Advances in Medical Education and Practice

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Evaluating the Applicability of Advanced Large Language Models in Laboratory Medicine Test Questions: A Comparative Performance Study

Wenzheng Han[1,*], Wenkai Zhu[1,*], Gang Feng[1,*], Yankang Wang[1], Guang Chen[2], Huan Zhou[3], Bin Quan[4], Qiwen Wu[1], Jianghua Yang[4], Kai Jin[5], Shaoneng Tao[6], Xiaoning Li[1], Qing Chen[6]

[1]Department of Clinical Laboratory, The First Affiliated Hospital, Wannan Medical College, Wuhu, Anhui, People's Republic of China; [2]Department of Pediatrics, The First Affiliated Hospital, Wannan Medical College, Wuhu, Anhui, People's Republic of China; [3]School of Laboratory Medicine, Wannan Medical College, Wuhu, Anhui, People's Republic of China; [4]Department of Infectious Diseases, The First Affiliated Hospital, Wannan Medical College, Wuhu, Anhui, People's Republic of China; [5]Department of Eye Center, The Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, Zhejiang, People's Republic of China; [6]Department of Nuclear Medicine, The First Affiliated Hospital, Wannan Medical College, Wuhu, Anhui, People's Republic of China

*These authors contributed equally to this work

Correspondence: Qing Chen, Department of Nuclear Medicine, The First Affiliated Hospital, Wannan Medical College, Wuhu, Anhui, People's Republic of China, Email chenq1104@wnmc.edu.cn

Background: While large language models (LLMs) show promise in medical education, their comprehensive performance in specialized domains like medical laboratory science remains inadequately assessed.

Purpose: This study aimed to evaluate advanced LLMs on medical laboratory questions, assessing accuracy, natural language generation (NLG) quality, reasoning performance, and efficiency.

Methods: We conducted a multi-faceted evaluation of three advanced LLMs (DeepSeek-R1, Gemini-2.5 Pro, GPT-5), benchmarking them against medical laboratory scientists and earlier ChatGPT versions. The evaluation utilized 493 questions sourced from the internal Medical Laboratory Test Bank of Wannan Medical College. These questions comprised both knowledge-based and reasoning-based single- and multiple-choice types (SCQs and MCQs). Performance was measured by accuracy, Macro-F1, response time, NLG scores (ROUGE-L, METEOR), and structured logical reasoning assessment. Appropriate statistical tests (including χ^2, Wilcoxon, ANOVA, and non-parametric alternatives) with post-hoc corrections were applied to determine significance.

Results: DeepSeek-R1's accuracy on total questions was 78.3%, nearing the 79.3% of the higher-performing senior expert. Notably, it excelled at complex reasoning-based MCQ, demonstrating an advantage over senior experts with an accuracy of 64.4%, compared to 58.7% (SMLS-1) and 56.7% (SMLS-2). While ChatGPT-5 was the fastest model, DeepSeek-R1 exhibited intermediate efficiency, aligning with human experts on SCQ but requiring more time for MCQ. In terms of NLG, DeepSeek-R1 consistently achieved the highest scores, with ROUGE-L scores of 0.36 ± 0.14 (Total Q), 0.33 ± 0.15 (SCQ), and 0.38 ± 0.13 (MCQ), and METEOR scores of 0.53 ± 0.19 (Total Q), 0.40 ± 0.17 (SCQ), and 0.63 ± 0.14 (MCQ). Furthermore, it significantly outperformed all other LLMs in logical reasoning comprehensiveness. A critical strength was its consistent integration of key negative findings, vital for diagnosis.

Conclusion: DeepSeek-R1 approaches or even surpasses senior expert performance in certain tasks, showing strong potential as an effective tool for education and assessment despite slower processing times.

