Journal of Multidisciplinary Healthcare

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Assessing DeepSeek-R1 for Clinical Decision Support in Multidisciplinary Laboratory Medicine

Qinpeng Li*, Lili Zhan*, Xinjian Cai

Department of Clinical Laboratory Medicine, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Shenzhen, Guangdong, People's Republic of China

*These authors contributed equally to this work

Correspondence: Lili Zhan, Department of Clinical Laboratory Medicine, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Baohe Road No. 113, Longgang District, Shenzhen, Guangdong, 518116, People's Republic of China, Email zhanlily@163.com

Background: Recent advancements in artificial intelligence (AI), particularly with large language models (LLMs), are transforming healthcare by enhancing diagnostic decision-making and clinical workflows. The application of LLMs like DeepSeek-R1 in clinical laboratory medicine demonstrates potential for improving diagnostic accuracy, supporting decision-making, and optimizing patient care.

Objective: This study evaluates the performance of DeepSeek-R1 in analyzing clinical laboratory cases and assisting with medical decision-making. The focus is on assessing its accuracy and completeness in generating diagnostic hypotheses, differential diagnoses, and diagnostic workups across diverse clinical cases.

Methods: We analyzed 100 clinical cases from *Clinical Laboratory Medicine Case Studies*, which includes comprehensive case histories and laboratory findings. DeepSeek-R1 was queried independently for each case three times, with three specific questions regarding diagnosis, differential diagnoses, and diagnostic tests. The outputs were assessed for accuracy and completeness by senior clinical laboratory physicians.

Results: DeepSeek-R1 achieved an overall accuracy of 72.9% (95% CI [69.9%, 75.7%]) and completeness of 73.4% (95% CI [70.5%, 76.2%]). Performance varied by question type: the highest accuracy was observed for diagnostic hypotheses (85.7%, 95% CI [81.2%, 89.2%]) and the lowest for differential diagnoses (55.0%, 95% CI [49.3%, 60.5%]). Notable variations in performance were also seen across disease categories, with the best performance observed in genetic and obstetric diagnostics (accuracy 93.1%, 95% CI [84.0%, 97.3%]; completeness 86.1%, 95% CI [76.4%, 92.3%]).

Conclusion: DeepSeek-R1 demonstrates potential for a decision-support tool in clinical laboratory medicine, particularly in generating diagnostic hypotheses and recommending diagnostic workups. However, its performance in differential diagnosis and handling specific clinical nuances remains limited. Future work should focus on expanding training data, integrating clinical ontologies, and incorporating physician feedback to improve real-world applicability. DeepSeek-R1 and the new versions under development may be promising tools for non-medical professionals and professionals in medical laboratory diagnoses.

Keywords: DeepSeek-R1, artificial intelligence, large language model, clinical decision support systems, clinical laboratory medicine

Introduction

The application of artificial intelligence (AI) in healthcare is transforming diagnostic methodologies,[1,2] enabling streamlined laboratory processes and more efficient clinical workflows.[2,3] Among AI technologies, large language models (LLMs), such as Generative Pre-trained Transformers (GPT), have attracted significant attention for their ability to perform complex natural language processing tasks with contextual understanding.[4] These models are being rapidly adapted across multiple domains due to their versatility and linguistic fluency.

In medicine, LLMs are being increasingly explored for their potential to support clinical decision-making, assist in diagnosis, and facilitate information delivery.[5] ChatGPT, a widely used LLM, has been applied in clinical tasks such as drug development, image recognition, data analysis, medical report enhancement, and literature review.[6–12] Several studies have evaluated its capacity to answer multiple-choice medical questions[13] and provide explanatory reasoning with

Received: 22 May 2025
Accepted: 30 July 2025
Published: 12 August 2025

accuracy comparable to healthcare professionals.[14] Furthermore, LLMs can help reduce unnecessary clinical visits by offering accessible and understandable medical information,[15–17] thus potentially improving patient engagement and outcomes.[18,19] However, their performance in specialized fields such as laboratory medicine remains underexplored, highlighting the need for focused evaluation.

Despite these advancements, clinical decision-making is often influenced by cognitive bias and complex patient conditions. LLMs can draw on a wide range of medical knowledge and use algorithms to simulate clinicians' clinical diagnostic and therapeutic thinking, providing users with medical information and diagnostic and therapeutic recommendations.[13,20–22] However, these high-performance LLMs typically come with high computational demands and expensive subscription fees. DeepSeek-R1, a new and affordable LLM, has gained global attention for its ability to perform complex tasks, including logical reasoning, mathematical computations, and code generation while maintaining contextual coherence in multi-turn conversations.[23–25] Moreover, Deepseek's open-source and free nature significantly reduces deployment costs, making it an attractive option for hospitals and healthcare facilities. Its flexibility enables easy integration into hospital systems, facilitating the broader adoption of advanced diagnostic and therapeutic support, especially in resource-limited settings. Despite these strengths, its capability to support healthcare decision-making, akin to ChatGPT, remains under investigation.

