MEDAGENTSARENA: Benchmarking Thinking Models and Agent Frameworks for Complex Medical Reasoning

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have shown impressive performance on existing medical reasoning benchmarks. This high performance makes it increasingly difficult to meaningfully evaluate and differentiate advanced methods. We present MEDAGENTSARENA, a carefully curated benchmark that focuses on challenging medical questions where current models still struggle. Drawing from seven established medical datasets, our benchmark addresses three key limitations in existing evaluations: (1) the prevalence of straightforward questions where even base models achieve high performance, (2) inconsistent sampling and evaluation protocols across studies, and (3) lack of systematic analysis of the interplay between performance, cost, and inference time. Through experiments with various base models and reasoning methods, we demonstrate that the latest thinking models, DEEPSEEK R1 and OPENAI O3, exhibit exceptional performance in complex medical reasoning tasks. Additionally, advanced search-based agent methods also perform effectively in handling intricate medical queries. Our benchmark and evaluation framework are publicly available at https://anonymous.4open.science/r/MedAgentArena.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in medical natural language processing tasks, from answering clinical questions to assisting in diagnostic processes (Singhal et al., 2025; Jin et al., 2022; Chen et al., 2023; Zhou et al., 2023; Gao et al., 2024). However, as shown in Figure 1, even OPENAI O3, GPT-4O and CLAUDE 3.5 SONNET struggle with complex medical scenarios that require deep domain expertise and multi-step reasoning (Xu et al., 2024; Fan et al., 2025; Shi et al., 2024).

To enhance LLMs' medical reasoning capabilities, researchers have proposed various approaches. As summarized in Table 1, these methods range

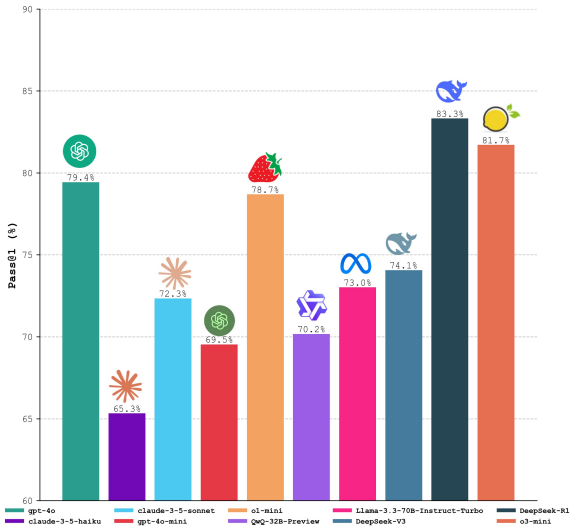

Figure 1: **Performance analysis of large language models on medical tasks.** Overall **Pass@1** accuracy comparison across models in zero-shot setting. The score is an average of results on all test sets of seven datasets (MedQA, PubMedQA, MedMCQA, MedBullets, MMLU, MMLU-Pro, and AfriMedQA).

from general-purpose techniques like CHAIN-OF-THOUGHT (COT) and SELF-CONSISTENCY (SC) (Wei et al., 2022; Wang et al., 2022) to domain-specific frameworks such as MEDPROMPT (Chen et al., 2024b). While these traditional approaches provide modest improvements, recent evidence suggests that agent-based methods, or "agent workflows," demonstrate superior performance. Methods like MEDAGENTS (Tang et al., 2023) and MDAGENTS (Kim et al., 2024) leverage multi-agent collaboration frameworks to achieve more robust medical reasoning. However, with the advent of advanced thinking models like OPENAI O3-MINI and DEEPSEEK R1, as well as the development of search-based agent frameworks, it remains an open question how these models perform in medical reasoning tasks.

Several critical challenges impede the evaluation of those thinking models and more effective agent frameworks for medical reasoning. **(a)** As shown in Appendix A Table 4, while existing medical reasoning datasets are extensive, many are derived from

Method	Description
CHAIN-OF-THOUGHT (Wei et al., 2022)	Elicits reasoning in large language models
SELF-CONSISTENCY (Wang et al., 2022)	Improves chain of thought reasoning in language models
MEDPROMPT (Chen et al., 2024b)	Multi-round prompting with ensemble voting for medical question answering
MULTI-PERSONA (Wang et al., 2023)	Task-solving agent through multi-persona self-collaboration
SELF-REFINE (Madaan et al., 2024)	Iterative refinement with self-feedback
MEDAGENTS (Tang et al., 2023)	Collaborative multi-agent framework for zero-shot medical decision making
MDAGENTS (Kim et al., 2024)	Dynamic multi-agent collaboration framework for medical reasoning
AFLOW (Zhang et al., 2024)	Automating agentic workflow generation
SPO (Xiang et al., 2025)	Self-supervised prompt optimization

Table 1: **Overview of Methods.** Survey of methods used for medical reasoning and question answering. General-purpose methods, domain-specific methods, and search-based methods are shown.

US medical licensing examinations and contain a substantial proportion of straightforward questions. On these easier questions, even base LLMs achieve high performance (see Table 3, "FULL" columns), making it difficult to assess the improvements brought by advanced agent frameworks meaningfully. **(b)** Furthermore, since most existing datasets were designed for evaluating smaller models in the pre-LLM era, they often contain thousands of questions, leading current agent-based studies to arbitrarily subsample around 300 questions for evaluation (Tang et al., 2023). This inconsistent sampling across different works, coupled with varying evaluation protocols, makes fair comparison challenging. **(c)** Additionally, there exists a complex interplay between performance, computational costs, and inference time that current benchmarks fail to systematically capture.

This landscape motivates our work on MEDAGENTSARENA, a benchmark for evaluating LLMs' medical reasoning capabilities. Drawing from seven established medical datasets (MedQA, PubMedQA, MedMCQA, MedBullets, MMLU, MMLU-Pro, and AfriMedQA), we curate a challenging subset of questions that specifically test advanced reasoning capabilities. As shown in Figure 2, we select questions that current models find particularly challenging, allowing for a more nuanced evaluation of model performance. Additionally, recognizing the importance of data quality in model evaluation, we analyze potential data leakage, as presented in Appendix C Figure 5.

