JMIR Publications
The Leading eHealth Publisher

Evaluation of the Performance of 3 Large Language Models in Clinical Decision Support: A Comparative Study Based on Actual Cases

Xueqi Wang, Haiyan Ye, Sumian Zhang, Mei Yang, Xuebin Wang

Submitted to: Journal of Medical Internet Research
on: May 23, 2024

Preprint
JMIR Publications

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Preprint JMIR Publications

Evaluation of the Performance of 3 Large Language Models in Clinical Decision Support: A Comparative Study Based on Actual Cases

Xueqi Wang[1] MD, MSc; Haiyan Ye[1] MD, MSc; Sumian Zhang[1] MD, MSc; Mei Yang[1] MD, MSc; Xuebin Wang[1] MD, PhD

[1]Department of Critical Care Medicine, Shanghai East Hospital Tongji University School of Medicine Shanghai CN

Corresponding Author:
Xuebin Wang MD, PhD
Department of Critical Care Medicine, Shanghai East Hospital
Tongji University School of Medicine
No.150, Jimo Road, Pudong New Area
Shanghai
CN

Abstract

Background: Generative large language models (LLMs) are increasingly integrated into the medical field. However, their actual efficacy in clinical decision-making remains partially unexplored.

Objective: This study evaluated the diagnostic and therapeutic capabilities of 3 LLMs (ChatGPT-4, Gemini and Med-Go) in addressing real clinical cases.

Methods: This study involved 134 clinical cases spanning 9 medical disciplines. The LLMs evaluated were ChatGPT-4, Gemini and Med-Go. Each LLM was required to provide suggestions for diagnosis, diagnostic criteria, differential diagnosis, examination and treatment for every case. Responses were scored by 2 experts using a predefined rubric.

Results: In overall performance among the models, Med-Go achieved the highest median score (37.5, IQR 31.9-41.5), while Gemini recorded the lowest (33.0, IQR 25.5-36.6), showing significant statistical difference among the 3 LLMs ($p < 0.001$). Analysis revealed that responses related to differential diagnosis were the weakest, while those pertaining to treatment recommendations were the strongest. Med-Go displayed notable performance advantages in gastroenterology, nephrology, and neurology.

Conclusions: The findings show that all 3 LLMs achieved over 60% of the maximum possible score, indicating their potential applicability in clinical practice. However, inaccuracies that could lead to adverse decisions underscore the need for caution in their application. Med-Go's superior performance highlights the benefits of incorporating specialized medical knowledge into LLMs training. It is anticipated that further development and refinement of medical LLMs will enhance their precision and safety in clinical use.

(JMIR Preprints 23/05/2024:60796)
DOI: https://doi.org/10.2196/preprints.60796

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

Original Manuscript

Preprint
JMIR Publications

Evaluation of the Performance of 3 Large Language Models in Clinical Decision Support: A Comparative Study Based on Actual Cases

Introduction

The integration of Artificial Intelligence (AI) in the medical field represents a significant paradigm shift, offering new possibilities for enhancing diagnostic and therapeutic processes [1 2]. Emerging as one of the most innovative AI applications, Generative Large Language Models (LLMs) hold immense promise for revolutionizing diverse aspects of healthcare [3]. These advanced models can facilitate clinical decision-making by supporting a diverse range of applications, including aiding diagnostic and therapeutic processes [4], providing health-related question answering [5], and optimizing patient care pathways [6].

As a leading example of LLMs, ChatGPT-4 [7], developed by OpenAI, has been trained on a massive dataset of text and code to excel in generating human-quality text, translating languages, and engaging in open-ended conversations [8]. The capabilities have increasingly captured interest regarding their potential role in medicine. Recent studies have shown that ChatGPT-4 can pass medical exams across various countries [9 10 11] and effectively answer medical questions across multiple disciplines [12 13 14]. Additionally, it has demonstrated strengths in composing medical documents, further underlining its utility in healthcare settings [15 16].

The landscape of LLMs is rapidly evolving, with a surge of new models emerging post-ChatGPT. Gemini [17], from Google AI, showcases versatility in creative text generation, coding, and answering questions informatively. Med-Go [18], developed jointly by the Institute of Software Chinese Academy of Sciences and East Hospital Affiliated To Tongji University, employs a transformer architecture trained with extensive medical corpora to provide tailored medical advice. LLMs are emerging as powerful tools in healthcare, demonstrating the potential to generate informed medical advice and propose healthcare solutions. This newfound potential is revolutionizing healthcare with its transformative capabilities [19].

Researchers have dedicated extensive efforts to evaluating the effectiveness of LLMs in clinical applications. In an investigation into the evaluation ChatGPT’s effectiveness across the full spectrum of clinical workflow, it reveals that ChatGPT can perform with notable accuracy, particularly in final diagnosis [20]. A recent comparative study by Wilhelm employed a straightforward question-based approach to evaluate the efficacy of various LLMs in generating medical content across 3 specialties. They uncovered Claude's superior performance in a mean score comparison [21] . A comparative analysis evaluated the performance of several LLMs, including ChatGPT, Google Bard, and Microsoft Bing Chat, in their ability to support evidence-based dentistry. While ChatGPT-4 exhibited the highest efficacy, all models were prone to errors such as inaccuracies and outdated information [22].

The potential of LLMs to aid clinical decision-making is attracting growing research focus. However, assessments based on structured dialogues may not fully capture the complexity the real-world clinical cases. This study assessed 3 LLMs, including 2 general-purpose models—ChatGPT and Gemini—and 1 specialized model, Med-Go, which had been fine-tuned with medical expertise. Using real clinical case texts, we assessed their ability to generate diagnoses, suggest examinations, and propose treatment plans. This comparative analysis aimed to illuminate the potential and limitations of using LLMs in clinical decision-making and discusses whether there is a necessity to develop specialized medical LLMs to enhance their utility and accuracy in healthcare settings.

Methods

Overview

This study involved 134 actual clinical cases from 9 medical disciplines, including Cardiology (12), Respiratory (13), Gastroenterology (39), Nephrology (13), Neurology (11), Endocrinology (6), Gynaecology (12), Pediatrics (12), and Others (16). The "Others" category encompasses cases related to toxicology, orthopedics, dermatology, hematology and others. The cases were derived from classic teaching cases as well as typical clinical cases. All patient-identifying information was removed to ensure confidentiality. Cases were posed to 3 generative LLMs: ChatGPT-4, Gemini, and Med-Go. Each case provided to the LLMs included a patient history, physical examination findings, laboratory results, and, when applicable, a textual description of imaging studies.

Each LLM was informed that the questioning would be conducted in Chinese, and all clinical case texts and queries were provided in Chinese, requiring responses in the same language. Each clinical case was presented using professional terminology, with medical terms appearing exactly as they would in a professional context, without any simplification or modification. Each question was posed only once by a single author to each LLM, with no follow-up questions, rephrasing, or additional explanations provided. At the end of each case text, the LLMs were prompted with the following instruction: "Please analyze the above medical record and provide a diagnosis, diagnostic criteria, differential diagnosis, further examinations, and a treatment plan."