This study aims to evaluate DeepSeek-R1's performance in clinical laboratory medicine, specifically assessing its ability to analyze case reports, generate diagnostic hypotheses, suggest differential diagnoses, and recommend appropriate diagnostic workups.

Methods

Study Design

We conducted a study to evaluate the diagnostic, differential diagnostic, and next-step examination analysis capabilities of DeepSeek-R1 based on medical case history. The case analysis reports were from the textbook *Clinical Laboratory Medicine Case Studies*, edited by esteemed experts Tiesheng Zheng and Yan Li. This textbook is primarily used by senior students in medical laboratory technology programs and also serves as a clinical reference for practicing laboratory physicians.

A total of 100 case reports were analyzed. Each case includes a concise summary of medical history, presenting symptoms, physical signs, and auxiliary examination results. Although structured for educational purposes, these cases are derived from anonymized real-world clinical cases, which were professionally reviewed and curated by experienced clinicians to reflect representative diagnostic patterns and reasoning pathways.

In this study, each case was independently submitted to DeepSeek-R1 three separate times using standardized prompts, and the model's responses were carefully recorded. By referencing the case analysis results, we assessed the accuracy and completeness of the responses provided by DeepSeek-R1.

Study Data

Disease-related data from 100 case reports were entered into DeepSeek-R1 three times independently; three questions were asked: 1. What is the most likely diagnosis for this patient based on the history and physical examination? What is the diagnostic basis? (Diagnosis Hypothesis and Basis, Dx Hypothesis & Basis); 2. What differential diagnoses should be considered? (Differential Diagnoses, DDx); 3. What diagnostic tests should be performed to confirm the diagnosis? (Diagnostic Workup, Dx Workup). The outputs from DeepSeek-R1 were recorded for further analysis to assess the consistency and accuracy of its responses across multiple queries for the same case.

Assessment of Outputs

Three senior clinical laboratory physicians independently evaluated the accuracy and completeness of DeepSeek-R1's responses. All evaluators are from the Department of Clinical Laboratory Medicine, Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, and each has 5–10 years of clinical diagnostic experience in laboratory medicine, with expertise in hematology, biochemical diagnostics, and molecular/genetic testing.

The Likert scale for evaluating accuracy using a six-point scale can be described as follows: (1) Completely incorrect response; (2) More errors than correct elements; (3) Balance between correct and incorrect elements; (4) More correct elements than errors; (5) Nearly completely correct response; (6) Completely correct response.

The three-point Likert scale for evaluating completeness can be expressed as: (1) Incomplete answer, addressing only some aspects of the question and missing important parts; (2) Sufficient answer covering all necessary aspects of the question; (3) Comprehensive response covering all aspects of the question with additional information or context beyond expectations.

Statistical Methods

Statistical analyses were performed using SPSS (version 25.0; IBM Corp). The accuracy and completeness scores were reported as mean ± standard error (SEM). The percentage of correct answers for each type of question was calculated based on the number of cases where the score was 5 or 6 (accuracy) and 2 or 3 (completeness). The Kruskal–Wallis test was used for omnibus group comparisons of non-normally distributed data. Wilson score confidence intervals were calculated for each group to estimate the accuracy rates with uncertainty quantification. Pearson's chi-square test was employed to assess statistical significance between different groups. A significance level of $P < 0.05$ was considered statistically significant.

Results

Overview of DeepSeek-R1's Performance

The accuracy and completeness scores for the 100 cases analyzed are presented in Figure 1, with a detailed list of the scores in Supplementary Table 1. These cases were categorized into seven distinct groups: Organ-specific Pathology & Diagnostics (ORG), Hematologic Evaluation & Management (HEM), Metabolic & Endocrine Testing (MET), Infectious Diseases & Antimicrobial Testing (INF-ANT), Genetic & Obstetric Diagnostics (GEN-OB), Immune-mediated Disorders (IMM) and Oncology Laboratory Diagnostics (ONC). The distribution of cases across these categories was as follows: ORG (18%), HEM (24%), MET (20%), INF-ANT (16%), GEN-OB (8%), IMM (7%) and ONC (7%).

Performance by Question Type

DeepSeek-R1 was evaluated based on three specific queries for each case: (1) the most likely diagnosis based on the clinical history and examination (Dx Hypothesis & Basis), (2) differential diagnoses to consider (DDx), and (3) recommended diagnostic tests (Dx Workup).

