Critical Care

RESEARCH Open Access

A large language model improves clinicians' diagnostic performance in complex critical illness cases

Xintong Wu[1†], Yu Huang[1*†] and Qing He[1]

Abstract

Background Large language models (LLMs) have demonstrated potential in assisting clinical decision-making. However, studies evaluating LLMs' diagnostic performance on complex critical illness cases are lacking. We aimed to assess the diagnostic accuracy and response quality of an artificial intelligence (AI) model, and evaluate its potential benefits in assisting critical care residents with differential diagnosis of complex cases.

Methods This prospective comparative study collected challenging critical illness cases from the literature. Critical care residents from tertiary teaching hospitals were recruited and randomly assigned to non-AI-assisted physician and AI-assisted physician groups. We selected a reasoning model, DeepSeek-R1, for our study. We evaluated the model's response quality using Likert scales, and we compared the diagnostic accuracy and efficiency between groups.

Results A total of 48 cases were included. Thirty-two critical care residents were recruited, with 16 residents assigned to each group. Each resident handled an average of 3 cases. DeepSeek-R1's responses received median Likert grades of 4.0 (IQR 4.0–5.0; 95% CI 4.0–4.5) for completeness, 5.0 (IQR 4.0–5.0; 95% CI 4.5–5.0) for clarity, and 5.0 (IQR 4.0–5.0; 95% CI 4.0–5.0) for usefulness. The AI model's top diagnosis accuracy was 60% (29/48; 95% CI 0.456–0.729), with a median differential diagnosis quality score of 5.0 (IQR 4.0–5.0; 95% CI 4.5–5.0). Top diagnosis accuracy was 27% (13/48; 95% CI 0.146–0.396) in the non-AI-assisted physician group versus 58% (28/48; 95% CI 0.438–0.729) in the AI-assisted physician group. Median differential quality scores were 3.0 (IQR 0–5.0; 95% CI 2.0–4.0) without and 5.0 (IQR 3.0–5.0; 95% CI 3.0–5.0) with AI assistance. The AI model showed higher diagnostic accuracy than residents, and AI assistance significantly improved residents' accuracy. The residents' diagnostic time significantly decreased with AI assistance (median, 972 s; IQR 570–1320; 95% CI 675–1200) versus without (median, 1920 s; IQR 1320–2640; 95% CI 1710–2370).

Conclusions For diagnostically difficult critical illness cases, DeepSeek-R1 generates high-quality information, achieves reasonable diagnostic accuracy, and significantly improves residents' diagnostic accuracy and efficiency. Reasoning models are suggested to be promising diagnostic adjuncts in intensive care units.

Keywords Generative artificial intelligence, Reasoning models, Critical care, Diagnostic dilemmas

Introduction

Large language models (LLMs), a subclass of artificial intelligence (AI), have become increasingly prevalent in healthcare applications, including medical education, research, and clinical care [1, 2]. Healthcare professionals have shown increasing interest in assessing the role

†Xintong Wu and Yu Huang contributed equally to this work.

*Correspondence:
Yu Huang
huangyu0120er@126.com
[1] Department of Intensive Care Medicine, Affiliated Hospital of Southwest Jiaotong University, The Third People's Hospital of Chengdu, Chengdu, Sichuan, China

of LLM chatbots in clinical practice [1, 2]. Recent studies demonstrated that LLMs without specialized training could pass all three stages of the United States Medical Licensing Exam (USMLE) [3]. Furthermore, LLMs have demonstrated the capacity to generate high-quality, empathetic, and readable responses to patient questions compared to physicians [4–6], and can generate largely accurate information for diverse medical queries developed by physicians across multiple specialties [7, 8]. These findings suggest LLMs may assist physicians with more complex clinical decision-making, such as differential diagnosis in challenging cases.

Diagnostic dilemmas commonly occur in critically ill patients due to the complexity and diversity of clinical presentations, making differential diagnosis particularly challenging as it requires comprehensive knowledge and the ability to integrate complex information [9]. In intensive care units (ICUs), rapid and accurate diagnosis is crucial for making timely, effective decisions to improve patient outcomes [9]. A recent study demonstrated that LLMs such as ChatGPT-4o and ChatGPT-4o-mini achieved high accuracy and consistency in answering critical care medicine questions at a European examination level [10]. However, this study focused on examination-level questions which might not fully reflect the complexity and depth of understanding required in real-world clinical practice. Another study found that both ChatGPT-3.5 and ChatGPT-4.0 generated hallucinations (misinformation delivered with high confidence) and lacked consistency when asked the same question multiple times across 50 core critical care topics [11]. Therefore, a knowledge gap exists regarding the effectiveness of such LLMs in aiding differential diagnosis for complex ICU cases.

