scientific reports

Check for updates

OPEN

Benchmarking large language models against clinicians across hospital levels in cardiovascular decision-making: a cross-sectional vignette-based study

Zixi Zhang[1], Yingxu Ma[1], Yichao Xiao[1], Yongguo Dai[2], Qiuzhen Lin[1], Chan Liu[3], Cancan Wang[4], Tao Tu[1]✉ & Qiming Liu[1]✉

Large language models (LLMs) have showed strong performance on standardized medical examinations, yet their comparative clinical relevance against human clinicians remains limited. This study benchmarked the performance of DeepSeek-R1 and ChatGPT 4.0 against cardiovascular clinicians from different hospital levels in China. We conducted a cross-sectional, vignette-based assessment consisting of 100 standardized cardiovascular multiple-choice questions covering four competency domains: clinical reasoning (CR), frontier updates (FU), basic memory (BM), and emergency decision (ED). Thirty clinicians from six hospitals (three primary and three tertiary) were compared with two LLMs. Each question was executed five times per model, and run-to-run consistency was evaluated. Mean differences (LLM – clinician) with 95% confidence intervals (CIs) were estimated using nonparametric bootstrap resampling (10,000 iterations). Clinicians achieved a mean total score of 69.7 ± 7.9, whereas DeepSeek-R1 and ChatGPT-4.0 scored 97 and 95, respectively. The mean total score differences were + 27.3 points (95% CI 24.4–30.1) for DeepSeek-R1 and + 25.3 points (22.4–28.1) for ChatGPT 4.0. Both models outperformed clinicians in CR, FU, BM, and ED. Run-to-run agreement was high (DeepSeek-R1 κ = 0.73; ChatGPT 4.0 κ = 0.76). LLMs substantially outperformed clinicians in knowledge- and decision-based tasks while approaching clinician-level performance in CR. These findings suggest that LLMs may complement clinical expertise and enhance diagnostic consistency across hospital levels.

Keywords Artificial intelligence, DeepSeek-R1, ChatGPT 4.0, Clinical decision-making, China

Abbreviations

AI	Artificial intelligence
BM	Basic memory
ChatGPT	Chat Generative Pretrained Transformer
CI	Confidence interval
CR	Clinical reasoning
ED	Emergency decision
FU	Frontier updates
LLM	Large language model
SD	Standard deviation

[1]Department of Cardiology, The Second Xiangya Hospital, Central South University, 139 Renmin Road, Furong District, Changsha City, Hunan Province, People's Republic of China. [2]Department of Pharmacy, Xiangya Hospital, Central South University, Changsha City, Hunan Province, People's Republic of China. [3]Department of International Medicine, The Second Xiangya Hospital, Central South University, Changsha City, Hunan Province, People's Republic of China. [4]The First Detention Area, Central Hospital of Hunan Provincial Prison Administration, Changsha City, Hunan Province, People's Republic of China. ✉email: tutaotodd@csu.edu.cn; qimingliu@csu.edu.cn

The integration of artificial intelligence (AI) into clinical practice has increasingly reshaped medical knowledge dissemination, diagnostic accuracy, and clinical decision support[1,2]. Among the most transformative AI technologies, large language models (LLMs), such as OpenAI's Chat Generative Pretrained Transformer (ChatGPT 4.0), have demonstrated impressive performance on standardized medical assessments, achieving high accuracy and rapid response in knowledge retrieval and vignette-based diagnostic tasks[3–5]. By leveraging extensive biomedical corpora and frequently updated guidelines, these models can assist with evidence synthesis and information access. Nevertheless, several limitations persist, including variable performance in complex clinical reasoning (CR), reduced adaptability in resource-limited environments, and unresolved ethical concerns regarding accountability and reliability in high-stakes or emergency contexts[6,7]. These issues underscore the need of rigorous, task-specific evaluations before LLMs are integrated into clinical workflows.

Despite rapid advances in AI applications, empirical evidence regarding the performance of emerging models such as DeepSeek-R1 remains limited[8–10]. While domain-specific pretraining and fine-tuning have improved model accuracy and contextual understanding, most existing work has emphasized algorithmic optimization rather than direct benchmarking against clinicians using structured, domain-specific evaluations[11]. In addition, disparities in clinical expertise between primary hospitals and tertiary centers raise questions about whether LLMs could mitigate, or inadvertently exacerbate, inequalities in diagnostic performance and decision quality.

