Infection and Drug Resistance

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Evaluation of Large Language Models in Infectious Disease Decision-Making: From Examination to Clinical Practice

Dandan Wu[1,*], Keying Chen[1,*], Xiaocui Wu[1], Jiongfei Jin[1,2], Feng Xu[1]

[1]Department of Infectious Diseases, The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang, People's Republic of China; [2]Department of Infectious Diseases, the Affiliated Yangming Hospital of Ningbo University, Ningbo, Zhejiang, People's Republic of China

*These authors contributed equally to this work

Correspondence: Feng Xu, Department of Infectious Diseases, The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang, People's Republic of China, Email xufeng99@zju.edu.cn

Purpose: To evaluate the performance of large language models (LLMs) in infectious disease–specific clinical reasoning and decision-making.

Patients and Methods: A comprehensive evaluation of four widely used LLMs—DeepSeek-R1, ChatGPT-5, Grok 3, and Gemini 2.5 Flash—was conducted using a dual assessment framework that combined examination-based questions with real-world clinical cases. LLMs performance was compared with that of infectious disease residents. Examination outcomes were assessed using accuracy and score rates, while clinical case responses were evaluated by expert reviewers using predefined Likert-scale criteria. In addition, doctors' independent clinical decisions were compared with those supported by LLMs to assess the potential value of human–AI collaboration.

Results: Across examination-based assessments, LLMs performed comparably to infectious disease residents, with no significant differences observed in accuracy or score rates ($p = 0.54$). LLMs showed a trend toward better performance on low-order, knowledge-based questions ($p = 0.34$), whereas doctors tended to perform better on simple case-based questions ($p = 0.74$), particularly those requiring higher-order clinical reasoning ($p = 0.10$). In real-world clinical case evaluations (n=10), LLM-generated responses achieved high ratings for accuracy, completeness, individualization, safety, and readability, with comparable performance across models ($p > 0.05$). Importantly, doctors' decision-making supported by LLMs showed a trend toward improved accuracy compared with independent decisions ($p = 0.19$). Notably, for completeness, doctors supported by LLMs achieved significantly higher scores compared with both doctors alone ($p = 0.0015$) and the LLMs alone ($p = 0.0010$). Nevertheless, clinically meaningful errors occurred in certain high-risk scenarios, underscoring the limitations of standalone LLM decision-making.

Conclusion: LLMs show substantial potential in infectious disease education and clinical decision support, particularly for knowledge-based tasks. However, their limitations in complex clinical reasoning underscore the necessity of clinician oversight. A human–AI collaborative approach appears to offer the greatest benefit, enhancing decision quality while maintaining clinical safety. Continued refinement of regulatory and medico-legal frameworks is critical to support the safe, ethical, and responsible deployment of LLMs in clinical practice.

Keywords: large language models, infectious diseases, clinical decision-making, human–AI collaboration

Introduction

In recent years, the application of artificial intelligence (AI) in medicine has expanded rapidly, particularly with the development of large language models (LLMs), which have introduced new possibilities for medical diagnosis, therapeutic recommendations, and clinical decision support. Through large-scale pretraining on diverse corpora and advanced deep learning architectures, LLMs have demonstrated substantial capabilities in natural language understanding, text generation, and complex reasoning tasks.[1] In medical contexts, these models can integrate extensive biomedical

https://doi.org/10.2147/IDR.S593763
Received: 5 February 2026
Accepted: 18 May 2026
Published: 12 June 2026

knowledge and emulate clinicians' diagnostic and therapeutic reasoning processes, thereby providing clinical information and decision support. Previous studies have shown that LLMs can serve as valuable tools to help medical students efficiently master complex medical knowledge systems,[2] and have been increasingly incorporated into clinical workflows, such as assisting disease diagnosis, generating personalized treatment plans, and optimizing disease management and screening strategies.[3–6] In the field of infectious diseases, LLMs have also shown considerable clinical potential. Models such as ChatGPT-5 and DeepSeek-R1 have reportedly passed specialty examinations in infectious diseases.[7] Multimodal LLMs integrating radiological imaging and clinical text have demonstrated improved diagnostic efficiency for COVID-19,[8] and LLMs have exhibited diagnostic and therapeutic potential in bloodstream infections.[9] Collectively, these advances suggest that LLMs may represent a significant driving force in the future development of medicine.

Several widely used large language models developed in different global contexts have demonstrated strong capabilities in medical applications, including DeepSeek-R1, ChatGPT-5, Grok 3, and Gemini 2.5 Flash (among others). Notably, DeepSeek-R1 is an open-source model released in January 2025. Unlike conventional LLMs optimized primarily through instruction tuning, DeepSeek-R1 adopts a pure reinforcement learning paradigm, enabling the autonomous emergence of reasoning behaviors—including self-reflection, verification, and adaptive strategy optimization—without reliance on manually annotated reasoning trajectories.[10] This design allows DeepSeek-R1 to achieve superior performance in verifiable reasoning tasks. Multiple studies have reported its promising performance in clinical medicine, particularly in diagnostic hypothesis generation and clinical reasoning.[11–13]

However, despite its immense potential, integrating LLMs into clinical practice remains fraught with significant challenges. The primary issue is the occurrence of hallucinations, where LLMs generate plausible but incorrect or misleading information.[14,15] Such errors frequently result from inadequate medical specialization within existing models and the absence of robust evaluation frameworks tailored specifically for clinical use. Moreover, recent studies have highlighted important limitations, including variability in accuracy, lack of standardized evaluation frameworks, and ongoing uncertainty regarding clinical applicability.[16,17] In the field of infectious diseases, prior evaluations have shown that although LLMs may assist antibiotic decision-making, their accuracy declines in complex real-world cases, leading to the conclusion that relying solely on LLMs for antibiotic selection currently carries significant clinical risks.[18] Similarly, studies assessing LLMs in infection control scenarios have identified substantial errors in clinical judgment and practical applicability, emphasizing that LLMs cannot replace trained infection control professionals.[19]

Taken together, these findings indicate that the clinical application of LLMs in infectious diseases warrants cautious and systematic evaluation. Therefore, this study aimed to evaluate the performance of four representative LLMs—DeepSeek-R1, ChatGPT-5, Grok 3, and Gemini 2.5 Flash—using practice questions from the China Health Professional Technical Qualification Examination (Infectious Diseases) and real-world clinical case management scenarios. We propose the hypothesis that although LLMs excel in knowledge-based tasks, their performance in complex clinical decision-making is limited; and human-machine collaboration can enhance overall decision quality. By integrating examination-based assessment with clinical decision analysis, this study seeks to comprehensively characterize the strengths and limitations of LLMs in infectious disease decision support, particularly with respect to their accuracy and reliability in clinical medicine.