Reasoning models represent a significant evolution of traditional LLMs, standing out for their ability to address complex problems through structured, sequential thinking processes [12]. DeepSeek-R1, a reasoning model released in January 2025 by DeepSeek, is an open-source model based on reinforcement learning techniques [13]. It has become the fastest-growing consumer application and demonstrates potential for greater utility in complex critical care scenarios compared to traditional LLMs. In the present study, we evaluated the diagnostic performance of DeepSeek-R1 on complex critical illness cases, compared the diagnostic accuracy and efficiency of critical care physicians with and without DeepSeek-R1 assistance for these cases, to assess the reasoning model's potential benefits in these scenarios.

Methods

We conducted this prospective comparative study in March 2025. As this study was based on published case reports from medical journals, ethics board approval was not required due to the lack of involvement of patient data. We followed the TRIPOD-LLM guideline for reporting [14].

Case selection

We searched for cases of critical illness published after December 2023 (the date of DeepSeek-R1's training) using the following method: We conducted a comprehensive literature search in PubMed using the combination of the keywords "case report," "case," "diagnosis," "differential diagnosis," "diagnostic challenge," "intensive care units," "critical illness," "multiple organ failure," "respiratory failure," "shock," "hypoxia," "respiratory insufficiency," "renal insufficiency," "dyspnea," "critical illness," and "clinical reasoning." We also reviewed similar articles from relevant journals or publications. Cases that were challenging to diagnose and involved a complexity of manifestations and processes were included. In the absence of objective criteria to define case complexity in the relevant literature, the eligibility of cases was determined by consensus between two authors (XW and YH). Typical cases that met the eligibility criteria were case challenges from the *New England Journal of Medicine* (NEJM). These cases are diagnostic dilemmas and have been used for medical education of complex diagnostic reasoning. We included recent published cases to ensure that the AI model had not been trained on these cases beforehand. We excluded cases on management reasoning, non-critical cases, and cases without final diagnoses, as determined by consensus between two authors (XW and YH).

AI model and prompting

We selected DeepSeek-R1, a reasoning model released in January 2025 and last trained in December 2023, for the present study. The website version was used, with a total params of 671B, a maximal token length of 128 K and a default temperature of 0.6. Each case's clinical summary, including the patient's history, physical examination findings, relevant investigation results and clinical course, was iteratively transformed into a standard prompt: "Act as an attending physician. A summary of the patient's clinical information will be presented, and you will use this information to predict the diagnosis. Describe the differential diagnoses and the rationale for each, listing the most likely diagnosis at the top: [case information]." Each case was run in an independent chat session to prevent the model from applying any "learning" to subsequent cases. No retrieval-augmented generation (RAG) was incorporated in the model search. The prompt was developed to encourage the AI model to generate consistent and inclusive responses, using another dataset of

cases that were published in 2023 and met our inclusion criteria.

Participants and study groups

As residents are most likely to require external diagnostic assistance, we recruited critical care medicine residents from six tertiary teaching hospitals. All residents were blinded to the study design and were unaware that their responses would be compared to AI models' responses. They were randomly assigned to either non-AI-assisted physician or AI-assisted physician group using stratified randomization based on years of critical care experience. All participants confirmed no prior exposure to the study cases. In the non-AI-assisted physician group, cases were randomly allocated to residents who were allowed to use traditional resources (PubMed, UpToDate, Medscape, etc.). The residents were provided the same descriptions that were input into the AI model and were asked to make a differential diagnosis list with the most likely diagnosis at the top for each case. In the AI-assisted physician group, the same cases were allocated randomly to the residents, and the residents received both case descriptions and corresponding AI model outputs. Traditional resources were also allowed in the AI-assisted physician group. Search results were limited to pre-December 2023 publications when traditional resources were used.

Outcomes

We evaluated the diagnostic accuracy according to top diagnosis accuracy and differential quality score (a previously published ordinal 5-point rating system based on accuracy and usefulness. A score of 5 is given for a differential including the exact diagnosis; 4: the suggestions included something very close, but not exact; 3: the suggestions included something closely related that might have been helpful; 2: the suggestions included something related, but unlikely to be helpful; 0: no suggestions were close to the target diagnosis) [15]. DeepSeek-R1's response quality was further evaluated in terms of completeness (the degree to which the information was medically comprehensive and offered broad diagnostic possibilities and additional details: 1 = very incomplete; 2 = incomplete; 3 = moderately complete; 4 = complete; 5 = very complete), clarity (conciseness of presentation and overall legibility: 1 = very unclear; 2 = unclear; 3 = somewhat clear; 4 = clear; 5 = very clear), and usefulness (the degree to which the response provided helpful and effective decision-critical information: 1 = not useful; 2 = slightly useful; 3 = moderately useful; 4 = useful; 5 = highly useful) using 5-point Likert scales. The outcomes were independently assessed by two authors (XW and YH), with disagreements resolved by a third author (QH). In the assessment of each outcome, a consensus score was determined for each case in each condition (AI model, non-AI-assisted physicians, and AI-assisted physicians). Diagnostic time was recorded and analyzed for the evaluation of diagnostic efficiency.