Scoring

Each case was crafted by a physician with unique answers and scoring rubric based on the actual circumstances. 2 experienced clinicians with over 5 years of clinical practice independently scored the responses provided by the LLMs, without knowing which model generated the answers. Scoring was conducted for each of the 5 evaluation categories—diagnosis, diagnostic criteria, differential diagnosis, examination, and treatment—each rated out of 10, resulting in a total possible score of 50 per case. A full score of 10 points was awarded only if the response was entirely correct, while entirely incorrect answers received 0. If an LLM's diagnosis was incorrect, no point was awarded for the diagnostic criteria. The clinicians recorded the scores for each question accordingly. During the scoring process, potential grammatical issues in the responses were overlooked. Scoring was based solely on the correctness of the content provided in the answers. The final score for each case was calculated as the average of the scores given by the 2 clinicians.

Model setting

In this study, we utilized 3 models: ChatGPT-4, Gemini and Med-Go. ChatGPT-4 and Gemini are general-purpose LLMs, both of which have been trained with medical texts to enhance their ability to handle medical queries. ChatGPT-4, part of the advanced generative pre-trained transformer series, has been trained on a diverse dataset that includes extensive medical literature, enabling it to provide relevant and context-aware responses. Gemini has undergone rigorous training with a comprehensive medical dataset and employs a blend of machine learning techniques to continuously improve its capabilities through real-time updates and evaluations. On the other hand, Med-Go is a specialized medical LLM, specifically designed and trained for medical applications. It uses a transformer-based architecture and is fine-tuned with a substantial corpus of medical texts in Chinese, making it highly effective at delivering accurate medical advice and uniquely adapted to handle complex medical scenarios.

This study did not involve any form of training or fine-tuning of the models. Our approach was to use the models as they are, utilizing their ability to generate responses to questions in their current state.

Data analysis

All data were compiled and summarized in Microsoft Excel 2021, version, 2403 (Microsoft Corp., Washington, DC, USA), with calculations for mean, median, quartiles, and standard deviation. Statistical analyses were subsequently conducted using IBM SPSS Statistics, version 25 (International Business Machines Corp., Armonk, NY, USA). The normality of data was assessed with the Kolmogorov-Smirnov test. For data not meeting normality assumptions, non-parametric tests were used to analyze score differences between groups. The significance threshold was set at $p < 0.05$. ChiPlot (https://www.chiplot.online/) (accessed on 15 April 2024) was used for data visualization.

Results

Overview

The overall score distribution is shown in Figure 1. In a comprehensive evaluation of overall question performance, Med-Go achieved the highest mean score at 35.9 out of a maximum 50 points, followed by ChatGPT-4 at 33.3, and Gemini at 31.0. Med-Go also led with a median score of 37.5 (IQR 31.9-41.5), compared to ChatGPT-4's 35.0 (IQR 28.5-39.5) and Gemini's 33.0 (IQR 25.5-36.6). The Kruskal-Wallis Test revealed statistically significant difference among the 3 LLMs ($p < 0.001$) (Table 1).

Figure 1. Distribution of overall scores for ChatGPT, Gemini, and Med-Go.

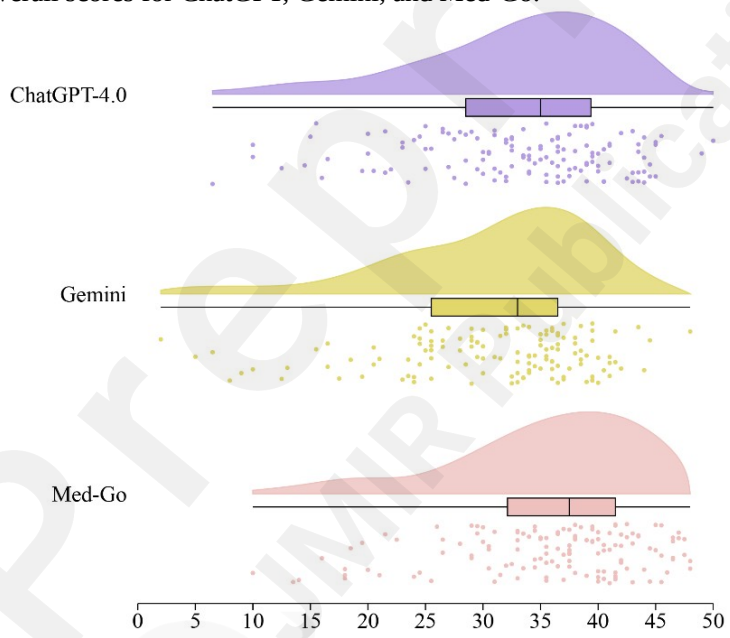


Table 1. Descriptive statistics and performance difference among 3 LLMs.

	ChatGPT-4	Gemini	Med-Go	p-value
Minimum	6.5	2.0	10.0	-
Maximum	50.0	48.0	48.0	-
Mean	33.3	31.0	35.9	-
Median (IQR)	35.0 (28.5-39.5)	33.0 (25.5-36.6)	37.5 (31.9-41.5)	< 0.001

IQR, interquartile range.

Since the scores for the 3 groups did not meet the assumption of normality, pairwise comparisons were conducted using the Mann-Whitney U test. These comparisons revealed that there is a statistical difference between the median scores of ChatGPT-4 and Gemini ($Z = -2.277$, $p = 0.023$), and statistically significant difference were noted between the ChatGPT-4 and Med-Go ($Z = -2.746$, $p = 0.006$), as well as between the Gemini and Med-Go ($Z=-5.130$, $p < 0.001$) (Table 2).

Table 2. Mann-Whitney U Test for the median scores for the answers provided by the 3 LLMs.

Group	*Z*-value	*p*-value
ChatGPT-4 vs. Gemini	-2.277	0.023
ChatGPT-4 vs. Med-Go	-2.746	0.006
Gemini vs. Med-Go	-5.130	< 0.001