Materials and Methods

Study Design and Ethics

This study consisted of two components: (1) evaluation of LLMs performance on standardized examination questions and (2) assessment of LLM-assisted clinical decision-making using real-world infectious disease cases (Supplementary Figure 1 in Supplementary material 1). The study was conducted at the Department of Infectious Diseases, the Second Affiliated Hospital of Zhejiang University School of Medicine, China, and was approved by the institutional ethics committee (Approval No. 2025–0671). All procedures complied with the Declaration of Helsinki. Given the retrospective nature of the study and the use of anonymized electronic medical record data only, the requirement for informed consent was waived by the institutional ethics committee.

Examination Questions and Classification

The National Health Professional Technical Qualification Examination is a national medical professional qualification examination administered by the Health Human Resources Development Center under the National Health Commission of the People's Republic of China, and successful completion is required for certification as an attending physician in infectious diseases in China. Question formats include A1/A2 (single-best-answer questions), A3/A4 (case-based question sets), B-type (matching questions), and comprehensive case analysis questions. To simulate examination conditions, 250 infectious disease questions were extracted from the official 2025 practice question book,[20] including 109 A1/A2 questions, 85 A3/A4 questions, 13 B-type questions, and 43 comprehensive case analysis questions. Following established frameworks, questions were classified as low-order (knowledge recall and basic application, including A1/A2, A3/A4 and B-type questions) or high-order (complex case analysis requiring advanced reasoning, including comprehensive case analysis questions).[21,22]

Clinical Case Design

Given the exploratory nature of this comparative evaluation, no formal a priori power analysis and sample size calculation were conducted. Ten clinical cases were constructed from real-world practice, encompassing central nervous system, pulmonary, abdominal, urinary tract, vertebral, ocular, and bloodstream infections, as well as viral, bacterial, fungal, rickettsial, and opportunistic infections (tuberculosis) (Supplementary Tables 1–10 in Supplementary material 2). All cases followed international guidelines. One attending physician drafted the case abstracts in Chinese and in English, which were subsequently reviewed and optimized by a second attending physician. Any discrepancies were resolved through discussion until agreement on accuracy was achieved. All clinical cases were fully de-identified prior to analysis.

LLMs Selection and Query Strategy

Four LLMs were evaluated using their official web-based interfaces (ChatGPT-5 https://chatgpt.com/, DeepSeek-R1 https://chat.deepseek.com/, Gemini 2.5 Flash https://gemini.google.com/, and Grok 3 https://grok.com/). The generation parameters (eg., temperature and other decoding settings) were not explicitly controlled in this study. All models were evaluated using their default settings via the respective official interfaces to reflect typical real-world usage. Interactions were conducted between October 1 and October 17, 2025. All interactions were conducted in the language in which each model demonstrates optimal performance, in order to minimize potential language-related performance bias. Specifically, English was used for interactions with ChatGPT-5, Gemini 2.5 Flash, and Grok 3, while Chinese was used for DeepSeek-R1. For transparency and reproducibility, the exact prompts, including wording, formatting requirements, and task instructions, are provided in Supplementary material 3. All prompts were standardized and applied consistently across all LLMs, without additional system-level customization or few-shot examples. Additionally, model-generated responses are also provided in Supplementary Tables 1–10 in Supplementary material 2. Translations were carried out with strict preservation of the original semantic content. During the examination-based questioning, questions of the same type were presented consecutively within a single conversation, whereas different question types were administered in separate conversation sessions to minimize interference from prior prompts on model responses. For case-based questioning, the conversation was reset after each case. Each large language model was allowed only a single response per question, with no post hoc modification or regeneration of the initial response permitted. All generated responses were recorded.

Data Collection and Scoring

The collection and organization of all LLM-generated outputs were performed by the question submitter. In order to compare between LLMs and human doctors, three resident doctors preparing for the National Health Professional Technical Qualification Examination were additionally invited to independently analyze both practice questions and clinical cases.

All responses were fully de-identified, randomly shuffled, and assigned anonymous labels before being submitted to expert reviewers who were not involved in earlier stages of the study, ensuring blinded evaluation. For examination-based questions, scoring was conducted through comparison with standardized reference answers, with accuracy calculated for lower-order questions and score rates for higher-order questions.

Clinical case responses were independently evaluated by three attending doctors with more than 10 years of clinical experience. To better characterize model-specific strengths across different clinical decision-making steps, we predefined three task stages: (1) the diagnostic stage, focusing on differential diagnosis generation based on initial clinical information; (2) the work-up stage, assessing the appropriateness of recommended diagnostic tests and investigations; and (3) the treatment stage, evaluating the suitability and safety of proposed management strategies. These stages were predefined based on real-world clinical decision-making processes according to clinical guidelines.

A five-point Likert scale encompassing five dimensions (accuracy, completeness, individualization, safety, and readability) was adapted from previously published LLMs evaluation studies and optimized for the present study[23,24] (Supplementary Table 11 in Supplementary material 4). Accuracy and completeness were assessed for both LLM- and human-generated responses, whereas individualization, safety, and readability were evaluated exclusively for LLMs outputs to reflect AI-specific characteristics. Each evaluator independently assessed all response sets using predefined criteria to minimize subjective bias, including four LLM groups, three doctor groups, and three doctor+LLM groups. Evaluators were instructed to focus primarily on clinical accuracy, safety and logical reasoning rather than writing fluency or rhetorical style during scoring, to reduce the impact of language polish. Median scores across the three evaluators were used as the final score for each case and dimension to reduce inter-rater variability. When comparing the LLMs, doctors, and doctors+LLMs groups, median scores within each group were used accordingly. As most scores clustered at 4 or 5, arithmetic means were calculated for visualization purposes, while all statistical analyses were performed on the original, unaggregated data.

Statistical Analysis

All statistical analyses were conducted using SPSS version 21.0 (IBM Corp., Armonk, NY, USA), with data visualization performed using GraphPad Prism version 8.0.1 (GraphPad Software, San Diego, CA, USA). Statistical methods were selected based on the data type and measurement structure. Specifically, the chi-square (χ^2) test was used to compare accuracy rates for low-order questions, while one-way analysis of variance (ANOVA) was applied to evaluate differences in score rates for high-order questions. Normality and homogeneity of variance assumptions for ANOVA were assessed using the Shapiro–Wilk test and Levene's test, respectively. As ordinal and paired data, Likert-scale scores were compared across groups using the Friedman test. Statistical significance was defined as a two-sided p value < 0.05.

Results

Performance on Examination-Based Questions

A total of 250 infectious diseases questions were included, comprising 207 low-order questions and 43 high-order questions.

