addition to assessing the model's reasoning process, we also evaluate the completeness of the keywords involved in its reasoning.

4 Experiments

In this section, we will show the detailed experimental setup, analyses, and results of different LLMs in our benchmark. The dataset is still being refined, and more results will be released soon.

4.1 Experimental Setup

Evaluation Models. To provide a comprehensive benchmark, We conduct evaluations on 11 advanced LLMs, comprising 7 general LLMs and 4 medical LLMs, including DeepSeeK R1 (Guo et al., 2025), DeepSeek V3 (Liu et al., 2024), GPT-4o (OpenAI, 2024a), o1-mini (OpenAI, 2024b) and Baichuan4-Turbo (Yang et al., 2023). It is important to note that our experiments are still ongoing and the results are being continuously refined, with additional evaluation data expected to be updated in the near future.

Implementation Details. Following previous work (Jiang et al., 2025), we leverage two types of prompts to guide the model to give answers: chain-of-thought prompts (Wei et al., 2022) and direct prompts. Chain-of-thought prompts have the following format: *'Give your answer in the following form with clear logic: Step1: Step2:.... . Therefore, the answer is \box{}.'* and direct prompts have the following format: *'Please answer the following question and end your answer in this format: Therefore, the answer is \box{}.'*. When calling LLMs, the temperature is set to 0.7, Top-P is set to 0.9, and max tokens is set to 1000.

4.2 Benchmarking Medical LLMs

We benchmarked several advanced language models on the MedReason-Dx dataset under two prompting settings: Chain-of-Thought (CoT) and Direct Answering, across both multiple-choice and open-ended formats. Overall, DeepSeek-R1 achieved the highest performance in the multiple-choice setting, with CoT prompting slightly outperforming direct answering (65.03% vs. 64.36%). However, in the open-ended setting, its performance reversed, with direct prompting yielding higher accuracy (42.39%) than CoT (40.14%). DeepSeek-V3 showed a similar trend with modest gains from direct answering in open-ended questions (37.02% vs. 33.56%). Interestingly, GPT-4o exhibited the largest gap in favor of direct prompting for open-ended questions (47.70% vs. 37.72%), while maintaining comparable results in multiple-choice settings. o1-mini demonstrated relatively balanced performance across settings, with a slight edge for direct prompting in both question types. In contrast, Baichuan4-Turbo underperformed across all configurations, with particularly low scores on open-ended questions, indicating a significant gap in step-by-step reasoning capabilities compared to stronger models. These results suggest that while CoT prompting can provide marginal gains in structured formats, direct answering may be more effective in complex open-ended clinical scenarios, particularly for stronger LLMs.

References

Cai, Y., Wang, L., Wang, Y., de Melo, G., Zhang, Y., Wang, Y., and He, L. (2024). Medbench: A large-scale chinese benchmark for evaluating medical large language models. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 17709–17717.

Chen, J., Wang, X., Ji, K., Gao, A., Jiang, F., Chen, S., Zhang, H., Song, D., Xie, W., Kong, C., et al. (2023). Huatuogpt-ii, one-stage training for medical adaption of llms. arXiv preprint arXiv:2311.09774.

Chiang, C.-H. and Lee, H.-y. (2023). Can large language models be an alternative to human evaluations? arXiv preprint arXiv:2305.01937.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH), 3(1):1–23.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948.

Han, T., Adams, L. C., Papaioannou, J.-M., Grundmann, P., Oberhauser, T., Löser, A., Truhn, D., and Bressem, K. K. (2023). Medalpaca–an open-source collection of medical conversational ai models and training data. arXiv preprint arXiv:2304.08247.

Jiang, D., Zhang, R., Guo, Z., Li, Y., Qi, Y., Chen, X., Wang, L., Jin, J., Guo, C., Yan, S., et al. (2025). Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. arXiv preprint arXiv:2502.09621.

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., and Szolovits, P. (2021). What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Applied Sciences, 11(14):6421.

Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., and Lu, X. (2019). Pubmedqa: A dataset for biomedical research question answering. arXiv preprint arXiv:1909.06146.

Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Mark, R. G., and Horng, S. (2019). Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data, 6(1):317.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. Scientific data, 3(1):1–9.

Karimi, S., Metke-Jimenez, A., Kemp, M., and Wang, C. (2015). Cadec: A corpus of adverse drug event annotations. Journal of biomedical informatics, 55:73–81.

Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). Genia corpus—a semantically annotated corpus for bio-textmining. Bioinformatics, 19(suppl_1):i180–i182.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4):1234–1240.

Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C.-H., Leaman, R., Davis, A. P., Mattingly, C. J., Wiegers, T. C., and Lu, Z. (2016). Biocreative v cdr task corpus: a resource for chemical disease relation extraction. Database, 2016.

Li, J., Wang, X., Wu, X., Zhang, Z., Xu, X., Fu, J., Tiwari, P., Wan, X., and Wang, B. (2023a). Huatuo-26m, a large-scale chinese medical qa dataset. arXiv preprint arXiv:2305.01526.

Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., and Zhang, Y. (2023b). Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. Cureus, 15(6).

