Portraits of Large Language Models: Deciphering the Taxonomy of Medical LLMs

Radha Nagarajan, Vanessa Klotzman, Midori Kondo, Sandip Godambe, Adam Gold, John Henderson, Steven Martel

Submitted to: JMIR Preprints
on: February 26, 2025

Table of Contents

Portraits of Large Language Models: Deciphering the Taxonomy of Medical LLMs

Radha Nagarajan[1]; Vanessa Klotzman[1]; Midori Kondo[2]; Sandip Godambe[1]; Adam Gold[1]; John Henderson[1]; Steven Martel[1]

[1] Children's Hospital of Orange County Orange US
[2] Fred Hutch Cancer Center Seattle US

Corresponding Author:
Radha Nagarajan

Children's Hospital of Orange County
1201 W La Veta Ave
Orange
US

Abstract

Background: Large Language Models (LLMs) continue to enjoy enterprise-wide adoption in healthcare while evolving in number, size, complexity, cost, and more importantly performance. Performance benchmarks play a critical role in their ranking across community leaderboards and subsequent adoption.

Objective: Given the small operating margins of healthcare organizations and growing interest in LLMs and conversational AI, there is an urgent need for objective approaches that can assist in identifying viable LLMs without compromising their performance. The objective of the present study is to generate a taxonomy portrait of medical LLMs (N = 33) whose domain-specific and domain non-specific multivariate performance benchmarks were available from Open-Medical LLM and Open LLM leaderboards on Hugging Face.

Methods: Hierarchical clustering of multivariate performance benchmarks is used to generate taxonomy portraits revealing inherent partitioning of the medical LLMs across diverse tasks. While domain-specific taxonomy is generated using nine performance benchmarks related to medicine from the Hugging Face Open-Medical LLM initiative, domain non-specific taxonomy is presented in tandem to assess their performance on a set of six benchmarks on generic tasks from the Hugging Face Open LLM initiative. Subsequently, non-parametric Wilcoxon Ranksum test and linear correlation is used to assess differential changes in the performance benchmarks between two broad groups of LLMs and potential redundancies between the benchmarks.

Results: Two broad families of LLMs with statistically significant differences (\alpha = 0.05) in performance benchmarks are identified for each of the taxonomies. Consensus in their performance on the domain-specific, and domain non-specific tasks revealed inherent robustness of these LLMs across diverse tasks. Subsequently, statistically significant correlations between performance benchmarks revealed inherent redundancies, indicating a subset of these benchmarks may be sufficient in assessing the domain-specific performance of medical LLMs.

Conclusions: Understanding the medical LLM taxonomies is an important step in identifying LLMs with similar performance while aligning with the needs, economics, and other demands of healthcare organizations. While the focus of the present study is on a subset of medical LLMs from the Hugging Face, enhanced transparency of performance benchmarks and economics across a larger family of medical LLMs is needed to generate more comprehensive taxonomy portraits accelerating their strategic and equitable adoption in healthcare. Clinical Trial: Not applicable

(JMIR Preprints 26/02/2025:72918)
DOI: https://doi.org/10.2196/preprints.72918

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Portraits of Large Language Models: Deciphering the Taxonomy of Medical LLMs

Radha Nagarajan[1*], Vanessa Klotzman[2], Midori Kondo[3], Sandip Godambe[1], Adam Gold[1], John Henderson[1], Steven Martel[1]

Rady Children's Health, Orange, CA, USA[1]
University of California Irvine, Irvine, CA, USA[2]
Fred Hutchinson Cancer Center, Seattle, WA, USA[3]

Abstract

Background: Large Language Models (LLMs) continue to enjoy enterprise-wide adoption in healthcare while evolving in number, size, complexity, cost, and more importantly performance. Performance benchmarks play a critical role in their ranking across community leaderboards and subsequent adoption.

Objective: Given the small operating margins of healthcare organizations and growing interest in LLMs and conversational AI, there is an urgent need for objective approaches that can assist in identifying viable LLMs without compromising their performance. The objective of the present study is to generate a taxonomy portrait of medical LLMs (N = 33) whose domain-specific and domain non-specific multivariate performance benchmarks were available from Open-Medical LLM and Open LLM leaderboards on Hugging Face.

Methods: Hierarchical clustering of multivariate performance benchmarks is used to generate taxonomy portraits revealing inherent partitioning of the medical LLMs across diverse tasks. While domain-specific taxonomy is generated using nine performance benchmarks related to medicine from the Hugging Face Open-Medical LLM initiative, domain non-specific taxonomy is presented in tandem to assess their performance on a set of six benchmarks on generic tasks from the Hugging Face Open LLM initiative. Subsequently, non-parametric Wilcoxon Ranksum test and linear correlation is used to assess differential changes in the performance benchmarks between two broad groups of LLMs and potential redundancies between the benchmarks.