The results revealed significant variation in DeepSeek-R1's performance depending on the question type. As shown in Figure 2A, DeepSeek-R1 demonstrated significantly higher accuracy for Dx Hypothesis & Basis compared to DDx ($P < 0.0001$) and Dx Workup ($P < 0.01$). Additionally, Dx Workup significantly outperformed DDx in accuracy ($P < 0.01$), highlighting a distinct hierarchy in the model's performance across diagnostic reasoning tasks.

Regarding completeness, as seen in Figure 2B, a similar trend was observed: Dx Hypothesis & Basis achieved the highest completeness, significantly outperforming DDx ($P < 0.0001$) and Dx Workup ($P < 0.001$). However, there was no statistically significant difference in completeness between DDx and Dx Workup ($P \geq .05$).

Moreover, we calculated the overall accuracy and completeness of DeepSeek-R1's outputs in all cases. We found that, on average, across all cases, DeepSeek-R1 achieved an average accuracy of 72.9% (95% CI [69.9%, 75.7%]) and an average completeness of 73.4% (95% CI [70.5%, 76.2%]). In addition, we calculated the percentage of correct answers for each of the three question types, covering both accuracy and completeness. As shown in Table 1 and Table 2, there was a statistically significant variation in DeepSeek-R1's performance across the question categories, with $P < 0.001$ for both accuracy and completeness. Among the question categories, Dx Hypothesis & Basis yielded the highest performance, achieving an accuracy of 85.7% and completeness of 87.7%. In contrast, DDx (Differential Diagnoses) exhibited the lowest performance, with only 55.0% accuracy and 55.3% completeness.