We performed a consistency check to assess the reliability of DeepSeek-R1. To assess, 16 cases were randomly selected from the included cases and were presented to the AI model 3 times in independent conversations. The model's response was considered consistent if it provided the same top diagnosis and a similar differential diagnosis list for 3 repetitions, while did not introduce or omit major information that would affect the response quality. The response for each case was assessed by two authors (XW and YH) independently, and was considered consistent if both the two authors scored consistent.

Statistical analysis

The top diagnosis accuracy was reported as percentages and analyzed using χ^2 test with Bonferroni correction. Differential quality scores and Likert scores were reported using medians and interquartile ranges (IQR). Kruskal–Wallis H test was used to determine differences in differential quality scores between groups with Dunn's post-hoc pairwise comparison, and Mann–Whitney U test was used to identify difference in diagnostic time between groups. The 95% confidence intervals (CI) for the top diagnosis accuracy, differential quality scores, Likert scores and diagnostic time were calculated. Inter-rater reliability was evaluated using intraclass correlation coefficient (ICC) calculated by a mean-rating, absolute agreement, two-way mixed-effects model. A p value of < 0.05 was considered statistically significant. Analyses and visualizations used Python 3.12 (SciPy 1.13.0, numpy 2.1.3, pandas 2.2.1, matplotlib 3.8.3, seaborn 0.13.2, pingouin 0.5.4).

Results

A total of 48 cases were included in the study, the final included cases were published in the NEJM, Mayo Clinic Proceedings, CHEST, Neurology et al. These selected cases were characterized by their diagnostic complexity and clinical relevance. Thirty-two critical care residents were recruited, with 16 residents assigned to each group. Each resident handled an average of 3 cases, with a maximum of 5 and minimum of 1. Participants' critical care experience (beyond trainee) ranged from 1 to 3 years.

A representative example of AI-generated response and differential diagnoses from each group is shown in Table 1. DeepSeek-R1 demonstrated strong performance across completeness, clarity, and usefulness, with median Likert scores of 4.0 (IQR 4.0–5.0; 95% CI 4.0–4.5), 5.0 (IQR 4.0–5.0; 95% CI 4.5–5.0), and 5.0 (IQR 4.0–5.0; 95%

Table 1 A representative example of AI-generated response and differential diagnoses from each group, along with subsequent scores

