Questioning Our Questions: How Well Do Medical QA Benchmarks Evaluate Clinical Capabilities of Language Models?

Siun Kim
Seoul National University Hospital
Seoul, South Korea
shiuhn95@snu.ac.kr

Hyung-Jin Yoon
Seoul National University Hospital
Seoul, South Korea
hjyoon@snu.ac.kr

Abstract

Recent advances in large language models (LLMs) have led to impressive performance on medical question-answering (QA) benchmarks. However, the extent to which these benchmarks reflect real-world clinical capabilities remains uncertain. To address this gap, we systematically analyzed the correlation between LLM performance on major medical QA benchmarks (e.g., MedQA, MedMCQA, PubMedQA, and MMLU medicine subjects) and clinical performance in real-world settings. Our dataset included 702 clinical evaluations of 85 LLMs from 168 studies. Benchmark scores demonsrated a moderate correlation with clinical performance (Spearman's $\rho = 0.59$), albeit substantially lower than inter-benchmark correlations. Among them, MedQA was the most predictive but failed to capture essential competencies such as patient communication, longitudinal care, and clinical information extraction. Using Bayesian hierarchical modeling, we estimated representative clinical performance and identified GPT-4 and GPT-4o as consistently top-performing models, often matching or exceeding human physicians. Despite longstanding concerns about the clinical validity of medical QA benchmarks, this study offers the first quantitative analysis of their alignment with real-world clinical performance.[1]

1 Introduction

The rapid advancement of large language models (LLMs), accelerated by the release of ChatGPT, has continued into 2025. Open-source models such as Llama 3.3, Phi-4, and DeepSeek-R1 are rapidly narrowing the performance gap with proprietary models (Grattafiori et al., 2024; Abdin et al., 2024; Guo et al., 2025). This progress is especially consequential in healthcare, where strigent privacy and security requirments frequently necessitate on-premise deployment (Faray de Paiva et al., 2025; Gupta and Pande, 2025).

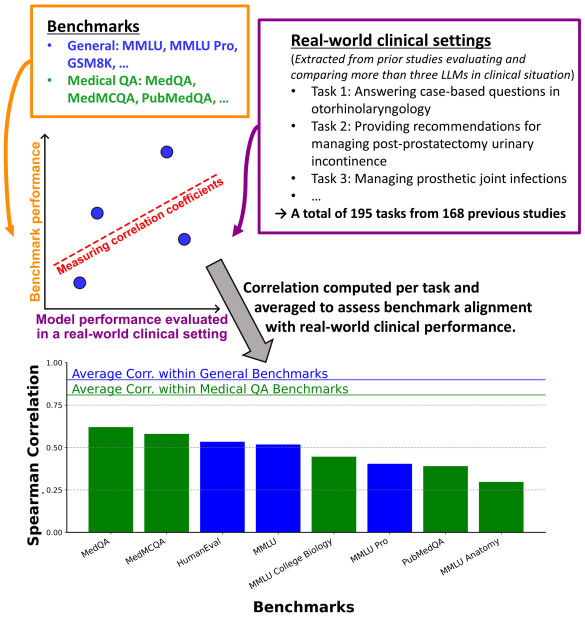

Figure 1: Overview of our study assessing the alignment between medical QA benchmarks and real-world clinical performance.

As LLMs attains expert-level performance on both general and medical QA benchmarks, the limitations of such benchmarks have become increasingly apparent. For instance, OpenAI's o1-preview achieved 96% on MedQA and 99% on MMLU Medical Genetics, outperforming human experts (Nori et al., 2024; Liévin et al., 2024). However, such benchmarks are thought to focus predominantly on static knowledge and structured reasoning, which may not fully reflect core competencies essential for clinical practice (Nori et al., 2023; Singhal et al., 2023), such as decision-making under uncertainty (Han et al., 2011), patient communication (Barry and Edgman-Levitan, 2012), and ethical reasoning (Kaldjian et al., 2005).

Although concerns over the limited clinical validity of existing medical QA benchmarks have

[1]The dataset and code are available at: https://github.com/SiunKim/questioning-medqa.

Proceedings of the 24th Workshop on Biomedical Language Processing (BioNLP 2025), pages 274–296

been raised, there remains a lack of systematic evidence. In this study, we address this gap through a comprehensive meta-analysis evaluating how effectively conventional medical QA benchmarks reflect the real-world clinical performance of LLMs (Figure 1).

Our key contributions are as follows:

- **Quantitative assessment of benchmark-clinical alignment**: We demonstrate a moderate correlation (Spearman's $\rho = 0.59$) between medical QA benchmarks and real-world clinical performance, highlighting significant limitations in the current evaluation practices.
- **Identification of clinical gaps in MedQA**: MedQA demonstrates strong alignment with core competencies such as treatment, clinical knowledge, and diagnosis. However, it fails to adequately assess essential aspects of real-world clinical practice, including patient communication, longitudinal care, and clinical information extraction.
- **Bayesian modeling of representative clinical performance**: Using hierarchical Bayesian models, we estimate the generalized clinical capabilities of LLMs, suggesting that models like GPT-4 and GPT-4o match or exceed human physician-level performance in real-world clinical settings.

2 Related Works

MedQA—based on the USMLE Step 1 and 2 exams—has emerged as a de facto benchmark in the medical domain, owing to its high-quality multiple-choice questions (MCQs) and comprehensive topical coverage (Jin et al., 2021). As a representative benchmark, improvements in MedQA performance have frequently been interpreted as a proxy for progress in medical LLMs (Singhal et al., 2025; Saab et al., 2024).

MedMCQA, derived from Indian medical entrance exams (AIIMS and NEET PG), complements MedQA by offering broader topical diversity and varied question types (Pal et al., 2022). In contrast, PubMedQA focused on biomedical literature comprehension by requiring models to infer answers from PubMed abstracts (Jin et al., 2019).

Despite their widespread use, these traditional medical benchmarks primarily assess factual recall and structured reasoning. They have been criticized for failing to evaluate essential aspects of practical clinical competence (Tang et al., 2023; Kim et al., 2025; Liu et al., 2024).

In response, recent datasets aim to capture the complexity of real-world clinical practice. Datasets like Medbullet (Chen et al., 2024), MedExQA (Kim et al., 2024), and MedXpertQA (Zuo et al., 2025) introduce open-ended questions, expert-written explanations, and multimodal data to facilitate more comprehensive evaluations. Furthermore, integrated evaluation frameworks like MedAgentBench (Tang et al., 2025) and MEDIC (Kanithi et al., 2024) encompass multiple clinical tasks and explicitly address ethical and safety concerns.

In parallel, agent-based evaluations have emerged to assess interactive and dynamic reasoning. For instance, MedQA-CS adopts OSCE-style clinical scenarios (Yao et al., 2024), while AgentClinic (Schmidgall et al., 2024) evaluates LLMs during simulated physician-patient dialogues.

Building on these developments, our study systematically examines the alignment between conventional medical QA benchmarks and real-world clinical evaluations. By identifying existing gaps, we aim to inform the design of future benchmarks that more accurately reflect practical clinical competencies.

3 Methods

To evaluate the extent to which existing medical QA benchmarks reflect real-world clinical performance, we analyzed 168 published studies that assessed at least three distinct language models in clinical settings. Benchmark scores on both medical QA and general-purpose benchmarks were collected and standardized to ensure comparability. To address the missing benchmark scores, multiple imputation was applied. Correlations between benchmark scores and clinical performance were calculated using rank-based methods weighted by sample size. Finally, we employed Bayesian hierarchical modeling to estimate each model's representative clinical capability.

3.1 Literature Review for Collecting Clinical Performance Data

We conducted a multi-stage literature review to identify studies evaluating LLM performance in real-world clinical settings (Figure 2). Using the Semantic Scholar API, we first retrieved articles published between January 1, 2023, and January 10, 2025, based on search queries designed to encom-

pass a wide range of clinical scenarios (Appendix A.1). Title-based filtering retained studies explicitly mentioning LLM-related terms, followed by DOI-based deduplication. Abstract and full-texts were retrieved via publisher and open-access APIs.

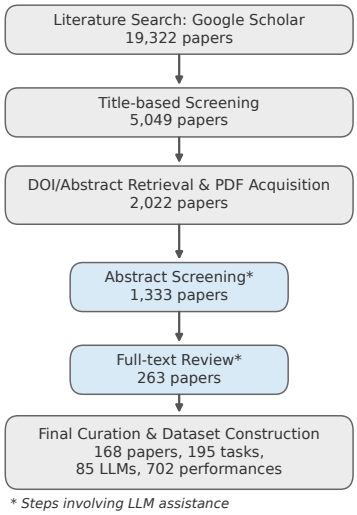

Figure 2: Flowchart of literature review for collecting LLM performances in real-world clinical settings.

Phi-4 model-assisted screening of abstracts and full-texts identified studies that reported performance for at least three distinct LLMs, enabling correlation analyses (AppendixA.2). Manual review was conducted to extract structured data, normalized model names (Appendix A.3) and classified evaluation settings (Appendix A.4).

3.2 LLM Performance Collection in Real-World Clinical Settings

To address overrepresentation issues caused by redundant evaluations of similar model abilities within a single study, we extracted one representative performance score for each task-model combination. Preference was given to the simplest inference setting (e.g., zero-shot without CoT). If multiple measures existed for the same therapeutic area and capabilities, we selected the most frequently used metric, or averaged scores, if no dominant measure was evident. Evaluations spanning multiple therapeutic areas or distinct capabilities were treated as a separate task.

Studies relying on readability metrics, inter-model correlation analyses, or with fewer than 20 evaluation samples were excluded. Encoder-based language models (e.g., BERT, RoBERTa) were also excluded because the study focused solely on autoregressive LLMs.

Performance scores were normalized to a 0–100 scale using min-max scaling. Metrics indicating better performance through lower values (e.g., proportion of biased answers) were inverted by subtracting from 100.

3.3 Benchmark Performance Collection

Benchmarks were divided into medical QA benchmarks and general benchmarks. Medical QA benchmarks included MedQA, MedMCQA, PubMedQA, and six MMLU medical subsets (Anatomy, Clinical Knowledge, College Biology, College Medicine, Medical Genetics, and Professional Medicine). General benchmarks consisted of MMLU, MMLU Pro, BBH, HumanEval, GSM8K, and MATH.