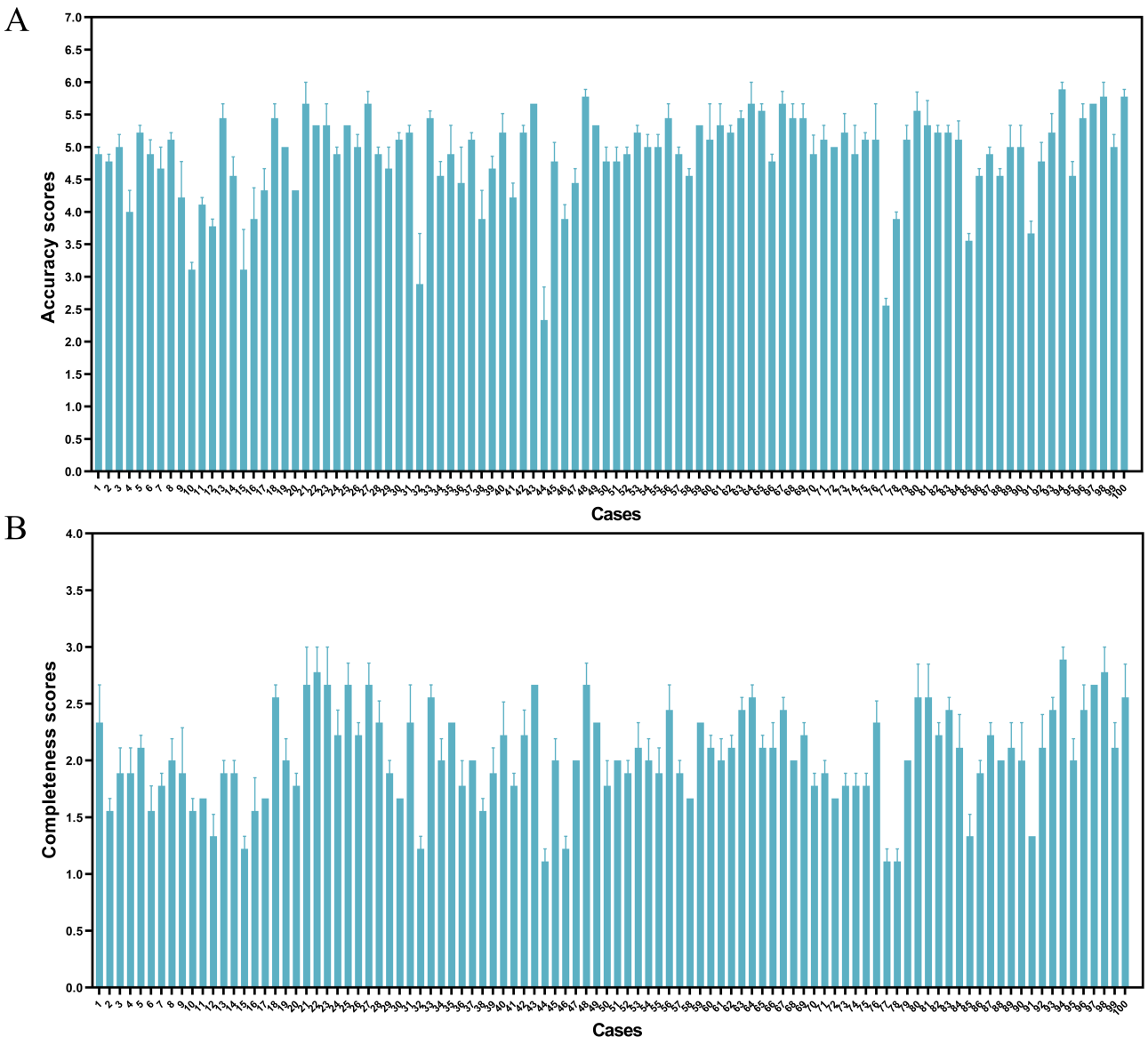

Figure 1 DeepSeek-R1's performance for 100 cases. (**A**) Accuracy scores for 100 cases. (**B**) Completeness scores for 100 cases. Error bars are 1 SEM of the mean.

Performance by Disease Category

The case analyses (n = 100) were categorized into seven distinct disease categories, allowing for an evaluation of DeepSeek-R1's performance not only at the case level but also at the disease category level. Figures 3A and B present the trends in the accuracy and completeness of DeepSeek-R1's responses across various disease types.

Across all disease categories, as shown in Figure 3A, DeepSeek-R1 achieved the highest accuracy scores in the GEN-OB category and the lowest in HEM, while there were no statistically significant differences between the groups ($P > 0.05$). In contrast, Figure 3B reveals that completeness scores varied significantly across disease types. GEN-OB demonstrated significantly higher completeness than several other disease categories, particularly ORG and MET ($P<0.001$).

Furthermore, as summarized in Tables 3 and 4, statistical analysis confirmed significant differences in the accuracy and completeness rates across the disease categories ($\chi^2=36.524$, $P<0.001$ for accuracy and $\chi^2=19.693$, $P=0.003$ for completeness). Across all categories, DeepSeek-R1 achieved correctness rates exceeding 60%, with GEN-OB attaining the highest accuracy (93.1%) and completeness (86.1%) rates, as detailed in Tables 3 and 4.

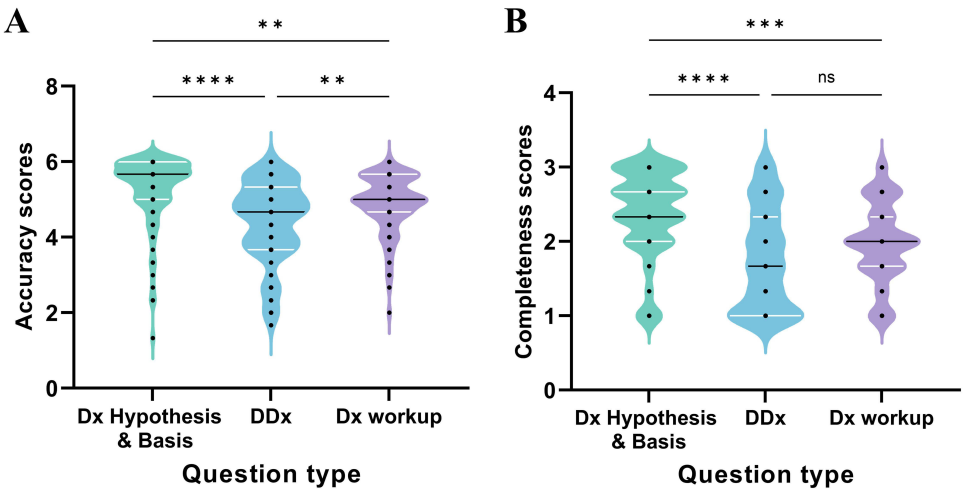

Figure 2 Comparison of Mean Scores in question type. (**A**) Comparisons of the accuracy performance of DeepSeek-R1 in question type. (**B**) Comparisons of the completeness performance of DeepSeek-R1 in question type.
Notes: [b]*P*-value calculated with Kruskal–Wallis test, ns = not significant ($P \geq .05$); **P <0.01; ***P <0.001, ****P <0.0001.
Abbreviations: [a]Dx Hypothesis & Basis, Diagnosis Hypothesis and Basis; DDx, Differential Diagnoses; Dx Workup: Diagnostic Workup.

Discussion

Principal Results

Similar predictive models in decision support have been explored in other clinical domains, such as advanced systems for heart disease detection[26] and AI-based diagnostic frameworks for diabetic retinopathy,[27] demonstrating the growing potential of LLMs and multimodal data integration in medical diagnostics.