Comparable Performance of LLMs and Doctors in Low-Order Questions

Across low-order questions, all LLMs demonstrated high accuracy, comparable to that of doctors. DeepSeek-R1 achieved the highest accuracy (191/207, 92.3%), followed by Gemini 2.5 Flash (186/207, 89.9%). Both models showed numerically higher accuracy than doctors (183/207, 88.4%), whereas Grok 3 (182/207, 87.9%) and ChatGPT-5 (181/207, 87.4%) showed slightly lower accuracy at a numerical level (Figure 1A). No statistically significant differences were observed among groups ($p = 0.54$).

These findings indicate that LLMs may be capable of handling examination tasks primarily involving knowledge recall and straightforward clinical application, achieving performance levels comparable to trained doctors.

Performance Gap Favoring Doctors in High-Order Questions

In contrast, performance differed in high-order questions, which require precise option selection under penalty-based scoring rules (see Supplementary material 3 for details). Doctors achieved the highest score rate (164.5/215, 76.5%), with numerically higher performance than all LLMs. Among LLMs, Grok 3 (143/215, 66.5%) achieved the highest score rate, followed by Gemini 2.5 Flash (133.41/215, 62.1%) and DeepSeek-R1 (131.49/215, 61.2%). ChatGPT-5 demonstrated the lowest performance (117.09/215, 54.5%) and was the only model that did not reach the conventional 60% passing threshold (Figure 1B).

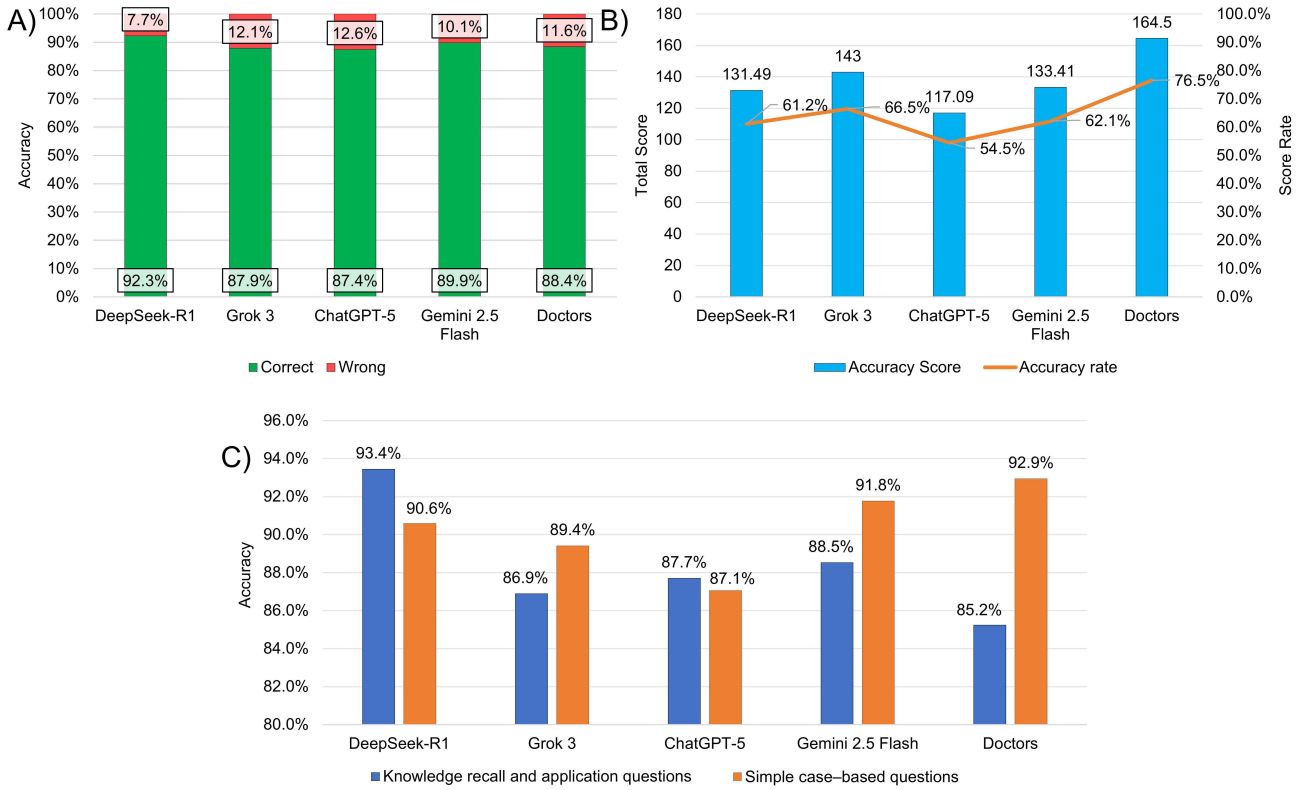

Figure 1 Performance comparison between LLMs and doctors on medical questions. (**A**) Accuracy of LLMs and doctors in answering low-order questions. Green bars indicate correct responses, and red bars indicate incorrect responses. (**B**) Score rates of LLMs and doctors in answering high-order questions. Blue bars indicate accuracy scores (left y-axis), and the Orange line represents the corresponding accuracy rates (right y-axis). (**C**) Differential performance of LLMs and doctors across subtypes of low-order questions. Blue bars represent knowledge recall and application questions, and orange bars represent simple case-based questions.

Although statistical significance was not reached ($p = 0.10$), the consistent numerical performance gap suggests that complex clinical reasoning tasks—requiring integration of diagnostic uncertainty, risk stratification, and therapeutic prioritization—remain a relative weakness for current LLMs.

Differential Strengths of LLMs and Doctors in Low-Order Question Subtypes

Low-order questions were further subdivided into two categories: knowledge recall and application questions (n = 122), including A1/A2 and B-type questions, and simple case-based questions (n = 85), corresponding to A3/A4 questions.

In the knowledge recall and application category, all LLMs displayed numerically higher accuracy than doctors (104/122, 85.2%). DeepSeek-R1 demonstrated the highest accuracy (114/122, 93.4%), followed by Gemini 2.5 Flash (108/122, 88.5%), ChatGPT-5 (107/122, 87.7%), and Grok 3 (106/122, 86.9%) (Figure 1C). Although no statistically significant differences were observed among groups ($p = 0.34$), these findings suggest a trend toward a relative advantage of LLMs in knowledge retention and retrieval.

In contrast, for simple case-based questions, doctors achieved numerically higher accuracy than all LLMs (79/85, 92.9%). Among the LLMs, Gemini 2.5 Flash showed the highest accuracy (78/85, 91.8%), followed by DeepSeek-R1 (77/85, 90.6%) and Grok 3 (76/85, 89.4%), whereas ChatGPT-5 demonstrated the lowest accuracy (74/85, 87.1%) (Figure 1C). Despite the absence of statistically significant differences ($p = 0.74$), these results indicate that LLMs performance remains numerically inferior to that of doctors in questions requiring logical reasoning.