Performance data were extracted from published articles, technical reports, and model cards. Additional web searches supplemented version-specific scores for widely used proprietary models (e.g., GPT, Claude, and Gemini).

We standardized benchmark performances by focusing on zero-shot without CoT. If multiple results were available, averages were used. If zero-shot data were unavailable, performance was estimated through linear regression using reported results under different inference settings, by considering few-shot examples and CoT usage as covariates.

3.4 Benchmark Performance Imputation

While complete benchmark data are ideal for reliable correlation analyses, missing values were inevitable as performance scores were compiled through literature review rather than direct evaluation. To systematically address missing values, we employed Multiple Imputation by Chained Equations (MICE), which leverages observed interdependencies within available data to estimate absent benchmark performances.

Before imputation, we confirmed the Missing at Random (MAR) assumption, a necessary condition for reducing bias in estimation. Two imputation techniques were tested: Random Forest (RF-MICE) for capturing non-linear dependencies and Bayesian Ridge (BR-MICE) for small datasets with collinearity. Imputation was performed separately for each benchmark category.

We validated imputation accuracy through masking test, randomly removing and subsequently estimating 10% of the observed values. To incorporate uncertainty, multiple imputations were conducted, and within- and between-imputation variance were

estimated (Appendix A.5). Based on validation results, we selected a final version of the imputed dataset for downstream analysis.

3.5 Correlation Measurement

We evaluated correlations in two ways: benchmark-to-benchmark and benchmark-to-clinical performance.

Benchmark-to-benchmark correlation were calculated based on the performance scores of models that were evaluated on both. This analysis allowed us to identify redundant benchmarks, assess the quality of benchmark datasets, and set a correlation baseline for subsequent benchmark-to-clinical correlation analyses.

Benchmark-to-clinical correlations were computed at the evaluation task level, weighted logarithmically by evaluation sample size to reflect varying reliability across studies. Analyses utilized imputed benchmark scores primarily, with non-imputed data serving as sensitivity checks.

Although we measured rank-based (Spearman's rank correlation coefficient and Kendall's tau) and linear-based (Pearson's correlation coefficient and Lin's concordance correlation coefficient) metrics, primary analyses used Spearman's rank and Kendall's tau due to their suitability for handling diverse evaluation scoring scales without assuming linear relationships. Linear correlations were calculated but used only as reference points.

3.6 Bayesian Modeling

To estimate representative clinical performance for each language model independent of task-specific biases, we employed Bayesian hierarchical modeling. Given the limited number of model evaluated per task (average 3.6 models), individual task effects could not be directly estimated. Instead, task-related variations were approximated using metadata attributes including task type, data source, and evaluation methods. Therapeutic areas were excluded due to inconsistent categorization and unclear impact on performance (Appendix A.6.1). Furthermore, models for which performance data were available for fewer than three distinct tasks were excluded to enhance the reliability of model-specific performance estimates, which served as proxies for general clinical competence.

To further assess the robustness of the model-specific estimates, connectivity measures were calculated. Higher connectivity indicates stronger support from direct and indirect comparisons across models, thereby resulting in more stable and accurate performance estimates (Appendix A.6.2).

Table 1: Summary of clinical performance dataset and evaluation settings.

Category	Count (%)
Total samples	702 (100.0)
Task type	
Diagnosis	183 (26.1)
Clinical Knowledge	182 (25.9)
Overall Management	111 (15.8)
Answering to Patients	83 (11.8)
Information Extraction	61 (8.7)
Treatment	48 (6.8)
Other	34 (4.8)
Data source	
Clinical Vignettes	271 (38.6)
Quizzes	160 (22.8)
Board Examination	114 (16.2)
FAQs	74 (10.5)
Other	83 (11.8)
Therapeutic area	
General Medicine	154 (21.9)
Oncology	77 (11.0)
Ophthalmology	60 (8.5)
Orthopedics & Musculoskeletal	58 (8.3)
Emergency Medicine	53 (7.5)
Neuropsychiatric	53 (7.5)
Others	247 (35.2)
Evaluation type	
MCQs	463 (66.0)
Human Rating	239 (34.0)

4 Results and Discussion

4.1 Clinical Performance Dataset

Our dataset comprised 702 clinical performance evaluations from 168 studies covering 195 distinct clinical tasks. Evaluations involved 85 LLMs, predominantly from GPT (51.7%), LLaMA (10.3%), and Gemini (8.8%) families. Task types included diagnosis (26.1%), clinical knowledge assessment (25.9%), and overall patient management (15.8%). Data sources were primarily clinical vignettes (38.6%) and quizzes (22.8%), with evaluations conducted through MCQs (66.0%) and expert human ratings (34.0%) (Table 1).

4.2 Benchmark Performance Imputation

The benchmark dataset contained a notable proportion of missing values: 42.4% for medical

Table 2: Average correlation coefficients of medical QA benchmarks with other benchmarks. The highest score in each column is **bold**, and the second highest is underlined.

Medical QA Benchmarks	Spearman		Kendall	
	Medical QA	General	Medical QA	General
MedQA	0.809	**0.867**	0.664	**0.703**
MedMCQA	0.808	0.855	0.651	0.693
MMLU Medical Genetics	0.835	0.748	0.684	0.607
MMLU Clinical Knowledge	**0.851**	0.820	**0.714**	0.664
MMLU College Medicine	0.822	0.784	0.683	0.618
MMLU Professional Medicine	0.849	0.789	0.705	0.632
MMLU College Biology	0.819	0.666	0.672	0.522
MMLU Anatomy	0.703	0.558	0.571	0.449
PubMedQA	0.484	0.441	0.333	0.318
Average	**0.787**	**0.725**	**0.675**	**0.576**

QA benchmarks (9 benchmarks, 138 models, 715 scores) and 40.6% for general benchmarks (6 benchmarks, 126 models, 449 scores).

Imputation accuracy, assessed through masking tests, indicated RF-MICE outperformed BR-MICE. Specifically, RF-MICE achieved lower mean absolute error (MAE=2.04) and higher R^2 (0.98) on medical QA benchmarks (Table 9). Variance analysis of multiple imputations further supported RF-MICE due to lower total variance and improved stability (Table 10). Consequently, RF-MICE was utilized to generate the final imputed dataset.

4.3 Benchmark-to-Benchmark Correlation

Medical QA benchmarks showed strong internal correlations overall, with MMLU Clinical Knowledge and MMLU Professional Medicine exhibiting particularly high correlations with other medical QA benchmarks (Table 2). This is likely due to their broad content coverage, encompassing topics found in other MMLU medical subjects, thereby forming a high-correlation block (Figure 6).

In contrast, PubMedQA and MMLU Anatomy showed weaker correlations with other medical QA benchmarks. PubMedQA's lower correlations may stem from its distinct task formulation, which is more aligned with biomedical summarization rather than clinical reasoning (Jin et al., 2019). Similarly, MMLU Anatomy's lower correlations likely reflect its narrower content scope compared to other benchmarks.

MedQA and MedMCQA demonstrated the highest correlations with general benchmarks among medical QA benchmarks (Table 2, Figure 6). This suggests that these two datasets not only assess domain-specific knowledge but also required a broad set of reasoning skills, many of which overlapped with general benchmarks.

Within general benchmarks, BBH (Spearman's 0.891) and MMLU (0.853) exhibited the strongest correlations with medical QA benchmarks (Table 11). This result also indicates logical reasoning capabilities and broad domain knowledge are closely linked to solve medical problems. In contrast, mathematics-focused benchmarks (i.e., GSM8K and MATH) displayed weaker correlations, highlighting the distinct types of reasoning involved in medical contexts.

4.4 Benchmark-to-Clinical Performance Correlation

MedQA showed the strongest correlation with real-world clinical performance, outperforming general benchmarks in capturing actual clinical competency (Spearman's 0.588, Kendall's 0.520; Figure 3A). However, correlation strength was notably lower than inter-benchmark correlations (0.675–0.787; Table 2). These results suggest MedQA remains the most representative current benchmark for clinical tasks, although its ability to predict comprehensive clinical performance remains limited.

Further analysis across evaluation settings highlighted MedQA's strengths and limitations (Figure 4). MedQA performed well predicting clinical competency in tasks involving treatment, clinical knowledge, and diagnosis (Figure 4A). In contrast, it showed significantly weaker correlations in patient communication, overall patient management, and information extraction.

Similarly, while MedQA strongly correlated with performance derived from board examination-

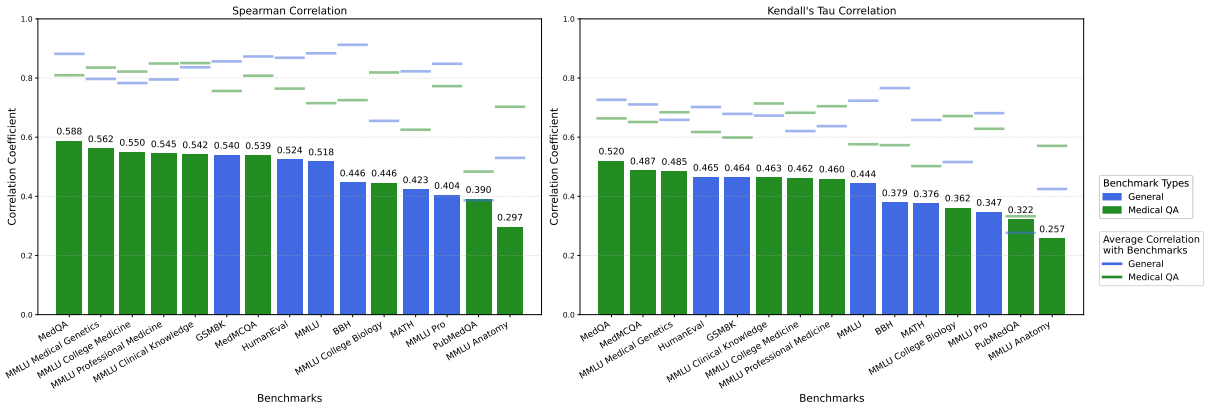

Figure 3: Comparison of correlation coefficients between benchmark and clinical performance.

style datasets, correlations with evaluations based on clinical vignette or FAQ, which closely resemble real-world clinical practice, were considerably lower (Figure 4B). These findings suggest that while MedQA reliably evaluates core medical knowledge and reasoning skills, it does not adequately reflect the broader competencies required in real-world clinical practice.