To address these gaps, we conducted a cross-sectional, vignette-based multiple-choice assessment to compare the performance of DeepSeek-R1 and ChatGPT-4.0 with that of practicing cardiovascular clinicians. The assessment encompassed four predefined domains—CR, frontier updates (FU), basic memory (BM), and emergency decision (ED)—representing key cognitive competencies in cardiovascular care. The primary objective was to quantify performance differences between LLMs and clinicians in total and domain-specific scores. Secondary objectives were to evaluate whether these differences varied by hospital level (primary versus tertiary) and professional title (junior, intermediate, senior), and to examine robustness through sensitivity analyses that addressed run-to-run variability in model responses. This study provides an empirical benchmark for the domain-specific capabilities of LLMs relative to clinicians across hierarchical healthcare systems and informs potential use cases for augmenting clinical decision-making.

Methods

Study design

This cross-sectional, vignette-based assessment was conducted from December 2024 to April 2025 across six hospitals in China, including three primary hospitals and three comprehensive tertiary research hospitals. The study was approved by the Ethics Committee of the Second Xiangya Hospital, Central South University, and was determined to be exempt from ethical review and individual informed consent. The study adhered to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines.

Item development and standardized testing

An independent adjudication committee comprising three board-certified senior cardiologists selected 100 single-best-answer, four-option multiple-choice questions (options A–D) from the Hunan Provincial Standardized Cardiovascular Question Bank, an accredited repository used for licensing and continuing-education examinations (Supplementary File: Sample Assessment Questions). A stratified random sampling approach was used to ensured balanced coverage of the four competency domains: CR (44 items), FU (14 items), BM (31 items), and ED (11 items). Each question had one verified correct answer, and option order was randomized to minimize positional bias. For each respondent (clinician or LLM), domain scores were defined as the number of correctly answered items, and the composite total score was the number of correct responses across all 100 items (range 0–100).

Thirty cardiovascular clinicians—ten per professional title (junior, intermediate, senior)—were recruited evenly across the six hospitals (three primary and three tertiary). All participants completed the examination under proctored conditions. Responses were collected by independent researchers, anonymized, and verified by the adjudication committee.

LLMs testing and response protocol

Each of the 100 multiple-choice items was independently submitted to DeepSeek-R1 and ChatGPT-4.0 using identical prompts. To evaluate model stability, each item was executed in five independent runs in new sessions. In the primary analysis, an item was considered correct only if all five runs produced the correct response (5/5); any inconsistency ($\leq 4/5$) was considered incorrect. To evaluate robustness, a proportional-credit sensitivity analysis assigned fractional credit proportional to the number of correct outputs (e.g., $3/5 = 0.6$). All LLM-generated responses were anonymized as Model 1 and Model 2, stripped of identifiers, and returned to the adjudication committee for verification of response validity (ensuring single-option outputs and interpretable responses), after which scoring was performed automatically.

Classification of hospitals and professional titles

Primary hospitals were defined as Level I or Level II, non-university-affiliated institutions serving county- or city-level populations. Comprehensive research hospitals were Level III tertiary centers affiliated with universities or located in provincial capitals that provide subspecialty care and academic training.

Professional titles were categorized as junior, intermediate, or senior according to the National Health Technical Qualification Examination system in China. Promotion is determined by cumulative experience, examination performance, and verified clinical workload (Supplementary Table 1). All clinicians held valid practice licenses and up-to-date continuing-education credentials. Title classification was cross-checked against institutional records before inclusion.

Group	Number	Total score (mean ± SD)	CR	FU	BM	ED	Time (min)
Clinicians–overall	30	69.7 ± 7.9	37.4 ± 2.7	7.3 ± 2.5	18.1 ± 2.4	6.9 ± 2.0	41.1 ± 9.4
Junior clinicians	10	60.3 ± 2.3	34.3 ± 1.8	5.9 ± 1.2	15.5 ± 1.1	4.6 ± 0.7	49.4 ± 10.1
Intermediate clinicians	10	72.1 ± 4.1	37.8 ± 0.9	7.6 ± 2.6	19.2 ± 1.2	7.5 ± 1.0	40.1 ± 5.2
Senior clinicians	10	76.8 ± 4.5	40.2 ± 0.6	8.3 ± 2.8	19.6 ± 2.0	8.7 ± 1.0	33.7 ± 4.4
Primary hospitals	15	66.7 ± 6.4	36.8 ± 3.2	5.4 ± 1.1	17.9 ± 1.9	6.7 ± 1.7	44.6 ± 10.8
Tertiary hospitals	15	72.7 ± 8.4	38.1 ± 2.1	9.1 ± 2.0	18.3 ± 2.8	7.2 ± 2.2	37.5 ± 6.4
DeepSeek-R1 (LLM)	–	97	43	13	30	11	–
ChatGPT-4.0 (LLM)	–	95	42	12	30	11	–

Table 1. Descriptive performance of clinicians and LLMs across professional titles and hospital levels. BM, basic memory; CR, clinical reasoning; ED, emergency decision; FU, frontier updates; LLM, large language model; SD, standard deviation.