Results: Two broad families of LLMs with statistically significant differences (α = 0.05) in performance benchmarks are identified for each of the taxonomies. Consensus in their performance on the domain-specific, and domain non-specific tasks revealed inherent robustness of these LLMs across diverse tasks. Subsequently, statistically significant correlations between performance benchmarks revealed inherent redundancies, indicating a subset of these benchmarks may be sufficient in assessing the domain-specific performance of medical LLMs.

Conclusion: Understanding the medical LLM taxonomies is an important step in identifying LLMs with similar performance while aligning with the needs, economics, and other demands of healthcare organizations. While the focus of the present study is on a subset of medical LLMs from the Hugging Face,

enhanced transparency of performance benchmarks and economics across a larger family of medical LLMs is needed to generate more comprehensive taxonomy portraits accelerating their strategic and equitable adoption in healthcare.

*Radha Nagarajan, PhD
Rady Children's Health
1201 W. La Veta Ave, Orange, CA, 92868, USA
Email: Radha.Nagarajan@choc.org

Introduction

Large language models (LLMs) continue to show considerable promise and growth in healthcare [1]. Popular healthcare LLM applications fall under three broad task categories, namely: *clinical tasks, documentation tasks, medical research and education tasks* [2]. Specific LLM healthcare applications include (*i*) virtual health assistants and language translation [3], (*ii*) summarization of clinical narratives and ambient listening [4, 5], (*iii*) patient education [6] and (*iv*) clinical trial matching [7]. More importantly, LLMs have continued to evolve in numbers, size, complexity, costs, and performance, impacting their adoption [8]. A recent perspective discussed three broad LLM implementation pathways (*Training from Scratch, Fine-Tuned Pathway, Out of the Box Pathway*) along with risks, benefits, and economics across four major cloud service providers for their equitable adoption in healthcare [9]. The study also elucidated the essential ingredients such as digital data, infrastructure, workforce, ethics, and regulatory aspects that can significantly impact LLM implementations. While helpful, these three pathways represent broad categorizations of LLM implementations and do not necessarily incorporate objective metrics impacting their adoption such as *performance benchmarks* that accompany LLM leaderboards [10]. This study investigates performance benchmarks used to evaluate and rank LLMs across *Open LLM* [11] and *Open Medical LLM* [12] leaderboards at Hugging Face [13, 14]. Hugging Face has witnessed increasing visibility and adoption by the Generative AI and LLM community over the years. Its structured and transparent approach enables enhanced reproducibility of the reported metrics and implementation, widespread collaboration between experts, unbiased comparison of the different models, and selection of the LLMs based on the performance, needs and affordability. This study elucidates the *domain-specific* (**DS**) taxonomy of LLMs using publicly reported performance on medical tasks (i.e., *medical LLMs*) from Hugging Face. *Domain non-specific* (**DN**) taxonomy of these medical LLMs from the Open LLM initiative is also discussed. Given the low-operating margins [15] across healthcare organizations, the taxonomies are expected to assist in choosing the right LLM from performance, needs, and parameter standpoints, with the latter having a direct nexus to infrastructure and implementation costs, hence LLM economics [9]. The taxonomies were generated by hierarchical clustering of the domain-specific and domain non-

specific multivariate performance benchmarks that assess the task-specific and generic capabilities of these LLMs. Subsequently, two broad groups of medical LLMs with markedly different performance benchmark profiles is discussed. Potential redundancies between the performance benchmarks across the DS and DN taxonomies are also elucidated.

Table 1. Domain-specific performance benchmarks of medical LLMs from Hugging Face Open Medical LLM initiative.