MMLU Medical Genetics, College Medicine, Professional Medicine, Clinical Knowledge, and MedMCQA displayed moderately high correlations with clinical performance, outperforming general benchmarks (Figure 3A). Conversely, PubMedQA and MMLU Anatomy consistently underperformed, indicating their limited suitability as representative clinical evaluation tools (Figures 9, 10).

4.5 Representative Clinical Performances Estimated through Bayesian Modeling

Representative clinical performances of 59 language models were robustly estimated using Bayesian hierarchical modeling across 717 performance samples. Model convergence was strong, indicated by effective sample sizes (ESS) above 300 and R-hat values below 1.02.

Among evaluated models, GPT-4 and GPT-4o consistently demonstrated the highest clinical performance, often exceeding the average performance of medical professionals (labeled as 'human - doctor') and substantially outperforming both smaller open-source models and other proprietary models (Figure 5). The strong and consistent performance of the GPT family is further supported by newly developed medical benchmark studies (Olatunji et al., 2024; Yao et al., 2024; Zuo et al., 2025), which similarly highlight their superior clinical reasoning capabilities.

Proprietary models (purple) generally outperformed open-source models (orange, Figure 5), suggesting that commercially optimized systems remain more reliable in clinical settings—though this conclusion may shift with the rapid progress of open-source LLMs in 2025.

Within the open-source category, Llama-3.1-8B-instruct was the only model to surpass the minimum threshold set for human-level performance (labeled as 'human - cut-off'). Notably, however, its lower connectivity implies that this performance estimate should be interpreted with caution due to high uncertainty.

Notably, language models fine-tuned for the medical domain (marked with a star, ★) did not show substantial improvements over general-purpose models like Llama, despite having comparable model sizes (Figure 5). This may be due to overfitting to the specific characteristics of their training datasets—typically composed of structured medical QA corpora or textbook-style materials—which could limit their generalizability in practical clinical contexts (Olatunji et al., 2024). These findings are consistent with previous results showing that biomedical models often underperform on newer, more complex benchmarks, and support concerns regarding their sensitivity to dataset-specific biases and limitations (Olatunji et al., 2024; Yao et al., 2024).

5 Conclusion

This study demonstrates that existing medical QA benchmarks posses only a moderate capacity to predict real-world clinical performance. Among them, MedQA showed the strongest correlation with clinical performance but was still insufficient for evaluating practical clinical competencies such as patient

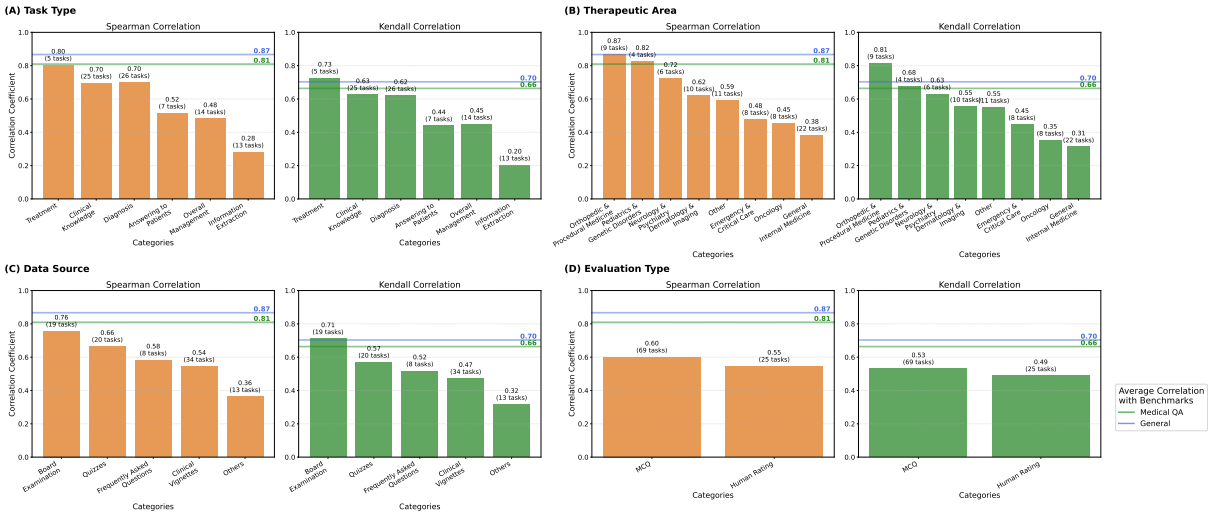

Figure 4: Comparison of correlations between MedQA performance and individual outcomes measured in real-world clinical settings across different task types, therapeutic areas, data sources, and evaluation methods.

interaction, longitudinal patient management, and clinical information extraction. Bayesian hierarchical modeling further revealed that proprietary models—particularly GPT-4 and GPT-4o—consistently outperformed open-source counterparts and, across many versions, matched or exceeded the performance of human experts in real-world clinical settings. Notably, despite longstanding concerns regarding the validity of medical QA benchmarks, this study provides the first systematic and quantitative evidence evaluating the alignment between medical QA benchmarks and actual clinical performance.

Limitations

This study has several limitations. First, our analysis is based on published studies, which inevitably lag behind ongoing LLM advancements due to publication delays. Consequently, it does not account for recent developments in LLMs, such as the emergence of reasoning-based LLMs (Guo et al., 2025).

Second, although several medical benchmarks have been introduced to better assess multifaceted capabilities (Kim et al., 2024; Yao et al., 2024; Zuo et al., 2025), we could not obtain sufficient model performance results on these datasets to conduct correlation analyses. To support future research, we make our clinical performance dataset available and encourage its use in validating how well these newly proposed medical benchmarks reflect the complexity of real-world medical tasks.

Lastly, despite employing statistical methods to address missing data and selection biases, our findings are inherently constrained by the incompleteness and potential biases of literature-derived data.

Acknowledgments

This work was supported by the Institute of Information Communications Technology Planning Evaluation (IITP)-Innovative Human Resource Development for Local Intellecutallization program grant funded by the Korea goverment (MSIT) (IITP-2024-RS-2024-00441407). This research was supported by a grant of the Korea Health Technology RD Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health Welfare, Republic of Korea (grant number: RS-2023-KH136520).

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Michael J Barry and Susan Edgman-Levitan. 2012. Shared decision making—the pinnacle of patient-centered care. *New England Journal of Medicine*, 366(9):780–781.

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

Lisle Faray de Paiva, Gijs Luijten, Behrus Puladi, and Jan Egger. 2025. How does deepseek-r1 perform on usmle? *medRxiv*, pages 2025–02.

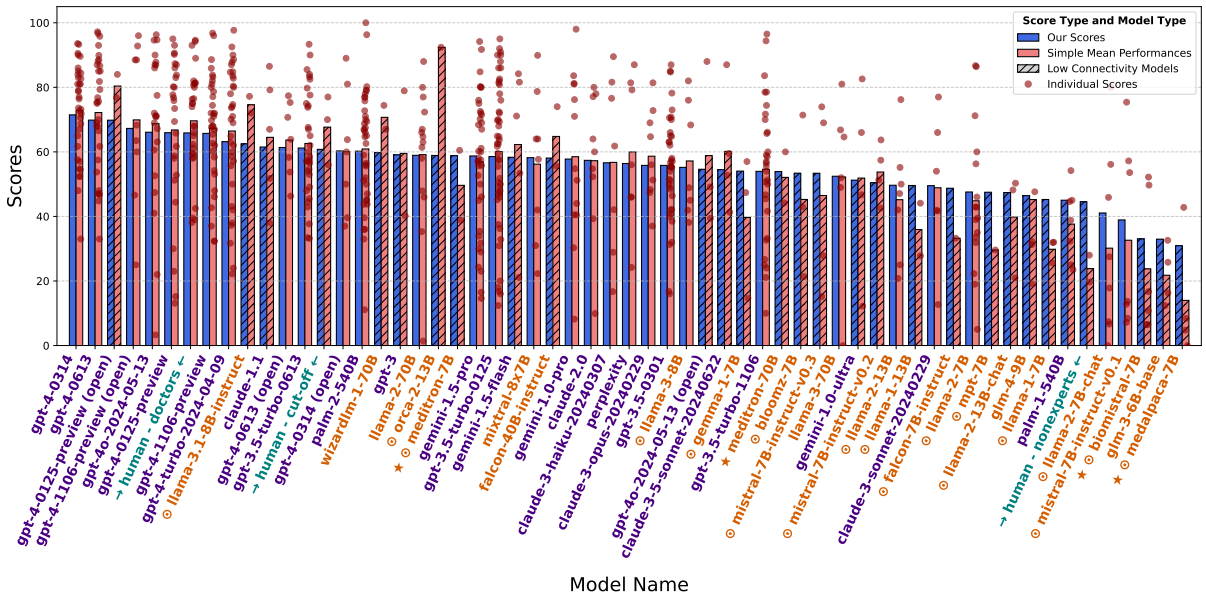

Figure 5: Representative clinical performance estimated via Bayesian modeling. Proprietary models appear in purple text while open-source models are shown in orange. Medical domain fine-tuned models are marked with a star (★) and small language models with 13B parameters or fewer display a circle prefix (⊙).

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Gaurav Kumar Gupta and Pranal Pande. 2025. Llms in disease diagnosis: A comparative study of deepseek-r1 and o3 mini across chronic health conditions. *arXiv preprint arXiv:2503.10486*.

Paul KJ Han, William MP Klein, and Neeraj K Arora. 2011. Varieties of uncertainty in health care: a conceptual taxonomy. *Medical Decision Making*, 31(6):828–838.

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*.

Lauris C Kaldjian, Robert F Weir, and Thomas P Duffy. 2005. A clinician's approach to clinical ethical reasoning. *Journal of general internal medicine*, 20:306–311.