Performance Analysis of LLMs Across 5 Assessment Categories

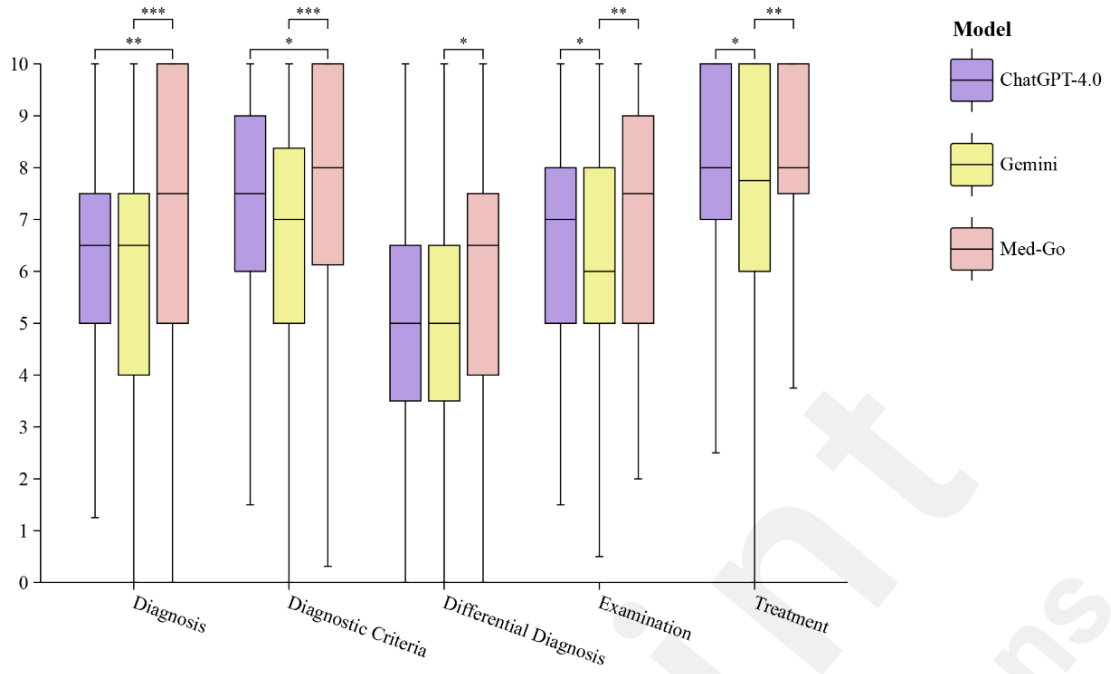
Across the 5 assessment categories, the 3 LLMs generally performed best in treatment recommendations, with median scores of 8.0 for ChatGPT-4, 7.5 for Gemini, and 8.0 for Med-Go. In contrast, their weakest performance was in differential diagnosis, with median scores of 5.0 for both ChatGPT-4 and Gemini, and 6.5 for Med-Go (Table 3). As illustrated in Figure 2, there were notable differences in the performance of each LLM across all categories. Med-Go demonstrated superior performance in diagnosis, showing a statistically significant difference from ChatGPT-4 ($p = 0.008$) and Gemini ($p < 0.001$). Similarly, in diagnostic criteria, Med-Go outperformed the others, with statistical difference from ChatGPT-4 ($p = 0.033$) and Gemini ($p < 0.001$). In differential diagnosis, although all 3 LLMs performed relatively poorly, Med-Go still showed a statistical difference over Gemini ($p = 0.011$). For the further examinations, statistical difference was noted between ChatGPT-4 and Gemini ($p = 0.019$), as well as between Med-Go and Gemini ($p = 0.003$). Treatments also highlighted statistical difference, with both ChatGPT-4 and Med-Go showing superior performance compared to Gemini ($p = 0.041$ and $p = 0.005$) (Table 3).

Table 3. The median scores for the answers provided by the 3 LLMs to 5 assessment categories.

Project	Model	Median (IQR)	*p*-value	
			vs. Gemini	vs. Med-Go
Dx.	ChatGPT-4	6.5 (5.0-7.5)	0.255	0.008
	Gemini	6.5 (4-7.5)	-	<0.001
	Med-Go	7.5 (5.0-10.0)	-	-
DC.	ChatGPT-4	7.5 (7.0-9.0)	0.080	0.033
	Gemini	7.0 (5.0-8.5)	-	<0.001
	Med-Go	8.0 (6.0-10.0)	-	-
DDx.	ChatGPT-4	5.0 (3.5-6.5)	0.552	0.050
	Gemini	5.0 (3.5-6.5)	-	0.011
	Med-Go	6.5 (4-7.5)	-	-
Ex.	ChatGPT-4	7.0 (5.0-8.0)	0.019	0.310
	Gemini	6.0 (5.0-8.0)	-	0.003
	Med-Go	7.5 (5.0-9.0)	-	-
Tx.	ChatGPT-4	8.0 (6.3-10.0)	0.041	0.427
	Gemini	7.5 (6.0-10.0)	-	0.005
	Med-Go	8.0 (7.5-10.0)	-	-

Dx., Diagnosis; DC., Diagnostic Criteria; DDx., Differential Diagnosis; Ex., Examination; Tx., Treatment; IQR, interquartile range; Repeated LLMs' comparisons were indicated by "-".

Figure 2. 3 LLMs performance on 5 assessment categories. Med-Go outperformed ChatGPT-4 and Gemini significantly in diagnosis and diagnostic criteria. Although all LLMs struggled with differential diagnosis, Med-Go still performed better than Gemini. In examination and treatment, both ChatGPT-4 and Med-Go showed superior performance compared to Gemini ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).



Performance Analysis of LLMs Across 9 Disciplines

The results detail performance variations among the 3 LLMs across 9 different medical disciplines. ChatGPT-4 scored highest in the "Others" category (Median 37.8, IQR 31.8-43.8) and nephrology (Median 37.5, IQR 34-40.0) but had the lowest scores in pediatrics (Median 28.5, IQR 17.5-35.4). Gemini achieved its highest scores in neurology (Median 36.5, IQR 32.0-39.0), followed by "Others" (Median 35.3, IQR 31.9-39.3) and, like ChatGPT-4, performed worst in pediatrics (Median 24.5, IQR 17.0-37.8). Med-Go excelled in neurology (Median 41.5, IQR 40.0-47.0), with strong scores in respiratory (Median 41.0, IQR 30.8-42.8), but also scored lowest in pediatrics (Median 29.3, IQR 24.6-39.8), indicating a common challenge in pediatric cases across 3 LLMs (Table 4).

In the comparison within the same discipline, statistically significant difference was observed in gastroenterology between the median scores of Med-Go and Gemini (p = 0.008) (Figure 3A). In nephrology, Med-Go also showed statistical difference compared to Gemini (p = 0.029) (Figure 3B). In neurology, statistical difference was noted between Med-Go and both ChatGPT-4 (p = 0.015) and Gemini (p = 0.016) (Figure 3B). In other disciplines, no statistical difference weas observed among the 3 LLMs, suggesting a more uniform performance across those fields.