Description	Source
• **MedQA [Medical Question and Answer]: Consists of multiple-choice questions (11450 questions in the development set and 1273 questions in the test set) from the United States Medical License Exam for benchmarking the LLMs general medical knowledge and reasoning skills on United States Medical Licensure.**	**MQA** [16]
• **MedMCQA [Medical Multiple-Choice Question and Answer]: Consists of multiple-choice questions (187000 questions in the development set and 6100 questions in the test set) related to the Indian medical entrance exam (AIIMS/NEET). As with MedQA, MedMCQA is used to benchmark the LLMs general medical knowledge and reasoning ability as it pertains to the Indian medical entrance exam.**	**MCQA** [17]
• **MMLU Anatomy: [Massive Multitask Language Understanding, Anatomy]: MMLU subset consists of multiple-choice questions (135 questions) for benchmarking the knowledge of the LLM on human anatomy.**	**ANAT** [18]
• **MMLU Clinical Knowledge [Massive Multitask Language Understanding, Clinical Knowledge]: MMLU subset consists of multiple-choice questions (265 questions) for benchmarking the clinical knowledge and decision-making skills of the LLM.**	**CLIN** [18]
• **MMLU College Biology [Massive Multitask Language Understanding, College Biology]: MMLU subset consists of multiple-choice questions (144 questions) for benchmarking the knowledge on college biology.**	**BIOL** [18]
• **MMLU College Medicine [Massive Multitask Language Understanding, College Medicine] MMLU subset consists of multiple-choice questions (173 questions) for benchmarking the college-level medical knowledge**	**CMED** [18]
• **MMLU Medical Genetics [Massive Multitask Language Understanding, Medical Genetics] MMLU subset consists of 100 questions related to medical genetics.**	**GEN** [18]
• **MMLU Professional Medicine [Massive Multitask Language Understanding, Professional Medicine] MMLU subset consists of multiple-choice questions (272 questions) for benchmarking the LLM on knowledge required for medical professionals.**	**PMED** [18]
• **PUBMEDQA [PUBMED Question & Answer] Closed-domain data set comprising expert-labeled questions (500 questions in the development set and 500 questions in the test set) for benchmarking the LLMs ability to comprehend and reason biomedical literature.**	**PUBM** [19]

The domain-specific benchmarks considered include (a) those that assess the medical question and answering capabilities and reasoning skills related to medical licensing exams, (b) a series of subject and domain-specific evaluation broadly under massive multitask language understanding, and (c) LLMs ability to comprehend and reason biomedical literature. A detailed description of the domain-specific benchmark along with the references is shown in **Table 1**. The domain non-specific benchmarks were also retrieved for the medical LLMs

through the Open LLM initiative. The benchmarks included (a) those that assess the LLMs ability to follow verifiable instructions, (b) chain of thought prompting, (c) mathematical problem-solving skills, (d) graduate level reasoning capabilities across diverse subjects, (e) multistep reasoning abilities, and (f) multitask language understanding on challenging reasoning-based questions. **Table 2** shows a detailed description of the domain non-specific benchmarks and the references.

Table 2. Domain non-specific performance benchmarks from Hugging Face Open LLM initiative.

Description	Source
• **IFEval [Instruction Following Evaluation]: Benchmarks LLMs ability to follow verifiable instructions using 25 distinct types of verifiable instructions and 500 prompts, with each prompt containing at least one verifiable instruction.**	**IFEV** [20]
• **BBH [Big Bench Hard]: Benchmarks the performance of LLMs on 23 challenging BIG Bench tasks (BBH) where prior LLMs failed to outperform an average-human rater. Emphasized the importance of chain-of-thought prompting.**	**BBH** [21]
• **MATH [Math] Benchmarks the mathematical problem-solving ability of the LLM using 12500 mathematics competition problems.**	**MATH** [22]
• **GPQA [Graduate Level Google Proof Q & A]: Benchmarks LLMs using 448 multiple choice questions generated by experts in areas such as biology, physics, and chemistry.**	**GPQA** [23]
• **MuSR [Multistep Soft Reasoning]: Benchmarks LLMs ability on complex multistep reasoning instances and long-range (~1000 words) free text narratives from real-world domains.**	**MUSR** [24]
• **MMLU Pro [Multitask Language Understanding Pro]: Benchmarks the reasoning and language comprehension abilities of LLMs across diverse domains by incorporating challenging, reasoning-focused question, and expanding the choice of the original MMLU from four to ten.**	**MPRO** [25]

Table 3. List of (N = 33) LLMs used in medicine (medical LLMs) from Hugging Face along with the abbreviations.