Praveen K Kanithi, Clément Christophe, Marco AF Pimentel, Tathagata Raha, Nada Saadi, Hamza Javed, Svetlana Maslenkova, Nasir Hayat, Ronnie Rajan, and Shadab Khan. 2024. Medic: Towards a comprehensive framework for evaluating llms in clinical applications. *arXiv preprint arXiv:2409.07314*.

Jonathan Kim, Anna Podlasek, Kie Shidara, Feng Liu, Ahmed Alaa, and Danilo Bernardo. 2025. Limitations of large language models in clinical problem-solving arising from inflexible reasoning. *arXiv preprint arXiv:2502.04381*.

Yunsoo Kim, Jinge Wu, Yusuf Abdulle, and Honghan Wu. 2024. Medexqa: Medical question answering benchmark with multiple explanations. *arXiv preprint arXiv:2406.06331*.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3).

Fenglin Liu, Zheng Li, Hongjian Zhou, Qingyu Yin, Jingfeng Yang, Xianfeng Tang, Chen Luo, Ming Zeng, Haoming Jiang, Yifan Gao, et al. 2024. Large language models in the clinic: a comprehensive benchmark. *arXiv preprint arXiv:2405.00716*.

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*.

Harsha Nori, Naoto Usuyama, Nicholas King, Scott Mayer McKinney, Xavier Fernandes, Sheng Zhang, and Eric Horvitz. 2024. From medprompt to o1: Exploration of run-time strategies for medical challenge problems and beyond. *arXiv preprint arXiv:2411.03590*.

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.

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

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.

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

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, Daniel Shao, Jiwoong Sohn, Jiapeng Chen, Jiayi Zhang, Jinyu Xiang, Fang Wu, Yilun Zhao, Chenglin Wu, Wenqi Shi, et al. 2025. Medagentsbench: Benchmarking thinking models and agent frameworks for complex medical reasoning. *arXiv preprint arXiv:2503.07459*.

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*.

Zonghai Yao, Zihao Zhang, Chaolong Tang, Xingyu Bian, Youxia Zhao, Zhichao Yang, Junda Wang, Huixue Zhou, Won Seok Jang, Feiyun Ouyang, et al. 2024. Medqa-cs: Benchmarking large language models clinical skills using an ai-sce framework. *arXiv preprint arXiv:2410.01553*.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*.

A Supplementary Methods

A.1 Search Query for Literature Review

Our literature search followed a systematic approach to identify studies at the intersection of large language models (LLMs) and medical applications. The search queries were structured using three essential components: LLM-related terms, medical terms, and evaluation terms (Table 3).

Each query was formulated as:

[LLM Term] AND [Medical Term] AND [Evaluation Term]

where the medical terms were drawn from either MedQA-related categories (e.g., "medical question answering", "clinical reasoning") or clinical application categories (e.g., medical specialties, clinical documents, diseases, and procedures). The search was restricted to publications from 2022 to 2025 to ensure coverage of recent developments.

This combinatorial approach balanced coverage and precision, ensuring that retrieved papers addressed all three aspects of our research focus (Table 4).

A.2 Screening Process for Collecting Performance data of LLM in Real-World Clinical Settings

We conducted abstract screening and full-text review based on LLM to refine selection process and alleviated burden of manual curation. The LLM utilized for this process was Phi-4 (14.7B), Q4_K_M quantized, based on the Ollama framework (as of March 18, 2025). The model was deployed locally on a single RTX 4080 GPU.

A.3 Model Name Normalization

We normalized language model names by categorizing them into proprietary (Table 7) and open-source models (Table 8). For proprietary models, specific model names were often unspecified in papers, as they were accessed via APIs. In such cases, we assumed the most recent model available at the research time: three months before the received date for journal papers and six months before publication for conference papers. For open-source models, normalization was based on explicitly stated model names, versions, and parameter sizes in billions (Table 8). If these details were insufficient, we excluded the model from analysis.

Table 3: Search Query Components for LLM Applications in Medical Research.

Query Component	Terms
LLM Terms	"large language model", "language model", "GPT-4", "ChatGPT"
MedQA Terms	"medical question answering", "USMLE", "MedQA", "medical benchmark", "clinical reasoning"
Clinical Application Terms	*Medical Specialties (31 terms):* "internal medicine", "surgery", "pediatrics", "obstetrics", "gynecology", ... *Surgery Settings (10 terms):* "surgery", "pediatric surgery", "breast surgery", "colorectal surgery", "neurosurgery", ... *Clinical Settings (6 terms):* "emergency department", "icu", "operating room", "outpatient", "primary care", "trauma center" *Clinical Documents (11 terms):* "electronic health record", "clinical notes", "discharge summary", "medical history", "radiology report", ... *Common Diseases (34 terms):* "breast cancer", "lung cancer", "colorectal cancer", "prostate cancer", "leukemia", "lymphoma", ... *Clinical Procedures (6 terms):* "chemotherapy", "radiation therapy", "transplantation", "dialysis", "ventilation", "ecmo" *Age-Specific Care (7 terms):* "newborn care", "child development", "growth disorders", "birth defects", "falls prevention", "memory disorders", "polypharmacy management" *Special Populations (6 terms):* "maternal health", "prenatal care", "postpartum care", "women's health", "social determinants", "medical ethics"
Evaluation Terms	"evaluation", "accuracy", "benchmark", "validation", "application"

Table 4: Query Construction Pattern and Examples.

Query Pattern: [LLM Term] AND [Medical Term] AND [Evaluation Term]
Medical Term Selection: Either [MedQA Terms] OR [Clinical Application Terms] based on research focus
Example Queries: *With MedQA Terms:* 1. large language model" AND medical question answering" AND evaluation" 2. ChatGPT" AND clinical reasoning" AND benchmark" *With Clinical Application Terms:* 3. GPT-4" AND electronic health record" AND validation" 4. language model" AND internal medicine" AND application" 5. GPT-4" AND breast cancer" AND "accuracy"

A.4 Classification of Evaluation Settings

We categorized the evaluation settings for language models in clinical contexts based on four key criteria: (1) task type, (2) therapeutic area, (3) data source, and (4) evaluation method.

Task Type Task types represent the core capabilities being assessed, classified into the following six categories:

- **Clinical Knowledge:** General assessment of fundamental clinical knowledge within a given specialty, without a specific focus on diagnosis, treatment, or prevention.
- **Treatment:** Evaluation of the model's ability to recommend and assess treatment plans based on a given clinical scenario.
- **Diagnosis:** Determining the correct diagnosis based on the provided patient information.
- **Answering to Patients:** Providing responses to common patient inquiries or explaining clinical conditions in plain language understandable by non-experts.
- **Overall Management:** Beyond diagnosis and treatment, evaluating long-term patient management and decision-making.
- **Information Extraction:** Extracting specific clinical information from given texts.

Table 5: Abstract Screening Prompt Template.

Prompt for abstract screening

Please analyze the following research paper's title and abstract to extract information about LLM performance evaluation in clinical settings. Present your analysis in the following structured format, maintaining exact quotes where possible. Start your response with "ANALYSIS_START" **and end with** "ANALYSIS_END". **INPUT REQUIRED:**

- Title: [*paper title*]
- Abstract: [*paper abstract*]

TASK: Analyze the title and abstract to extract the following information:

1. `PAPER_TYPE`: Classify the paper as one of the following:

- **"Clinical LLM Performance Evaluation - Original"**: Paper that conducts new experiments to evaluate LLM performance in clinical tasks and reports original performance metrics/results.
- **"Clinical LLM Performance Review"**: Paper that summarizes or analyzes existing LLM clinical performance evaluations without conducting new experiments or reporting new performance data.
- **"Non-Clinical LLM Evaluation"**: Paper not related to clinical LLM performance evaluation.

2. `MODELS`: Extract all LLM models mentioned in the abstract.

Format: ["model1", "model2", ...]

Return empty list if no specific models are mentioned.

3. `MULTIPLE_MODELS_USAGE`: For papers classified as "Clinical LLM Performance Evaluation - Original" only.

Format: true/false/NA

- **true**: Paper clearly evaluates multiple LLMs.
- **false**: Paper clearly focuses on single LLM evaluation.
- **NA**: For non-original clinical LLM evaluation papers.

4. `HUMAN_GROUPS`: Extract all medical professional groups that underwent the same evaluation tasks as the models for direct performance comparison.

Format: ["group1", "group2", ...]

Return empty list if no human groups underwent direct performance comparison.

5. `EVALUATION_TASKS`: Extract all clinical evaluation tasks.

Format:

```
[{"task_name_extractive": "exact task name",
  "task_name_abstractive": "standardized name",
  "task_description": "exact quote",
  "metrics_extractive": ["metric1", "metric2"],
  "metrics_abstractive": ["std_metric1", "std_metric2"]}, ...]
```

6. `PERFORMANCE_RESULTS`: Extract all reported performance metrics.

Format:

```
[{"value": "exact value with units",
  "metric": "exact metric name",
  "subject": "model/human group name"}, ...]
```

Table 6: Abstract Screening Prompt Template (continued).