Regional Disease Prevalence May Influence Error Patterns of LLMs

Error analysis of low-order questions revealed that tuberculosis accounted for the largest proportion of errors across all groups, including both LLMs and doctors. Except for DeepSeek-R1 (7/16, 43.8%) and doctors (12/24, 50.0%),

https://doi.org/10.2147/IDR.S593763

tuberculosis-related errors exceeded 60% of total errors in most LLMs trained primarily on Western corpora, including Gemini 2.5 Flash (14/21, 66.7%), ChatGPT-5 (17/26, 65.4%), and Grok 3 (16/25, 64.0%) (Figure 2; Supplementary Table 12 in Supplementary material 5). Given the substantial differences in tuberculosis epidemiology between China and Western countries, this finding highlights the potential influence of regional disease prevalence and training data distribution on LLMs reasoning performance.

Performance in Real-World Clinical Case Management

Overall High Performance with Model-Specific Strengths

In real-world clinical case management, expert reviewers evaluated LLM-generated clinical decisions using a five-dimensional Likert scale. Overall, all LLMs demonstrated high performance across accuracy, completeness, individualization, safety, and readability, with no statistically significant differences observed among models ($p > 0.05$). Gemini 2.5 Flash showed a numerical trend toward slightly higher overall performance, with higher mean scores for accuracy and completeness (accuracy: 4.60 ± 0.56; completeness: 4.70 ± 0.47). Its safety and readability scores (both 4.70 ± 0.53) were similar to those of ChatGPT-5 and DeepSeek-R1. In contrast, Grok 3 showed the lowest score in accuracy (4.33 ± 0.96), safety (4.53 ± 0.97) and readability (4.53 ± 0.90), whereas DeepSeek-R1 had the lowest scores for individualization (4.50 ± 0.63). Additionally, both Grok 3 (4.53 ± 0.68) and DeepSeek-R1 (4.53 ± 0.57) shared the similar lowest score in completeness (Figure 3A; Supplementary Table 13 in Supplementary material 6).

When clinical decision-making was examined by task stage, model-specific strengths were observed. For diagnostic decision-making, Grok 3 achieved the highest scores in completeness and individualization (both 4.80 ± 0.60), while Gemini 2.5 Flash demonstrated the highest accuracy score (4.60 ± 0.49) and safety score (4.80 ± 0.40). Readability scores were highest for Gemini 2.5 Flash (4.80 ± 0.40) and Grok 3 (4.80 ± 0.60). In comparison, DeepSeek-R1 consistently received the lowest scores across all five dimensions—accuracy (4.30 ± 0.90), completeness (4.30 ± 0.64), individualization (4.40 ± 0.66), and safety (4.50 ± 0.67)—with readability (4.60 ± 0.49) comparable to that of ChatGPT-5 (Figure 3B; Supplementary Table 13 in Supplementary material 6). No statistically significant differences were detected among LLMs ($p > 0.05$).

Regarding the diagnostic work-up decision-making, Gemini 2.5 Flash achieved the highest accuracy score (4.70 ± 0.48), whereas DeepSeek-R1 attained the highest readability score (4.80 ± 0.42). The scores for completeness, individualization, and safety were comparable between Gemini 2.5 Flash and DeepSeek-R1 (all 4.70 ± 0.48). By contrast, Grok 3 yielded the lowest scores across all five dimensions, with accuracy, completeness, individualization,

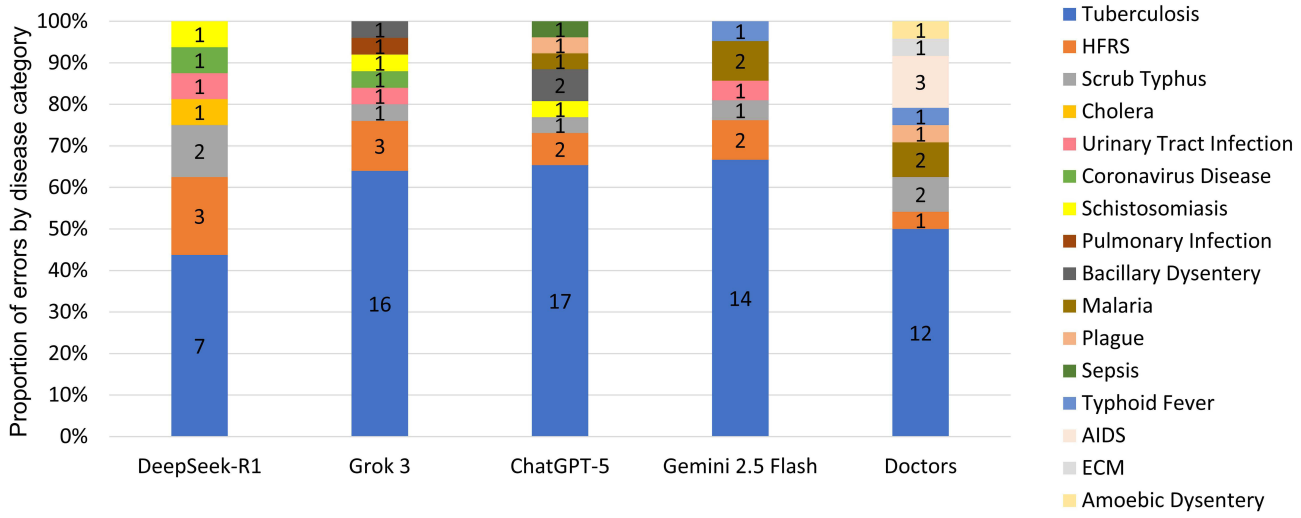

Figure 2 Distribution of disease categories among errors in low-order questions generated by LLMs and doctors. Stacked bar charts show the proportional distribution of disease categories. Each colored segment represents one infectious disease category, with the corresponding count labeled inside each segment. The cumulative percentage for each group equals 100%.