LLM	Abbreviation	LLM	Abbreviation
mistralai/Mistral-7B-Instruct-v0.1	MIS-7BI	VAGOsolutions/SauerkrautLM-Gemma-7b	GEMS-7B
mistralai/Mistral-7B-v0.1	MIS-7B	VAGOsolutions/Llama-3-SauerkrautLM-8b-Instruct	LM3S-8BI
EleutherAI/pythia-2.8b	PYT-2.8B	openai-community/gpt2-xl	GPT-XL
EleutherAI/gpt-neo-2.7B	GPTN-2.7B	openai-community/gpt2	GPT-2
lmsys/vicuna-7b-v1.5	VIC-7B	HuggingFaceH4/zephyr-7b-beta	ZEP-7B
abacusai/Llama-3-Smaug-8B	LM3S-8B	tiiuae/falcon-7b-instruct	FAL-7BI
abacusai/Liberated-Qwen1.5-14B	QW-14B	tiiuae/falcon-7b	FAL-7B
HPAI-BSC/Llama3-Aloe-8B-Alpha	LM3A-8B	NousResearch/Nous-Hermes-2-Mistral-7B-DPO	MISH-7BD
google/gemma-2b	GEM-2B	NousResearch/Hermes-2-Pro-Mistral-7B	MISH-7B
google/gemma-1.1-7b-it	GEM-7BI	CohereForAI/aya-23-8B	AYA-8B

google/recurrentgemma-2b	GEMR-2B	upstage/SOLAR-10.7B-Instruct-v1.0	SOL-10.7B
google/gemma-7b	GEM-7B	01-ai/Yi-1.5-9B-32K	YI-9BK
microsoft/phi-1_5	PHI-1.5	01-ai/Yi-1.5-9B	YI-9B
TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T	TLM-1.1B	lightblue/suzume-llama-3-8B-multilingual	L3S-8B
Qwen/Qwen1.5-7B	QWN-7B	meta-llama/Meta-Llama-3-8B-Instruct	L3M-8BI
Qwen/Qwen1.5-7B-Chat	QWN-7BC	meta-llama/Meta-Llama-3-8B	L3M-8B
stabilityai/stablelm-2-1_6b	STA-6B		

Methods

Medical LLMs with domain-specific and domain non-specific benchmarks were retrieved from Hugging Face Open Medical LLM [12] and Open LLM [11] leaderboards, **Table 3**. While Hugging Face has several contributions from the AI open-source community, the present study restricted the medical LLMs to those contributed by organizations, large teams, and commercial LLMs while excluding individual contributions, resulting in (N = 33) medical LLMs, **Table 3**. LLM taxonomies were generated using hierarchical clustering of the DS and DN multivariate performance benchmarks with Euclidean distance and average linkage [26]. DS taxonomy was generated based on the nine performance benchmarks, **Table 1**, whereas DN taxonomy was generated based on the six normalized performance benchmarks, **Table 2**. The performance benchmarks were scaled to zero-mean and unit variance prior to clustering to minimize the impact of the range that can vary across the performance benchmarks. Color-coded dendrograms were subsequently used to generate visualizations of the performance benchmark profiles of the respective taxonomies. Two clusters with statistically significant differential changes in performance benchmark profiles were identified for the DS and DN taxonomy. Statistically significant differential changes were investigated using the Wilcoxon-Ranksum non-parametric test (α = 0.01). Subsequently, Pearson correlation and scatter plots were used to elucidate potential redundancies between the performance benchmarks for the DS and DN taxonomies. Since correlation estimates can be deceptive under a sparse distribution of data points about the linear trend, scatter plots of the pair-wise benchmarks are also provided for visualization. Bonferroni correction was used to control for family-wise error rates in the statistical tests for differential changes and correlation.