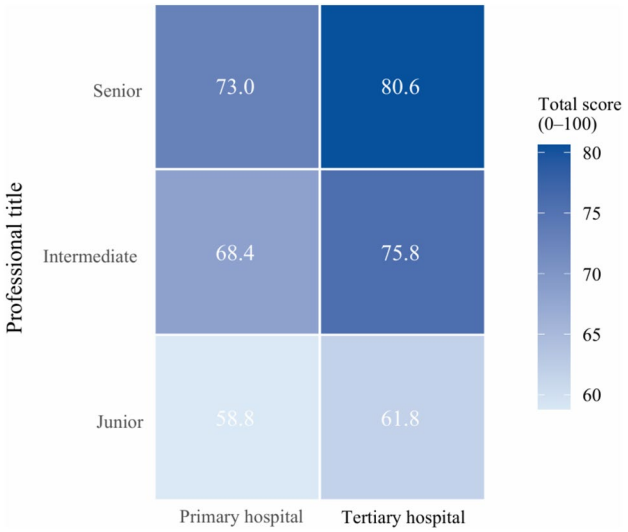

Fig. 1. Total score distributions of clinicians by professional title and hospital level. Color intensity represents the mean total score (0–100 scale) within each stratum of professional title and hospital level.

Statistical analysis

Continuous variables are presented as mean ± standard deviation (SD). For each domain (CR, FU, BM, and ED) and for the total score, mean differences between LLMs and clinicians (LLM–clinician) with 95% confidence intervals (CIs) were estimated using non-parametric bootstrap resampling of clinician means (10,000 iterations), treating LLM scores as fixed comparators. Two-sided bootstrap *P* values were calculated for all comparisons. Stratified bootstrap analyses were performed by hospital level and professional title to assess the consistency of LLM–clinician differences across subgroups. Robustness was evaluated using a proportional-credit sensitivity analysis that assigned fractional credit based on the number of correct outputs across five independent LLM executions. Run-to-run stability was quantified using percent agreement and Fleiss' κ. All analyses were conducted using R version 4.4.0 (R Foundation for Statistical Computing, Vienna, Austria). A two-sided $P < 0.05$ was considered statistically significant.

Results

Baseline performance among clinicians

Significant differences were observed in both total and domain-specific scores across professional titles (Table 1). Senior clinicians achieved the highest total scores (76.8 ± 4.5), followed by intermediate (72.1 ± 4.1) and junior clinicians (60.3 ± 2.3). Similar gradients were observed in CR, BM, and ED.

When stratified by hospital level, clinicians from tertiary hospitals achieved higher total scores (72.7 ± 8.4) than those from primary hospitals (66.7 ± 6.4), with the largest inter-hospital difference observed in FU. Completion times were shorter in tertiary hospitals (37.5 ± 6.4 min) than in primary hospitals (44.6 ± 10.8 min). Figure 1 illustrates these baseline distributions across professional titles and hospital levels.

Overall performance comparison between LLMs and clinicians

Both LLMs demonstrated higher performance than clinicians across all evaluated domains (Table 2). Clinicians achieved a mean total score of 69.7 ± 7.9, whereas DeepSeek-R1 and ChatGPT 4.0 scored 97 and 95, respectively.