2 MedAgentsArena

MEDAGENTSARENA is a carefully curated benchmark specifically designed to evaluate complex medical reasoning tasks. Drawing from seven established medical datasets (MedQA (Jin et al., 2021), PubMedQA (Jin et al., 2019), MedMCQA (Pal et al., 2022), MedBullets (Chen et al., 2024a), MMLU (Hendrycks et al., 2020), MMLU-Pro (Wang et al., 2024), and AfriMedQA (Olatunji et al., 2024)), we systematically construct a challenging subset that focuses on more complex reasoning scenarios. As shown in Table 2, these source datasets vary significantly in size (from 174 to 2,816 questions), average token length (18.7 to 316.1), and number of options (3 to 10), providing diverse evaluation contexts. Our hard set selection process follows three key criteria:

(1) Model Performance Distribution As visualized in Figure 2, we analyze the proportion of models that correctly answer each question (k/N ratio). Questions where less than 50% of models provide correct answers (left of the dashed line in Figure 2) are categorized as hard candidates. This ensures our benchmark focuses on truly challenging questions that current models struggle with.

(2) Question Diversity We maintain the relative proportion of questions from each source dataset to ensure broad coverage of medical knowledge domains. Specifically, from each dataset, we select: MedQA: 94 questions (particularly from USMLE-style complex scenarios); PubMedQA: 78 questions (focusing on biomedical research comprehension); MedMCQA: 156 questions (from specialized medical entrance exams); MedBullets: 82 questions (emphasizing clinical diagnosis); MMLU: 86 questions (covering broad medical concepts); MMLU-Pro: 64 questions (targeting advanced medical knowledge); AfriMedQA: 34 questions (incorporating global healthcare contexts).

(3) Reasoning Depth We prioritize questions that require multi-step reasoning (selected and verified by four M.D. students), as evidenced by the performance gap between base models and agent-based approaches. As shown in Table 3, while models achieve high accuracy on the FULL set (e.g., GPT-4o: 87.8% on MedQA), their performance drops significantly on our HARD subset (e.g., GPT-4o: 35.0% on MedQA-Hard), confirming the increased difficulty.

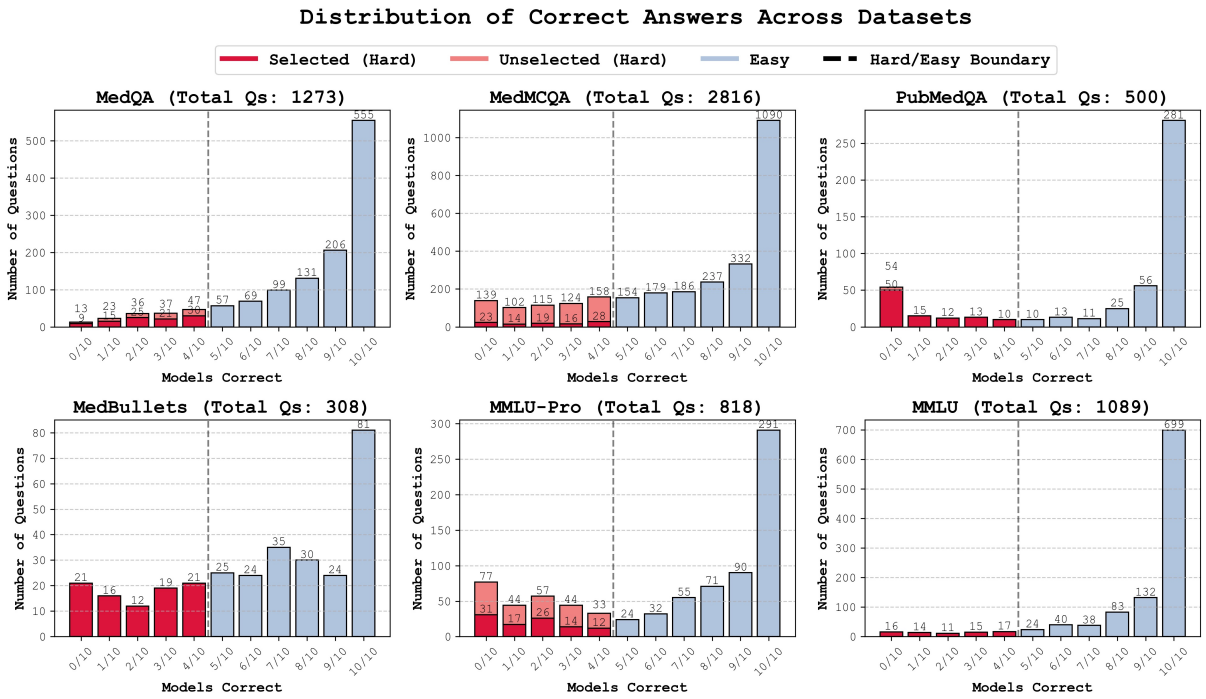

Figure 2: **Distribution of model performance across six medical datasets (MedQA, MedMCQA, PubMedQA, MedBullets, MMLU-Pro, and MMLU).** Each subplot shows the number of questions answered correctly by different proportions of models (x-axis: k/N, where k is the number of correct models and N is the total number of models). Questions are categorized as either hard (left of the dashed line, $< 50\%$ of models correct) or easy (right of the dashed line, $\geq 50\%$ of models correct), with selected questions highlighted in darker shades. The total question count for each dataset is indicated in the subplot titles. AfriMedQA is in Appendix D.

In summary, firstly we evaluate each candidate question across multiple model architectures (as shown in Table 3) to ensure the difficulty is architecture-independent. Secondly, using MELD (Memorization affects Levenshtein Detector), we analyze potential data leakage. As shown in Figure 5, we maintain low similarity scores (20-40%) between model outputs and question text, suggesting questions test genuine reasoning rather than memorization. Finally, the final question set is reviewed by four medical professionals (M.D. students) to verify clinical relevance and reasoning complexity. The resulting MEDAGENTSARENA benchmark contains 594 questions with an average token length of 134.8.