Abbreviations: HFRS, Hemorrhagic Fever with Renal Syndrome; ECM, Epidemic Cerebrospinal Meningitis; AIDS, Acquired Immunodeficiency Syndrome.

https://doi.org/10.2147/IDR.S593763

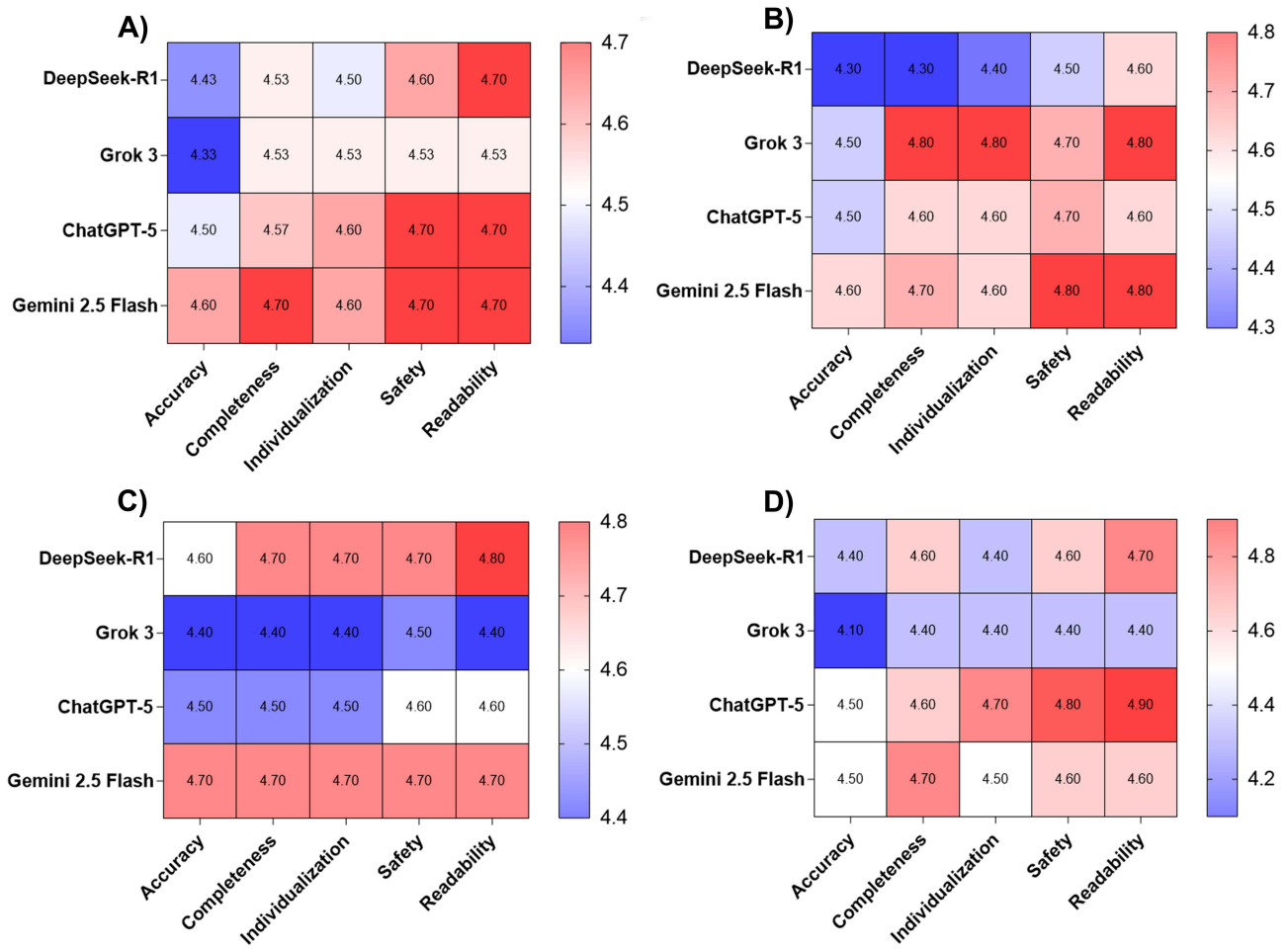

Figure 3 Multidimensional Likert-scale assessment of LLMs performance in clinical case management. Mean Likert-scale scores (1–5) across four LLMs and five dimensions (Accuracy, Completeness, Individualization, Safety, Readability) are shown, with red indicating higher scores and blue indicating lower scores. (**A**) Overall Assessment, (**B**) Assessment of Diagnosis, (**C**) Assessment of Diagnostic Work-up, (**D**) Assessment of Treatment Plan.

and readability each at 4.40 ± 0.70, and safety at 4.50 ± 0.71 (Figure 3C; Supplementary Table 13 in Supplementary material 6). No statistically significant differences were observed among LLMs ($p > 0.05$).

In treatment decision-making, Gemini 2.5 Flash and ChatGPT-5 achieved similarly high accuracy scores (4.50 ± 0.71 and 4.50 ± 0.53, respectively), while Gemini 2.5 Flash attained the highest completeness score (4.70 ± 0.48). ChatGPT-5 demonstrated the highest scores in individualization (4.70 ± 0.48), safety (4.80 ± 0.42), and readability (4.90 ± 0.32). In contrast, Grok 3 showed the lowest performance across all dimensions, with accuracy at 4.10 ± 1.20, completeness at 4.40 ± 0.70, and both safety and readability at 4.40 ± 1.26. Its individualization score (4.40 ± 0.97) was comparable to that of DeepSeek-R1 (4.40 ± 0.70). (Figure 3D; Supplementary Table 13 in Supplementary material 6). No statistically significant differences were identified among LLMs ($p > 0.05$).

Clinically Meaningful Errors Highlight Limitations of Standalone LLM Decision-Making

Overall, all LLMs demonstrated relatively high accuracy in clinical diagnostic and therapeutic decision-making; however, notable errors were still observed. In Case 1 (Supplementary Table 1 in Supplementary material 2), which involved a patient with post-influenza pneumonia and a history of chronic obstructive pulmonary disease, the top three differential diagnoses generated by DeepSeek-R1 and Grok 3 did not include invasive pulmonary aspergillosis. Notably, Influenza-associated pulmonary aspergillosis (IAPA) has been widely reported in the literature,[25] with IAPA likely affecting 10–20% of critically ill influenza patients and being associated with overall mortality rates of approximately 50%.[26]

In this patient, despite antiviral and antibacterial therapy, the clinical course worsened, accompanied by the development of multiple patchy pulmonary infiltrates, which should raise strong suspicion for invasive pulmonary

https://doi.org/10.2147/IDR.S593763

aspergillosis. More concerningly, even after additional diagnostic evidence was provided—including a markedly elevated Aspergillus IgG level (500.00 AU/mL, normal range <120 AU/mL) and bronchoalveolar lavage fluid Next-Generation Sequencing detecting *Aspergillus fumigatus* sequences (read count: 12)—Grok 3 still responded: "Elevated IgG and NGS sequences suggest possible colonization or chronic infection, but negative GM and low sequence numbers make immediate antifungal treatment (e.g., voriconazole) unnecessary unless clinical worsening or CT findings (e.g., cavities) suggest CPA." These findings underscore that, at the current stage, the use of LLMs as standalone tools for medical decision-making remains unsafe and requires close supervision and critical review by experienced clinicians.