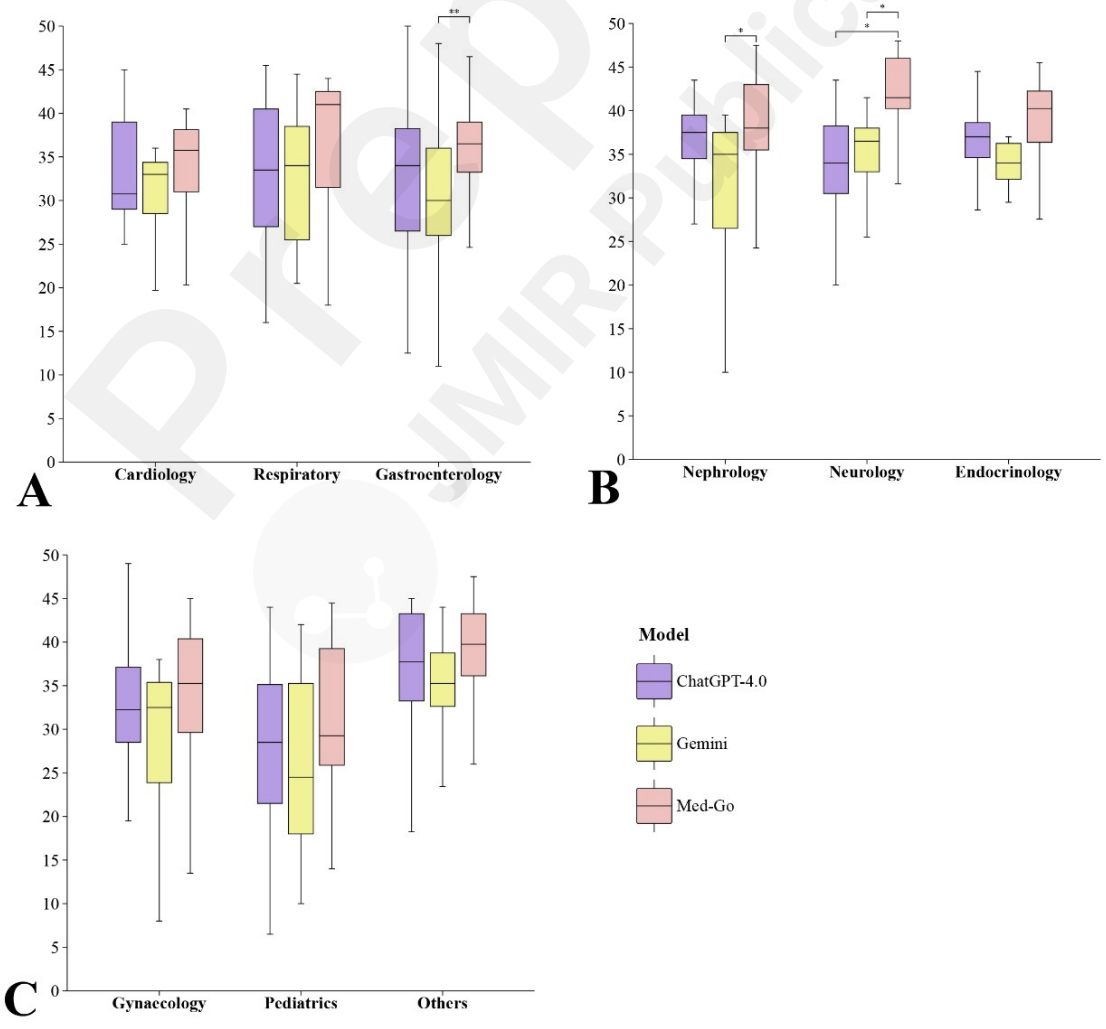
Table 4. The median scores for the answers provided by the 3 LLMs to 9 disciplines.

Subject	Model	Median (IQR)	*p*-value	
			vs. Gemini	vs. Med-Go
Card.	ChatGPT-4	30.8 (28.0-39.0)	0.435	0.524
	Gemini	33.0 (26.5-35.1)	-	0.069
	Med-Go	35.8 (31.0-39.4)	-	-
Resp.	ChatGPT-4	33.5 (25.5-41.8)	0.797	0.369
	Gemini	34 (25.0-39.0)	-	0.123
	Med-Go	41.0 (30.8-42.8)	-	-
GI.	ChatGPT-4	34 (25.5-38.5)	0.299	0.169
	Gemini	30.0 (25.5-36.0)	-	0.008
	Med-Go	36.5 (33.0-39.5)	-	-
Neph.	ChatGPT-4	37.5 (34-40.0)	0.085	0.281
	Gemini	35.0 (25.3-37.8)	-	0.029
	Med-Go	38.0 (35.5-44.3)	-	-

	ChatGPT-4	34 (30.0-40.0)	0.758	0.015
Neuro.	Gemini	36.5 (32.0-39.0)	-	0.016
	Med-Go	41.5 (40.0-47.0)	-	-
	ChatGPT-4	37.0 (31.3-40.4)	0.558	0.662
Endo.	Gemini	34 (31.4-36.7)	-	0.298
	Med-Go	40.3 (32.3-43.3)	-	-
	ChatGPT-4	32.3 (28.5-38.4)	0.506	0.623
Gyn.	Gemini	32.5 (23.6-36.1)	-	0.216
	Med-Go	35.3 (22.9-42.1)	-	-
	ChatGPT-4	28.5 (17.5-35.4)	0.779	0.487
Peds.	Gemini	24.5 (17.0-37.8)	-	0.297
	Med-Go	29.3 (24.6-39.8)	-	-
	ChatGPT-4	37.8 (31.8-43.8)	0.258	0.417
Others	Gemini	35.3 (31.9-39.3)	-	0.059
	Med-Go	39.8 (32.4-44.8)	-	-

Card., Cardiology; Resp., Respiratory; GI., Gastroenterology; Neph., Nephrology; Neuro., Neurology; Endo., Endocrinology; Gyn., Gynaecology; Peds., Pediatrics; IQR, interquartile range; Repeated LLMs' comparisons were indicated by "-".

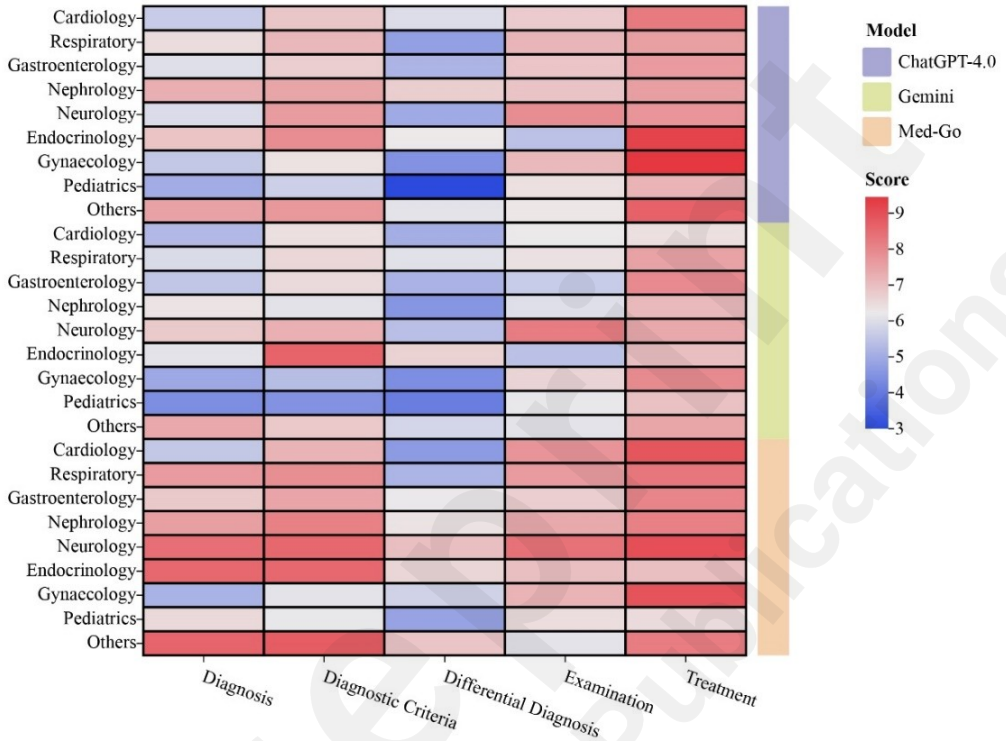
Figure 3. Performance of the 3 LLMs across 9 medical disciplines. In gastroenterology (A) and nephrology (B), Med-Go's median scores showed significant difference from Gemini's. In neurology, Med-Go showed significant difference compared to both ChatGPT and Gemini (B). In other disciplines, the 3 LLMs exhibited no statistical difference ($^{*}p < 0.05$, $^{**}p < 0.01$).