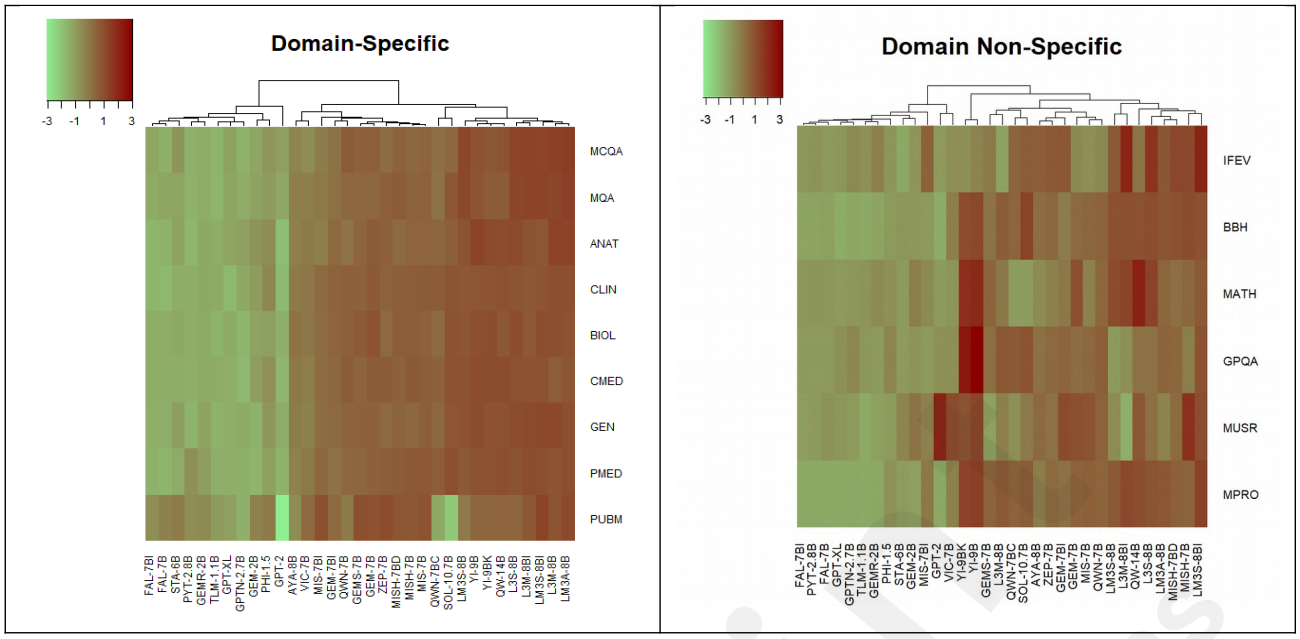

Figure 1 Dendrogram representing the scaled performance benchmark profiles for the domain-specific (left) and domain non-specific (right) taxonomies from hierarchical clustering of the (N = 33) medical LLMs is shown. The color-key represents variation in the performance benchmarks and increases from green to red.

Results

DS and DN taxonomies of the (N = 33) medical LLMs generated by hierarchical clustering revealed two well-separated clusters high (H) and low (L) with markedly different median benchmark profiles, **Fig. 2**. For the DS taxonomy, the H cluster comprised 22 LLMs, whereas the L cluster had 11 LLMs. Wilcoxon-Ranksum tests revealed statistically significant differences between these clusters across all nine benchmarks after controlling for family-wise error rate (i.e., $\alpha = 0.01/9 \sim 0.001$). LLMs in the H cluster were (MIS-7BI, MIS-7B, VIC-7B, LM3S-8B, QW-14B, LM3A-8B, GEM-7BI, GEM-7B, QWN-7B, QWN-7BC, GEMS-7B, LM3S-8BI, ZEP-7B, MISH-7BD, MISH-7B, AYA-8B, SOL-10.7B, YI-9BK, YI-9B, L3S-8B, L3M-8BI, L3M-8B) whereas those in the L cluster were (PYT-2.8B, GPTN-2.7B, GEM-2B, GEMR-2B, PHI-1.5, TLM-1.1B, STA-6B, GPT-XL, GPT-2, FAL-7BI, FAL-7B). The L cluster with a relatively lower median performance profile consisted primarily of LLMs with smaller numbers of parameters. While earlier studies [27, 28] emphasized the impact of parameters on LLM performance, the present findings reiterated the empirical findings from a domain-specific standpoint.

For the DN taxonomy, the H and the L clusters comprised of 20 LLMs and 13 LLMs, respectively. Wilcoxon-Ranksum tests revealed statistically significant differences between these clusters for the performance benchmarks (IFEV, BBH, MATH, GPQA, MPRO) but not MUSR after controlling for family-wise error rate (i.e., $\alpha = 0.01/6 \sim 0.002$). The corresponding boxplots are shown in **Fig. 2**. LLMs in the H cluster were (MIS-7B, LM3S-8B, QW-14B, LM3A-8B, GEM-7BI, GEM-7B, QWN-7B, QWN-7BC, GEMS-7B, LM3S-8BI, ZEP-7B, MISH-7BD, MISH-7B, AYA-8B, SOL-10.7B, YI-9BK, YI-9B, L3S-8B, L3M-8BI, L3M-8B) whereas those in the L cluster were (MIS-7BI, PYT-2.8B, GPTN-2.7B, VIC-7B, GEM-2B, GEMR-2B, PHI-1.5,