Liao, Y., Meng, Y., Liu, H., Wang, Y., and Wang, Y. (2023). An automatic evaluation framework for multi-turn medical consultations capabilities of large language models. arXiv preprint arXiv:2309.02077.

Liévin, V., Hother, C. E., Motzfeldt, A. G., and Winther, O. (2024). Can large language models reason about medical questions? Patterns, 5(3).

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al. (2024). Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437.

Lu, Q., Dou, D., and Nguyen, T. (2022). Clinicalt5: A generative language model for clinical text. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 5436–5443.

Luo, L., Ning, J., Zhao, Y., Wang, Z., Ding, Z., Chen, P., Fu, W., Han, Q., Xu, G., Qiu, Y., et al. (2024). Taiyi: a bilingual fine-tuned large language model for diverse biomedical tasks. Journal of the American Medical Informatics Association, 31(9):1865–1874.

Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E. P., and Rajpurkar, P. (2023). Med-flamingo: a multimodal medical few-shot learner. In Machine Learning for Health (ML4H), pages 353–367. PMLR.

Nori, H., King, N., McKinney, S. M., Carignan, D., and Horvitz, E. (2023). Capabilities of gpt-4 on medical challenge problems. arXiv preprint arXiv:2303.13375.

OpenAI (2024a). Hello gpt-4o. https://openai.com/index/hello-gpt-4o/.

OpenAI (2024b). Learning to reason with llms. https://openai.com/index/learning-to-reason-with-llms/.

Peng, C., Yang, X., Chen, A., Smith, K. E., PourNejatian, N., Costa, A. B., Martin, C., Flores, M. G., Zhang, Y., Magoc, T., et al. (2023). A study of generative large language model for medical research and healthcare. NPJ digital medicine, 6(1):210.

Romanov, A. and Shivade, C. (2018). Lessons from natural language inference in the clinical domain. arXiv preprint arXiv:1808.06752.

Saab, K., Tu, T., Weng, W.-H., Tanno, R., Stutz, D., Wulczyn, E., Zhang, F., Strother, T., Park, C., Vedadi, E., et al. (2024). Capabilities of gemini models in medicine. arXiv preprint arXiv:2404.18416.

Shi, X., Xu, J., Ding, J., Pang, J., Liu, S., Luo, S., Peng, X., Lu, L., Yang, H., Hu, M., et al. (2023). Llm-mini-cex: Automatic evaluation of large language model for diagnostic conversation. arXiv preprint arXiv:2308.07635.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. (2023). Large language models encode clinical knowledge. Nature, 620(7972):172–180.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S. R., Cole-Lewis, H., et al. (2025). Toward expert-level medical question answering with large language models. Nature Medicine, pages 1–8.

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., and Stojnic, R. (2022). Galactica: A large language model for science. arXiv preprint arXiv:2211.09085.

Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., Eide, D., Funk, K., Katsis, Y., Kinney, R., et al. (2020). Cord-19: The covid-19 open research dataset. ArXiv, pages arXiv–2004.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems 35.

Wu, C., Zhang, X., Zhang, Y., Wang, Y., and Xie, W. (2023). Pmc-llama: Further finetuning llama on medical papers. arXiv preprint arXiv:2304.14454, 2(5):6.

Wu, J., Deng, W., Li, X., Liu, S., Mi, T., Peng, Y., Xu, Z., Liu, Y., Cho, H., Choi, C.-I., et al. (2025). Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. arXiv preprint arXiv:2504.00993.

Xia, F., Li, B., Weng, Y., He, S., Liu, K., Sun, B., Li, S., and Zhao, J. (2022). Lingyi: medical conversational question answering system based on multi-modal knowledge graphs. arXiv preprint arXiv:2204.09220.

Xu, J., Lu, L., Yang, S., Liang, B., Peng, X., Pang, J., Ding, J., Shi, X., Yang, L., Song, H., et al. (2023). Medgpteval: A dataset and benchmark to evaluate responses of large language models in medicine. arXiv preprint arXiv:2305.07340.

Yang, A., Xiao, B., Wang, B., Zhang, B., Bian, C., Yin, C., Lv, C., Pan, D., Wang, D., Yan, D., et al. (2023). Baichuan 2: Open large-scale language models. arXiv preprint arXiv:2309.10305.

Yang, S., Zhao, H., Zhu, S., Zhou, G., Xu, H., Jia, Y., and Zan, H. (2024). Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In Proceedings of the AAAI conference on artificial intelligence, volume 38, pages 19368–19376.

Yang, X., Chen, A., PourNejatian, N., Shin, H. C., Smith, K. E., Parisien, C., Compas, C., Martin, C., Flores, M. G., Zhang, Y., et al. (2022). Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. arXiv preprint arXiv:2203.03540.

Zhang, H., Chen, J., Jiang, F., Yu, F., Chen, Z., Li, J., Chen, G., Wu, X., Zhang, Z., Xiao, Q., et al. (2023). Huatuogpt, towards taming language model to be a doctor. arXiv preprint arXiv:2305.15075.