Performance Analysis of LLMs Across All Types

Figure 4 displays the performance of the 3 LLMs across 9 medical disciplines and 5 question categories, using average scores. In comparisons among the LLMs, Med-Go consistently scored higher, while Gemini's scores were relatively lower. Across different assessment categories, scores were lowest for differential diagnosis and highest for treatment. The lowest score was observed in ChatGPT-4's differential diagnosis in pediatrics, and the highest in ChatGPT-4's treatment in gynaecology.

Figure 4. The comparison of scores for 3 LLMs—ChatGPT-4, Gemini, and Med-Go—across 9 medical disciplines and 5 question categories.



Discussion

Major Findings

This study evaluated the diagnostic and management capabilities of 3 LLMs: ChatGPT-4, Gemini, and Med-Go. All 3 models achieved an overall accuracy exceeding 60% on real clinical cases, demonstrating a foundational proficiency in clinical tasks. However, the potential for inaccurate responses remains a challenge. Notably, Med-Go, specifically fine-tuned with medical expertise, outperformed both ChatGPT-4 and Gemini, with statistically significant difference.

The performance of the 3 LLMs across 5 question categories was best in treatment recommendations. During the scoring by 2 experienced clinicians, it was noted that for time-sensitive conditions such as tension pneumothorax or visceral organ perforation, the LLMs were able to provide direct prompts like "perform immediately". Importantly, under correct diagnostic circumstances, the LLMs rarely suggested harmful treatments. This tendency led to more conservative recommendations, such as the cautious suggestion of procedures like bone marrow biopsies in hematological cases. This conservative approach in clinical applications is not necessarily negative, as the implementation of invasive procedures should ideally be proposed by professional doctors.

It is important to note that in the treatment question section, the assessment focused solely on the correctness of the treatment plans, not considering medication dosages. Consequently, when scoring,

the precise dosing details were not evaluated. However, the researchers observed that none of the 3 LLMs mentioned specific treatment medication dosages in their responses without prompts. This omission likely relates to the content of the training texts. To enable LLMs to provide detailed medication plans, it would be necessary to employ more directive prompts during questioning. Additionally, incorporating more professional medical guidelines and pharmacological texts into the training materials could further refine the models' capabilities in this aspect.

The weakest performance of the LLMs was observed in differential diagnosis, which may reflect deficiencies in the training datasets regarding the interconnectedness of diseases or an intrinsic limitation of the models in simulating the nuanced decision-making process of clinical diagnoses. This highlights a critical area for improvement in training LLMs to better handle the complexity of clinical diagnostics and underscores the importance of integrating these tools with professional oversight in clinical settings.

Across various medical disciplines, each of the 3 LLMs exhibited strengths in different areas, yet all performed poorly in pediatrics, particularly when the patient's age was explicitly stated at the start of each case. This underperformance likely reflects a gap in the training process, where distinctions between adult and pediatric diseases were not adequately captured. Statistically significant differences were noted in the performance of the LLMs in gastroenterology, nephrology, and neurology. However, no significant difference was observed in other disciplines, suggesting that for certain disease areas with distinct features, the general medical knowledge embedded in universal LLMs can be effectively utilized. This highlights the potential for these technologies to serve as assistive tools, even without extensive specialization. Nonetheless, for high-risk environments or specialized fields, specifically training models to enhance their performance is crucial.

The Importance of Developing Specialized Medical LLMs

The notable variance in performance across different medical disciplines and categories highlights the limitations of general LLMs when faced with complex medical scenarios that require specific knowledge and nuanced understanding. Med-Go’s superior performance underscores the value of specialized training and fine-tuning using domain-specific datasets. This model not only performed better overall but also demonstrated a better grasp of intricate medical details necessary for accurate diagnostics and effective treatment plans.

Previous research has demonstrated the advantages of developing specialized medical LLMs. For instance, Google’s Med-PaLM2 [23], a medical LLM, was fine-tuned and optimized using multiple medical QA training datasets. This fine-tuning significantly enhanced the model's medical reasoning capabilities. Similarly, ChatDoctor [24], a model fine-tuned on medical domain knowledge atop the LLM LLaMA, outperformed ChatGPT in terms of accuracy and recall in medical patient dialogues.

This suggests that LLMs, when trained with extensive medical literature and adjusted for particular clinical contexts, can achieve higher accuracy and reliability. It underlines the potential for these models to become more useful tools in clinical settings when they are tailored to specific medical knowledge and practices.

Challenges in Medical Applications of LLMs

Despite the significant potential of LLMs to assist in various medical tasks, their application in healthcare still faces critical challenges concerning accuracy and safety. The fact that even the best-performing LLM in this study, Med-Go, as well as the others, still face issues with inaccuracies points to inherent limitations of current LLM technologies in handling complex medical information. This might be due to the variability and subtleties of medical data, which require not just factual

recall but deep understanding and reasoning.

Moreover, LLMs lack the ability to understand context in the way human clinicians do, which is crucial for making nuanced medical judgments. They are also unable to perform physical examinations or consider emotional and non-verbal cues from patients, which are often vital in making accurate medical assessments.

The integration of LLMs into clinical practice must therefore be approached with caution. Robust validation processes, ongoing monitoring, and the development of clear guidelines for use are essential to address these challenges. It is also crucial to ensure that the final clinical decisions always lie with trained healthcare providers who can interpret LLM's outputs with professional judgment and a deep understanding of patient context.

Limitations

A limitation of our study might be that each question was posed without additional prompts. The lack of explanations for specific medical terminology or additional contextual information could have influenced the LLMs' outputs. However, since all 3 LLMs were tested under these consistent conditions, this factor is less significant for cross-model comparisons. Nonetheless, in practical applications, providing LLMs with more information and context could potentially enhance the accuracy of their conclusions [25]. Additionally, the clinical cases used in this study were historical cases with established diagnoses and optimized treatment plans. In contrast, the complexity of clinical environments often means that patient diagnosis and treatment evolve over time. Future research could explore how LLMs perform in dynamic clinical situations, providing a clearer reflection of real-world conditions and furthering the understanding of LLM capabilities in ongoing medical contexts.

Conclusion

This study indicates that while LLMs demonstrate potential in aiding medical diagnosis and treatment recommendations, those tailored with specific medical training outperform general-purpose models. This underscores the critical role of specialized medical training in enhancing the effectiveness of LLMs in clinical settings. Future applications of LLMs in healthcare should involve the development of specialized models for particular specialties or diseases, refining their use in specific clinical contexts to provide more precise diagnostic and therapeutic information.

Conflicts of Interest

None declared.

Data Availability

Any data not appearing in this paper are available from the corresponding author upon reasonable request.

Author contributions

Xuebin Wang conceptualized the study and developed the methodology. HY and SZ provided resources and scored all the responses. MY and Xueqi Wang performed all the experiments and curated the data. Xueqi Wang performed the analysis and wrote the original draft. HY and SZ contributed to the review and editing of the manuscript. Xueqi Wang and HY are shared first authors.