Domain	Clinician	DeepSeek-R1	Mean difference (95% CI)	*P* value	ChatGPT 4.0	Mean difference (95% CI)	*P* value
CR	37.4 ± 2.7	43	+ 5.6 (4.4–6.9)	< 0.001	42	+ 4.6 (3.4–5.8)	< 0.001
FU	7.3 ± 2.5	13	+ 5.7 (3.8–7.6)	< 0.001	12	+ 4.7 (2.8–6.6)	< 0.001
BM	18.1 ± 2.4	30	+ 11.9 (9.2–14.5)	< 0.001	30	+ 11.9 (9.2–14.5)	< 0.001
ED	6.9 ± 2.0	11	+ 4.1 (2.9–5.3)	< 0.001	11	+ 4.1 (2.9–5.3)	< 0.001
Total	69.7 ± 7.9	97	+ 27.3 (24.4–30.1)	< 0.001	95	+ 25.3 (22.4–28.1)	< 0.001

Table 2. Comparative performance of LLMs and clinicians across total and domain scores. Mean differences (LLM – clinician) and 95% CIs were obtained from 10,000 bootstrap iterations using clinician means. A *P* value < 0.05 indicated a significant difference. BM, basic memory; CI, confidence interval; CR, clinical reasoning; ED, emergency decision; FU, frontier updates; LLM, large language model; SD, standard deviation.

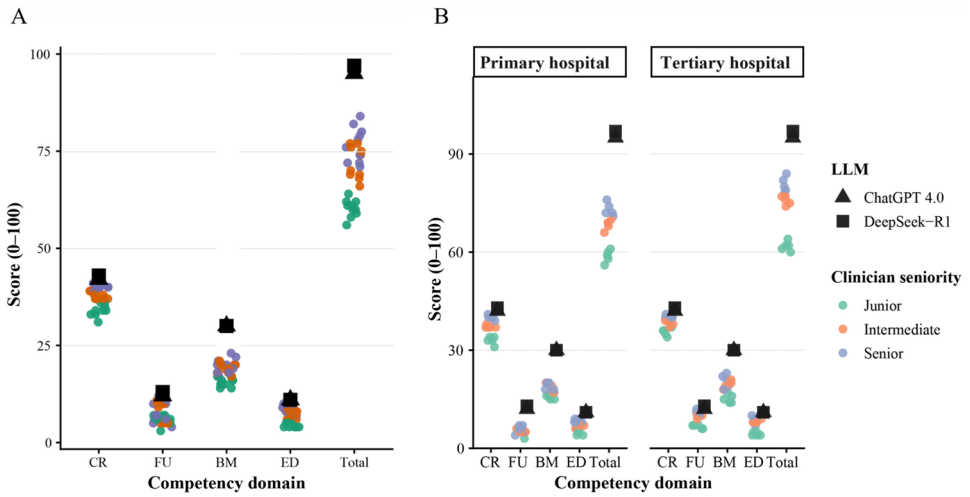

Fig. 2. Absolute scores of clinicians and LLMs across competency domains. (**A**) Absolute scores for all clinicians and two LLMs across four competency domains and the total score; (**B**) Scores stratified by hospital level (primary vs. tertiary). Abbreviations: BM, basic memory; CR, clinical reasoning; ED, emergency decision; FU, frontier updates; LLM, large language model.

Bootstrap analyses revealed mean total score differences of + 27.3 points (95% CI 24.4–30.1) for DeepSeek-R1 and + 25.3 points (95% CI 22.4–28.1) for ChatGPT 4.0 (both $P < 0.001$).

Across specific domains, both LLMs outperformed clinicians in CR, FU, BM, and ED (all $P < 0.001$). Figure 2A visualizes individual-level clinician performance alongside both LLMs, demonstrating substantial dispersion among clinicians and the consistently superior scores of the LLMs. When stratified by hospital level, Fig. 2B further demonstrates that LLM advantages persisted in both primary and tertiary hospitals, with similar relative positions across all domains.

Stratified analysis by hospital level and professional title

Stratified bootstrap analyses confirmed higher LLM performance across all professional title groups and hospital levels (Table 3). In primary hospitals, the mean total score of DeepSeek-R1 exceeded clinicians by + 30.3 (95% CI 27.3–33.5) points, compared with + 24.3 (95% CI 20.3–28.5) points in tertiary hospitals. For ChatGPT 4.0, the corresponding differences were + 28.3 (95% CI 25.3–31.5) and + 22.3 (95% CI 18.3–26.5).

When stratified by professional title, the largest differences were observed among junior clinicians (+ 36.7 for DeepSeek-R1 and + 34.7 for ChatGPT 4.0), followed by intermediate clinicians (+ 24.9 and + 22.9) and senior clinicians (+ 20.2 and + 18.2) (all $P < 0.001$). Domain-specific stratified findings are summarized in Supplementary Table 2.