Prompt for abstract screening (continued)

INPUT EXAMPLE:

```
Title: A Comparison of LLMs in Clinical Triage: Brief Study
Abstract: We evaluated ChatGPT and GEMINI for triaging complex maxillofacial trauma cases at
a referral center. Using 10 standardized cases, we compared LLM recommendations against
center guidelines. Results showed ChatGPT achieved 70% accuracy in examinations while
GEMINI reached 50%. Additional metrics included diagnosis accuracy scores (GEMINI: 3.30,
ChatGPT: 2.30) and recommendation relevance (GEMINI: 2.90, ChatGPT: 3.50).
```

EXAMPLE OUTPUT:

```
ANALYSIS_START

<PAPER_TYPE>Clinical LLM Performance Evaluation - Original</PAPER_TYPE>

<MODELS>["ChatGPT", "GEMINI"]</MODELS>

<MULTIPLE_MODELS_USAGE>true</MULTIPLE_MODELS_USAGE>

<HUMAN_GROUPS>[]</HUMAN_GROUPS>

<EVALUATION_TASKS>
[{"task_name_extractive": "triaging complex maxillofacial
trauma cases",
  "task_name_abstractive": "clinical trauma triage assessment",
  "task_description": "triaging complex maxillofacial trauma cases at a referral center",
  "metrics_extractive": ["accuracy in examinations", "diagnosis accuracy scores",
  "recommendation relevance"],
  "metrics_abstractive": ["examination accuracy", "diagnostic performance",
  "recommendation quality"]}]
</EVALUATION_TASKS>

<PERFORMANCE_RESULTS>
[{"value": "70%", "metric": "accuracy in examinations", "subject": "ChatGPT"},
 {"value": "50%", "metric": "accuracy in examinations", "subject": "GEMINI"},
 {"value": "2.30", "metric": "diagnosis accuracy scores", "subject": "ChatGPT"},
 {"value": "3.30", "metric": "diagnosis accuracy scores", "subject": "GEMINI"},
 {"value": "3.50", "metric": "recommendation relevance", "subject": "ChatGPT"},
 {"value": "2.90", "metric": "recommendation relevance", "subject": "GEMINI"}]
</PERFORMANCE_RESULTS>

ANALYSIS_END
```

Now analyzing the following paper:

- Title: [*paper title*]
- Abstract: [*paper abstract*]

Prompt for full-text review

Please analyze the research paper to extract information about LLM performance evaluation in clinical settings. Present your analysis in the following structured format, maintaining exact quotes where possible. Start your response with "ANALYSIS_START" **and end with** "ANALYSIS_END".
REQUIRED:

- Title: [*paper title*]
- Full Text: [*full paper text*]

TASK: Extract the following structured information from the paper:
1. `PAPER_TYPE`: Classify the paper as one of the following:

- **"Clinical LLM Performance Evaluation - Original"**: Paper that conducts new experiments to evaluate LLM performance in clinical tasks and reports original performance metrics/results.
- **"Clinical LLM Performance Review"**: Paper that summarizes or analyzes existing LLM clinical performance evaluations without conducting new experiments or reporting new performance data.
- **"Non-Clinical LLM Evaluation"**: Paper not related to clinical LLM performance evaluation.

Note: If the paper is not classified as "Clinical LLM Performance Evaluation - Original", return empty values for all subsequent sections.
2. `BIBLIOGRAPHIC_DATES`: Extract the paper's submission and publication dates.

Format:

```
{"received_date": "YYYY-MM-DD",
 "accepted_date": "YYYY-MM-DD",
 "published_date": "YYYY-MM-DD"}
```

3. `CLINICAL_DOMAIN`: Extract the clinical specialty and context information.

Format:

```
{"specialty": "primary clinical specialty field",
 "disease_treatment": "specific diseases or treatments in focus",
 "mesh_terms": ["relevant MeSH term 1", "relevant MeSH term 2"]}
```

4. `MODELS`: Extract all LLM models mentioned in the paper.

Format:

```
[{"common_name": "most frequently used name in paper",
  "full_name": "complete name including version",
  "base_model": "base model name if fine-tuned, NA if not applicable"}]
```

5. `EXPERIMENTAL_SETTINGS`: Extract LLM inference settings.

Format:

```
{"llm_inference_temperature": "0.x",
 "llm_inference_few_shot": "n-shot",
 "llm_inference_CoT": true/false}
```

Table 6: Full-Text Review Prompt Template.

Table 7: Full-Text Review Prompt Template (continued).

Prompt for full-text review (continued)

6. HUMAN_GROUPS: Extract all medical professional groups that underwent the same evaluation tasks as the models.

Format:

```
["group1", "group2"]
```

7. EVALUATION_TASKS: Extract all clinical evaluation tasks.

Format:

```
[{"task_name_extractive": "exact task name",
  "task_name_abstractive": "standardized task name",
  "reference_sentence": "exact quote describing the task",
  "metrics": [{"metric_name_extractive": "exact metric name from text",
               "metric_name_abstractive": "standardized metric name",
               "value_range": [min, max],
               "higher_better": true/false,
               "reference_sentence": "exact quote describing the metric"}],
  "sample_size": integer,
  "sample_size_reference_sentence": "exact quote mentioning sample size",
  "data_source_extractive": "exact quote of data source",
  "data_source_abstractive": "standardized description of data source"}]
```

8. PERFORMANCE_RESULTS: Extract all reported performance metrics.

Format:

```
[{"value": "exact performance value with units/confidence intervals",
  "metric": "exact metric name from EVALUATION_TASKS metrics_extractive",
  "subject": "model name or human group name",
  "reference_sentence": "exact quote reporting this result"}]
```

Table 7: Proprietary Language Models Release Timeline.

Company	Model Name	Release Date	Normalized Name
OpenAI	**ChatGPT/GPT-3.5**	2022-12-30	gpt-3.5-turbo
	gpt-3.5-0301	2023-03-01	gpt-3.5-0301
	gpt-3.5-turbo-0613	2023-06-13	gpt-3.5-turbo-0613
	gpt-3.5-turbo-1106	2023-11-06	gpt-3.5-turbo-1106
	gpt-3.5-turbo-0125	2024-01-25	gpt-3.5-turbo-0125
	GPT-4		
	gpt-4-0314	2023-03-14	gpt-4-0314
	gpt-4-0613	2023-06-13	gpt-4-0613
	gpt-4-1106-preview	2023-11-06	gpt-4-1106-preview
	gpt-4-0125-preview	2024-01-25	gpt-4-0125-preview
	gpt-4-turbo-2024-04-09	2024-04-09	gpt-4-turbo-2024-04-09
	GPT-4o		
	gpt-4o updates	2024-05-13	gpt-4o-2024-05-13
	gpt-4o updates	2024-08-06	gpt-4o-2024-08-06
	gpt-4o updates	2024-11-20	gpt-4o-2024-11-20
	GPT-4o Mini	2024-07-18	gpt-4o-mini-2024-07-18
Microsoft	**Bing Chat**	2023-02-07	
	Rebranded as Copilot	2023-09-21	Based on latest GPT models
	Bing Chat integration	2023-11-15	Based on latest GPT models
	Copilot upgrade	2024-05-20	Based on latest GPT models
Claude	**Claude 1**		
	Claude 1.0/Claude 1.1	2023-03-14	claude-1.0/claude-1.1
	Claude 1.2	2023-08-09	claude-1.2
	Claude 1.3	2023-04-18	claude-1.3
	Claude 2		
	Claude 2.0	2023-17-11	claude-2.0
	Claude 2.1	2023-11-21	claude-2.1
	Claude 3		
	Claude 3 Haiku	2024-03-07	claude-3-haiku-20240307
	Claude 3 Sonnet	2024-02-29	claude-3-sonnet-20240229
	Claude 3 Opus	2024-02-29	claude-3-opus-20240229
	Claude 3.5		
	Claude 3.5 Sonnet	2024-06-20	claude-3-5-sonnet-20240622
	Claude 3.5 Haiku	2024-10-22	claude-3-5-haiku-20241022
	Claude 3.5 Opus	2024-10-22	claude-3-5-opus-20241022
	Claude 3.5 update	2024-12-03	claude-3.5-sonnet-20241203
Google	**Bard**	2023-03-21	lamda
	Bard upgrade	2023-05-10	palm-2
	Gemini 1.0		
	Gemini 1.0 Nano	2023-12-06	gemini-1.0-nano
	Gemini 1.0 Pro	2023-12-06	gemini-1.0-pro
	Gemini 1.0 Ultra (Advanced)	2023-12-06	gemini-1.0-ultra
	Gemini 1.5		
	Gemini 1.5 Flash (Basic)	2024-02-15	gemini-1.5-flash
	Gemini 1.5 Pro	2024-05-23	gemini-1.5-pro-001
	Gemini 1.5 Pro update	2024-09-24	gemini-1.5-pro-002
	Gemini 2.0		
	Gemini 2.0	2025-01-22	gemini-2.0-flash-001
	Gemini 2.0 Flash + Thinking	2025-01-22	Not used
Cohere	**Command**	2024-02-07	command
	Command R	2024-06-04	command-r
	Command R+	2024-04-30	command-rplus
	Command R-08-2024	2024-08-28	command-r-2408
	Command R+ 08-2024	2024-08-29	command-rplus-2408

Therapeutic Area For normalization and analysis purposes, we predefined 22 therapeutic areas, including cardiology, oncology, dentistry, and emergency medicine. Depending on the analytical objec-

Table 8: Open-Source Language Models Release Timeline.

Model Brand	Model Name	Release Date	Normalized Name
LLaMA (Meta)	**LLaMA 1**	2023-03	llama-1-7B, llama-1-13B, llama-1-30B, llama-1-65B
	LLaMA 2	2023-07	llama-2-7B, llama-2-13B, llama-2-70, llama-2-7B-chat, llama-2-13B-chat
	LLaMA 3	2024-04-18	llama-3-8B, llama-3-70B
	LLaMA 3.1	2024-07-23	llama-3.1-8B, llama-3.1-70B, llama-3.1-405B
	LLaMA 3.2	2024-10	llama-3.2-1B, llama-3.2-3B, llama-3.2-11B, llama-3.2-90B
	LLaMA 3.3	2024-12	llama-3.2-70B, llama-3.2-405B
Phi (Microsoft)	**Phi-1**	2023-06	phi-1-1.3B
	Phi-1.5	2023-11	phi-1.5-1.3B
	Phi-2	2024-02	phi-2-2.7B
	Phi-3		
	Phi-3 Mini	NA	phi-3-mini-3B
	Phi-3 Small	NA	phi-3-small-7B
	Phi-3 Medium	NA	phi-3-medium-14B
	Phi-3.5	2024-09	phi-3.5-3.8B
	Phi-4	2025-01-20	Not used
Gemma (DeepMind)	**Gemma 2B**	2024-02-21	gemma-1-2B
	Gemma 7B	2024-02-21	gemma-1-7B
	Gemma 1.1	2024-04-05	gemma-1.1
	Gemma 2 (9B, 27B)	2024-06-27	gemma-2-9B, gemma-2-27B
	Gemma 2 (2B)	2024-07-31	gemma-2-2B
Qwen (Alibaba)	**Qwen-7B**	2023-08-03	qwen-7B
	Qwen-14B	2023-09-25	qwen-14B
	Qwen-72B	2023-11-30	qwen-72B
	Qwen-2-7B-instruct	2024-05-16	qwen-2-7B-instruct
	Qwen-2-72B-instruct	2024-10-18	qwen-2-72B-instruct
	Qwen Max	2025-01-29	qwen-max
Mistral	**Mistral 7B**		
	mistral-7B-instruct-v0.1	2023-09-27	mistral-7B-instruct-v0.1
	mistral-7B-instruct-v0.2	2023-10	mistral-7B-instruct-v0.2
	mistral-7B-instruct-v0.3	2023-11	mistral-7B-instruct-v0.3
	Mistral Medium	2023-12	mistral-medium-2312
	Mixtral 8x7B	2023-12-09	mixtral-8x7B
	Mixtral 8x22B	2024-04-10	mixtral-8x22B
	Mistral Large	2024-02-26	mistral-large-2402
	Mistral Small	NA	mistral-small-2402
	Mistral Large 24.07	2024-07-24	mistral-large-2707
Medical Domain Fine-tuned	**ClinicalCamel-1-70B**	NA	clinicalcamel-1-70B
	Med42-70B	NA	med42-70B
	BioMistral-7B	NA	biomistral-7B
	Meditron	NA	meditron-7B
	MedLlama	NA	medllama-1-2

tives, these areas were further grouped into broader categories.