Final diagnosis	Case information	DeepSeek-R1's responses	Differential diagnoses from non-AI-assisted physicians	Differential diagnoses from AI-assisted physicians	Likert scores for AI's responses (Completeness/Clarity/Usefulness)[a]	Differential quality scores (DeepSeek-R1/non-AI-assisted physicians/AI-assisted physicians)[b]
Icteric leptospirosis	A 37-year-old male patient presented with fever, myalgia, jaundice, and hypoxemia, following a 9-day progressive illness characterized initially by severe fatigue, exhaustion, and generalized weakness that rendered him bedridden. Seven days prior to admission, he developed high-grade fever (39.4 °C/102.9°F) accompanied by headache, limb pain and stiffness, anorexia, and nausea without abdominal pain, vomiting, or diarrhea. While his fever and headache resolved after 2 days, his myalgia worsened with the onset of dark-colored urine. Initial evaluation at an emergency clinic 3 days before admission revealed negative respiratory viral panel testing (severe acute respiratory syndrome coronavirus 2, respiratory syncytial virus, influenza A/B), and he was managed conservatively with hydration advice. Due to persistent symptoms and emerging jaundice noted 2 days later, he was referred to the emergency department where initial assessment showed normal oxygen saturation (100% on room air) but revealed leukocytosis (white blood cell count 17,900/μL), thrombocytopenia (34,000/μL), acute kidney injury (creatinine 3.0 mg/dL), and marked hyperbilirubinemia (total bilirubin 15.9 mg/dL, direct > 10 mg/dL). His condition deteriorated rapidly by hospital day 3 with hemodynamic instability (blood pressure 71/60 mmHg), worsening hypoxemia, progressive renal dysfunction (creatinine 3.8 mg/dL), anemia (hemoglobin 10.3 g/dL from 15.7 g/dL), and escalating hyperbilirubinemia (20.9 mg/dL), necessitating transfer for higher-level care. Upon transfer, the patient reported persistent severe fatigue and dyspnea without respiratory symptoms, with notable exposure history including recent woodland/riverine activities with insect bites and a partner's similar self-limited febrile illness. Physical examination demonstrated tachypnea (42/min) with accessory muscle use, hypoxemia (91% on 4L oxygen), marked icterus, and bilateral crackles without hepatosplenomegaly or rash. Laboratory findings showed progressive leukocytosis (21,700/μL with neutrophilia), thrombocytopenia (67,000/μL), worsening renal impairment (creatinine 3.36 mg/dL, blood urea nitrogen 59 mg/dL), and profound conjugated hyperbilirubinemia (total/direct bilirubin 26.1 mg/dL) with mild transaminitis (aspartate aminotransferase 44 U/L, alanine aminotransferase 41 U/L). Imaging revealed bilateral pulmonary consolidations with tree-in-bud pattern, mediastinal lymphadenopathy, and characteristic hepatic "starry sky" appearance without biliary obstruction. The patient was initiated on continuous positive airway pressure respiratory support and broad-spectrum antimicrobial therapy including intravenous vancomycin, cefepime, and oral doxycycline	Most Likely Diagnosis: Leptospirosis (Weil's Syndrome) Diagnostic Basis: Epidemiological Exposure: Recent activities near rivers and wooded areas with insect bites, consistent with typical transmission routes of Leptospira (skin/mucosal contact with contaminated water) Clinical Triad: Fever and Myalgia: Early high fever, severe muscle pain (especially in the calves, though not specified), and headache Jaundice and Hepatic Injury: Markedly elevated direct bilirubin (> 10 mg/dL) with mild aspartate aminotransferase/alanine aminotransferase elevation (suggesting cholestasis predominant over hepatocellular damage) Renal Injury: Acute kidney injury (creatinine 3.36 mg/dL), positive urinary bilirubin Multisystem Involvement: Respiratory: Hypoxemia, tachypnea, computed tomography findings of diffuse pulmonary consolidations and ground-glass opacities (possible pulmonary hemorrhage, a severe complication of leptospirosis) Hematologic: Thrombocytopenia (34,000/μL), leukocytosis with neutrophilia Partner's Similar Symptoms: Suggests shared exposure (e.g., contaminated water or environment) Imaging: "Starry sky" sign on hepatic ultrasound (periportal edema/inflammation), consistent with leptospiral hepatitis Differential Diagnoses and Rationale: 1. Hantavirus Pulmonary Syndrome Supporting Evidence: Zoonotic transmission (exposure to rodent excreta, aligns with woodland activity) Rapidly progressive respiratory failure (hypoxemia, bilateral infiltrates), thrombocytopenia, and acute kidney injury Contradictions: Hantavirus typically lacks severe jaundice or direct hyperbilirubinemia Simultaneous illness in the partner is rare unless both were exposed to rodents 2. Rickettsial Infections (e.g., Query fever or Typhus) Supporting Evidence: Fever, headache, myalgia, and thrombocytopenia Query fever can cause hepatitis (granulomatous) and pneumonitis Contradictions: Query fever rarely presents with severe jaundice or acute renal failure Absence of typical rash (e.g., eschar in scrub typhus or maculopapular rash in epidemic typhus) 3. Viral Hepatitis (e.g., Hepatitis E Virus) Supporting Evidence:	Leptospirosis; Sepsis (caused by hypervirulent Klebsiella pneumoniae and Escherichia coli); ANCA-associated vasculitis; Allergic disease; Lymphoma	Leptospirosis	4/5/5	5/5/5

Table 1 (continued)

Final diagnosis	Case information	DeepSeek-R1's responses	Differential diagnoses from non-AI-assisted physicians	Differential diagnoses from AI-assisted physicians	Likert scores for AI's responses (Completeness/ Clarity/ Usefulness)[a]	Differential quality scores (DeepSeek-R1/ non-AI-assisted physicians/ AI-assisted physicians)[b]
		Jaundice, elevated liver enzymes, and cholestasis Hepatitis E is waterborne, potentially linked to environmental exposure Contradictions: Viral hepatitis typically features marked alanine aminotransferase/ aspartate aminotransferase elevation (often > 1000 U/L), whereas this case shows only mild transaminitis Respiratory involvement and thrombocytopenia are uncommon 4. Malaria (Plasmodium falciparum Malaria) Supporting Evidence: Fever, hemolytic jaundice (indirect bilirubin predominance), and thrombocytopenia Acute kidney injury (e.g., blackwater fever) Contradictions: No travel history to endemic regions, and no reported cyclical fever or positive blood smears Pulmonary infiltrates and respiratory failure are atypical 5. Autoimmune Disorders (e.g., systemic lupus erythematosus or Autoimmune Hepatitis) Supporting Evidence: Multisystem involvement (hepatic, renal, pulmonary) Contradictions: No prior autoimmune history or serologic markers (e.g., antinuclear antibodies, anti-smooth muscle antibodies) Acute onset and exposure history favor infectious etiology				

AI Artificial intelligence

[a] The completeness scale was a 5-point Likert scale (with 1 indicating very incomplete; 2, incomplete; 3, moderately complete; 4, complete; and 5, very complete), the clarity scale was a 5-point Likert scale (with 1 indicating very unclear; 2, unclear; 3, somewhat clear; 4, clear; and 5, very clear), and the usefulness scale was a 5-point Likert scale (with 1 indicating not beneficial; 2, slightly beneficial; 3, moderately beneficial; 4, beneficial; 5, highly beneficial). These three dimensions were used to evaluate the performance of AI responses