Sensitivity and consistency analyses

Sensitivity analyses using proportional-credit scoring confirmed the robustness of the primary results (Supplementary Table 3), with both DeepSeek-R1 and ChatGPT-4.0 maintaining statistically significant advantages over clinicians across all domains ($P < 0.001$). Run-to-run consistency analyses demonstrated high output stability for both models. As shown in Supplementary Table 4, DeepSeek-R1 generated identical responses across all five executions for 97 of 100 items, while ChatGPT-4.0 did so for 95 items, corresponding to percent agreements of 97.0% and 96.0%, respectively. Fleiss' κ coefficients indicated substantial reliability for both models (DeepSeek-R1: $\kappa = 0.73$; ChatGPT-4.0: $\kappa = 0.76$). Domain-specific consistency, summarized in

Subgroup	Number	Model	Clinician mean	LLM score	Mean difference (95% CI)	*P* value
Hospital level						
Primary	15	DeepSeek-R1	66.7	97	+ 30.3 (27.3–33.5)	< 0.001
		ChatGPT-4.0	66.7	95	+ 28.3 (25.3–31.5)	< 0.001
Tertiary	15	DeepSeek-R1	72.7	97	+ 24.3 (20.3–28.5)	< 0.001
		ChatGPT-4.0	72.7	95	+ 22.3 (18.3–26.5)	< 0.001
Professional title						
Junior	10	DeepSeek-R1	60.3	97	+ 36.7 (35.4–38.0)	< 0.001
		ChatGPT-4.0	60.3	95	+ 34.7 (33.4–36.0)	< 0.001
Intermediate	10	DeepSeek-R1	72.1	97	+ 24.9 (22.4–27.4)	< 0.001
		ChatGPT-4.0	72.1	95	+ 22.9 (20.4–25.4)	< 0.001
Senior	10	DeepSeek-R1	76.8	97	+ 20.2 (17.6–22.8)	< 0.001
		ChatGPT-4.0	76.8	95	+ 18.2 (15.6–20.8)	< 0.001

Table 3. Stratified LLM – clinician mean differences by hospital level and professional title. LLM scores are fixed (identical across subgroups). Differences estimated via stratified bootstrap resampling (10,000 iterations). A *P* value < 0.05 indicated a significant difference. CI, confidence interval; LLM, large language model.

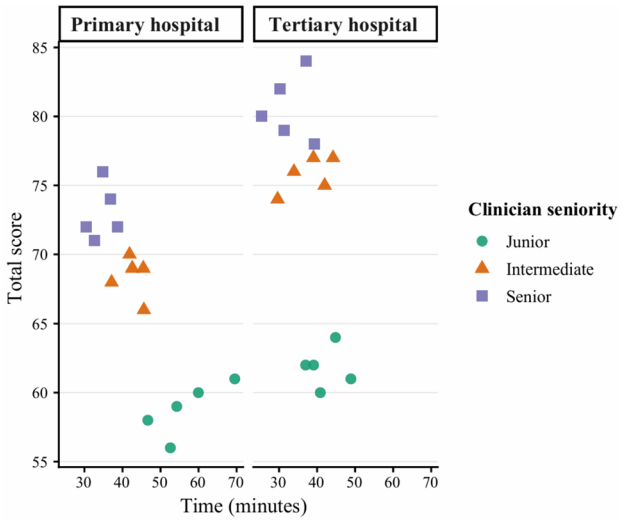

Fig. 3. Time–accuracy relationship among clinicians by hospital level.

Supplementary Table 5 and Supplementary Fig. 1, showed that 5/5 identical-response rates exceeded 90% across all domains for both models, with the highest reproducibility observed in ED and CR.

Efficiency–performance correlation among clinicians

Time–accuracy relationships showed distinct patterns across hospital levels (Fig. 3). Tertiary-hospital clinicians tended to complete the assessment more rapidly while maintaining higher total scores, suggesting greater efficiency of clinical decision-making. In contrast, clinicians from primary hospitals demonstrated wider dispersion in both time and accuracy.

Discussion

This study provides a structured comparison of LLMs and clinicians across hierarchical hospital levels in cardiovascular medicine in China. DeepSeek-R1 and ChatGPT 4.0 exceeded clinician performance on the total score and on CR, BM, FU, and ED. These findings extend beyond technical benchmarking and highlight the public health potential of LLMs to strengthen diagnostic capacity, reduce gaps in medical knowledge, and mitigate inequities in care delivery.