Data Source The source of evaluation data was classified into four types:

- **Board Examinations:** Questions derived from professional board certification exams used to assess medical expertise.
- **Quizzes:** Clinical questions sourced from textbooks, medical societies, or online educational platforms, excluding board exams.
- **Frequently Asked Questions:** Questions reflecting common patient inquiries in clinical settings.
- **Clinical Vignettes:** Case-based questions developed using real patient data, publicly avail-

able case reports, or LLM-generated simulated patient scenarios.

Evaluation Method Evaluation methods were divided into two main categories:

- **Multiple-Choice Questions (MCQs):** Assessing correctness based on predefined answer choices.
- **Human Rating:** Clinical experts rating model-generated responses according to structured evaluation guidelines. This includes both closed-ended rating systems with predefined criteria and open-ended assessments.

A.5 Benchmark Performance Imputation

The MICE framework was configured with optimized settings to ensure stable imputation. Missing values were initially imputed using the median of observed values, followed by a maximum of 50 iterative updates with a convergence tolerance of 1×10^{-6}

For BR-MICE, posterior sampling was enabled, and each missing variable was modeled using all available predictors. The base random seed was set to 42, with independent seeds assigned for multiple imputations.

For RF-MICE, 100 trees were used with bootstrap sampling enabled. The maximum tree depth was set to 15, and feature selection per split followed the square root of the total number of features. BR-MICE regularization parameters were optimized iteratively, with convergence determined via evidence maximization.

To estimate the variance of imputed values, we computed both the within-imputation variance (W) and between-imputation variance (B) across m independent imputations. The total variance (T) was calculated using Rubin's rules:

$$W = \frac{1}{m} \sum_{j=1}^{m} S_j^2$$

$$B = \frac{1}{m-1} \sum_{j=1}^{m} (\bar{Q}_j - \bar{Q})^2$$

$$T = W + \left(1 + \frac{1}{m}\right) B$$

where S_j^2 is the variance of the j-th imputed dataset, $\bar{Q}_j$ is the mean of the j-th imputation, and $\bar{Q}$ is the overall mean of all imputations. The final imputed values were obtained by taking the median of all imputations to ensure robustness against extreme values.

A.6 Bayesian Modeling

We implemented our hierarchical Bayesian model usingNumPyro (v0.17.0) with a JAX (v0.5.0) backend. Posterior inference was conducted via the No-U-Turn Sampler (NUTS), utilizing 1,000 warmup iterations and 2,000 sampling iterations across 8 parallel chains. We assessed convergence using the Gelman-Rubin diagnostic ($\hat{R}$) and effective sample size. The model specification is as follows:

A.6.1 Model Structure

We formulate our hierarchical Bayesian model as follows: We begin by specifying half-normal hyperpriors for the standard deviations that govern the variability of different components in our model:

$$\begin{aligned}
\sigma_{\text{model}} &\sim \text{HalfNormal}(1) \\
\sigma_{\text{obs}} &\sim \text{HalfNormal}(1) \\
\sigma_{\text{type}} &\sim \text{HalfNormal}(1) \\
\sigma_{\text{source}} &\sim \text{HalfNormal}(1) \\
\sigma_{\text{eval}} &\sim \text{HalfNormal}(1)
\end{aligned}$$

These hyperpriors control the variation in model effects, observation noise, task type effects, data source effects, and evaluation method effects, respectively.

The model effects component captures the inherent performance capabilities of each language model:

$$\begin{aligned}
\mu_{\text{model}} &\sim \text{Normal}(0, 1) \\
\beta_{\text{model},j} &\sim \text{Normal}(\mu_{\text{model}}, \sigma_{\text{model}})
\end{aligned}$$

where $j = 1, 2, \ldots, n_{\text{models}}$, and $\beta_{\text{model},j}$ represents the effect of the j-th model. The parameter μ_{model} serves as a global mean for model effects.

We model three task-related components:

$$\begin{aligned}
\beta_{\text{type},k} &\sim \text{Normal}(0, \sigma_{\text{type}}) \\
\beta_{\text{source},l} &\sim \text{Normal}(0, \sigma_{\text{source}}) \\
\beta_{\text{eval},m} &\sim \text{Normal}(0, \sigma_{\text{eval}})
\end{aligned}$$

where:

- $k = 1, 2, \ldots, n_{\text{task\_types}}$, with $\beta_{\text{type},k}$ representing the effect of the k-th task type
- $l = 1, 2, \ldots, n_{\text{data\_sources}}$, with $\beta_{\text{source},l}$ representing the effect of the l-th data source

- $m = 1, 2, \ldots, n_{\text{evaluation\_methods}}$, with $\beta_{\text{eval},m}$ representing the effect of the m-th evaluation method

Each task-related effect is centered at zero, reflecting our assumption that these effects represent deviations from an average difficulty level.

The predicted performance for each data point i is given by:

$$\mu_i = \beta_{\text{model,model}[i]} + \beta_{\text{type,type}[i]} + \beta_{\text{source,source}[i]} + \beta_{\text{eval,eval}[i]} + \epsilon_\mu$$

where model$[i]$, type$[i]$, source$[i]$, and eval$[i]$ are the indices for model, task type, data source, and evaluation method for data point i, respectively, and $\epsilon_\mu \sim \text{Normal}(0, 0.1)$ represents additional noise in the prediction process.

Finally, we model the observed performance metrics using a normal likelihood:

$$y_i \sim \text{Normal}(\mu_i, \sigma_{\text{obs}})$$

where y_i is the observed performance metric for data point i on the normalized scale.

A.6.2 Centrality Measurement

Quantifying the connectivity of models within the evaluation network is essential for understanding their role and influence. Some models exhibit weak connections to other major models, meaning they contribute useful information to Bayesian modeling but have limited relevance for downstream analysis. By computing centrality scores, we classified models based on their connectivity and excluded the lower 50% from downstream evaluations.

The evaluation network was represented as a bipartite graph $G = (V, E)$, where the vertex set V consisted of two disjoint subsets: models and tasks. An edge $(m, t) \in E$ was formed if and only if model m was evaluated on task t. This structure provided a basis for analyzing connectivity patterns and assessing the relative importance of models within the evaluation framework.

Model connectivity was quantified using three centrality measures. Degree centrality (C_D) captured the number of direct connections a model had, normalized by the maximum possible connections:

$$C_D(v) = \frac{deg(v)}{|V| - 1}$$

where $deg(v)$ represents the number of edges incident to node v. Between-ness centrality (C_B) measured how often a model served as a bridge along the shortest paths between other nodes:

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where σ_{st} is the total number of shortest paths from node s to node t, and $\sigma_{st}(v)$ is the number of those paths passing through node v. Closeness centrality (C_C) assessed how close a model was to all other nodes in the network:

$$C_C(v) = \frac{|V| - 1}{\sum_{u \neq v} d(v, u)}$$

where $d(v, u)$ is the shortest-path distance between nodes v and u. To integrate these measures into a single ranking, a combined connectivity score was computed by summing the three normalized centrality values:

$$\text{Combined Score}(m) = C_D(m) + C_B(m) + C_C(m)$$

Models were then ranked based on their combined scores, and those below the P_{th} percentile were classified as low-connectivity models:

$$\text{Low Connectivity}(m) = \begin{cases} \text{True} & \text{if Combined-} \\ & \text{Score}(m) < P_{th} \\ \text{False} & \text{otherwise} \end{cases}$$

where P_{th} was set at the 50th percentile, identifying the bottom 50% of models as low-connectivity. For downstream analysis, only high-connectivity models were retained. This ensured that subsequent evaluations focused on models with strong integration within the network while still utilizing all available information in Bayesian modeling.

A.7 Correlation Measurement

To evaluate the relationship between LLM performance on different benchmarks and in clinical settings, we computed four correlation measures: Pearson's correlation coefficient, Spearman's rank correlation coefficient, Kendall's tau, and Lin's concordance correlation coefficient (CCC). Among these, Spearman's and Kendall's correlations were used as the primary measures, as they better capture rank-based relationships given the diversity of evaluation methodologies.

Pearson's correlation coefficient (r) measures the strength of the linear relationship between two continuous variables. It is computed as:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

where x_i and y_i are individual data points, and $\bar{x}$ and $\bar{y}$ are their respective means.

Spearman's rank correlation coefficient (ρ) assesses the monotonic relationship between two variables by comparing their rank orders rather than raw values. It is given by:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

where d_i is the rank difference for each pair of observations, and n is the number of observations.

Kendall's tau (τ) quantifies the ordinal association between two variables based on concordant and discordant pairs:

$$\tau = \frac{C - D}{\frac{1}{2}n(n-1)}$$

where C is the number of concordant pairs, and D is the number of discordant pairs.

Lin's CCC (ρ_c) evaluates both correlation and agreement between two variables by incorporating measures of precision and accuracy:

$$\rho_c = \frac{2r\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$

where r is Pearson's correlation coefficient, σ_x and σ_y are standard deviations, and μ_x and μ_y are means of the two variables.