TLM-1.1B, STA-6B, GPT-XL, GPT-2, FAL-7BI, FAL-7B). In line with earlier empirical observations, LLMs in the L cluster with relatively lower median performance benchmarks were predominantly smaller in size. There were also marked consensus in the distribution of the LLMs between the L (11 LLMs) and H (20 LLMs) clusters between the DS and DN taxonomies, revealing the inherent robustness of the LLMs across generic as well as domain specific tasks. The variance for a majority of the performance benchmarks was markedly higher for the H cluster as opposed to the L cluster across DS and DN taxonomies indicating considerable heterogeneity in the performance of the LLMs in the H cluster. For the DS taxonomy, the model with the highest performance benchmarks (MCQA, MQA, ANAT, CLIN, BIOL, CMED, GEN, PMED, PUBM) was different implementations of LLaMA-3 with 8 billion parameters. For the DN taxonomy, there were considerable differences in the LLMs with the highest performance across the various benchmarks (MATH: Qwen1.5 with 14 billion parameters; IFEV and GPQA: LLaMA-3 with 8 billion parameters; BBH: SOLAR with 10.7 billion parameters; MUSR: GPT-2 with 355 million parameters; MPRO: Yi-1.5 with 9 billion parameters).

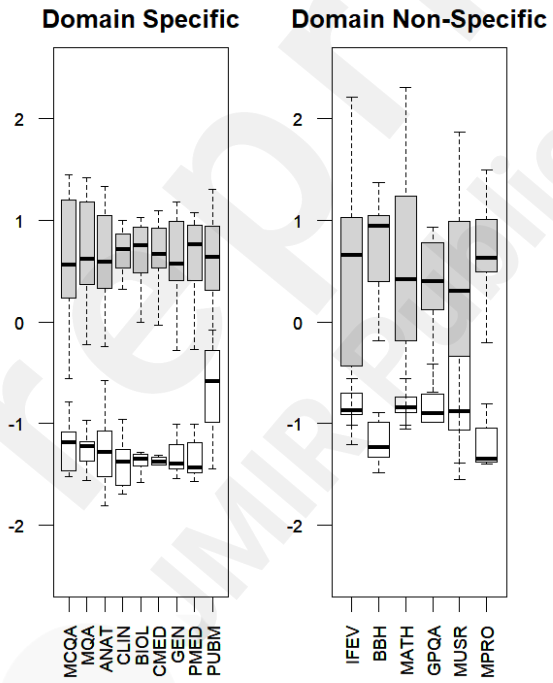

Figure 2 Box-whisker plots of the scaled performance benchmark profiles corresponding to H (gray) and L (white) clusters of the domain-specific (left) and the domain non-specific (right) taxonomies. Outliers are not shown for clarity.

Pearson-correlation was used to determine statistically significant correlations between pair-wise performance benchmarks of the DS and DN taxonomies, **Fig. 3**. For the DS taxonomy, the significance level was chosen as (i.e., $\alpha = 0.01/36 \sim 0.0002$) after controlling for family-wise error rate. The linear trend was especially pronounced between MCQA, MQA and MMLUs for various subjects. A possible explanation is LLMs that perform well on the different medically related MMLU subject-wise benchmarks (ANAT, CLIN, BIOL, CMED, GEN, PMED) may also

perform well on comprehensive medical exams (MCQA, MQA). While the linear correlation between the MMLU subject-wise benchmarks was statistically significant, the scatter plots revealed instances of sparse distribution of points along the linear trend line for some of these pairs (e.g., BIOL, PMED), challenging reliable correlation estimates, **Fig. 3**. The clustering of points about the linear trend line may in fact align with earlier observations of statistical differences, **Fig. 2**, between two broad groups of LLMs within the DS taxonomy, **Figs. 1** and **2**. The magnitude of performance benchmark (PUBM), that assesses the LLMs ability to comprehend and reason biomedical literature, did not exhibit a strong correlation with others as reflected by lack of a clear linear trend in the scatter plots, **Fig. 3**. In contrast to DS taxonomy, the correlation structures between the performance benchmarks was noisy in the case of DN taxonomy, as reflected by the scatter plots in **Fig. 3**. While there were instances of DN performance benchmark pairs (MPRO, BBH) with significant correlation, the redundancy was markedly lower than that observed in the case of DS taxonomy. Therefore, the DN performance benchmarks may assess the LLMs complementary characteristics.