Keywords: large language models, medical laboratory science, natural language generation, deepseek-R1, complex reasoning

Introduction

The education and training of medical laboratory professionals are critical to ensuring the accuracy and reliability of diagnostic testing, which forms the cornerstone of modern clinical decision-making.[1,2] A key objective in this field is to cultivate advanced competencies, including the interpretation of complex laboratory data and the synthesis of diagnostic reasoning. Traditional assessment of these skills relies heavily on human experts, which can be limited by scalability,

https://doi.org/10.2147/AMEP.S576651
Received: 25 October 2025
Accepted: 9 February 2026
Published: 24 February 2026

subjectivity, and resource constraints. The emergence of large language models (LLMs) presents a transformative opportunity to augment educational and assessment methodologies by providing intelligent, automated evaluation that can mimic expert judgment.[3–5]

Currently, mainstream LLMs include DeepSeek, Gemini, and the GPT series. These models are built upon the Transformer architecture and have been pre-trained on vast corpora comprising trillions of tokens from diverse sources, including academic literature, medical textbooks, clinical guidelines, and publicly available web text. Each model has undergone extensive instruction tuning and alignment optimization to enhance their conversational and reasoning capabilities. Notably, DeepSeek-R1 incorporates a mixture-of-experts (MoE) design and is explicitly optimized for complex reasoning tasks through reinforcement learning from human feedback (RLHF), which may explain its superior performance in multi-step logical inference. Gemini benefits from strong multimodal integration capabilities. The GPT series, while achieving the fastest response times, may rely more on heuristic pattern matching rather than deep analytical reasoning. The varying performance across models can be attributed to differences in architectural focus, training objectives, and the depth of domain-specific knowledge embedded during pre-training and fine-tuning. A growing body of empirical research provides substantiating evidence by evaluating the performance of LLMs in various medical contexts. For example, an analysis of 100 clinical records demonstrated that ChatGPT-4o exhibited superior accuracy and clinical safety compared to ChatGPT-3.5 and ERNIE Bot in assessing intervention necessity and recommending treatment regimens, with its recommendations showing stronger adherence to specialist consensus and maintained robustness even with incomplete datasets.[6] Similarly, in a complex bilingual ophthalmology assessment, DeepSeek-R1 outperformed contemporary models such as Gemini 2.0 Pro, OpenAI o1, and o3-mini, particularly in management-related questions, underscoring its utility in clinical reasoning tasks.[7] Further supporting evidence comes from a study involving 86 respiratory pathogen reports, where GPT models, including versions 4o, o1, and o1-mini, effectively identified several major error types with high accuracy and strong concordance with expert evaluations, highlighting their role in enhancing laboratory quality assurance and diagnostic support.[8]

However, the effective application of general-purpose LLMs in the highly specialized domain of medical laboratory science faces significant challenges.[9–11] While some models demonstrate formidable knowledge recall,[12,13] their ability to execute the nuanced logical reasoning required for differential diagnosis often remains unproven. Furthermore, comprehensive evaluations that simultaneously assess model performance on quantitative accuracy (eg, correct answers), qualitative explanation quality (eg, semantic coherence and fluency), and diagnostic logic against established human expert benchmarks are still lacking. Moreover, factors such as computational load and architectural typology lead to significant disparities in the response times of various LLMs. The response times of these models in processing specific medical laboratory test questions also remain poorly understood. These combined gaps hinder the reliable application of LLMs in medical laboratory education.

To address this gaps, we conducted a rigorous comparative evaluation. First, we benchmarked the performance of the latest advanced LLMs, including DeepSeek-R1, Gemini-2.5 Pro, and GPT-5, against that of junior and senior medical laboratory scientists (JMLS and SMLS) on diverse medical laboratory examinations. The question set comprised both knowledge-based and reasoning-based multiple-choice questions (in single-select and multi-select formats). These questions were sourced from the Medical Laboratory Test Bank of Wannan Medical College, an internal resource used for student assessment in medical laboratory science programs. This benchmarking aimed not only to investigate whether these models can achieve expert-level accuracy but also to characterize and compare their response times, an aspect currently poorly understood. Second, and more critically, we expanded the evaluation beyond accuracy to include a comprehensive assessment of the quality of natural language text generation and logical reasoning capabilities. This comparative analysis encompassed both the aforementioned latest models and established traditional ChatGPT models (such as GPT-3, GPT-4o, GPT-4-mini, and GPT-4.1). By integrating these quantitative and qualitative analyses, this study aims to provide a holistic understanding of the applicability and limitations of these advanced LLMs in medical laboratory education.