B Supplementary Results

B.1 Benchmark Performance Imputation

Table 9 presents the imputation accuracy of benchmark models evaluated through a masking test across both general and medical QA domains.

Table 9: Imputation accuracy on masking test for benchmark performances.

Model	MAE	RMSE	R^2
General			
RandomForest	4.21	8.17	0.89
BayesianRidge	5.63	8.14	0.89
Medical QA			
RandomForest	2.04	3.18	0.98
BayesianRidge	4.14	6.81	0.90

Table 10: Within- and between- variance results from multiple imputation.

Category	Within	Between	Total
General Benchmarks (Overall Variance: 481.1)			
RandomForest	1.1	204.5	215.7
(% of Overall)	0.2%	42.5%	44.8%
BayesianRidge	100.6	189.4	299.5
(% of Overall)	20.9%	39.4%	62.3%
Medical Benchmarks (Overall Variance: 390.5)			
RandomForest	2.1	43.2	47.5
(% of Overall)	0.5%	11.1%	12.2%
BayesianRidge	67.5	40.0	109.5
(% of Overall)	17.3%	10.2%	28.0%

B.2 Benchmark-to-Benchmark Correlation

Table 11 summarizes the average correlation coefficients between general benchmarks and other benchmarks, providing a comparative view across domains and correlation metrics.

Figures 6, 7, and 8 further illustrate the internal correlations within each domain and the cross-domain relationships.

B.3 Benchmark-to-Clinical Performance Correlation

Figures 9 and 10 present the correlations between benchmark performance and language model performance in real-world clinical settings, with and without imputed benchmark scores. The results are reported using four correlation measures—Pearson correlation coefficient, Spearman rank correlation coefficient, Kendall's tau, and Lin's CCC—to ensure robustness from multiple statistical perspectives.

Table 11: Average correlation coefficients of general benchmarks with other benchmarks. The highest score in each column is **bold**, and the second highest is underlined.

General Benchmark	Spearman		Kendall	
	Medical QA	General	Medical QA	General
MMLU	0.853	0.715	0.690	0.576
MMLU Pro	0.851	**0.773**	0.679	**0.628**
BBH	**0.891**	0.725	**0.736**	0.573
HumanEval	0.838	0.764	0.671	0.618
GSM8K	0.816	0.756	0.645	0.599
MATH	0.785	0.625	0.618	0.502
Average	**0.839**	**0.726**	**0.673**	**0.583**

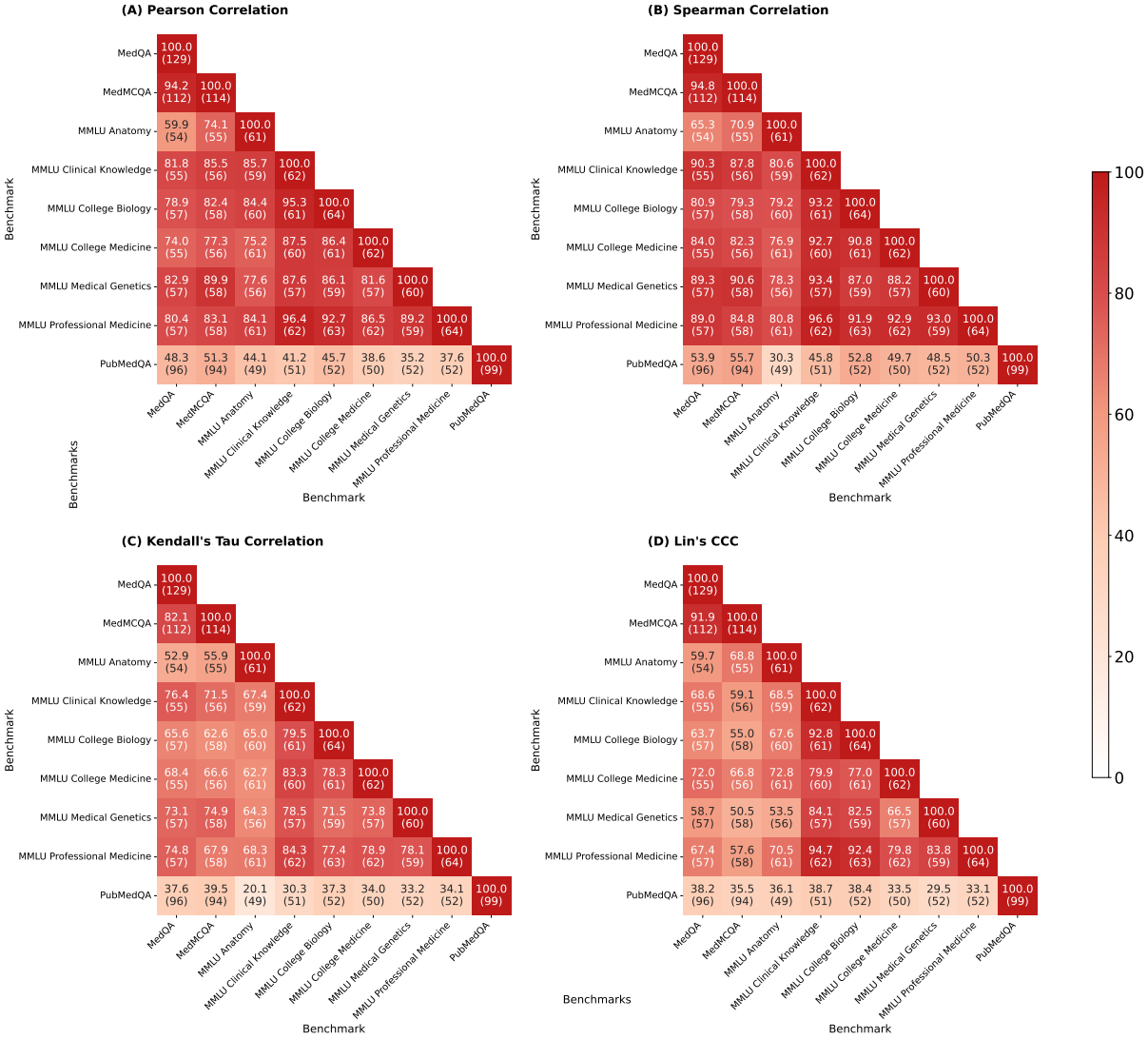

Figure 6: Correlations within medical QA benchmarks.

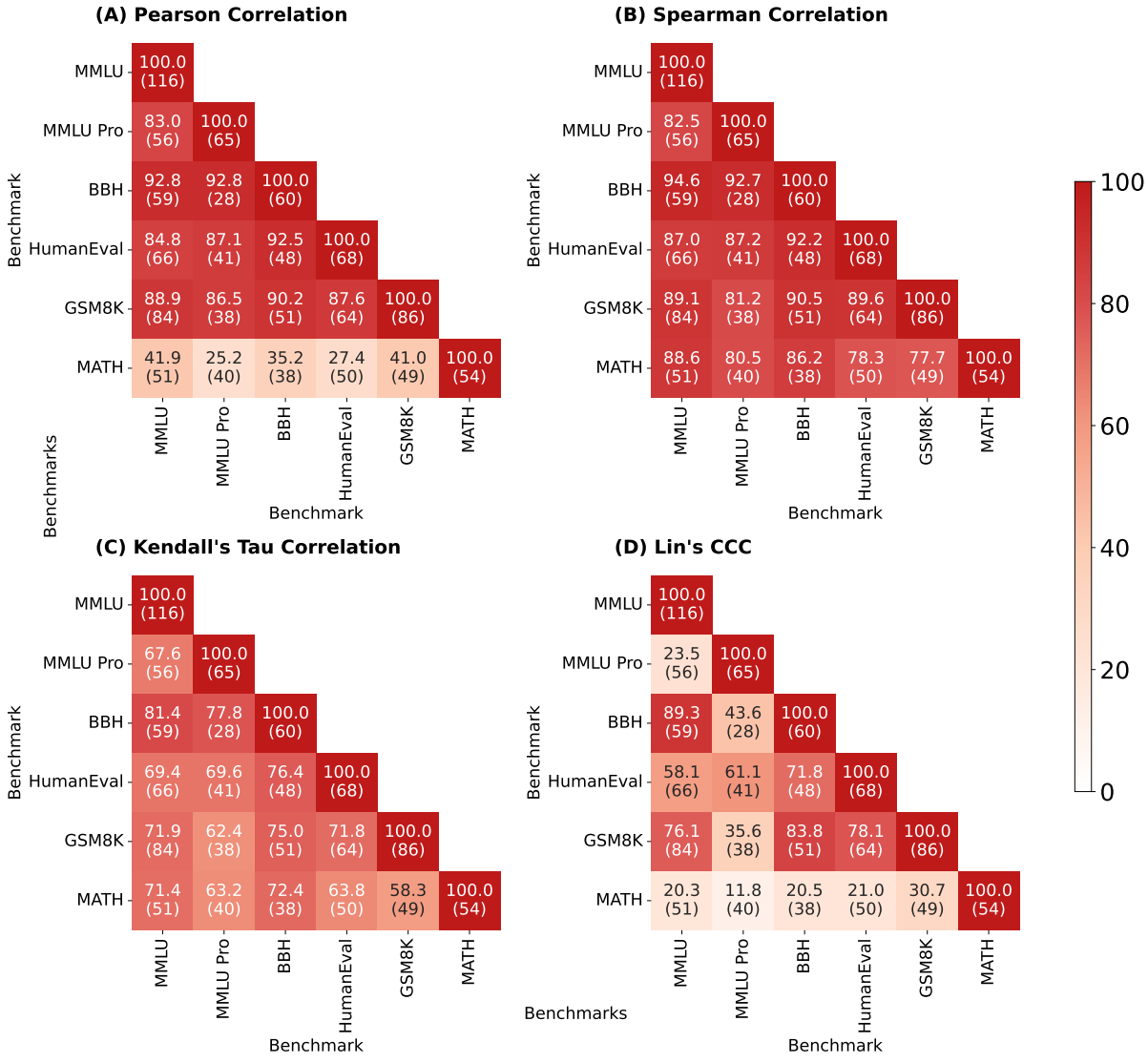

Figure 7: Correlations within general benchmarks.