[b] The differential quality score (a previously published ordinal 5-point rating system based on accuracy and usefulness. A score of 5 is given for a differential including the exact diagnosis; 4, the suggestions included something very close, but not exact; 3, the suggestions included something closely related that might have been helpful; 2, the suggestions included something related, but unlikely to be helpful; 0, no suggestions were close to the target diagnosis) was use to evaluated diagnostic accuracy of each group

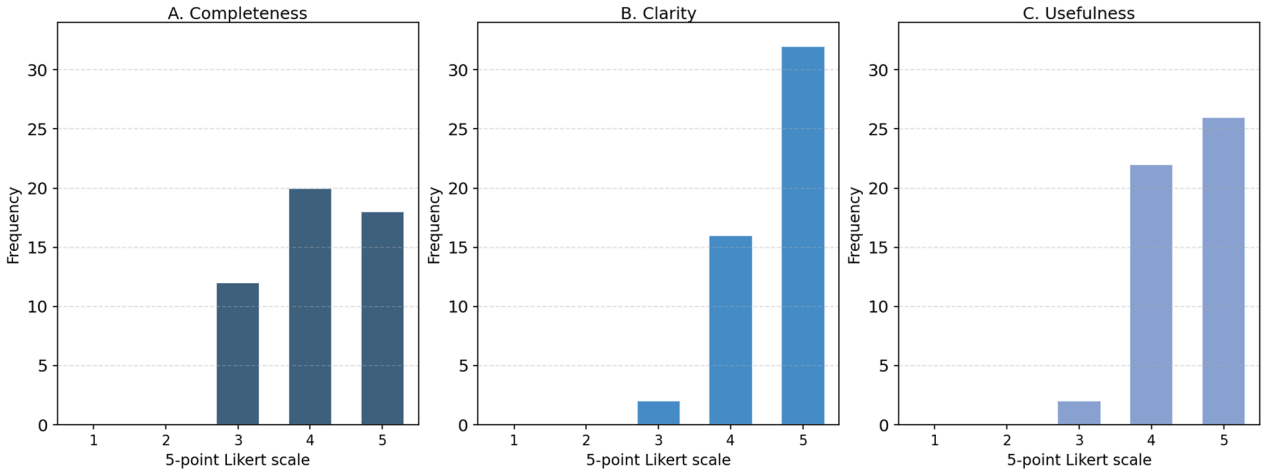

Fig. 1 Completeness, clarity, usefulness ratings for DeepSeek-R1's responses to cases. Bar charts were shown for the final consensus ratings by 3 evaluators using 5-point Likert scores. **A** The completeness rating was shown. **B** The clarity rating was shown. **C** The usefulness rating was shown

Table 2 Comparisons of top diagnostic accuracy and differential quality score between groups

	DeepSeek-R1	Non-AI-assisted physicians	AI-assisted physicians	*p* value[a]		
				Overall[b]	DeepSeek-R1 versus non-AI-assisted physicians	Non-AI-assisted physicians versus AI-assisted physicians
Top diagnosis accuracy (% (n/N, 95% CI))	60 (29/48, 0.456–0.729)	27 (13/48, 0.146–0.396)	58 (28/48, 0.438–0.729)	0.001	0.003	0.006
Differential quality score (median (IQR, 95% CI))	5.0 (4.0–5.0, 4.5–5.0)	3.0 (0–5.0, 2.0–4.0)	5.0 (3.0–5.0, 3.0–5.0)	0.004	0.005	0.036

AI Artificial intelligence, *IQR* Interquartile ranges, *CI* Confidence interval

[a] *p* values for top diagnosis accuracy were calculated using Chi-square test, with Bonferroni-corrected pairwise comparisons, *p* values for differential quality score were obtained from Kruskal–Wallis H test, followed by Dunn's post-hoc tests with Bonferroni adjustment

[b] This refers to comparisons among the three groups

CI 4.0–5.0), respectively (Fig. 1). The AI model's median response time was 48 s (IQR 39–57; 95% CI 45–50).

DeepSeek-R1 achieved 60% (29/48; 95% CI 0.456–0.729) top diagnosis accuracy, including the final diagnosis in its differential for 68% (33/48) of cases, and the median differential quality score was 5.0 (IQR 4.0–5.0; 95% CI 4.5–5.0) (Table 2). Without AI assistance, the residents' top diagnoses agreed with final diagnoses in 27% (13/48; 95% CI 0.146–0.396) of cases, and the median differential quality score was 3.0 (IQR 0–5.0; 95% CI 2.0–4.0) (Table 2). DeepSeek-R1 demonstrated a significantly higher diagnostic accuracy compared with the residents. With AI assistance, the residents' top diagnosis accuracy improved to 58% (28/48; 95% CI 0.438–0.729), while their median differential quality score improved to 5.0 (IQR 3.0–5.0; 95% CI 3.0–5.0), showing the AI model's significant enhancement of the residents' diagnostic accuracy (Table 2). No significant difference was found in diagnostic accuracy between DeepSeek-R1 and AI-assisted residents (Table 2). The two primary scorers (XW and YH) demonstrated excellent interrater reliability for differential quality scores (ICC, 0.960; 95% CI 0.946–0.971), and moderate reliability for Likert scores of completeness (ICC, 0.679; 95% CI 0.496–0.804), clarity (ICC, 0.640; 95% CI 0.442–0.778) and usefulness (ICC, 0.565; 95% CI 0.343–0.728), respectively. In the assessment of consistency of the DeepSeek-R1, 13 out of 16 responses were graded consistent, demonstrating the reliability of the AI model with the standard prompt.