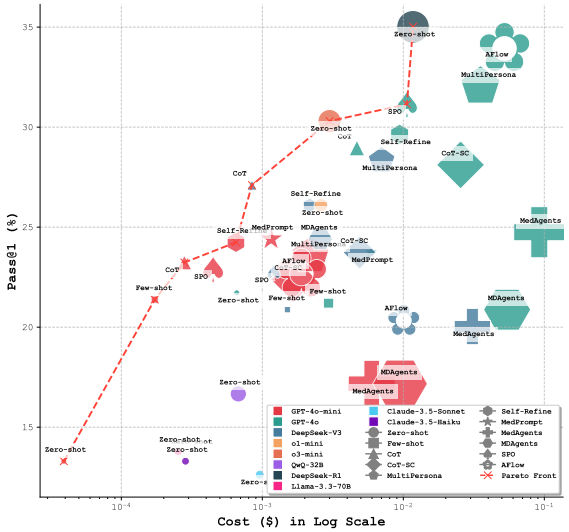

Figure 3: **Performance analysis of agents and models on MEDAGENTSARENA.** Cost-performance trade-off analysis showing **Pass@1** accuracy versus cost per sample (in log scale), with marker sizes indicating inference time. Different markers represent various prompting methods , while colors distinguish different models. The Pareto frontier (red dashed line) indicates optimal cost-performance trade-offs.

3 Experiments

3.1 Experimental Setup

We conduct experiments to evaluate both base models and reasoning methods. For base models, we consider both closed-source and open-source variants. The closed-source models include GPT-4O, GPT-4O-MINI, CLAUDE-3.5-SONNET, CLAUDE-3.5-HAIKU, O1-MINI, and O3-MINI, while the open-source models comprise DEEPSEEK-V3, DEEPSEEK-R1, LLAMA-3.3-70B, and QWQ-32B. In terms of reasoning methods, we evaluate 11 different approaches spanning five categories: (1) baseline methods (ZERO-SHOT, FEW-SHOT, CHAIN-OF-THOUGHT, & SELF-CONSISTENCY), (2) advanced general prompting methods (MULTI-PERSONA, SELF-

Method	MedQA			PubMedQA			MedMCQA			MedBullets			MMLU			MMLU-Pro			AfriMedQA		
Base Model	4o-m	4o	DS	4o-m	4o	DS	4o-m	4o	DS	4o-m	4o	DS	4o-m	4o	DS	4o-m	4o	DS	4o-m	4o	DS
Zero-shot	19.0	35.0	15.0	9.0	8.0	13.0	16.0	27.0	22.0	9.0	18.0	15.7	13.7	28.8	12.3	12.0	15.0	12.0	15.6	21.9	6.2
Few-shot	28.0	27.0	21.0	19.0	22.0	18.0	25.0	28.0	25.0	22.5	21.3	18.0	27.4	17.8	19.2	9.0	6.0	26.0	18.8	**34.4**	12.5
CoT	23.0	39.0	30.0	13.0	13.0	15.0	24.0	31.0	31.0	18.0	21.3	19.1	26.0	30.1	32.9	38.0	**42.0**	36.0	15.6	18.8	25.0
CoT-SC	20.0	39.0	25.0	13.0	7.0	13.0	26.0	31.0	24.0	15.7	22.5	16.9	26.0	28.8	31.5	36.0	**42.0**	32.0	15.6	21.9	28.1
MultiPersona	30.0	45.0	32.0	16.0	13.0	24.0	22.0	**34.0**	31.0	18.0	29.5	19.1	21.9	34.2	27.4	35.0	40.0	33.3	15.6	21.9	**34.4**
Self-Refine	24.0	42.0	30.0	13.0	9.0	17.0	27.0	30.0	26.0	19.1	29.2	23.6	26.0	34.2	24.7	39.0	37.0	36.0	15.6	21.9	21.9
MedPrompt	31.0	35.0	15.0	15.0	12.0	13.0	27.0	26.0	21.0	15.7	22.5	10.1	24.7	26.0	15.1	34.0	23.0	12.0	18.8	18.8	18.8
MedAgents	24.0	43.0	32.0	12.0	15.0	13.0	22.0	30.0	26.0	15.7	27.0	19.1	24.7	28.8	23.3	8.0	8.0	7.0	12.5	18.8	21.9
MDAgents	22.0	36.0	44.0	23.0	11.0	15.0	16.0	22.0	27.0	14.6	21.3	23.6	17.8	24.7	23.3	9.0	8.0	11.0	18.8	31.2	31.2
SPO	19.0	31.0	22.0	25.0	**31.0**	18.0	20.0	30.0	28.0	22.5	29.2	15.7	19.2	32.9	24.7	32.0	36.0	31.0	18.8	21.9	12.5
AFlow	30.0	**48.0**	28.0	15.0	18.0	18.0	25.0	31.0	22.0	15.7	**34.8**	16.9	24.7	**38.4**	16.4	29.0	37.0	21.0	12.5	28.1	15.6

Table 2: **Performance heatmap by reasoning methods and data sets.** All the tasks are evaluated on the Hard set. For each dataset, three base models are used in order: GPT-4o-mini, GPT-4o, and DeepSeek-V3. Accuracy is in %. **The best values** and the second-best values are highlighted.