The most notable finding was the consistent and large performance gap between LLMs and clinicians in primary hospitals, particularly among junior physicians. This pattern reflects wider structural disparities within health systems, where resources, training opportunities, and continuing education are concentrated in tertiary institutions[12]. In middle-income countries, the limited availability of experienced specialists in lower-level hospitals remains a determinant of delayed diagnosis, suboptimal treatment, and excess cardiovascular mortality[13,14]. By achieving high accuracy on standardized decision tasks, LLMs such as DeepSeek-R1 and ChatGPT 4.0 may function as intellectual equalizers, providing clinicians in under-resourced settings with

immediate access to current, evidence-based recommendations. If deployed equitably, AI-assisted decision support could help narrow the expertise gap between primary and tertiary care and support goals aligned with China's tiered medical system and the World Health Organization's universal health coverage agenda. However, this opportunity is accompanied by risk. Without coordinated policy support, LLMs may preferentially benefit well-funded tertiary centers with stronger digital infrastructure, thereby widening existing disparities[15]. Equitable integration requires deliberate public health planning, including investment in digital infrastructure, cloud access, broadband connectivity, and clinician training in AI literacy[16]. Policy makers should consider AI inclusion policies analogous to essential medicines lists to ensure that AI-based decision tools are treated as public goods rather than privileges restricted to large hospitals. The impact of AI in healthcare will depend not only on algorithmic accuracy but also on fairness and inclusiveness in deployment.

Beyond immediate clinical utility, this study also points to inefficiencies in traditional medical education. Performance gradients across professional titles and hospital levels indicate that seniority alone does not guarantee up-to-date knowledge. In rapidly evolving fields such as cardiology, knowledge half-lives are shortening, while access to continuing education remains uneven. LLMs can serve as dynamic, relatively low-cost platforms for continuous professional development by delivering real-time updates, clinical simulations, and personalized learning analytics[17]. Such tools may support a shift from time-based, hierarchical credentialing to competency-based medical education anchored in objective, adaptive assessment. When integrated responsibly, AI can broaden access to high-level CR exercises for clinicians in rural and community hospitals, supporting public health goals to enhance primary care quality and reduce avoidable referrals and expenditures[18]. Embedding LLM-based knowledge tools within national training programs could also help modernize continuing medical education, making it more responsive, scalable, and equitable across regions.

The observed time–accuracy differences between clinicians in primary and tertiary hospitals underscore workflow inefficiencies. LLMs produced accurate answers with minimal time burden, suggesting potential to reduce cognitive load and improve throughput in resource-limited environments[19]. When integrated into electronic health records or telehealth platforms, LLMs could help standardize initial triage, generate patient-specific summaries, and support adherence to evidence-based pathways. These functions may allow clinicians to focus on complex judgment, patient communication, and preventive care. At the population level, redistribution of routine tasks represents a pragmatic public health strategy. By automating low-complexity decision support and information retrieval, AI can increase the effective productivity of the existing workforce, which is particularly relevant in cardiovascular care where clinician-to-patient ratios are low and case complexity is high[20]. This scalability positions LLMs as operational assets for health systems seeking to improve both efficiency and quality.

Widespread adoption must proceed with appropriate governance and ethical oversight. Strong performance in knowledge-based domains does not resolve contextual limitations, including the absence of situational awareness, moral reasoning, and accountability. Overreliance on algorithmic guidance may distance decisions from patient narratives and values[21]. Transparent model auditing, traceable reasoning outputs, and clear accountability frameworks are therefore necessary. From a policy perspective, regulatory authorities may consider credentialing standards for AI similar to those for drugs and devices, including model validation, periodic recertification, and post-deployment monitoring for bias or performance degradation[22]. Data privacy protections and explainability requirements are also essential to preserve patient trust[23]. The human–AI interface should remain grounded in shared decision-making, with AI serving as a consultant rather than an arbiter.

From a public health perspective, the findings support integration of LLMs as scalable tools to improve diagnostic quality and medical education. Potential benefits include: (1) reduced variability in clinical decision-making across institutions and regions; (2) faster dissemination of new evidence and guidelines; (3) support for training and upskilling in lower-level hospitals; and (4) greater efficiency without expansion of the workforce. These outcomes are consistent with health system strengthening and equitable access to high-quality care. The results also provide empirical evidence that AI systems can complement existing human resources, enhancing accuracy and efficiency when implemented appropriately. Future work should focus on maximizing accessibility while maintaining strong governance to ensure that these technologies contribute to equitable, high-quality, and ethically responsible healthcare delivery.