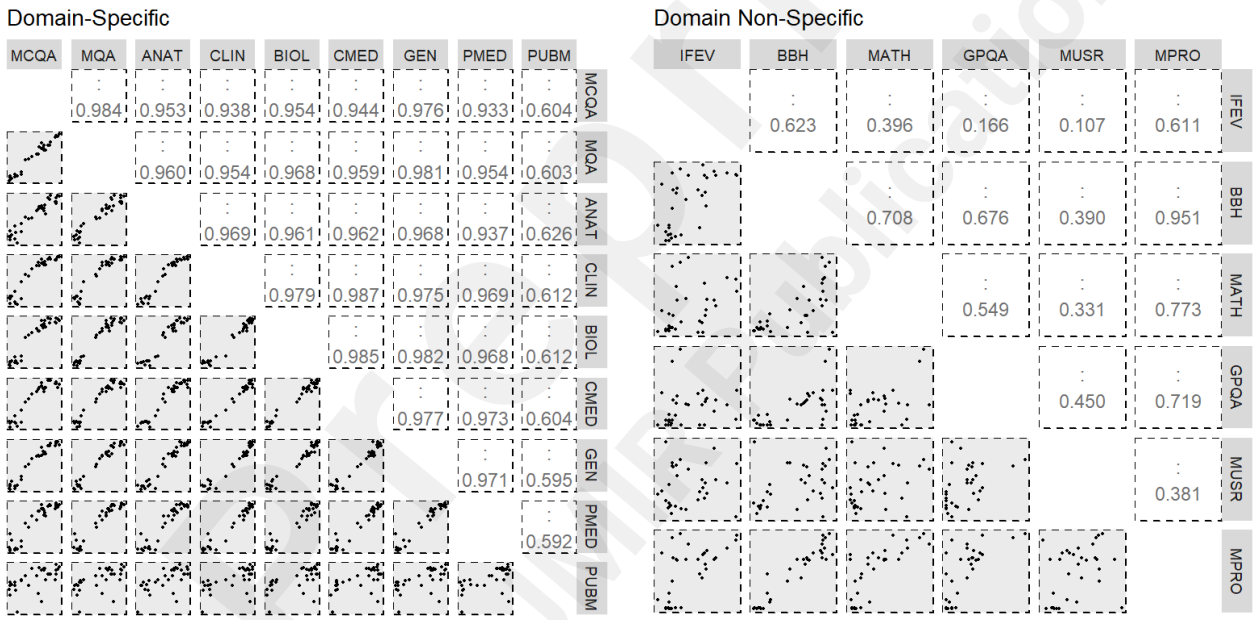

Figure 3 Scatter plots (gray panels) representing the correlation structure between the performance benchmarks corresponding to DS (left) and DN (right) taxonomies are shown. The pair-wise linear correlation (white panels) of the corresponding (row, column) pairs of performance benchmarks are shown in the upper triangle. The diagonals represent auto-correlation, hence not presented.

Discussion

LLMs continue to witness increasing adoption in healthcare. Popular LLM healthcare applications broadly fall under conversational AI and include chatbots, summarization, and translation tools. A critical characteristic that impacts LLM adoption is their performance. Several performance benchmarks have been proposed in the literature for objective assessment of domain-specific and domain non-specific LLMs. Open-source AI platforms such as Hugging Face provide critical performance benchmarks of widely used LLMs

through Open Medical LLM and Open LLM initiatives with enhanced transparency for an unbiased assessment of these models. These performance benchmarks are routinely updated across leaderboards directly impacting LLM adoption by the broader community including healthcare.

In this study, taxonomy portraits of medical LLMs were generated from domain-specific and domain non-specific performance benchmarks in conjunction with hierarchical clustering. The hierarchical cluster revealed two broad clusters whose performance benchmarks were statistically significant, indicating a natural partition of the medical LLMs based on their performance. As with some of the earlier empirical studies on LLM scaling laws, the cluster with lower performance mostly consisted of smaller LLMs. Interestingly, there was a significant overlap between the DS and DN taxonomies, revealing the inherent robustness of these LLMs across diverse tasks. Pair-wise correlation signatures of the performance benchmarks also revealed significant redundancies between the performance benchmarks for the domain specific taxonomy in contrast to domain non-specific taxonomy. Eliminating redundant evaluations can assist in lowering the costs while maintaining accuracy in model selection. Therefore, understanding the hierarchical organization and grouping of medical LLMs into clusters and inherent redundancies between the performance benchmarks is expected to assist from cost, performance, affordability standpoints, and providing preliminary cues to optimal LLM implementation pathways. While the present study focused on medical LLMs from Hugging Face whose performance benchmarks were available for domain-specific and domain non-specific tasks, a more detailed analysis including contemporary performance benchmarks in conjunction with a larger family of medical LLMs is required to generate comprehensive taxonomy portraits that could assist in their successful and strategic adoption. However, such an effort would implicitly demand enhanced transparency on various LLM characteristics such as performance, economics, as well as details on their implementation across on-prem and cloud infrastructures.

Acknowledgements

RN conceived the presented idea and carried out the analysis. All authors contributed to the writing and review of the manuscript.