Model	MedQA		PubMedQA		MedMCQA		MedBullets		MMLU		MMLU-Pro		AfriMedQA	
	Full	Hard	Full	Hard	Full	Hard	Full	Hard	Full	Hard	Full	Hard	Full	Hard
GPT-4o-mini	73.4	19.0	76.2	9.0	66.0	16.0	53.6	9.0	84.3	13.7	57.5	12.0	75.9	15.6
GPT-4o	87.8	35.0	79.2	8.0	76.6	27.0	70.5	18.0	91.3	28.8	69.1	15.0	**81.6**	21.9
DeepSeek-V3	79.3	15.0	73.6	13.0	74.3	22.0	61.0	15.7	89.7	12.3	64.7	12.0	75.9	6.2
o1-mini	89.9	41.0	77.4	**15.0**	73.2	25.0	73.1	39.3	90.7	34.2	67.8	9.0	78.7	15.6
o3-mini	**92.7**	**52.0**	**79.6**	13.0	77.1	25.0	**82.1**	**50.6**	93.4	34.2	70.0	14.0	77.0	18.8
QwQ-32B	71.3	18.0	77.2	10.0	67.6	18.0	49.7	6.7	86.9	13.7	68.9	32.0	69.5	15.6
DeepSeek-R1	92.0	47.0	76.2	12.0	**81.9**	**34.0**	79.2	43.8	**95.0**	**41.1**	**79.6**	**37.0**	79.3	**28.1**
Llama-3.3-70B	76.8	12.0	77.8	13.0	71.4	18.0	61.7	19.1	85.2	11.0	61.7	10.0	76.4	12.5
Claude-3.5-Sonnet	77.7	17.0	76.4	9.0	68.8	11.0	56.5	11.2	86.9	16.4	64.2	13.0	75.9	9.4
Claude-3.5-Haiku	63.4	13.0	73.8	12.0	62.9	23.0	49.4	7.9	79.7	11.0	57.5	11.0	70.7	15.6

Table 3: **Performance heatmap by base models and data sets.** For each task, accuracy values are in percentages, with separate columns for Full and Hard. **The best values** and the second-best values are highlighted.

Refine), (3) medical-specific prompting methods (MedPrompt), (4) medical-multi-agent-based methods (MedAgents & MDAgents), and (5) search-based multi-agent methods (SPO, & AFlow). All experiments are conducted using identical prompts and evaluation protocols across models for a fair comparison. [1] [2]

3.2 Performance Analysis

Our cost-performance analysis in Figure 3 reveals several important patterns in medical reasoning capabilities. First, while model size generally correlates positively with performance, it also leads to significant increases in computational cost, as indicated by the marker sizes representing inference time. Despite their specialized design for medical scenarios, agent-based methods like MedAgents and MDAgents did not perform better than the latest thinking models, with even heavier computational overhead.

The comparison of reasoning methods in Table 2 demonstrates the superiority of agent-based approaches to challenging medical questions. Chain-of-thought with Self-Consistency (CoT-SC) shows moderate improvements over bare CoT, with average gains of 2-3% across datasets. However, domain-specific methods like MedPrompt show mixed results. They perform well on specific datasets but lack consistency across different medical tasks. The base-model-level analysis in Table 3 reveals a striking performance gap between Full and Hard sets. For instance, GPT-4o's performance drops from 87.8% to 35.0% on MedQA, validating the effectiveness of our Hard set selection criteria in identifying truly challenging medical questions. Interestingly, o3-mini achieves the best overall performance on Hard sets. Additionally, open-source models, particularly DeepSeek-R1, demonstrate competitive performance compared to their closed-source counterparts.

4 Conclusion

Through MedAgentsArena, we demonstrate that thinking models, such as DeepSeek-R1 and OpenAI o3-mini, consistently outperform others in complex medical reasoning tasks. Additionally, advanced search-based agent methods also perform effectively.

[1] Methods requiring multi-round reasoning (e.g., MedAgents) use two rounds of inference per query.

[2] Multi-agent approaches (e.g., Multi-Persona) utilize three distinct agent roles per inference.

Limitations

While MEDAGENTSARENA provides a rigorous benchmark for evaluating medical reasoning capabilities, several important limitations remain:

First, our benchmark primarily focuses on medical question-answering tasks based on educational resources, which may not fully reflect the complexity and nuance of real-world clinical scenarios. A more comprehensive evaluation would require incorporating real-world clinical cases, physician-patient dialogues, and diagnostic decision-making processes.

Second, despite our efforts to validate the benchmark using MD students, we lack systematic verification of model outputs by practicing clinicians. This raises concerns about model-generated reasoning paths' reliability and alignment with established medical knowledge. Future work should establish a more rigorous verification framework involving domain experts to assess answer correctness and the validity of reasoning steps and potential hallucinations.

Third, while we demonstrate strong performance from agent-based approaches, we have only evaluated a limited set of reasoning architectures. The field would benefit from exploring a broader range of methods, including hybrid approaches combining symbolic and neural reasoning and techniques explicitly designed for verifiable medical reasoning.

Fourth, our evaluation metrics focus primarily on accuracy and cost but do not assess other crucial aspects like verifiability, robustness to distribution shifts, and calibration of model confidence. Future work should develop more comprehensive evaluation frameworks that can measure these critical dimensions of medical AI systems.

Finally, while our work demonstrates the effectiveness of multi-agent and ensemble approaches in medical reasoning, we have only scratched the surface of potential ensemble strategies. Sophisticated ensemble methods like step-wise verification, task-wise verification, and dynamic agent collaboration could potentially yield even better performance. For instance, verifying intermediate reasoning steps through model consensus, utilizing heterogeneous model combinations, or implementing adaptive voting strategies based on agent expertise remain unexplored. Future research could investigate:

(1)More sophisticated voting and aggregation strategies beyond simple majority voting. (2) Adaptive ensemble methods that dynamically adjust agent weights based on task characteristics. (3)Hierarchical ensemble approaches that combine both step-wise and task-wise verification. (4)Methods for increasing response diversity through systematic prompt variation and temperature tuning. (5)Integration of expert knowledge to guide ensemble selection and verification.

While our current approach shows promising results, we lack a thorough theoretical understanding of why specific ensemble configurations outperform others in medical reasoning tasks. A more systematic study of ensemble properties - such as diversity, correlation, and calibration - could guide the development of more effective medical reasoning systems.