This study demonstrates that DeepSeek-R1, an advanced large language model (LLM), holds considerable potential as a decision-support tool in clinical laboratory medicine. With an overall accuracy of 72.9% (95% CI [69.9%, 75.7%]) and completeness of 73.4% (95% CI [70.5%, 76.2%]), DeepSeek-R1 performs reasonably well in generating diagnostic hypotheses and recommending appropriate diagnostic workups. Notably, DeepSeek-R1 showed the highest accuracy and completeness in generating diagnostic hypotheses, highlighting its potential utility in providing initial diagnoses based on patient history and laboratory findings. This aligns with previous findings where LLMs have been recognized for their strengths in diagnostic reasoning and knowledge retrieval.[28]

Regarding the diagnostic output, DeepSeek-R1 can formulate personalized diagnoses for individual cases, utilizing a highly condensed clinical history. For instance, in a case involving a patient with a history of total gastrectomy for

Table 1 Comparison of Accuracy in Question Type

Question Type	Correctness Rate	95% Wilson CI	χ^2	*P*
Dx Hypothesis and Basis	85.7% (257/300)	[81.2%, 89.2%]	77.335	<0.001
DDx	55.0% (165/300)	[49.3%, 60.5%]		
Dx Workup	78.0% (234/300)	[73.0%, 82.3%]		

Table 2 Comparison of Completeness in Question Type

Question Type	Correctness Rate	95% Wilson CI	χ^2	*P*
Dx Hypothesis and Basis	87.7% (263/300)	[83.5%, 90.9%]	83.893	<0.001
DDx	55.3% (166/300)	[49.7%, 60.9%]		
Dx Workup	77.3% (232/300)	[72.3%, 81.7%]		

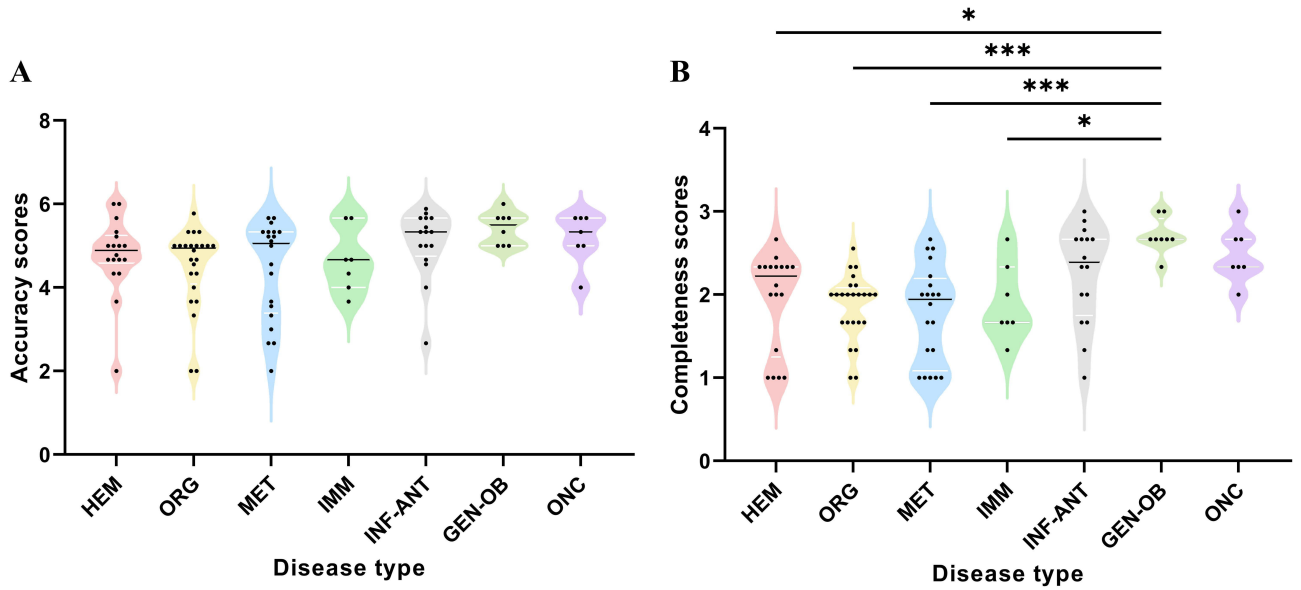

Figure 3 Comparison of Mean Scores in disease type. (**A**) Comparisons of the average accuracy performance of DeepSeek-R1 in disease type.(**B**) Comparisons of the average completeness performance of DeepSeek-R1 in disease type.
Note: [b] *P*-value calculated with Kruskal–Wallis test, **P* <0.05, ****P* <0.001.
Abbreviations: [a]HEM, Hematologic Evaluation & Management; ORG, Organ-specific Pathology & Diagnostics; MET, Metabolic & Endocrine Testing; IMM, Immune-mediated Disorders; INF-ANT, Infectious Diseases & Antimicrobial Testing; GEN-OB, Genetic & Obstetric Diagnostics; ONC, Oncology Laboratory Diagnostics.