Methods

Design of the Study and Research Setting

A comprehensive set of medical laboratory test questions was curated for this evaluation, comprising a total of 493 items categorized by format and cognitive demand. The question pool included 261 single-choice questions (SCQs) and 232 multiple-choice questions (MCQs). In terms of cognitive complexity, the items were classified into knowledge-based and reasoning-based types, with the latter requiring advanced analytical and diagnostic skills. To facilitate a focused assessment of logical reasoning, a diagnostic subset of 74 questions (61 SCQs and 13 MCQs) was extracted from the full set. This study was designed as a comparative evaluation of three state-of-the-art large language models (LLMs), including DeepSeek-R1, Gemini-2.5 Pro, and ChatGPT-5, in the context of medical laboratory testing. Their performance was assessed based on Accuracy, Macro-F1 score, and Efficiency, and benchmarked against a human cohort of four medical laboratory scientists. This cohort consisted of two senior experts (SMLS-1, SMLS-2) with 8–10 years of experience and two junior scientists (JMLS-1, JMLS-2) with 1–2 years of experience. All participants were aware of the study's objective but were blinded to the specific question source, content, and the answers or performance of other participants (both human and LLMs). Each expert completed the evaluation independently in a controlled setting. Additionally, the same three LLMs were evaluated on Natural Language Generation (NLG) quality and logical reasoning capability in comparison to several established ChatGPT variants, including o3, 4o, o4-mini, and 4.1. The overall design and workflow of the study are summarized in Figure 1.

Question Source and Selection Criteria

The study utilized 493 questions sourced from the Medical Laboratory Test Bank of Wannan Medical College. This internal educational resource, regularly updated and reviewed by faculty experts, is used for student assessment in medical laboratory science programs. The selected questions covered core areas including clinical chemistry, hematology, microbiology, immunology, molecular diagnostics, and laboratory management. They comprised both knowledge-based questions (assessing recall and comprehension) and reasoning-based questions (requiring analysis, synthesis, and diagnostic judgment). Questions were included only if they were clearly formulated and clinically relevant, possessed a correct answer along with an explanatory rationale; conversely, questions were excluded if they were ambiguous or poorly structured, relied on outdated clinical guidelines or obsolete technologies, or contained known errors. To minimize selection bias, questions were drawn randomly from each domain and cognitive category, ensuring proportional representation. The 493 questions are presented in Supplemental Table 1.

Prompting Strategy and Input

The prompts were crafted to provide the results of medical laboratory test questions (Prompt I and II) (Supplemental Table 2). Responses were generated by DeepSeek-R1 (DeepSeek, Hangzhou, China), Gemini-2.5 Pro (Google, CA, USA), and ChatGPT models (version 5, o3, 4o, o4-mini, and 4.1; OpenAI, CA, USA) in accordance with the provided prompts. To ensure impartiality and consistency, each question was processed by the models in isolation. The conversation history was cleared after each query to prevent cross-contamination from previous interactions. Additionally, all questions were converted to plain text before submission to minimize output variations caused by inherent formatting differences, such as bold and/or italic emphasis on certain words (eg, "not"). This precaution was taken to prevent any potential influence that visual emphasis might have on model processing.

Methodology for Time Assessment

For human evaluators, the time required to review each report was recorded using a digital stopwatch. Specifically, we instructed each medical laboratory scientist that, upon receiving a question, they should immediately press a button to start the timer, proceed to solve it, and then press the same button again upon completing their answer. A concealed stopwatch running in the background automatically captured the interval between the two button presses. We measured the LLM's response time by pasting each question into the dialog interface, starting a timer when clicking "submit", and stopping it once the answer was fully generated. We defined this response time as the wall-clock latency from prompt