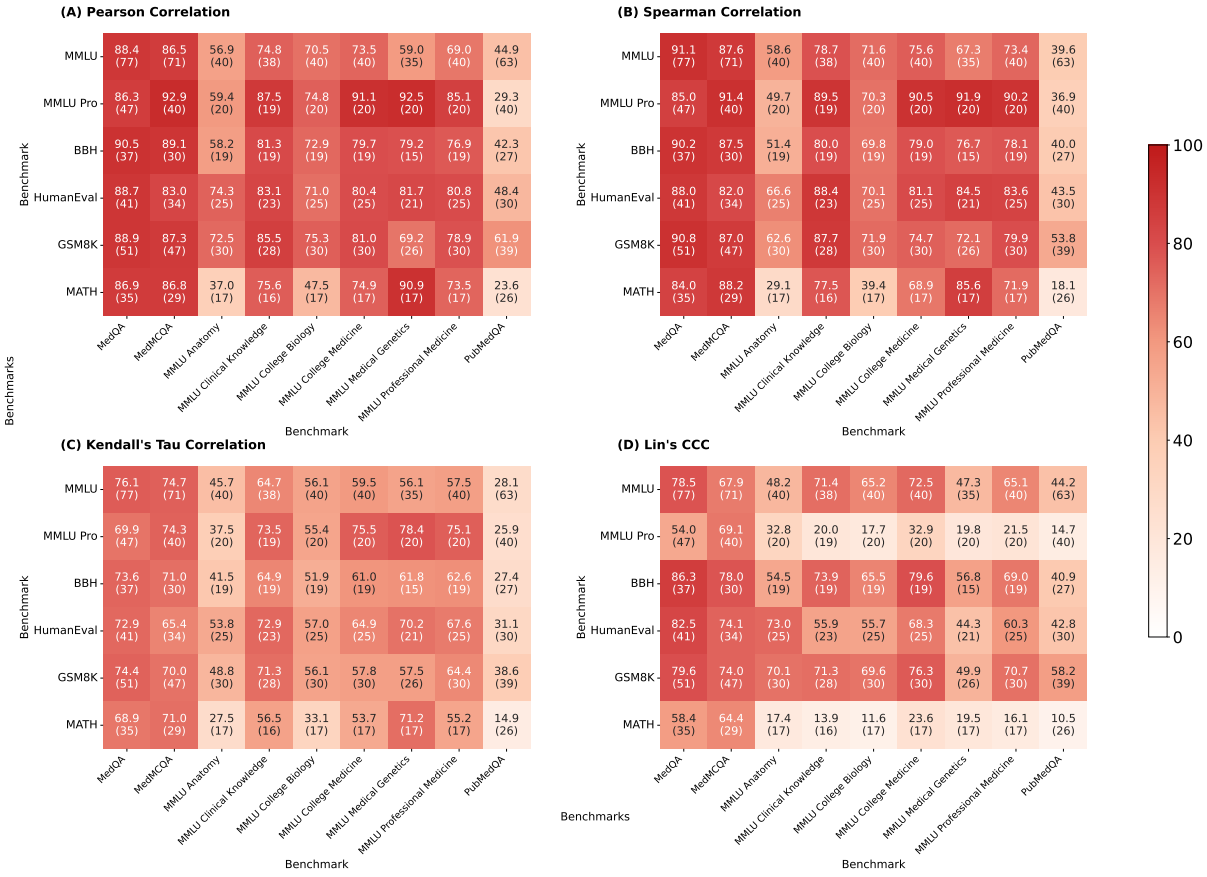

Figure 8: Correlations between general and medical QA benchmarks.

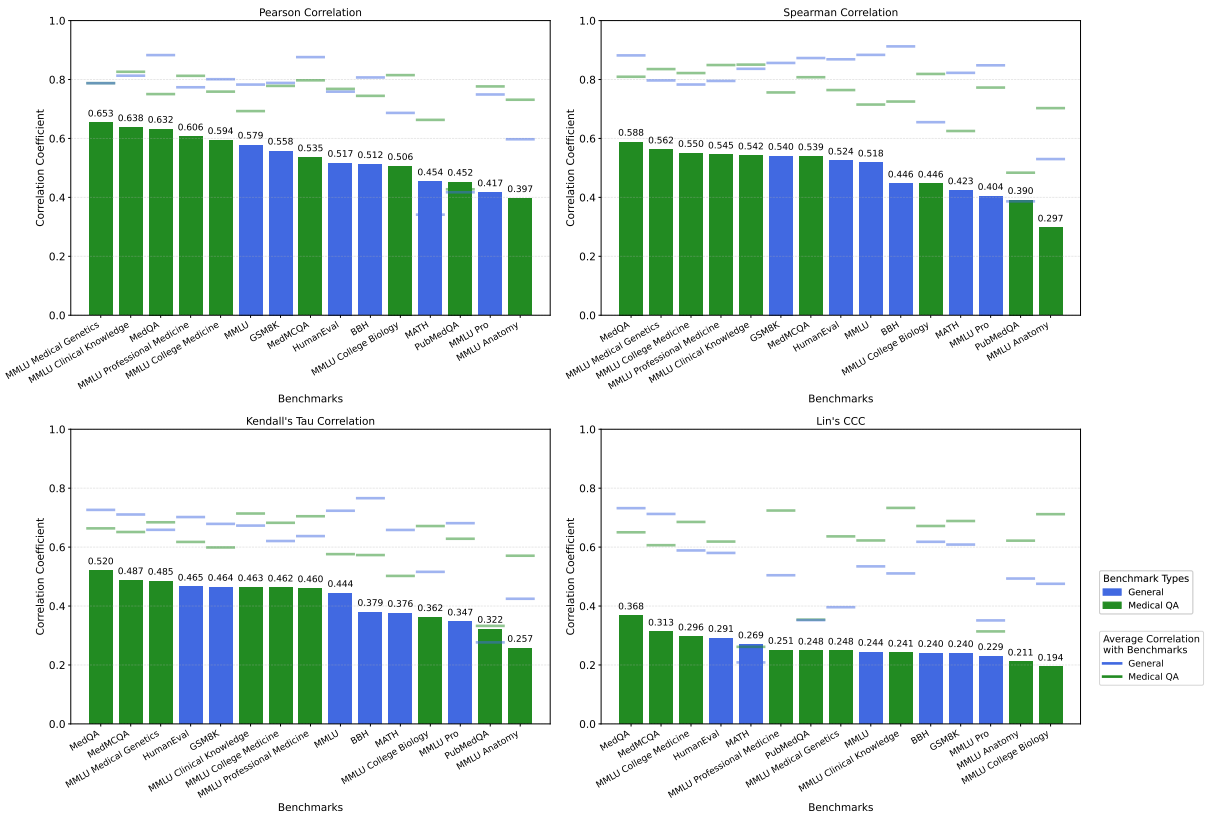

Figure 9: Correlations between benchmark performance and language model performance in real-world clinical settings with imputed benchmark scores.

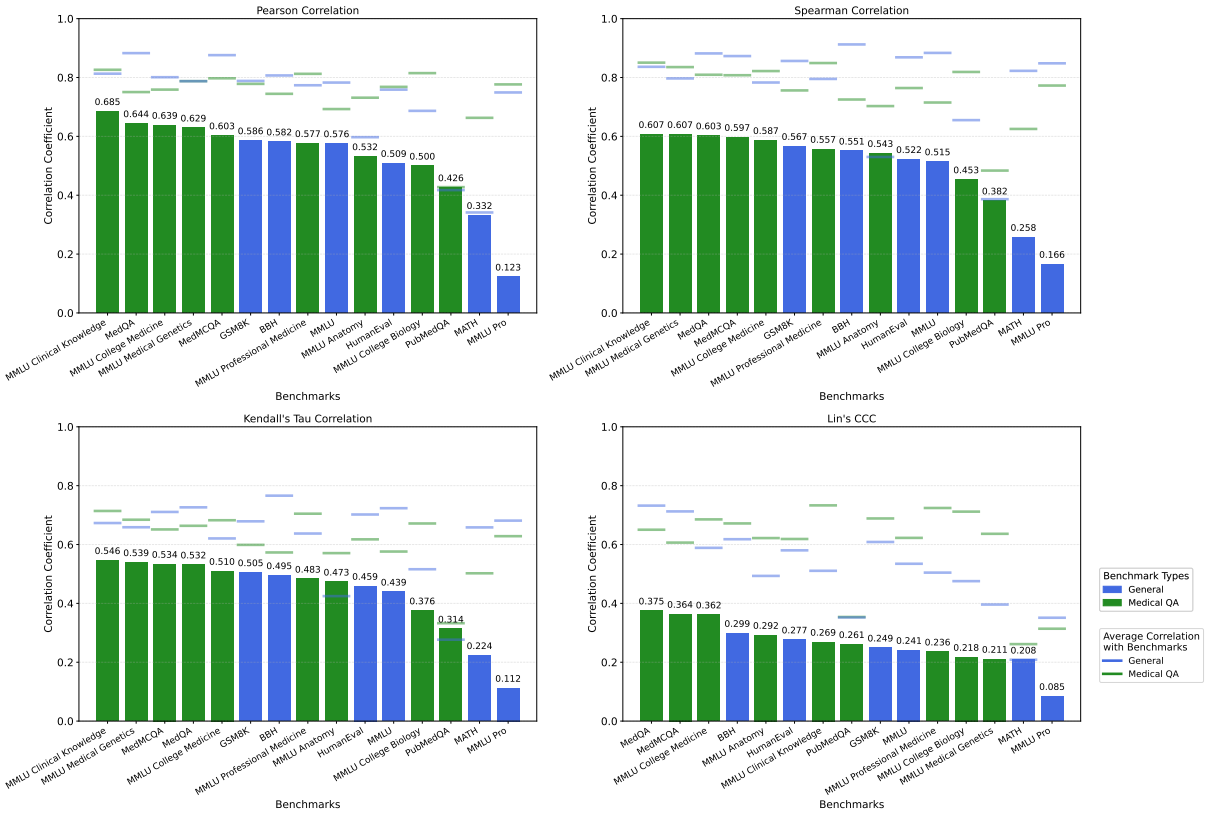

Figure 10: Correlations between benchmark performance and language model performance in real-world clinical settings with non-imputed benchmark scores.