gastric cancer, along with laboratory results indicating MCV 115 fL, MCH 37.4 pg, pancytopenia (low WBC/PLT), and an elevated reticulocyte percentage (10.42%), which suggests ineffective erythropoiesis, DeepSeek-R1 accurately identified megaloblastic anemia. However, in terms of differential diagnosis, the model exhibited minor biases. For example, in the case of suspected myeloma, clinicians would expect a differential diagnosis that includes plasma cell tumors and plasma cell leukemia. In contrast, DeepSeek-R1 may suggest conditions like myelodysplastic syndromes and

Table 3 Comparison of Accuracy in Disease Type

Question Type	Correctness Rate	95% Wilson CI	χ^2	*P*
HEM	61.1% (99/162)	[53.4%, 68.3%]	36.524	<0.001
ORG	75.0% (162/216)	[68.8%, 80.3%]		
MET	66.1% (119/180)	[59.2%, 73.0%]		
IMM	74.6% (47/63)	[62.7%, 83.7%]		
INF-ANT	75.0% (108/144)	[67.9%, 82.1%]		
GEN-OB	93.1% (67/72)	[84.0%, 97.3%]		
ONC	85.7% (54/63)	[75.0%, 92.3%]		

Table 4 Comparison of Completeness in Disease Type

Question Type	Correctness Rate	95% Wilson CI	χ^2	*P*
HEM	65.4% (106/162)	[57.8%, 72.3%]	19.693	0.003
ORG	78.2% (169/216)	[72.3%, 83.2%]		
MET	66.1% (119/180)	[59.2%, 73.0%]		
IMM	76.2% (48/63)	[63.7%, 85.5%]		
INF-ANT	76.4% (110/144)	[69.5%, 83.3%]		
GEN-OB	86.1% (62/72)	[76.4%, 92.3%]		
ONC	74.6% (47/63)	[62.7%, 83.7%]		

metastatic osteomas. This highlights the need for clinical examiners to exercise caution and verify the veracity of information provided by DeepSeek-R1, as it has occasionally yielded unreliable or inaccurate outputs.[29]

At the diagnostic workup level, DeepSeek-R1 performed well in recommending appropriate tests for most conditions. In most cases, the model prescribed more specific tests based on the diagnosis, which would help further refine the diagnosis or determine the type of condition. However, there were instances where DeepSeek-R1 recommended a broader array of tests, typically due to insufficient detail in the case analysis provided, which suggests that when using DeepSeek-R1 for diagnostic workup suggestions, it is crucial to provide a thorough and accurate medical history to allow for more personalized test recommendations.

Summary

In summary, DeepSeek-R1 demonstrates the capability to generate diagnostic hypotheses based on patient symptoms, history, and laboratory findings. However, its ability to personalize diagnoses, while promising, still requires refinement. The cases analyzed in this study primarily represent classic and commonly encountered diseases, which are widely used in clinical laboratory teaching to support pattern recognition training. These cases—though structured—are based on anonymized real-world clinical scenarios, as documented in authoritative textbooks. They are typically designed to provide students with a clear understanding of specific disease patterns, facilitating the development of a pattern recognition framework for differential diagnosis. From this perspective, DeepSeek-R1's performance in pattern recognition and diagnostic workup is commendable, as it has access to an extensive knowledge base, thereby minimizing diagnostic omissions.

Limitations and Challenges

DeepSeek-R1 has the potential for medical advances but has potential limitations. The accuracy of the text generated depends on the training of the model, which may lead to misinformation or misleading interpretations, although this can be corrected by training with multiple dialogue training, which also requires some discernment on the part of the user.[30,31] Furthermore, DeepSeek's training data may not fully encompass the latest medical advancements, as the current cut-off date for DeepSeek's progress is October 2023. Like other LLMs, such as ChatGPT, DeepSeek is susceptible to hallucinations, omissions, and incorrect responses, which may lead to contradictory or misleading conclusions. These challenges underscore the importance of verifying the generated content before clinical application. Ethical considerations, such as data privacy and security, are paramount in the broader context of LLM usage for scientific research and clinical decision-making.[32,33] Anonymizing patient data to train LLMs is essential to comply with privacy regulations and protect patient confidentiality.[34] Moreover, researchers and clinicians must ensure that biomedical data is securely collected, stored, and utilized.

From a methodological perspective, this study has several limitations. First, the dataset comprised 100 cases derived from structured, textbook-based clinical case materials used in medical laboratory education in China. Although these cases are based on anonymized real-world scenarios, their curated nature may not fully capture the complexity and heterogeneity of actual clinical practice. Second, all three physician evaluators were from the same institution, which may introduce institutional bias in the assessment of DeepSeek-R1's outputs. This limitation has been acknowledged, and future studies should include a larger and more diverse panel of evaluators from multiple institutions and specialties to improve objectivity and enable inter-rater reliability analysis. Lastly, this study did not include a direct comparison with other state-of-the-art large language models, such as GPT-4o. Future work should incorporate comparative benchmarking and real-world clinical data to better contextualize DeepSeek-R1's performance within the broader landscape of AI-driven diagnostic tool.