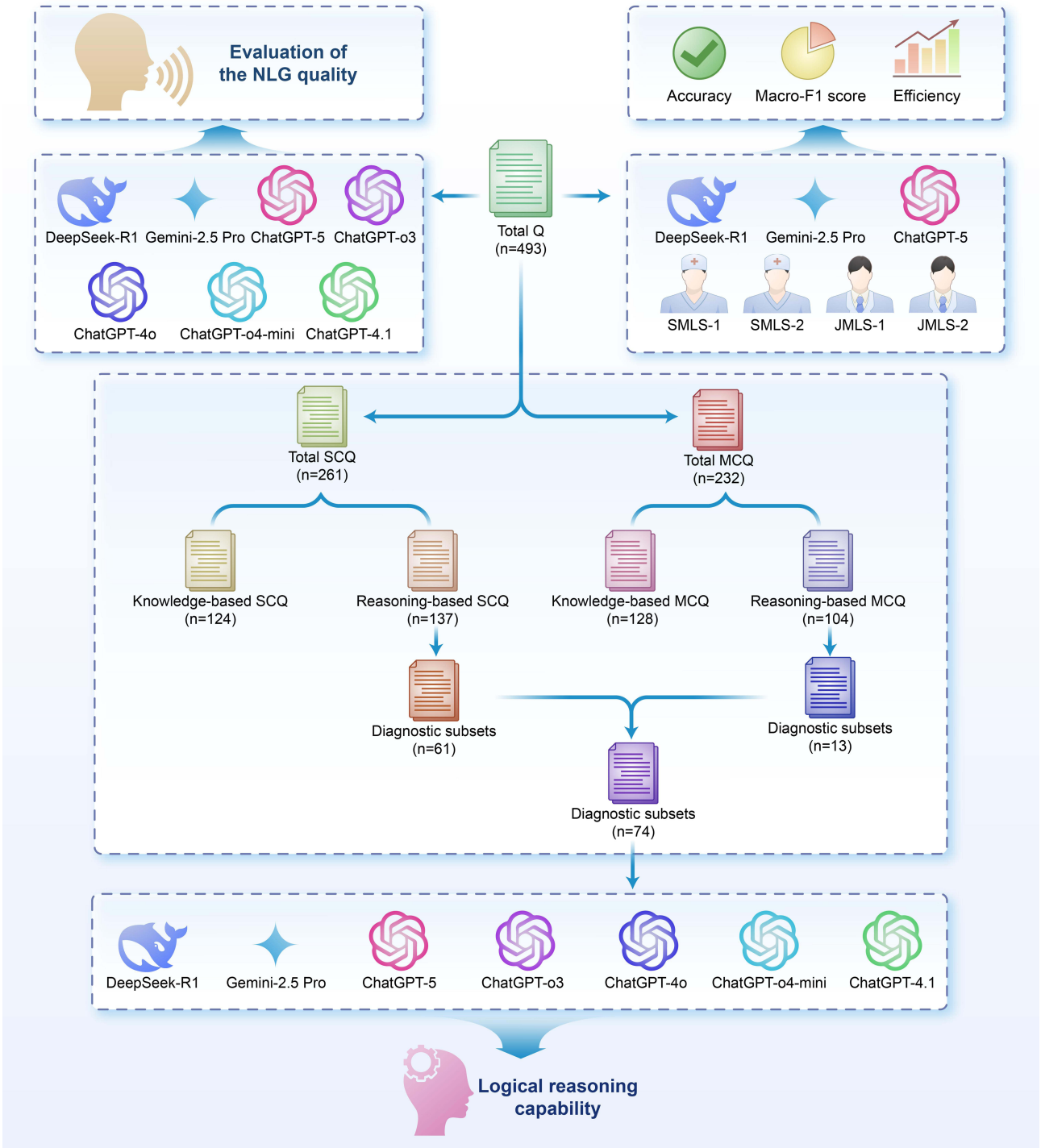

Figure 1 Study design flowchart. A total of 493 medical laboratory questions were collected and categorized by question type into single-choice questions (SCQs) (n = 261) and multiple-choice questions (MCQs) (n = 232). These were further classified into: knowledge-based SCQs (n = 124) and knowledge-based MCQs (n = 128); reasoning-based SCQs (n = 137) and reasoning-based MCQs (n = 104). Three state-of-the-art large language models (LLMs), including DeepSeek-R1, Gemini-2.5 Pro, and ChatGPT-5, were tasked with answering the full set of questions. Their performance was evaluated based on accuracy, Macro-F1 score, and efficiency, and compared against the performance of four human evaluators: junior and senior medical laboratory scientists (JMLS and SMLS). To further assess the natural language generation (NLG) quality and logical reasoning capability of the LLMs, several traditional ChatGPT models, including GPT-o3, GPT-4o, GPT-o4-mini, and GPT-4.1, were also presented with the same questions. Response outcomes from all models were recorded to enable a comprehensive comparative analysis between the state-of-the-art and traditional LLMs.

https://doi.org/10.2147/AMEP.S576651

submission to complete response receipt, measured on the client side under identical network conditions. For ChatGPT-5, whose response time is contingent upon configurable parameters such as verbosity and reasoning effort, both were set to a medium level for all evaluations.