References

1. Wang F, Preininger A. AI in Health: State of the Art, Challenges, and Future Directions.

Yearb Med Inform. 2019/08/16 2019;28(01):016-026. doi:10.1055/s-0039-1677908
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. Jan 2019;25(1):44-56. doi:10.1038/s41591-018-0300-7
3. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med*. Jan 2022;28(1):31-38. doi:10.1038/s41591-021-01614-0
4. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of Cardiovascular Disease Prevention Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence Model. *JAMA*. 2023;329(10):842-844. doi:10.1001/jama.2023.1044
5. Betzler BK, Chen H, Cheng CY, et al. Large language models and their impact in ophthalmology. *Lancet Digit Health*. Dec 2023;5(12):e917-e924. doi:10.1016/s2589-7500(23)00201-7
6. Liu S, Wright AP, Patterson BL, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *J Am Med Inform Assoc*. Jun 20 2023;30(7):1237-1245. doi:10.1093/jamia/ocad072
7. OpenAI. Introducing ChatGPT. Accessed March 25, 2024. https://openai.com/blog/chatgpt
8. Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report. *arXiv preprint arXiv:230308774*. 2023;
9. Gilson A, Safranek CW, Huang T, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med Educ*. Feb 8 2023;9:e45312. doi:10.2196/45312
10. Yanagita Y, Yokokawa D, Uchida S, Tawara J, Ikusaka M. Accuracy of ChatGPT on Medical Questions in the National Medical Licensing Examination in Japan: Evaluation Study. *JMIR Form Res*. Oct 13 2023;7:e48023. doi:10.2196/48023
11. Wang X, Gong Z, Wang G, et al. ChatGPT Performs on the Chinese National Medical Licensing Examination. *J Med Syst*. Aug 15 2023;47(1):86. doi:10.1007/s10916-023-01961-0
12. Grünebaum A, Chervenak J, Pollet SL, Katz A, Chervenak FA. The exciting potential for ChatGPT in obstetrics and gynecology. *Am J Obstet Gynecol*. Jun 2023;228(6):696-705. doi:10.1016/j.ajog.2023.03.009
13. Potapenko I, Boberg-Ans LC, Stormly Hansen M, Klefter ON, van Dijk EHC, Subhi Y. Artificial intelligence-based chatbot patient information on common retinal diseases using ChatGPT. *Acta Ophthalmol*. Nov 2023;101(7):829-831. doi:10.1111/aos.15661
14. Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol*. Jul 2023;29(3):721-732. doi:10.3350/cmh.2023.0089
15. Singh S, Djalilian A, Ali MJ. ChatGPT and Ophthalmology: Exploring Its Potential with Discharge Summaries and Operative Notes. *Semin Ophthalmol*. Jul 2023;38(5):503-507. doi:10.1080/08820538.2023.2209166
16. Zhou Z. Evaluation of ChatGPT's Capabilities in Medical Report Generation. *Cureus*. Apr 2023;15(4):e37589. doi:10.7759/cureus.37589
17. Google. Welcome to the Gemini era. Accessed March 29, 2024. https://deepmind.google/technologies/gemini/#introduction
18. Med-Go, Go For Changes. Accessed March 25, 2024. https://www.med-go.cn/
19. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. *J Med Syst*. Mar 4 2023;47(1):33. doi:10.1007/s10916-023-01925-4
20. Rao A, Pang M, Kim J, et al. Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study. *J Med Internet Res*. Aug 22 2023;25:e48659. doi:10.2196/48659
21. Wilhelm TI, Roos J, Kaczmarczyk R. Large Language Models for Therapy

Recommendations Across 3 Clinical Specialties: Comparative Study. *J Med Internet Res*. Oct 30 2023;25:e49324. doi:10.2196/49324
22. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the Performance of Generative AI Large Language Models ChatGPT, Google Bard, and Microsoft Bing Chat in Supporting Evidence-Based Dentistry: Comparative Mixed Methods Study. *J Med Internet Res*. Dec 28 2023;25:e51580. doi:10.2196/51580
23. Singhal K, Tu T, Gottweis J, et al. *Towards Expert-Level Medical Question Answering with Large Language Models*. 2023.
24. Yunxiang L, Zihan L, Kai Z, Ruilong D, You Z. *ChatDoctor: A Medical Chat Model Fine-tuned on LLaMA Model using Medical Domain Knowledge*. 2023.
25. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. *Nips '22*. 2024;

Abbreviations
AI: artificial intelligence
Card.: Cardiology
DC.: Diagnostic Criteria
DDx.: Differential Diagnosis
Dx.: Diagnosis
Endo.: Endocrinology
Ex.: Examination
GPT: generative pretrained transformer
GI.: Gastroenterology
Gyn.: Gynaecology
IQR: interquartile range
LLMs: large language models
Neph.: Nephrology
Neuro.: Neurology
Peds.: Pediatrics
Resp.: Respiratory
Tx.: Treatment