As LLMs continue to evolve, they may soon outperform human reasoning in certain areas, as evidenced by recent studies suggesting that LLMs can achieve superhuman performance on human tests of reasoning, as discussed in *The Lancet*.[35] However, it is critical to remember that clinical decision-making requires comprehensive judgment, which can only be exercised by trained healthcare professionals. Thus, LLMs should be viewed as tools to support, rather than replace, clinical practice.

The accuracy and reliability of the content generated by LLMs are vital, as inaccuracies could have serious consequences, particularly for patients or trainees without sufficient medical expertise.[36] Therefore, researchers must rigorously validate the safety and efficacy of LLMs before their application in clinical diagnosis. This study demonstrates that DeepSeek-R1, in its current iteration, is a promising tool for clinical decision-making, although it requires further refinement to fully realize its potential.

Conclusions

This study demonstrates that DeepSeek-R1 has the potential to serve as an effective tool for clinical decision support in laboratory medicine. While its performance in generating initial diagnoses and suggesting diagnostic workups was promising, there are limitations in its ability to personalize differential diagnoses and handle specific clinical nuances. Beyond expanding training data with more diverse and recent case distributions, future development could benefit from incorporating structured medical ontologies—such as SNOMED CT and UMLS—to enhance semantic comprehension and domain-specific accuracy. Additionally, introducing physician-in-the-loop training mechanisms—such as feedback-guided fine-tuning—may help the model better align with expert decision-making processes.

While DeepSeek-R1 should not be viewed as a replacement for human clinical judgment, it may serve as a valuable adjunct, especially in resource-limited settings. Further comparative studies against established models like GPT-4o and real-world validation trials will be essential to define its role in clinical workflows and ensure safe, effective integration into healthcare systems.

Abbreviations

LLM, large language model; AI, Artificial intelligence; GPT, Generative Pretrained Transformer; Dx Hypothesis & Basis, Diagnosis Hypothesis and Basis; DDx, Differential Diagnoses; Dx Workup, Diagnostic Workup; SEM, standard error.

Data Sharing Statement

The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Ethics Declarations

This study did not involve human participants, and as such, it was not subject to the requirements for institutional review board (IRB) approval or informed consent. The research was conducted in accordance with ethical guidelines for non-human subject research.

The AI model used in this study, DeepSeek-R1, is an open-source large language model released under the MIT License by DeepSeek. This license permits free academic and commercial use, modification, and distribution, provided that the original license and copyright notices are retained. All model usage in this study adhered strictly to the terms of the MIT License. No additional permissions were required, and no personal, social media, or sensitive data were used.

Acknowledgments

The authors would like to express their sincere gratitude to LYT, ZL, and ZXY for their valuable assistance in scoring the case analysis. Their expertise and contributions were instrumental in the success of this study.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work. Qinpeng Li and Lili Zhan contributed equally to this work and share first authorship.

Funding

This work was supported by Sanming Project of Medicine in Shenzhen (No.SZSM202311002).

Disclosure

The authors declare that they have no competing interests.