LLMs Assistance Enhances Accuracy and Completeness of Physician Decision-Making

In real-world clinical case management, the accuracy and completeness of decisions from independent doctors and LLM-supported doctors were scored by an expert panel. These results were further compared with those of standalone LLMs. The results showed that doctors achieved a mean accuracy score of 4.50 ± 0.68 (diagnosis: 4.60 ± 0.70; auxiliary examination: 4.50 ± 0.71; treatment: 4.40 ± 0.70), which was slightly higher than that of LLMs alone (4.48 ± 0.61; diagnosis: 4.50 ± 0.67; auxiliary examination: 4.55 ± 0.50; treatment: 4.40 ± 0.70). When doctors were supported by LLMs, the accuracy score increased to 4.67 ± 0.55 (diagnosis: 4.70 ± 0.48; auxiliary examination: 4.80 ± 0.42; treatment: 4.50 ± 0.71) (Figure 4A; Supplementary Table 14 and Supplementary Table 15 in Supplementary material 6).

Regarding completeness, doctors achieved a mean score of 4.50 ± 0.57 (diagnosis: 4.80 ± 0.42; auxiliary examination: 4.30 ± 0.67; treatment: 4.40 ± 0.52), which was slightly lower than that of LLMs alone (4.60 ± 0.48; diagnosis: 4.60 ± 0.52; auxiliary examination: 4.60 ± 0.46; treatment: 4.60 ± 0.52). With LLMs assistance, doctors' completeness scores increased to 4.90 ± 0.31 (diagnosis, auxiliary examination, and treatment: each 4.90 ± 0.32) (Figure 4B; Supplementary Table 14 and Supplementary Table 15 in Supplementary material 6).

No statistically significant differences were observed in accuracy across groups (p = 0.19). However, for completeness, statistically significant differences were identified (p = 0.0006). Post hoc analysis demonstrated that doctors supported by LLMs achieved significantly higher completeness scores compared with both the doctors alone (p = 0.0015) and the LLMs alone (p = 0.0010). Qualitatively, doctors often provided concise, key, and correct decision-making recommendations, whereas LLMs tended to generate more comprehensive and detailed responses when thoroughness was required. Notably, when doctors made decisions with the support of LLMs, both accuracy (numerically) and completeness (significantly) improved. These findings suggest that LLMs can serve as an important adjunctive tool in clinical medical decision-making.

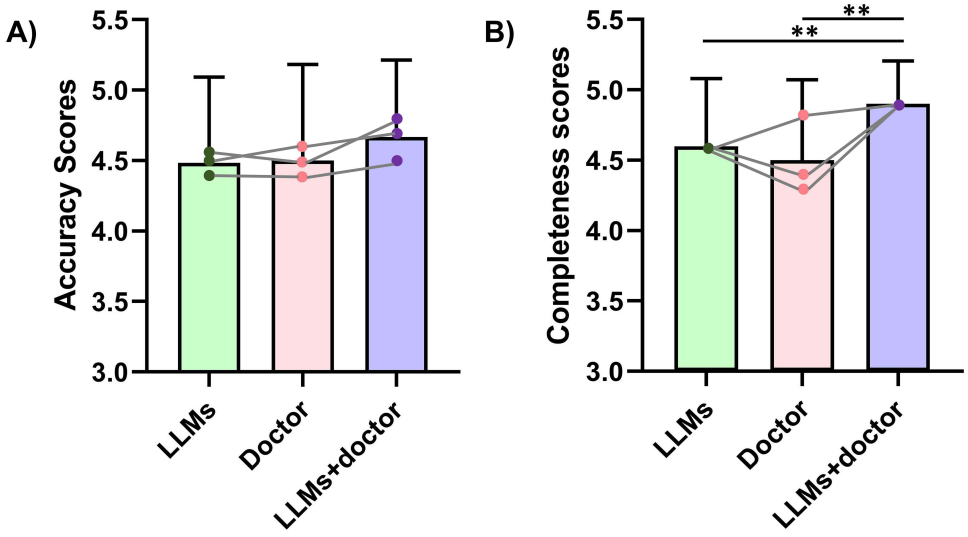

Figure 4 Assessment of accuracy and completeness in clinical case management across three groups: LLMs, doctors, and LLM-assisted doctors. (**A**) Accuracy, (**B**) Completeness. Bars show mean scores ± SD, connected points within each bar represent mean scores for diagnosis, auxiliary examination, and treatment planning. **p < 0.01.

Discussion

In recent years, the rapid advancement of artificial intelligence has drawn substantial attention to its potential applications in healthcare. Previous studies have reported encouraging results across a range of domains, including medical imaging, pathological analysis, epidemiological research, and clinical decision support.[27–30] However, systematic evaluations focused specifically on the diagnosis and management of infectious diseases remain limited. In this study, we combined practice questions from the China Health Professional Technical Qualification Examination with real-world clinical cases to evaluate the performance of four widely used LLMs in infectious disease management from two complementary perspectives: examination performance and clinical decision-making.

Overall, LLMs demonstrated strong performance in both examination-based questions and clinical case analyses, with performance generally comparable to that of doctors. However, variations across task types suggest that current LLMs are more effective in structured knowledge-based tasks than in complex clinical reasoning tasks. This pattern aligns with previous studies showing that LLMs perform well on standardized, fixed-answer tasks but struggle with complex clinical reasoning and individualized decision-making that rely on accumulated clinical experience.[21,22] Beyond independent performance, our findings also highlighted the potential value of LLMs within a human–AI collaborative framework. In complex cases, clinical decisions may be influenced by individual experience, knowledge structures, and cognitive biases. Meanwhile, healthcare resources are not evenly distributed worldwide. By integrating large-scale medical knowledge, structuring clinical reasoning, and highlighting potential omissions, LLMs may help mitigate gaps in clinical decision support, especially in resource-constrained environments. To achieve meaningful clinical utility, healthcare institutions should actively explore integrating LLMs into existing hospital information systems. Embedding LLMs into electronic health records (EHR) and clinical decision support systems (CDSS) can assist with diagnostic reasoning, auxiliary test selection, and treatment planning.[31] Technical compatibility, operational accessibility, and timely updates with guideline-based knowledge may facilitate the transition of LLMs from standalone research tools to practical assistants in routine clinical practice.