References

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2024a. Benchmarking large language models on answering and explaining challenging medical questions. *arXiv preprint arXiv:2402.18060*.

Qingyu Chen, Jingcheng Du, Yan Hu, Vipina Kuttichi Keloth, Xueqing Peng, Kalpana Raja, Rui Zhang, Zhiyong Lu, and Hua Xu. 2023. Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. *arXiv e-prints*, pages arXiv–2305.

Xuhang Chen, Shenghong Luo, Chi-Man Pun, and Shuqiang Wang. 2024b. MedPrompt: Cross-modal prompting for multi-task medical image translation. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 61–75. Springer.

Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Wang Siyuan, Zhongyu Wei, and Fei Huang. 2025. Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10183–10213.

Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. 2024. Empowering biomedical discovery with ai agents. *Cell*, 187(22):6125–6151.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

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

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. 2022. Biomedical question answering: a survey of approaches and challenges. *ACM Computing Surveys (CSUR)*, 55(2):1–36.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. MDAgents: An adaptive collaboration of llms in medical decision making. *arXiv preprint arXiv:2404.15155*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-Refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Tobi Olatunji, Charles Nimo, Abraham Owodunni, Tassallah Abdullahi, Emmanuel Ayodele, Mardhiyah Sanni, Chinemelu Aka, Folafunmi Omofoye, Foutse Yuehgoh, Timothy Faniran, et al. 2024. AfriMed-QA: A pan-african, multi-specialty, medical question-answering benchmark dataset. *arXiv preprint arXiv:2411.15640*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Haotian Sun, Hang Wu, Carl Yang, and May D Wang. 2024. Medadapter: Efficient test-time adaptation of large language models towards medical reasoning. *arXiv preprint arXiv:2405.03000*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2023. MedAgents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2023. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *arXiv preprint arXiv:2307.05300*.

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

Jinyu Xiang, Jiayi Zhang, Zhaoyang Yu, Fengwei Teng, Jinhao Tu, Xinbing Liang, Sirui Hong, Chenglin Wu, and Yuyu Luo. 2025. Self-supervised prompt optimization. *arXiv preprint arXiv:2502.06855*.

Shaochen Xu, Yifan Zhou, Zhengliang Liu, Zihao Wu, Tianyang Zhong, Huaqin Zhao, Yiwei Li, Hanqi Jiang, Yi Pan, Junhao Chen, et al. 2024. Towards next-generation medical agent: How o1 is reshaping decision-making in medical scenarios. *arXiv preprint arXiv:2411.14461*.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, et al. 2024. Aflow: Automating agentic workflow generation. *arXiv preprint arXiv:2410.10762*.

Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, et al. 2023. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*.

A Dataset Information

Our MEDAGENTSARENA benchmark draws from seven established medical datasets spanning diverse question types and formats. Table 4 provides a comprehensive overview of each source dataset:

MedQA consists of 1,273 questions from medical licensing examinations with an average token length of 167.1 and 4 multiple-choice options. These questions are designed to test clinical knowledge and diagnostic reasoning that would be expected of medical professionals.

PubMedQA contains 500 questions focused on biomedical research comprehension, with longer contexts (average 316.1 tokens) and 3 answer options. Questions are derived from PubMed abstracts, emphasizing scientific and research-oriented medical knowledge.

MedMCQA comprises 2,816 questions from AIIMS & NEET PG entrance exams, featuring concise questions (18.7 tokens on average) with four options. This represents a challenging set of specialized medical entrance exam questions.

MedBullets includes 308 questions from an online medical study platform, with detailed clinical scenarios (213.1 tokens on average) and five options per question. These questions emphasize practical clinical knowledge and decision-making.

Afrimed-QA contributes 174 questions that capture diverse medical scenarios from African healthcare contexts, with 30.0 tokens on average and five options. This dataset helps ensure geographical and cultural diversity in medical reasoning evaluation.

MMLU contains 1,089 questions covering both medical and broader academic domains, with moderate length (55.9 tokens) and four options. While not exclusively medical, it provides important cross-domain medical knowledge assessment.

MMLU-Pro features 818 questions with variable option count (3-10) and moderate length (57.4 tokens), offering more complex and challenging scenarios across medical and related domains.

Our MEDAGENTSARENA benchmark synthesizes these sources into a carefully curated set of 594 challenging questions, maintaining an average length of 134.8 tokens and preserving the variable option format (3-10 choices) of the source datasets. The questions are specifically selected to test advanced reasoning capabilities, with an emphasis on complex scenarios that require deep medical knowledge and multi-step reasoning.

B Cost-performance Analysis

B.1 Experimental Setup

To analyze cost-performance trade-offs, we followed a standardized evaluation protocol. For API-based models (GPT-4 and CLAUDE 3.5), we calculated costs using their published pricing rates based on total token usage (input + output). Based on their platform rates, we estimated costs for open-source models run on Together AI [3]. The total cost of experimentation was $226.17. We measured inference time as wall-clock time per sample, including prompt construction and model inference, with agent-based methods including their complete interaction cycles.

Our evaluation spans both closed-source models (GPT-4O, GPT-4O-MINI, CLAUDE-3.5-SONNET, CLAUDE-3.5-HAIKU, O1-MINI, O3-MINI) and open-source alternatives (DEEPSEEK-V3, DEEPSEEK-R1, LLAMA-3.3-70B, QWQ-32B). We further categorize models as thinking or non-thinking based on their prompt format and reasoning approach, as indicated by different markers in Figure 4.

B.2 Analysis of Cost-Performance Patterns

Our analysis in Figure 3 reveals complex trade-offs between model performance, computational cost, and inference time across medical domains. While larger models generally achieve better performance, they come with significant cost overhead, as evidenced by GPT-4O achieving 35.0% accuracy on MedQA-Hard but at approximately 10 times the cost per sample compared to GPT-4O-MINI (19.0% accuracy). Interestingly, open-source models demonstrate competitive performance at lower costs, with DEEPSEEK-R1 achieving comparable or better performance than some closed-source alternatives while maintaining lower operational costs, particularly evident in MedMCQA and MMLU-Pro tasks.