Furthermore, the residents' median diagnostic time decreased significantly from 1920 s (IQR 1320–2640; 95% CI 1710–2370) to 972 s (IQR 570–1320; 95% CI 675–1200) with AI assistance ($p = 0.000001$).

Discussion

In the present study, we assessed the diagnostic performance of a reasoning model, DeepSeek-R1, on challenging diagnostic dilemmas in critical illness. Furthermore, we prospectively evaluated the effects of DeepSeek-R1's assistance on improving the diagnostic accuracy and efficiency of critical care residents in such complex cases using a randomized design. Our findings demonstrate that for differential diagnosis of complex cases in critical care, DeepSeek-R1 can generate complete, clear, and clinically useful information, achieve reasonable accuracy, and improve critical care residents' diagnostic accuracy and efficiency.

Recent cross-sectional studies have shown that LLMs such as ChatGPT-4 and Google Bard could generate quality and empathetic responses to patient questions from either public online resources or physician-developed medical queries in various specialties such as ophthalmology, oncology and anesthesia, with performance comparable to physician responses [4–6, 16]. In the context of diagnostic tasks, recent studies have demonstrated comparable diagnostic performance for LLMs in cases involving retina, glaucoma, neurodegenerative disorders and general internal medicine [6, 17–19]. In the context of treatment tasks, a recent study has demonstrated the potential of AI models, including ChatGPT-4o, Gemini 2.0 and DeepSeek V3, to align with expert surgical recommendations for surgical diseases [20]. However, the urgency and complexity in these settings are far from comparable to clinical practice in ICUs. To the best of our knowledge, no prospective studies evaluating the effects of LLM assistance on diagnostic accuracy and efficiency have been reported in the literature. In the present study, all included cases represented challenging diagnostic dilemmas in ICUs, and the recruited physicians were critical care residents who face critically ill patients upon ICU admission and manage these particularly complex challenges initially without multidisciplinary team consultation. These study settings are more likely to capture the complexity of real-world critical care scenarios. Moreover, our randomized design enabled direct comparison of residents' diagnostic performance with versus without AI assistance under identical conditions. Therefore, our findings provide evidence with higher quality for the potential benefits of reasoning AI models as promising tools to assist critical care residents in making accurate and efficient diagnoses for complex critical illness cases.

In the present study, traditional online resources such as PubMed and UpToDate were permitted in both non-AI-assisted physician and AI-assisted physician groups. When comparing AI performance with that of human physicians, prior studies often allowed physicians to obtain answers without using references. However, in contemporary practice, traditional online tools have become basic daily adjuncts [21]. Therefore, evaluating the AI model's benefits under these conditions is warranted. Our study demonstrated that when traditional tools were available to both groups, the AI model improved diagnostic accuracy and significantly reduced diagnostic time, suggesting that reasoning AI models may serve as highly beneficial tools beyond traditional resources for differential diagnosis of complex cases.

In this study, the AI model outperformed human physicians in diagnostic accuracy. In addition to generating accurate differential diagnoses, the AI model provided comprehensive, clear, and clinically useful responses. DeepSeek-R1, a recently released advanced reasoning LLM, was selected for this study. Recent studies have shown a growing interest in leveraging LLM tools in clinical settings. ChatGPT, Google Bard/Gemini and Meta Llama accounted for the majority of the evaluations, however, studies evaluating the performance of reasoning models in clinical medicine are lacking [22]. In the present study, DeepSeek-R1 demonstrated diagnostic capabilities superior to those of AI models (e.g., ChatGPT-3.5, ChatGPT-4, and Google Bard) in prior studies that also evaluated LLMs' accuracy in complex diagnostic challenges [17, 20]. DeepSeek-R1 employs a reinforcement learning-based pretraining approach, enabling advanced reasoning development, including its self-reflection capability, which enables autonomous verification and optimization of logical reasoning, thereby enhancing performance on complex tasks [13]. These features may prove particularly valuable for complex clinical queries. Similarly, a recently developed phenotype-based natural language-processing model was shown to be more accurate in the diagnosis of rare diseases than physician experts [23]. Furthermore, as an entirely open-source model, DeepSeek-R1 is particularly advantageous for resource-limited healthcare settings, including ICUs, where free and adaptable solutions are required.