References

1. Shah, N.H., D. Entwistle, and M.A. Pfeffer, *Creation and adoption of large language models in medicine.* Jama, 2023. **330**(9): p. 866-869.
2. Denecke, K., et al., *Potential of Large Language Models in Health Care: Delphi Study.* Journal of Medical Internet Research, 2024. **26**: p. e52399.
3. Sezgin, E., *Redefining virtual assistants in health care: the future with large language models*. 2024, JMIR Publications Toronto, Canada. p. e53225.
4. Van Veen, D., et al., *Adapted large language models can outperform medical experts in clinical text summarization.* Nature medicine, 2024. **30**(4): p. 1134-1142.
5. Duggan, M.J., et al., *Clinician experiences with ambient scribe technology*

to assist with documentation burden and efficiency. JAMA Network Open, 2025. **8**(2): p. e2460637-e2460637.
6. Zaretsky, J., et al., *Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format.* JAMA network open, 2024. **7**(3): p. e240357-e240357.
7. Wornow, M., et al., *Zero-shot clinical trial patient matching with llms.* NEJM AI, 2025. **2**(1): p. AIcs2400360.
8. Klang, E., et al., *A strategy for cost-effective large language model use at health system-scale.* NPJ digital medicine, 2024. **7**(1): p. 320.
9. Nagarajan, R., et al., *Economics and Equity of Large Language Models: Health Care Perspective.* Journal of Medical Internet Research, 2024. **26**: p. e64226.
10. Hu, T. and X.-H. Zhou, *Unveiling LLM Evaluation Focused on Metrics: Challenges and Solutions.* arXiv preprint arXiv:2404.09135, 2024.
11. HuggingFace. *Open LLM Leaderboard: Performance Evaluation*. 2024; Available from: https://huggingface.co/docs/leaderboards/en/open_llm_leaderboard/about.
12. HuggingFace. *Hugging Face Open Medical LLM Leaderboard*. 2024; Available from: https://huggingface.co/blog/leaderboard-medicalllm.
13. Riedemann, L., M. Labonne, and S. Gilbert, *The path forward for large language models in medicine is open.* npj Digital Medicine, 2024. **7**(1): p. 339.
14. HuggingFace. *Hugging Face: Evaluate Documentation: A Quick Tour*. 2025; Available from: https://huggingface.co/docs/evaluate/main/en/a_quick_tour.
15. Ly, D.P., A.K. Jha, and A.M. Epstein, *The association between hospital margins, quality of care, and closure or other change in operating status.* Journal of general internal medicine, 2011. **26**: p. 1291-1296.
16. Shekar, S., et al., *People over trust AI-generated medical responses and view them to be as valid as doctors, despite low accuracy.* arXiv preprint arXiv:2408.15266, 2024.
17. Pal, A., L.K. Umapathi, and M. Sankarasubbu. *Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering*. in *Conference on health, inference, and learning*. 2022. PMLR.
18. Hendrycks, D., et al., *Measuring massive multitask language understanding.* arXiv preprint arXiv:2009.03300, 2020.
19. Jin, Q., et al., *Pubmedqa: A dataset for biomedical research question answering.* arXiv preprint arXiv:1909.06146, 2019.
20. Zhou, J., et al., *Instruction-following evaluation for large language models.* arXiv preprint arXiv:2311.07911, 2023.
21. Suzgun, M., et al., *Challenging big-bench tasks and whether chain-of-thought can solve them.* arXiv preprint arXiv:2210.09261, 2022.
22. Hendrycks, D., et al., *Measuring mathematical problem solving with the math dataset.* arXiv preprint arXiv:2103.03874, 2021.
23. Rein, D., et al., *Gpqa: A graduate-level google-proof q&a benchmark.* arXiv preprint arXiv:2311.12022, 2023.
24. Sprague, Z., et al., *Musr: Testing the limits of chain-of-thought with multistep soft reasoning.* arXiv preprint arXiv:2310.16049, 2023.
25. Wang, Y., et al., *Mmlu-pro: A more robust and challenging multi-task*

language understanding benchmark. arXiv preprint arXiv:2406.01574, 2024.
26. Johnson, R.A. and D.W. Wichern, *Applied Multivariate Statistical Analysis*. 2023: Pearson.
27. Kaplan, J., et al., *Scaling laws for neural language models.* arXiv preprint arXiv:2001.08361, 2020.
28. Thirunavukarasu, A.J., et al., *Large language models in medicine.* Nature medicine, 2023. **29**(8): p. 1930-1940.