The effectiveness of structured reasoning becomes apparent through the consistent outperformance of thinking models (marked with circles) compared to their non-thinking counterparts (marked with squares) at similar cost points. This advantage is particularly pronounced in complex tasks like MedBullets, where thinking models show 5-10% accuracy improvements. The Pareto frontier (red dashed line) in each subplot reveals op-

[3] https://www.together.ai/

Benchmark	Size	Avg Tokens	Options	Description
MedQA (Jin et al., 2021)	1273	167.1	4	Multiple choice questions from medical licensing exams
PubMedQA (Jin et al., 2019)	500	316.1	3	Questions based on PubMed abstracts
MedMCQA (Pal et al., 2022)	2816	18.7	4	Questions from AIIMS & NEET PG entrance exams
MedBullets (Chen et al., 2024a)	308	213.1	5	Questions from Medbullets online medical study platform
Afrimed-QA (Olatunji et al., 2024)	174	30.0	5	Diverse medical questions from African healthcare contexts
MMLU (Hendrycks et al., 2020)	1089	55.9	4	Questions covering medical, and other academic domains
MMLU-Pro (Wang et al., 2024)	818	57.4	3-10	Questions covering medical, and other academic domains
MEDAGENTSARENA	594	134.8	3-10	HARD subset across all datasets

Table 4: **Overview of Medical Question-Answering Datasets.** Survey of knowledge-based QA datasets curated from medical literature, professional journals, and educational resources. Traditional benchmarks, recently emerging benchmarks, and general purpose benchmarks are shown with corresponding colors.

timal cost-performance trade-offs, with O3-MINI frequently appearing as a Pareto-optimal solution. Agent-based methods also achieve efficient positioning on this frontier despite moderate computational costs, suggesting their effectiveness in balancing performance and resource utilization.

Different medical domains exhibit distinct cost-performance relationships. MedQA demonstrates steep performance gains with increased computational investment, while PubMedQA shows more gradual improvements, suggesting diminishing returns from larger models. Notably, AfriMedQA exhibits relatively consistent performance across cost ranges, indicating that domain knowledge may be more crucial than model scale in specialized medical contexts. These variations highlight the importance of considering domain-specific characteristics when selecting models and methods for medical reasoning tasks.

C Data Leakage Analysis

To ensure the reliability of our benchmark evaluation, we employ memorization effects Levenshtein Detector (MELD), initially introduced by Nori et al. (2023) to analyze potential data leakage across various datasets and language models. MELD quantifies memorization by measuring the similarity between model-generated text and the original question text, where higher similarity scores indicate a greater likelihood of memorization rather than genuine reasoning. MELD splits each question in the dataset into two halves, providing the first half is provided as context to the model, which is then tasked with generating the second half. The generated output is then compared to the unseen portion using the Levenshtein distance ratio, which measures the proportion of matching characters.

The analysis spans GPT-3.5/4, CLAUDE-3.5, and various open-source LLMs, providing comprehensive coverage of different model architectures and training approaches. MELD exhibits high precision but unknown recall, meaning that while a detected match strongly indicates memorization, the absence of a match does not guarantee that the data was not seen during training. For instance, Nori et al. (2023) report that GPT-4 reproduces SQuAD 2.0 questions with 99% character overlap in 17% of cases, highlighting significant memorization.

Interestingly, we observe consistent patterns across model families. Closed-source models (GPT-4, CLAUDE-3.5) generally show lower similarity scores compared to open-source alternatives, particularly on PubMedQA and MedMCQA datasets. This suggests potentially more robust generalization in their medical reasoning capabilities. The variation in similarity scores is notably smaller for AfriMedQA, likely due to its unique focus on African healthcare contexts that may be less represented in model training data.

The boxplot distributions also reveal outliers with significantly higher similarity scores (>60%), particularly in MedBullets and MMLU-Pro datasets. These cases often correspond to common medical terminology and standard descriptive phrases rather than wholesale memorization of question-answer pairs. To validate this interpretation, we manually reviewed a sample of high-similarity cases and found that the shared text primarily consisted of standard medical terminology and widely used clinical descriptions.

These findings support the robustness of our benchmark, indicating that model performance differences primarily reflect varying capabilities in medical reasoning rather than direct memorization of training examples. The consistently low similarity scores across diverse datasets and model architectures suggest that our HARD set selection process successfully identifies questions that require genuine reasoning rather than simple pattern

matching or memorization.

D AfriMedQA Dataset Infomation

Due to the page limit, we visualize the full distribution (Figure 6) of model performance across seven medical datasets in the appendix.