References

1. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. *Future Healthcare J*. 2021;8(2):e188–e194. doi:10.7861/fhj.2021-0095
2. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare*. 2023;11(6):887. doi:10.3390/healthcare11060887
3. Kleesiek J, Wu Y, Stiglic G, Egger J, Bian J. An opinion on chatgpt in health care—written by humans only. *J Nucl Med*. 2023;64(5):701–703. doi:10.2967/jnumed.123.265687
4. Qin L, Chen Q, Zhou Y, et al. A survey of multilingual large language models. *Patterns*. 2025;6(1). doi:10.1016/j.patter.2024.101118.
5. Betzler BK, C-Y C, Cheng C-Y, et al. Large language models and their impact in ophthalmology. *Lancet Digital Health*. 2023;5(12):e917–e924. doi:10.1016/S2589-7500(23)00201-7
6. OpenAI AJ, Adler S, Agarwal S, Ahmad L. GPT-4. *Technical Report*.
7. Ruksakulpiwat S, Kumar A, Ajibade A. Using ChatGPT in medical research: current status and future directions. *J Multidisciplinary Healthcare*. 2023;16:1513–1520. doi:10.2147/JMDH.S413470
8. Mann DL. Artificial Intelligence Discusses the Role of Artificial Intelligence in Translational Medicine. *JACC*. 2023;8(2):221–223.
9. Blanco-González A, Cabezón A, Seco-González A, et al. The role of AI in drug discovery: challenges, opportunities, and strategies. *Pharmaceuticals*. 2023;16(6):891. doi:10.3390/ph16060891
10. Salvagno M, Taccone FS, Gerli AG. Can artificial intelligence help for scientific writing? *Critical Care*. 2023;27(1).
11. Ueda D, Mitsuyama Y, Takita H, et al. Diagnostic Performance of ChatGPT from patient history and imaging findings on the diagnosis please quizzes. *Radiology*. 2023;308(1). doi:10.1148/radiol.231040.
12. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Frontiers in Artificial Intelligence*. 2023;6.
13. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172–180.
14. Tian S, Jin Q, Yeganova L, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Briefings Bioinf*. 2024;25(1). doi:10.1093/bib/bbae211.
15. Klasnja P, Pratt W. Healthcare in the pocket: mapping the space of mobile-phone health interventions. *J Biomed Informat*. 2012;45(1):184–198. doi:10.1016/j.jbi.2011.08.017
16. Mirzaei A, Aslani P, Luca EJ, Schneider CR. Predictors of health information–seeking behavior: systematic literature review and network analysis. *J Med Internet Res*. 2021;23(7):e21680. doi:10.2196/21680
17. Han J-W, Park J, Lee H. Analysis of the effect of an artificial intelligence chatbot educational program on non-face-to-face classes: a quasi-experimental study. *BMC Med Education*. 2022;22(1). doi:10.1186/s12909-022-03898-3
18. Gulati R, Nawaz M, Pyrsopoulos NT. Health literacy and liver disease. *Clin Liver Dis*. 2018;11(2):48–51. doi:10.1002/cld.690
19. Carusi A, Winter PD, Armstrong I, et al. Medical artificial intelligence is as much social as it is technological. *Nature Mach Intell*. 2023;5(2):98–100. doi:10.1038/s42256-022-00603-3
20. Rao A, Pang M, Kim J, et al. Assessing the Utility of ChatGPT throughout the entire clinical workflow: development and usability study. *J Med Internet Res*. 2023;25.
21. 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. doi:10.1001/jamainternmed.2023.1838
22. Levine DM, Tuwani R, Kompa B, et al. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model: an observational study. *Lancet Digital Health*. 2024;6(8):e555–e561. doi:10.1016/S2589-7500(24)00097-9
23. Dreyer J. China made waves with Deepseek, but its real ambition is AI-driven industrial innovation. *Nature*. 2025;638(8051):609–611.
24. Gibney E. China's cheap, open AI model DeepSeek thrills scientists. *Nature*. 2025;638(8049):13–14.
25. M-m P. Reflections on DeepSeek's breakthrough. *Natl Sci Rev*. 2025;12(3). doi:10.1093/nsr/nwaf001
26. Addison T, Bhadrashetty A. Advanced Predictive model for heart disease in clinical decision support systems. *Edraak*. 2023;2023:11–15. doi:10.70470/EDRAAK/2023/003
27. Yang Y, Wang H, Ji C, Niu Y. Artificial intelligence-driven diagnostic systems for early detection of diabetic retinopathy: integrating retinal imaging and clinical data. *Shifaa*. 2023;2023:83–90. doi:10.70470/SHIFAA/2023/010
28. Peng Y, Malin BA, Rousseau JF, et al. From GPT to DeepSeek: significant gaps remain in realizing AI in healthcare. *J Biomed Informat*. 2025;163.
29. Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science. *Nature*. 2023;614(7947):214–216. doi:10.1038/d41586-023-00340-6
30. Conroy G. How ChatGPT and other AI tools could disrupt scientific publishing. *Nature*. 2023;622(7982):234–236. doi:10.1038/d41586-023-03144-w
31. Logg JM, Minson JA, Moore DA. Algorithm appreciation: people prefer algorithmic to human judgment. *Organizational Beha Human Decis Processes*. 2019;151:90–103. doi:10.1016/j.obhdp.2018.12.005
32. Garry M, Chan WM, Foster J, Henkel LA. Large language models (LLMs) and the institutionalization of misinformation. *Trends Cognitive Sci*. 2024;28(12):1078–1088. doi:10.1016/j.tics.2024.08.007
33. Messeri L, Crockett MJ. Artificial intelligence and illusions of understanding in scientific research. *Nature*. 2024;627(8002):49–58. doi:10.1038/s41586-024-07146-0
34. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*. 2023;4(9):100804. doi:10.1016/j.patter.2023.100804
35. Rodman A, Topol EJ. Is generative artificial intelligence capable of clinical reasoning? *Lancet*. 2025;405(10480):689. doi:10.1016/S0140-6736(25)00348-4
36. AC F, Benefits MEVCS. Limits, and Risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med*. 2023;388(25):2399–2400.