Despite these promising prospects, notable safety challenges persist. Although LLMs generally provided accurate and comprehensive recommendations, they sometimes fail to recognize high-risk clinical scenarios that require careful judgment. In this study, several models did not promptly recognize the elevated risk of invasive pulmonary aspergillosis in patients with influenza, leading to diagnostic errors and delayed antifungal recommendations. Consistent with previous studies, challenges such as hallucinations, delayed knowledge updating, and limited contextual understanding persist,[32] and LLMs may also provide inaccurate responses with high confidence.[33] These limitations must be overcome before LLMs can be broadly used in clinical practice. Medico-legal responsibility is still poorly defined if AI-generated recommendations depart from clinical guidelines or bring about adverse clinical outcomes.[34] Current regulatory systems still have no consistent standards for the clinical use of generative AI. These challenges highlight the necessity of establishing robust governance rules and standardized validation processes before such tools are implemented in routine clinical care. Meanwhile, at the current stage, LLMs should act merely as auxiliary clinical decision-support tools and remain under close and constant clinician supervision to ensure safe, standardized and proper clinical use in real-world practice.

Error analysis of examination questions further indicated that LLMs performance is influenced by the disease spectrum represented in training data. Higher error rates were observed for tuberculosis-related questions, particularly among models trained predominantly on Western-language corpora. LLMs are typically trained on distinct textual corpora, with some relying heavily on Western biomedical literature while others incorporate broader regional and non-Western clinical resources. Given marked epidemiological and clinical practice differences between high- and low-burden tuberculosis settings,[35] such discrepancies may be partly attributable to differences in the underlying training data, which may affect how disease-related knowledge is weighted during training and applied in region-specific contexts.[36] In addition, variations in clinical guidelines, diagnostic pathways, and empirical treatment strategies across regions may further contribute to these disparities. Thus, LLMs performance depends not only on general capabilities but also on the representation of regional disease patterns and local clinical practices. Future development and deployment of LLMs in healthcare would benefit from the incorporation of representative localized medical data, region-specific clinical guidelines, and real-world case materials, as well as continuous feedback from medical professionals, to enhance safety, reliability, and clinical applicability in specific healthcare environments.

Finally, although overall performance among the evaluated LLMs was broadly comparable, model-specific tendencies were observed. DeepSeek-R1 showed higher accuracy on low-order examination questions, whereas Gemini 2.5 Flash achieved higher Likert-scale scores in clinical case analyses. These variations may reflect differences in training objectives, reinforcement learning strategies, and reasoning mechanisms.[37] However, the modest differences observed do not indicate clear superiority of any single model. Future work should therefore focus on task-oriented optimization and better integration of LLMs into routine clinical workflows, so as to fully realize their potential as safe, reliable aids for clinical decision-making.

Several limitations of this study should be acknowledged. First, these exam-style questions were released before the knowledge cutoff dates of the assessed LLMs. While we cannot definitively confirm whether this content was used for model training, the questions originate from printed publications with no official online editions. This greatly reduces the likelihood of them being included in public training corpora. Second, the number of real-world clinical cases included in this study was modest, which may have reduced the ability to detect small between-group differences. In addition, the cases were obtained from a single center and were subject to predefined selection criteria, which may not fully reflect the heterogeneity of routine clinical practice. To mitigate this, a structured case selection process was employed to include infections involving different organ systems and diverse pathogen types, aiming to enhance representativeness. Third, all expert evaluators were drawn from the same department within a single institution. While this helped maintain consistent clinical standards, it may also reflect similar institutional practice routines. Furthermore, although the evaluation criteria were adapted from previously published frameworks, they may not cover all facets of clinical reasoning or decision quality. Finally, inherent differences in language optimization and expressive ability across different LLMs may cause subjective rating bias, as models with more fluent language tend to receive higher ratings, even when their actual clinical reasoning performance is comparable.

Accordingly, future research should focus on validating these findings in larger multicenter cohorts and prospective real-world clinical studies across diverse healthcare settings. Particular attention should be given to endemic and region-specific infectious diseases to further explore how differences in training corpora and regional medical knowledge influence LLM performance. In addition, real-world studies evaluating the integration of LLMs into clinical workflows will be essential to assess their impact on diagnostic efficiency, clinical decision-making, and patient outcomes. Further efforts are also needed to establish standardized evaluation frameworks, optimize human–AI collaborative models, and improve alignment with regional clinical guidelines. Finally, continued refinement of regulatory and medico-legal frameworks will be critical to support the safe, ethical, and responsible deployment of LLMs in clinical practice.

Conclusion

Overall, LLMs demonstrated substantial potential in infectious disease diagnosis and clinical decision support. Human–AI collaboration raised doctors' overall completeness score from 4.50 to 4.90, corresponding to an 8.89% improvement. Although LLMs currently cannot replace clinicians in high-order clinical reasoning, their integration into a doctor–LLM collaborative model markedly boosted decision-making quality. Active exploration of LLM integration into routine clinical workflows, together with strengthened legal and regulatory governance of medical applications, will be essential to ensure safe, reliable, and efficient implementation of LLM-assisted healthcare delivery.

Acknowledgments

This research was funded by Noncommunicable Chronic Diseases-National Science and Technology Major Project (2025ZD0549100) and Key research and development program of Zhejiang Province (2025C02093). The authors declare that generative artificial intelligence tools (ChatGPT-5 and DeepSeek-R1) were used during manuscript preparation for language editing and improvement of clarity and readability. The AI tool was not used to generate scientific content, analyze data, or draw clinical conclusions. All content was reviewed and approved by the authors, who take full responsibility for the manuscript.

Disclosure

The author(s) report no conflicts of interest in this work.