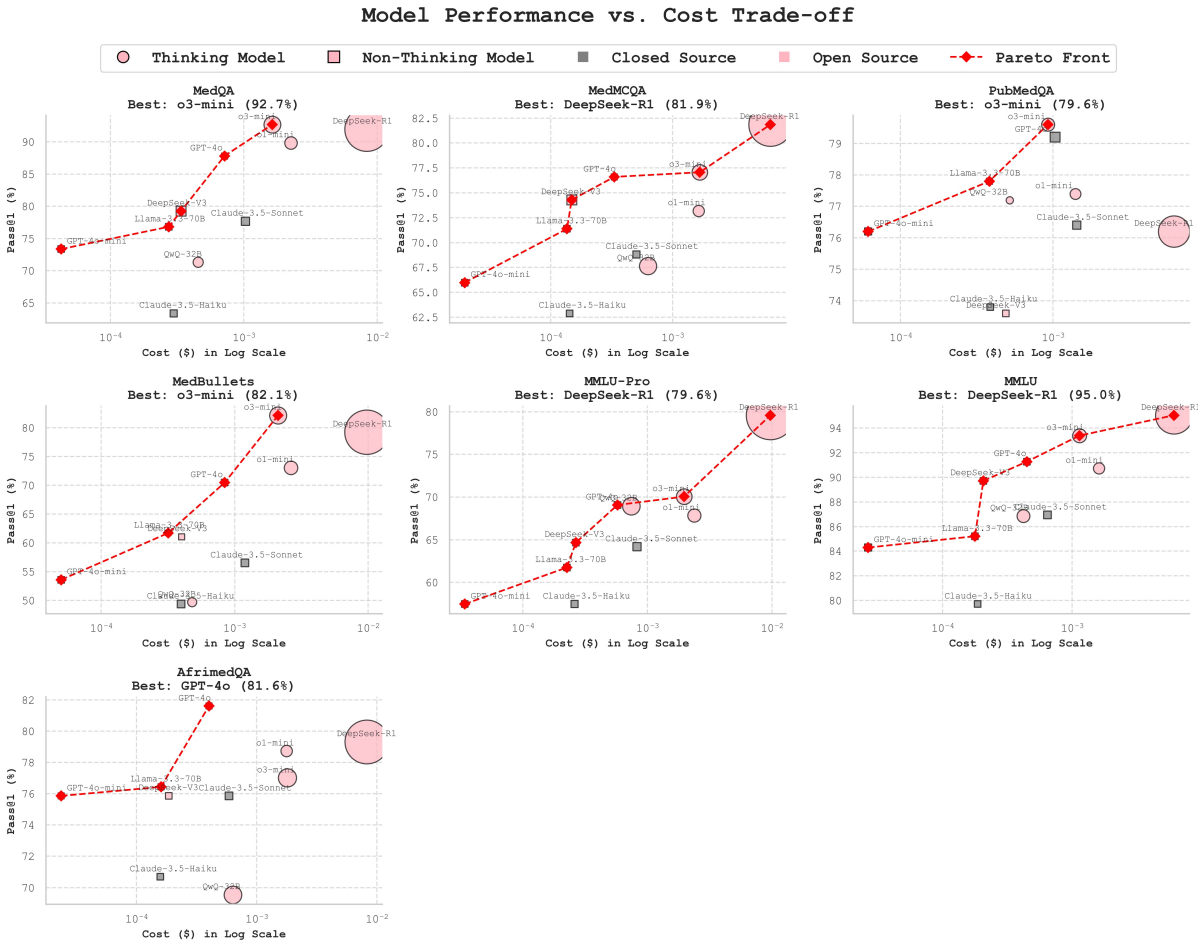

Figure 4: **Cost-performance analysis across seven medical datasets, comparing open and closed-source language models.** Each subplot shows Pass@1 accuracy (%) versus cost per sample (USD, log scale). Marker shapes distinguish thinking models from non-thinking models, while colors indicate open-source (blue) versus closed-source (red) models. Marker sizes represent inference time, and the red dashed line shows the Pareto frontier of optimal cost-performance trade-offs. Best-performing models and their accuracies are noted in subplot titles. The analysis includes models like GPT-4O, CLAUDE-3.5, DEEPSEEK, and LLAMA-3.3-70B, revealing distinct performance patterns across different medical domains.