A key challenge in applying LLMs to healthcare is their tendency to generate "hallucinations" [24, 25]. LLMs may confidently produce misinformation or exhibit cognitive biases, underscoring the risks of unsupervised reliance on these tools [11, 26, 27]. In our study, we employed the AI model as an adjunct for critical care residents. Although the AI model achieved higher diagnostic accuracy than clinicians in this study, its role as an assistive tool—rather than a standalone decision-maker—may help mitigate hallucination-related risks in clinical practice. However, this approach also introduces its own challenge. When human physicians apply the model's output

as one cue within their broader judgment processes and identify too many exceptions to the model's output, human-AI collaboration can sometimes result in suboptimal outcomes [28]. A recent study highlighted the heterogeneity of the effects of human-AI collaboration, the performance of human-AI collaboration loss when AI outperformed humans alone, and in tasks that involved making decisions [29]. These scenarios resembled those in the present study, wherein the AI model showed a higher top diagnosis accuracy compared to AI-assisted physicians (no significant difference) when outperformed standalone human physicians. Thus, future researches are recommended to explore for promising ways for improving human-AI collaboration.

Our study has several limitations. First, diagnostic accuracy heavily depends on the quality and specificity of clinical information [30], and all clinical data were provided by researchers. Potential subjectivity in outcome measures and the exclusion of certain diagnostic data (e.g., medical images) may lead to inconsistency to clinical practice. Additionally, in the present study, the completeness and accuracy of information entered into the AI chatbot serves as the basis of what the AI chatbot would use to assist with diagnosis. However, in clinical practice, case information gathered from history taking, physical examination, or other activities might be incomplete or incorrect. Second, AI responses were generated using standardized prompts and delivered by researchers, diverging from real-world scenarios where physicians interact dynamically with AI via personal devices. Third, all participants were critical care residents; no participants with differing training levels (e.g., attending physicians) were included. Diagnostic performance with AI assistance may vary across training levels. Fourth, we did not evaluate other reasoning models (e.g., ChatGPT-4, ChatGPT-o1 and DeepSeek V3), which may differ in accuracy and clinical utility compared to DeepSeek-R1. Noteworthy, the findings of recent benchmark studies evaluating the diagnostic performance of LLMs demonstrated that reasoning models such as DeepSeek-R1 and ChatGPT-o1 exhibited clearly superior performance compared to GPT-4 and GPT-3.5, and DeepSeek-R1 performed non-inferiorly compared to proprietary reasoning models such as ChatGPT-o1 and superiorly compared to another proprietary reasoning model, Gem2FTE [31, 32]. These findings highlight the potential of DeepSeek-R1 in clinical decision-making in the diagnostic context. Finally, assuming a power of 0.8 and an alpha of 0.05, a sample size of 35 was required for the comparisons of diagnostic performance between AI model and non-AI-assisted physicians and between AI-assisted physicians and non-AI-assisted physicians. However, in order to draw meaningful comparisons between AI model and AI-assisted physicians, a sample size of 227 was required for appropriateness. The sample size of this study is rather small for comparing diagnostic performance between AI model and AI-assisted physicians.

Conclusions

Above all, our findings suggest that reasoning models, such as DeepSeek-R1, are promising assistive tools for critical care residents facing challenging differential diagnoses. These findings warrant further research with larger sample sizes to evaluate the clinical adoption potential of reasoning models in real-world critical care decision-making.

Abbreviations

LLM	Large language model
AI	Artificial intelligence
USMLE	The United States Medical Licensing Exam
ICU	Intensive care unit
RAG	Retrieval-augmented generation
IQR	Interquartile ranges
CI	Confidence interval
ICC	Intraclass correlation coefficient

Acknowledgements

We thank Dr. Ying Qiao and Dr. Qianyun Cai for their efforts on recruitment of participants and data collection. We thank all the physicians who participated in the case studies.

Author contributions

YH contributed to the study concept and design, XW performed the initial literature research and recruited the participants. XW, YH, and QH analyzed and interpreted the data. XW contributed to the statistical analysis and graphing. YH drafted the initial manuscript, XW contributed to the manuscript sections related to the results obtained. QH and YH was responsible for the revision of the manuscript for important intellectual content. All authors read and approved the final manuscript.

Funding

No financial support.

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable as no patients were involved.

Consent for publication

Not applicable as no patients were involved.

Competing interests

The authors declare that they have no competing interests.

Received: 25 April 2025 Accepted: 24 May 2025
Published online: 06 June 2025

References

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930–40.

2. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023;388(13):1233–9.
3. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLoS Digit Health. 2023;2(2):e0000198.
4. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183(6):589–96.
5. Chen D, Parsa R, Hope A, et al. Physician and artificial intelligence chatbot responses to cancer questions from social media. JAMA Oncol. 2024;10(7):956–60.
6. Huang AS, Hirabayashi K, Barna L, Parikh D, Pasquale LR. Assessment of a Large language model's responses to questions and cases about glaucoma and retina management. JAMA Ophthalmol. 2024;142(4):371–5.
7. Goodman RS, Patrinely JR, Stone CA Jr, et al. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open. 2023;6(10):e2336483.
8. Anastasio MK, Peters P, Foote J, Melamed A, Modesitt SC, Musa F, Rossi E, Albright BB, Havrilesky LJ, Moss HA. The doc versus the bot: a pilot study to assess the quality and accuracy of physician and chatbot responses to clinical questions in gynecologic oncology. Gynecol Oncol Rep. 2024;55:101477.
9. James FR, Power N, Laha S. Decision-making in intensive care medicine—a review. J Intensive Care Soc. 2018;19(3):247–58.
10. Workum JD, Volkers BWS, van de Sande D, et al. Comparative evaluation and performance of large language models on expert level critical care questions: a benchmark study. Crit Care. 2025;29(1):72.
11. Balta KY, Javidan AP, Walser E, Arntfield R, Prager R. Evaluating the appropriateness, consistency, and readability of ChatGPT in critical care recommendations. J Intensive Care Med. 2025;40(2):184–90.
12. Xu F, Hao Q, Zong Z, et al. Towards large reasoning models: a survey of reinforced reasoning with large language models. arXiv. Preprint posted online January 25, 2025. https://doi.org/10.48550/arXiv.2501.09686
13. Mondillo G, Colosimo S, Perrotta A, Frattolillo V, Masino M. Comparative evaluation of advanced AI reasoning models in pediatric clinical decision support: ChatGPT O1 vs. DeepSeek-R1. medRxiv. Preprint posted online January 27, 2025. https://doi.org/10.1101/2025.01.27.25321169
14. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. 2025;31:60–9.
15. Bond WF, Schwartz LM, Weaver KR, Levick D, Giuliano M, Graber ML. Differential diagnosis generators: an evaluation of currently available computer programs. J Gen Intern Med. 2012;27(2):213–9.
16. Nguyen TP, Carvalho B, Sukhdeo H, et al. Comparison of artificial intelligence large language model chatbots in answering frequently asked questions in anaesthesia. BJA Open. 2024;10:100280.
17. Koga S, Martin NB, Dickson DW. Evaluating the performance of large language models: ChatGPT and Google Bard in generating differential diagnoses in clinicopathological conferences of neurodegenerative disorders. Brain Pathol. 2024;34(3):e13207.
18. Hirosawa T, Kawamura R, Harada Y, et al. ChatGPT-generated differential diagnosis lists for complex case-derived clinical vignettes: diagnostic accuracy evaluation. JMIR Med Inform. 2023;11:e48808.
19. Hadi A, Tran E, Nagarajan B, Kirpalani A. Evaluation of ChatGPT as a diagnostic tool for medical learners and clinicians. PLoS ONE. 2024;19(7):e0307383.
20. Seth I, Marcaccini G, Lim K, et al. Management of Dupuytren's disease: a multi-centric comparative analysis between experienced hand surgeons versus artificial intelligence. Diagnostics (Basel). 2025;15(5):587.
21. Kulkarni PA, Singh H. Artificial intelligence in clinical diagnosis: opportunities, challenges, and hype. JAMA. 2023;330(4):317–8.
22. 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.
23. Mao X, Huang Y, Jin Y, et al. A phenotype-based AI pipeline outperforms human experts in differentially diagnosing rare diseases using EHRs. NPJ Digit Med. 2025;8(1):68.
24. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare. 2023;11(6):887.
25. Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. Crit Care. 2023;27(1):120.
26. Griot M, Hemptinne C, Vanderdonckt J, Yuksel D. Large language models lack essential metacognition for reliable medical reasoning. Nat Commun. 2025;16(1):642.
27. Wang J, Redelmeier DA. Cognitive biases and artificial intelligence. NEJM AI. 2024;1(12):AIcs2400639.
28. Peringa IP, Cox EGM, Wiersema R, van der Horst ICC, Meijer RR, Koeze J. Human judgment error in the intensive care unit: a perspective on bias and noise. Crit Care. 2025;29(1):86.
29. Vaccaro M, Almaatouq A, Malone T. When combinations of humans and AI are useful: a systematic review and meta-analysis. Nat Hum Behav. 2024;8(12):2293–303.
30. Singh H, Giardina TD, Meyer AN, Forjuoh SN, Reis MD, Thomas EJ. Types and origins of diagnostic errors in primary care settings. JAMA Intern Med. 2013;173(6):418–25.
31. Sandmann S, Hegselmann S, Fujarski M, et al. Benchmark evaluation of DeepSeek large language models in clinical decision-making. Nat Med. 2025. https://doi.org/10.1038/s41591-025-03727-2.
32. 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. https://doi.org/10.1038/s41591-025-03726-3.

Publisher's Note