Limitations

This study has several limitations that should be considered when interpreting the findings. First, the cross-sectional design limits causal inference regarding whether access to LLM-based support would improve clinicians' performance over time or translate into better patient outcomes. Because clinicians and LLMs were assessed at a single time point in a controlled testing environment, the study cannot determine whether AI assistance causally enhances diagnostic accuracy or clinical decision quality in real-world practice. Second, the clinician cohort consisted of 30 physicians from six hospitals in China, which may not reflect broader variability in clinical expertise across regions, specialties, or health systems. Larger and more diverse samples are needed to improve external generalizability. Third, the study was confined to cardiovascular medicine and employed a vignette-based multiple-choice format. Although this approach ensures standardization, it does not fully capture real-world CR, communication, or decision-making under uncertainty. Fourth, only two LLMs—DeepSeek-R1 and ChatGPT-4.0—were evaluated. Performance may differ across architectures and future model iterations; continued benchmarking will be required. Fifth, assessments were conducted outside of routine clinical workflows. Future studies should examine LLM integration into electronic health record systems and evaluate its impact under realistic time pressures, data complexity, and clinician–AI interaction patterns. Sixth, reasoning ability was inferred from response accuracy rather than from explicit reasoning processes. Studies incorporating open-ended responses, reasoning-chain evaluation, and interpretability analyses may better characterize the qualitative nature of AI reasoning. Finally, this study did not evaluate human–AI collaboration. The extent to

which clinician–AI synergy might yield different performance patterns remains an important direction for future experimental and implementation research.

Conclusion

This study systematically evaluated the performance of two LLMs—DeepSeek-R1 and ChatGPT 4.0—compared with cardiovascular clinicians across four cognitive domains. Both models showed substantially higher overall and domain-specific accuracy, especially in factual knowledge, guideline-related content, and ED, while their CR performance approached that of senior clinicians. These findings suggest that LLMs may serve as useful adjuncts for standardized diagnostic tasks and medical training. Future studies should assess performance in real clinical workflows, explore longitudinal learning patterns, and examine how LLMs and clinicians can complement each other in decision-making processes.

Data availability

The datasets generated for this study are available on request to the corresponding author.

Received: 17 August 2025; Accepted: 11 December 2025
Published online: 15 December 2025