Preprint JMIR Publications

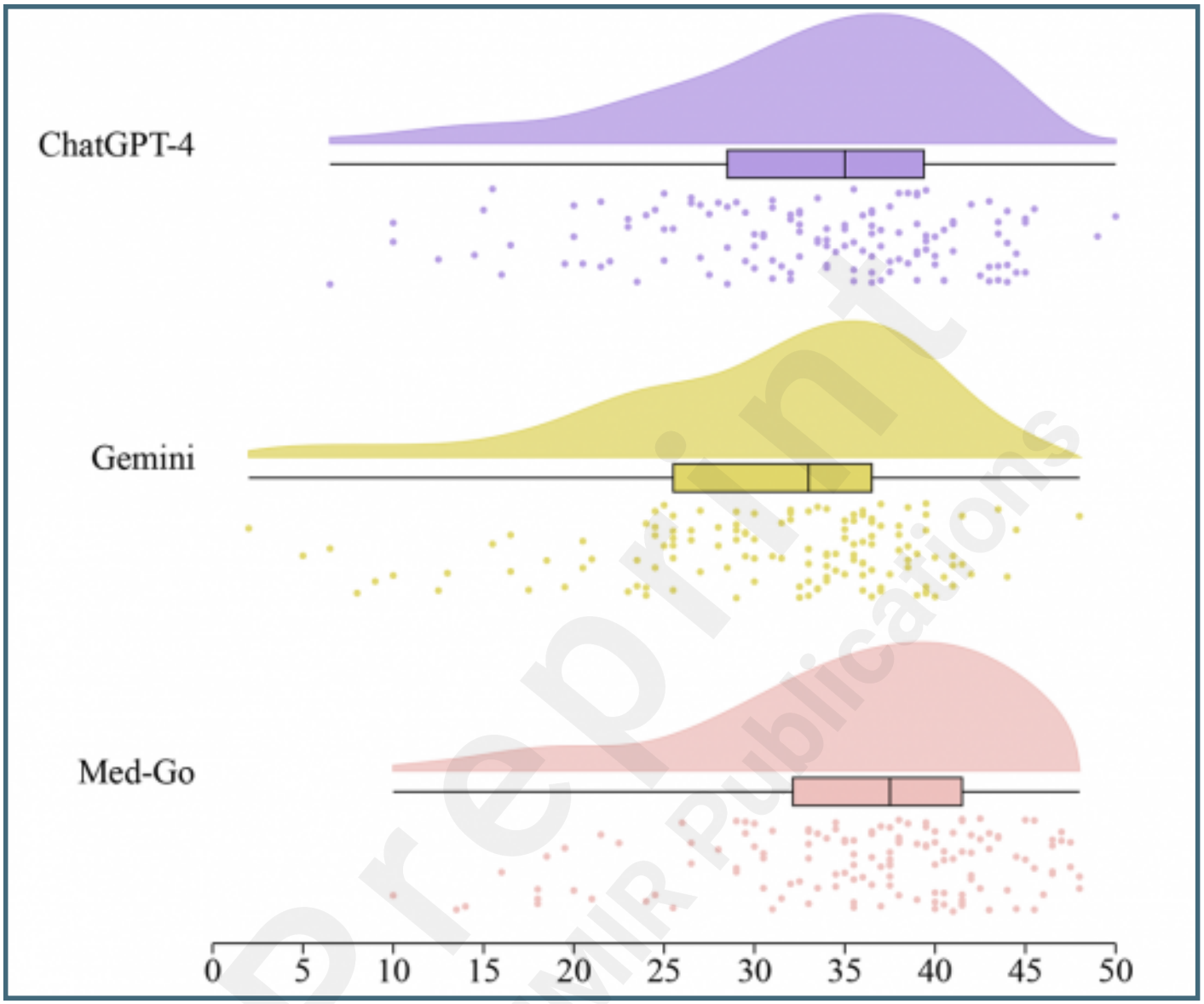
Supplementary Files

Preprint JMIR Publications

Figures

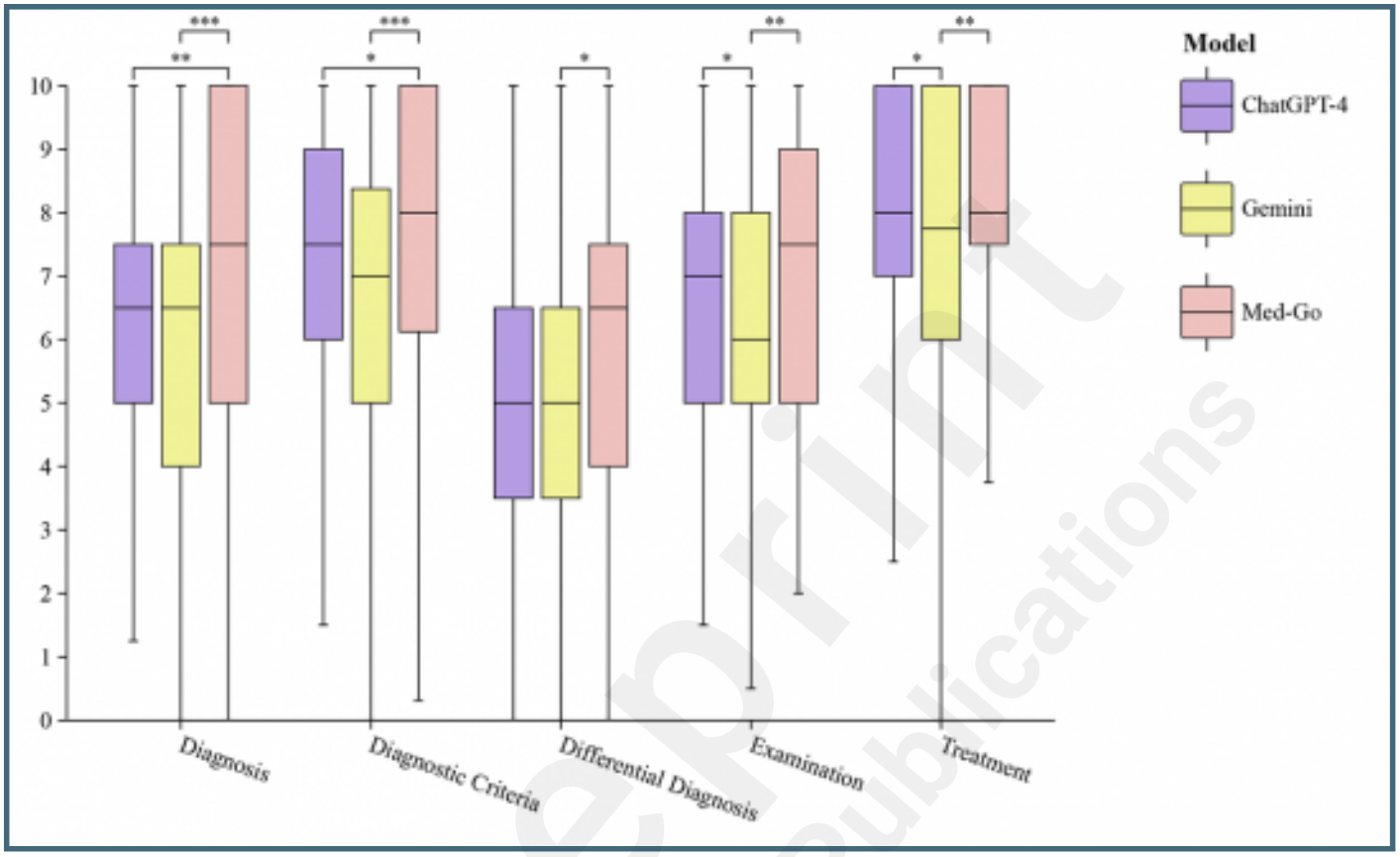
Preprint JMIR Publications

Distribution of overall scores for ChatGPT, Gemini, and Med-Go.