References

1. Meyer A, Riese J, Streichert T. Comparison of the performance of GPT-3.5 and GPT-4 with that of medical students on the written german medical licensing examination: observational study. *JMIR Med Educ*. 2024;10:e50965. doi:10.2196/50965
2. Lei Y-H, Chen -C-C, Shen C-J. Token-splitting improves GPT-4.1 performance on plastic surgery exams: implications for AI-Assisted medical education. *Med Educ Online*. 2025;30(1):2602788. doi:10.1080/10872981.2025.2602788
3. Shusterman R, Waters AC, O'Neill S, Bangs M, Luu P, Tucker DM. An active inference strategy for prompting reliable responses from large language models in medical practice. *NPJ Digit Med*. 2025;8(1):119. doi:10.1038/s41746-025-01516-2
4. Liu X, Liu H, Yang G, et al. A generalist medical language model for disease diagnosis assistance. *Nat Med*. 2025;31(3):932–942. doi:10.1038/s41591-024-03416-6
5. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Netw Open*. 2024;7(10):e2440969. doi:10.1001/jamanetworkopen.2024.40969
6. Oh Y, Park S, Byun HK, et al. Author Correction: LLM-driven multimodal target volume contouring in radiation oncology. *Nat Commun*. 2025;16(1):718. doi:10.1038/s41467-025-55963-2
7. Blecha Z, Jasinski D, Jaworski A, et al. Performance of GPT-4o and DeepSeek-R1 in the polish infectious diseases specialty exam. *Cureus*. 2025;17(4):e82870. doi:10.7759/cureus.82870
8. Bizel-Bizellot G, Galmiche S, Lelandais B, et al. Extracting circumstances of Covid-19 transmission from free text with large language models. *Nat Commun*. 2025;16(1):5836. doi:10.1038/s41467-025-60762-w
9. Maillard A, Micheli G, Lefevre L, et al. Can chatbot artificial intelligence replace infectious diseases physicians in the management of bloodstream infections? A prospective cohort study. *Clin Infect Dis*. 2024;78(4):825–832. doi:10.1093/cid/ciad632
10. Guo D, Yang D, Zhang H, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*. 2025;645(8081):633–638. doi:10.1038/s41586-025-09422-z
11. Sandmann S, Hegselmann S, Fujarski M, et al. Benchmark evaluation of DeepSeek large language models in clinical decision-making. *Nat Med*. 2025;31(8):2546–2549. doi:10.1038/s41591-025-03727-2
12. Tordjman M, Liu Z, Yuce M, et al. Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning. *Nat Med*. 2025;31(8):2550–2555. doi:10.1038/s41591-025-03726-3
13. Moell B, Sand Aronsson F, Akbar S. Medical reasoning in LLMs: an in-depth analysis of DeepSeek R1. *Front Artif Intell*. 2025;8:1616145. doi:10.3389/frai.2025.1616145
14. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med*. 2023;6(1):120. doi:10.1038/s41746-023-00873-0
15. Luo M-J, Pang J, Bi S, et al. Development and evaluation of a retrieval-augmented large language model framework for ophthalmology. *JAMA Ophthalmol*. 2024;142(9):798–805. doi:10.1001/jamaophthalmol.2024.2513
16. Shool S, Adimi S, Saboori Amleshi R, et al. A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Med Inform Decis Mak*. 2025;25(1):117. doi:10.1186/s12911-025-02954-4
17. Chen SF, Alyakin A, Seas A, et al. LLM-assisted systematic review of large language models in clinical medicine. *Nat Med*. 2026;32(3):1152–1159. doi:10.1038/s41591-026-04229-5
18. De Vito A, Geremia N, Bavaro DF, et al. Comparing large language models for antibiotic prescribing in different clinical scenarios: author's response. *Clin Microbiol Infect*. 2025;31(8):1410–1414. doi:10.1016/j.cmi.2025.04.041
19. Wong S-C, Chiu EK-Y, Chiu KH-Y, et al. Comparative evaluation and performance of large language models in clinical infection control scenarios: a benchmark study. *Healthcare*. 2025;13(20):2652. doi:10.3390/healthcare13202652
20. National Health Commission of the People's Republic of China. *Practice Questions for the China Health Professional Technical Qualification Examination: Infectious Diseases (2025 Edition)*. Beijing: People's Medical Publishing House; 2024.
21. He Q, Tan Z, Niu W, et al. From algorithms to operating room: can large language models master China's attending anesthesiology exam? A cross-sectional evaluation. *Int J Surg*. 2026;112(1):190–201. doi:10.1097/JS9.0000000000003406.
22. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. *Neurosurgery*. 2023;93(6):1353–1365. doi:10.1227/neu.0000000000002632
23. Wang Y-L, Tian L-C, Meng J-Y, et al. Evaluation of large language models in patient education and clinical decision support for rotator cuff injury: a two-phase benchmarking study. *BMC Med Inform Decis Mak*. 2025;25(1):289. doi:10.1186/s12911-025-03105-5
24. Li Q, Zhan L, Cai X. Assessing DeepSeek-R1 for clinical decision support in multidisciplinary laboratory medicine. *J Multidiscip Healthc*. 2025;18:4979–4988. doi:10.2147/JMDH.S538253
25. Shi J-X, Shao X-R, Wu P, et al. Clinical features of influenza-associated pulmonary aspergillosis: a retrospective multicenter cohort study. *Front Cell Infect Microbiol*. 2025;15:1648547. doi:10.3389/fcimb.2025.1648547
26. Feys S, Carvalho A, Clancy CJ, et al. Influenza-associated and COVID-19-associated pulmonary aspergillosis in critically ill patients. *Lancet Respir Med*. 2024;12(9):728–742. doi:10.1016/S2213-2600(24)00151-6
27. Jiang H, Wu M, Yu J, Xiao Y, Yin C. AI-powered epidemic control: deepseek's role in global health resilience. *J Transl Med*. 2025;23(1):1395. doi:10.1186/s12967-025-07082-1
28. Frascarelli C, Venetis K, Marra A, et al. Computational pathology in breast cancer: optimizing molecular prediction through task-oriented AI models. *NPJ Breast Cancer*. 2025;11(1):141. doi:10.1038/s41523-025-00857-1
29. Monkam P, Wang X, Liu S, et al. Deep learning-driven innovations in echocardiography: taxonomy, clinical impact, challenges, and opportunities. *Ann Biomed*. 2026;54(3):641–678. doi:10.1007/s10439-025-03944-3
30. Fu J, Fang M, Wu L, et al. Development, advancement, and clinical integration of artificial intelligence technology in gastric cancer. *Chin Med J*. 2025;138(24):3332–3350. doi:10.1097/CM9.0000000000003922

31. Schwab JH. AI-based medical decision support: exploring the data gap. *J Bone Joint Surg Am*. 2026;108(4):266–268. doi:10.2106/JBJS.25.01387
32. Zeng H, Liu G, Liu W, et al. Evaluation of large language models in the clinical management of multiple chronic conditions: a comparative study of DeepSeek-R1 and ChatGPT-o1. *Public Health*. 2025;249:106042. doi:10.1016/j.puhe.2025.106042
33. Gotta J, Le Hong QA, Koch V, et al. Large language models (LLMs) in radiology exams for medical students: performance and consequences. *Rofo*. 2025;197(9):1057–1067. doi:10.1055/a-2437-2067
34. Duffourc M, Gerke S. Generative AI in health care and liability risks for physicians and safety concerns for patients. *JAMA*. 2023;330(4):313–314. doi:10.1001/jama.2023.9630
35. Idayat M, von der Lippe E, Kozhekenova N, et al. Prevalence of tuberculosis in central Asia and Southern Caucasus: a systematic literature review. *Diagnostics*. 2025;15(18):2314. doi:10.3390/diagnostics15182314
36. Sourati Z, Ziabari AS, Dehghani M. The homogenizing effect of large language models on human expression and thought. *Trends Cogn Sci*. 2026: S1364-6613(26)00003–3. doi:10.1016/j.tics.2026.01.003
37. Zhou J, Li H, Chen S, et al. Large language models in biomedicine and healthcare. *npj Artif Intell*. 2025:1;44. doi:10.1038/s44387-025-00047-1