References

1. Yang, Y. et al. A survey of recent methods for addressing AI fairness and bias in biomedicine. *J. Biomed. Inform.* **154**, 104646 (2024).
2. Han Wang, M. et al. Applied machine learning in intelligent systems: Knowledge graph-enhanced ophthalmic contrastive learning with "clinical profile" prompts. *Front. Artif. Intell.* **8**, 1527010 (2025).
3. Akdogan, O. et al. Effect of a ChatGPT-based digital counseling intervention on anxiety and depression in patients with cancer: A prospective, randomized trial. *Eur. J. Cancer* **221**, 115408 (2025).
4. Sarraju, A. et al. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. *JAMA* **329**(10), 842–844 (2023).
5. Brigo, F. et al. Artificial intelligence (ChatGPT 4.0) vs. Human expertise for epileptic seizure and epilepsy diagnosis and classification in Adults: An exploratory study. *Epilepsy Behav.* **166**, 110364 (2025).
6. Li, R. et al. The artificial intelligence revolution in gastric cancer management: clinical applications. *Cancer Cell Int.* **25**(1), 111 (2025).
7. Al-Karawi, D. et al. A review of artificial intelligence in breast imaging. *Tomography* **10**(5), 705–726 (2024).
8. Gibney, E., Scientists flock to DeepSeek: How they're using the blockbuster AI model. *Nature* **2025**.
9. Liu, H. Global cooperation is crucial for DeepSeek and broader AI research. *Nature* **639**(8055), 577 (2025).
10. Peng, Y. et al. From GPT to DeepSeek: Significant gaps remain in realizing AI in healthcare. *J. Biomed. Inform.* **163**, 104791 (2025).
11. Scanzera, A. C. et al. Planning an artificial intelligence diabetic retinopathy screening program: A human-centered design approach. *Front. Med. (Lausanne)* **10**, 1198228 (2023).
12. Shen, Q. et al. Hospital pharmacists' knowledge of and attitudes towards the implementation of the National Essential Medicines System: A questionnaire survey in western China. *BMC Health Serv. Res.* **16**, 292 (2016).
13. Kruk, M. E. et al. High-quality health systems in the Sustainable Development Goals era: Time for a revolution. *Lancet Glob. Health* **6**(11), e1196–e1252 (2018).
14. Franco, M., Cooper, R. S., Bilal, U. & Fuster, V. Challenges and opportunities for cardiovascular disease prevention. *Am. J. Med.* **124**(2), 95–102 (2011).
15. Saeed, S. A. & Masters, R. M. Disparities in health care and the digital divide. *Curr. Psychiatry Rep.* **23**(9), 61 (2021).
16. Elikman, D., Anderson, R., Balish, M. & Zilbermint, M. Mitigating health disparities: Bridging the digital divide in modern health care. *South Med. J.* **118**(6), 330–332 (2025).
17. Wartman, S. A. & Combs, C. D. Medical education must move from the information age to the age of artificial intelligence. *Acad. Med.* **93**(8), 1107–1109 (2018).
18. Orton, M., Agarwal, S., Muhoza, P., Vasudevan, L. & Vu, A. Strengthening delivery of health services using digital devices. *Glob. Health Sci. Pract.* **6**(Suppl 1), S61–S71 (2018).
19. Yi, S. et al. Perspectives of digital health innovations in low- and middle-income health care systems from South and Southeast Asia. *J. Med. Internet Res.* **26**, e57612 (2024).
20. Ferrara, M., Bertozzi, G., Di Fazio, N., Aquila, I., Di Fazio, A., Maiese, A., Volonnino, G., Frati, P. & La Russa, R. Risk management and patient safety in the artificial intelligence era: A systematic review. *Healthcare (Basel)* **12**(5), (2024).
21. Chakraborty Samant, A., Tyagi, I., Vybhavi, J., Jha, H. & Patel, J. ChatGPT dependency disorder in healthcare practice: An editorial. *Cureus* **16**(8), e66155 (2024).
22. Ueda, D., Kakinuma, T., Fujita, S., Kamagata, K., Fushimi, Y., Ito, R., Matsui, Y., Nozaki, T., Nakaura, T., Fujima, N., Tatsugami, F., Yanagawa, M., Hirata, K., Yamada, A., Tsuboyama, T., Kawamura, M., Fujioka, T. & Naganawa, S. Fairness of artificial intelligence in healthcare: review and recommendations. *Jpn. J. Radiol.* **42**(1), (2023).
23. Oduoye, M. O. et al. Impacts of the advancement in artificial intelligence on laboratory medicine in low- and middle-income countries: Challenges and recommendations-A literature review. *Health Sci. Rep.* **7**(1), e1794 (2024).

Acknowledgements

In this study, LLMs (DeepSeek-R1 and ChatGPT 4.0) were used to compare their performance with human clinicians in clinical decision-making tasks. The use of AI tools was disclosed in the materials and methods section. However, LLMs do not meet authorship criteria and cannot assume legal or ethical responsibility for the manuscript. The authors are fully accountable for the manuscript's content, including any AI-generated portions.

Author contributions

Yongguo Dai and Cancan Wang collected the relevant information; Yichao Xiao and Yingxu Ma performed the statistical analysis; Chan Liu and Qiuzhen Lin performed the data visualization; Zixi Zhang wrote the manuscript, which was subsequently revised by Tao Tu and Qiming Liu. All the authors contributed to the discussion and approved the final manuscript.

Funding
This work was supported by the National Natural Science Foundation of China [No. 82070356, 81770337, 82470333, 82300357, 82200367], the Chinese Society of Cardiology's Foundation [No. CSCF2024B02], the Hunan Provincial Natural Science Foundation of China [No. 2021JJ30033, 2023JJ30791], the Key Project of Hunan Provincial Science and Technology Innovation [No. 2024JK2119], and the Clinical Medical Technology Innovation Guidance Project of Hunan Science and Technology Agency [No. 2021SK53519].

Declarations

Competing interests
The authors declare no competing interests.

Ethics approval
This study complied with the Declaration of Helsinki and relevant institutional guidelines. The protocol was reviewed by the Ethics Committee of Second Xiangya Hospital, Central South University and granted an exemption from formal human-subjects review and from patient informed consent because no individual-level patient data, patient reports, medical records, or biospecimens were collected or used. All participating clinicians provided informed consent prior to participation.

Consent to participate
All the authors participated in the study and made significant intellectual contributions to the manuscript.

Consent for publication
This manuscript is not currently under consideration for publication elsewhere, and the work reported will not be submitted for publication elsewhere until a final decision has been made as to its acceptability by the journal.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-32602-w.

Correspondence and requests for materials should be addressed to T.T. or Q.L.