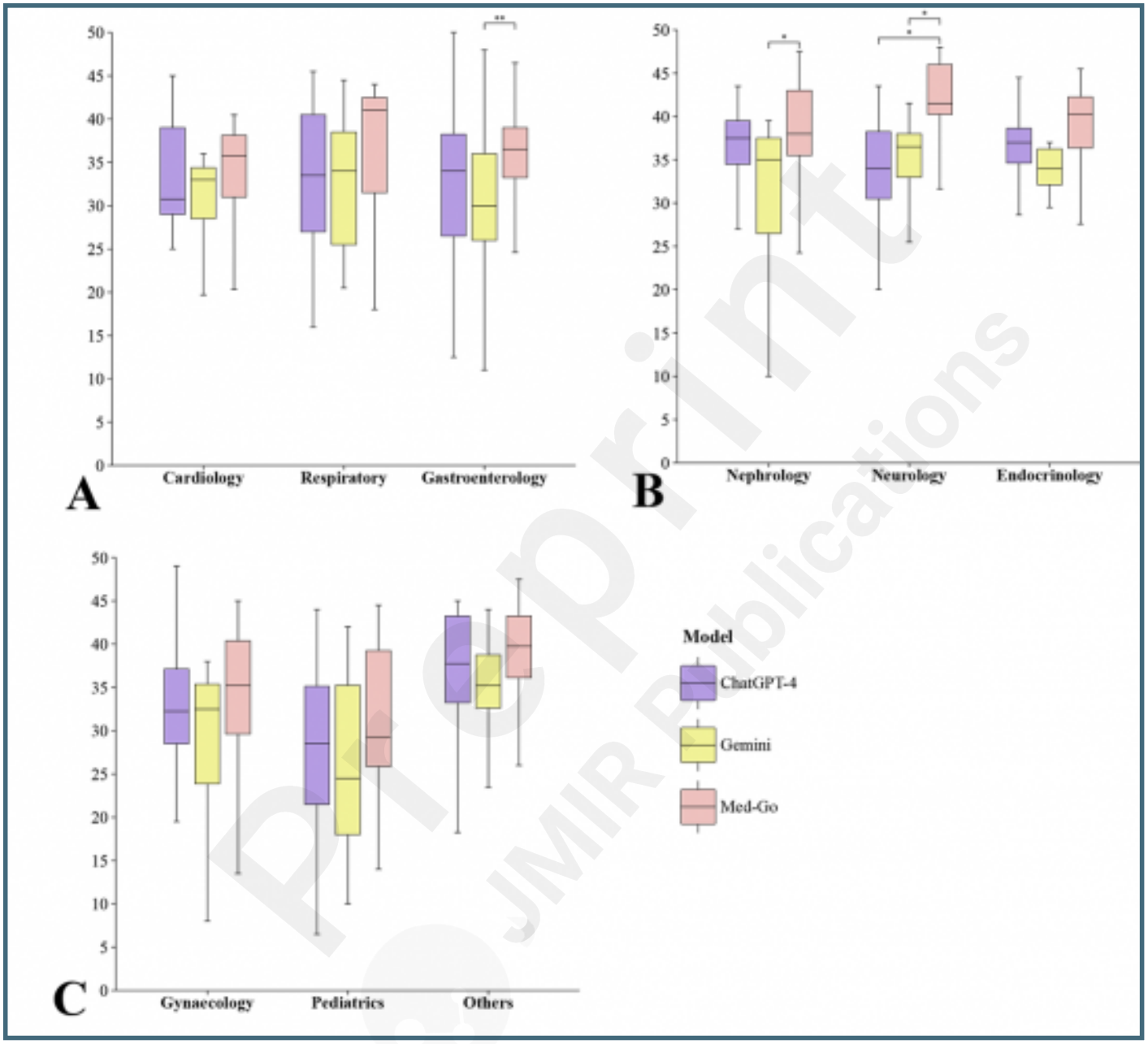
Preprint JMIR Publications

3 LLMs performance on 5 assessment categories. Med-Go outperformed ChatGPT-4 and Gemini significantly in diagnosis and diagnostic criteria. Although all LLMs struggled with differential diagnosis, Med-Go still performed better than Gemini. In examination and treatment, both ChatGPT-4 and Med-Go showed superior performance compared to Gemini (*p < 0.05, **p < 0.01, ***p < 0.001).



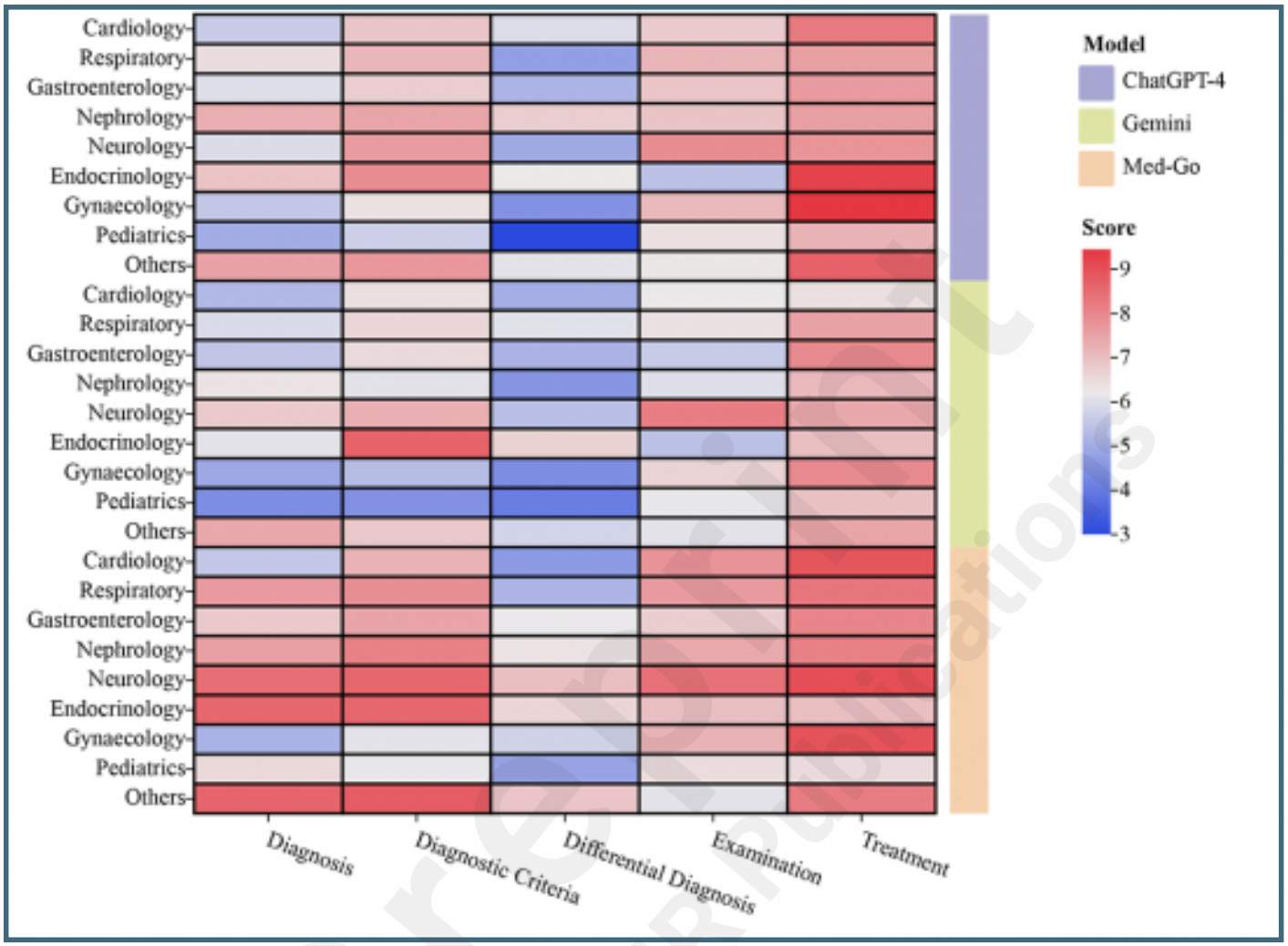
Preprint JMIR Publications

Performance of the 3 LLMs across 9 medical disciplines. In gastroenterology (A) and nephrology (B), Med-Go's median scores showed significant difference from Gemini's. In neurology, Med-Go showed significant difference compared to both ChatGPT and Gemini (B). In other disciplines, the 3 LLMs exhibited no statistical difference (*p < 0.05, **p < 0.01).



Preprint JMIR Publications

The comparison of scores for 3 LLMs—ChatGPT-4, Gemini, and Med-Go—across 9 medical disciplines and 5 question categories.



Preprint JMIR Publications

Multimedia Appendixes

Preprint JMIR Publications

Questions and the corresponding scores for each answer.
URL: http://asset.jmir.pub/assets/e3e250c9bff9639e855f198c9bb2b75a.docx

Preprint JMIR Publications