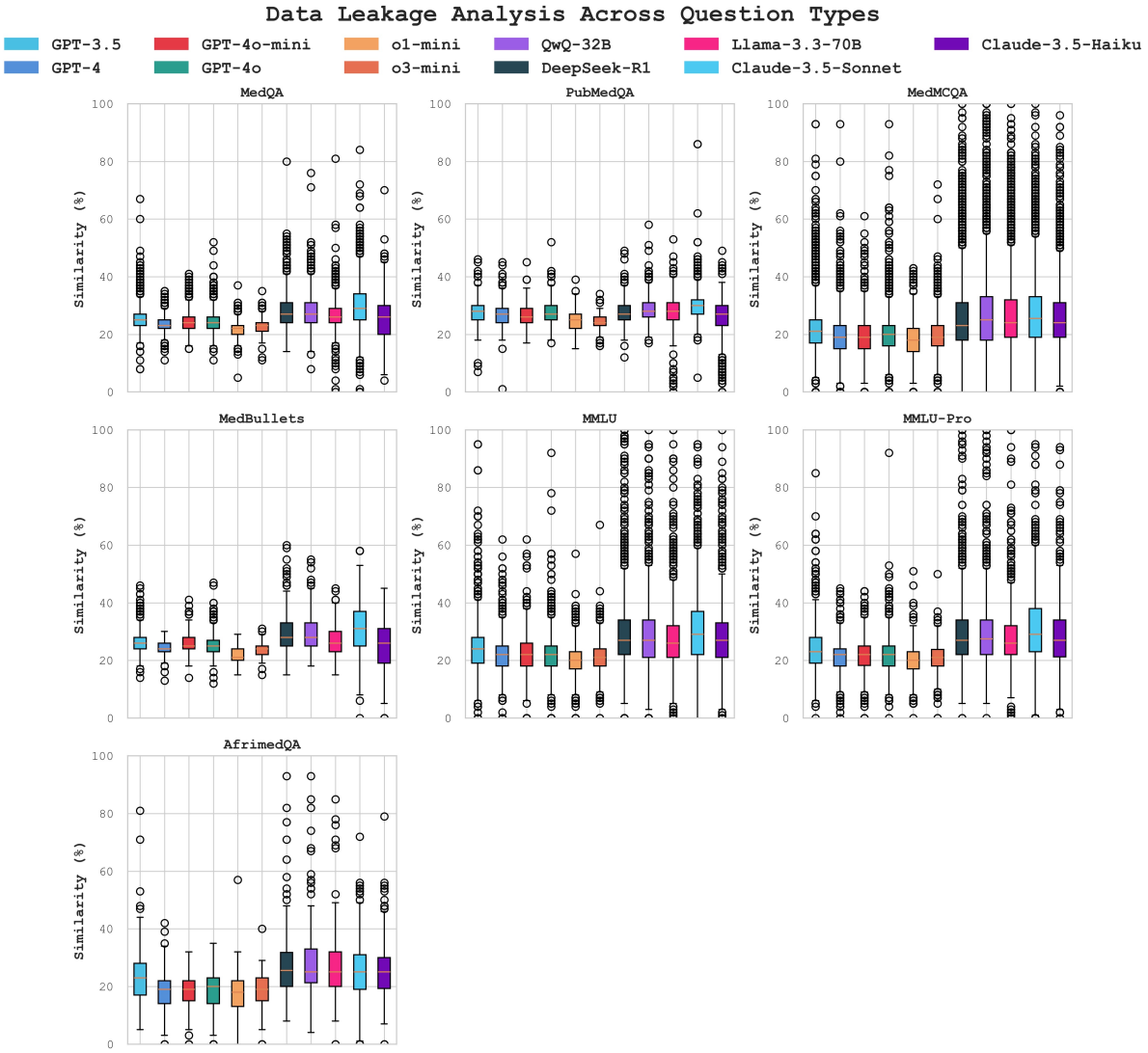

Figure 5: **Data leakage analysis across different medical question-answering datasets using our memorization affects Levenshtein detector (MELD).** Each boxplot shows the similarity percentage between the model-generated text and the original question text, with higher values indicating potential memorization. Models tested include GPT-3.5/4, CLAUDE-3.5, and various open-source LLMs. The analysis spans seven datasets: MEDQA, PUBMEDQA, MEDMCQA, MEDBULLETS, MMLU, MMLU-PRO, and AFRIMEDQA. Lower similarity scores (20-40%) suggest minimal data leakage, while higher scores may indicate potential memorization of training data.

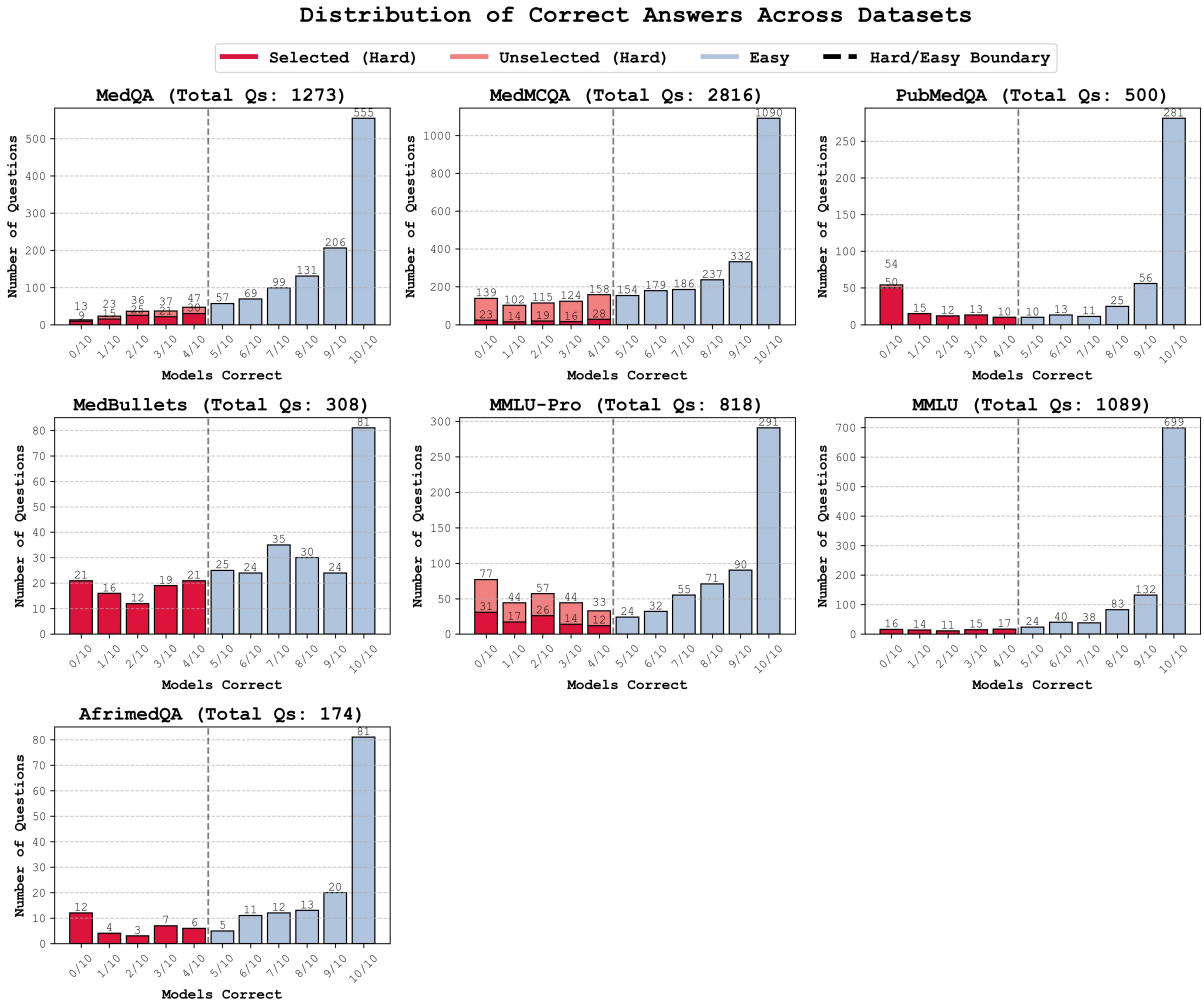

Figure 6: **Distribution of model performance across seven medical datasets (MedQA, MedMCQA, PubMedQA, MedBullets, MMLU-Pro, MMLU, and AfriMedQA).** Each subplot shows the number of questions answered correctly by different proportions of models (x-axis: k/N, where k is the number of correct models and N is the total number of models). Questions are categorized as either hard (left of the dashed line, $< 50\%$ of models correct) or easy (right of the dashed line, $\geq 50\%$ of models correct), with selected questions highlighted in darker shades. The total question count for each dataset is indicated in the subplot titles.