Evaluation of Diagnostic Reasoning Elements

For diagnostic questions, all LLM responses were evaluated using a structured framework that assessed five critical elements of diagnostic reasoning. Each element was scored as 1 point if fully included, or no point if absent or partially omitted. The elements and their scoring criteria are as follows: (A) Essential positive history (Score 1 point for inclusion of essential positive history). (B) Key positive findings (Score 1 point for inclusion of key positive test results, examination findings, and signs). (C) Essential negative history (Score 1 point for inclusion of essential negative history). (D) Key negative findings (Score 1 point for inclusion of key negative test results, examination findings, and signs). (E) Diagnostic conclusions (Score 1 point for inclusion of diagnostic conclusions and differential diagnoses). The presence of these elements was recorded by senior laboratory medicine professionals with over ten years of experience, who were blinded to model identities and performed all evaluations independently.

Statistical Analysis

Statistical analyses were conducted using SPSS Statistics version 22.0 (IBM Corp., Armonk, NY, USA), JASP (version 0.14.1; University of Amsterdam), and OriginPro (OriginLab, USA). Each LLM and human reader performance was assessed based on overall accuracy, defined as the proportion of correctly answered items. In addition, the Macro-F1 score for each LLM was computed as the unweighted harmonic mean of precision and recall across all response options. To assess the natural language generation (NLG) quality, model-generated explanations were compared against expert-authored reference answers using two established metrics: ROUGE-L (recall-oriented understudy for gisting evaluation-longest common subsequence) and METEOR (metric for evaluation of translation with explicit ordering).[14,15] These metrics evaluate the quality of each LLM's final output by comparing it to the ground truth reasoning for each item; the intermediate thinking process of the models were not included in this analysis. In short, ROUGE-L measures structural similarity and content overlap based on the longest common subsequence between model outputs and reference answers, while METEOR provides a holistic assessment of explanation quality by accounting for fluency, synonymy, and semantic alignment. For categorical data, Pearson's χ^2-tests, continuous χ^2-tests, or Fisher's exact tests were used as appropriate, and 95% confidence intervals were derived via the Clopper-Pearson method. Pairwise comparisons of macro-F1 scores were conducted using one-way ANOVA with Tukey's HSD post-hoc test for parametric analyses or the Kruskal–Wallis test with Dunn's test for non-parametric analyses, based on data distribution. Comparisons of reading times employed the Wilcoxon rank-sum test, with effect sizes measured by matched rank biserial correlation (interpreted on a scale from $|r| = 0.0$ to 1.0). The Kruskal–Wallis test was used to compare the significance of ROUGE-L and METEOR scores across the seven LLMs, followed by Dunn's test for post-hoc analysis. Statistical significance was defined as a two-sided p-value < 0.05, adjusted for multiple comparisons using the Bonferroni method unless otherwise specified.

Results

Accuracy Performance on Medical Laboratory Test Questions

We evaluated the accuracy of three LLMs, such as DeepSeek-R1, Gemini-2.5 Pro, and ChatGPT-5. They were compared against two SMLS and two JMLS readers. The evaluation covered multiple question types: the total set, SCQs, MCQs, along with their knowledge-based and reasoning-based subsets. Results are presented using radar charts of accuracy scores (Figure 2A) and pairwise statistical significance (p-values) derived from comparative analyses (Figure 2B–D). DeepSeek-R1 demonstrated high accuracy across all question categories, closely matching or even exceeding the performance of senior experts on specific questions. In total questions, the accuracy of the three LLMs was not statistically different from that of SMLS-1 and SMLS-2 (all $p > 0.007$), indicating expert-level competency. In contrast, the three LLMs showed significantly higher accuracy compared to the JMLS group (all $p < 0.001$) (Figure 2B). In the SCQs categories, the performance of DeepSeek-R1 and Gemini-2.5 Pro was comparable to that of SMLS readers yet

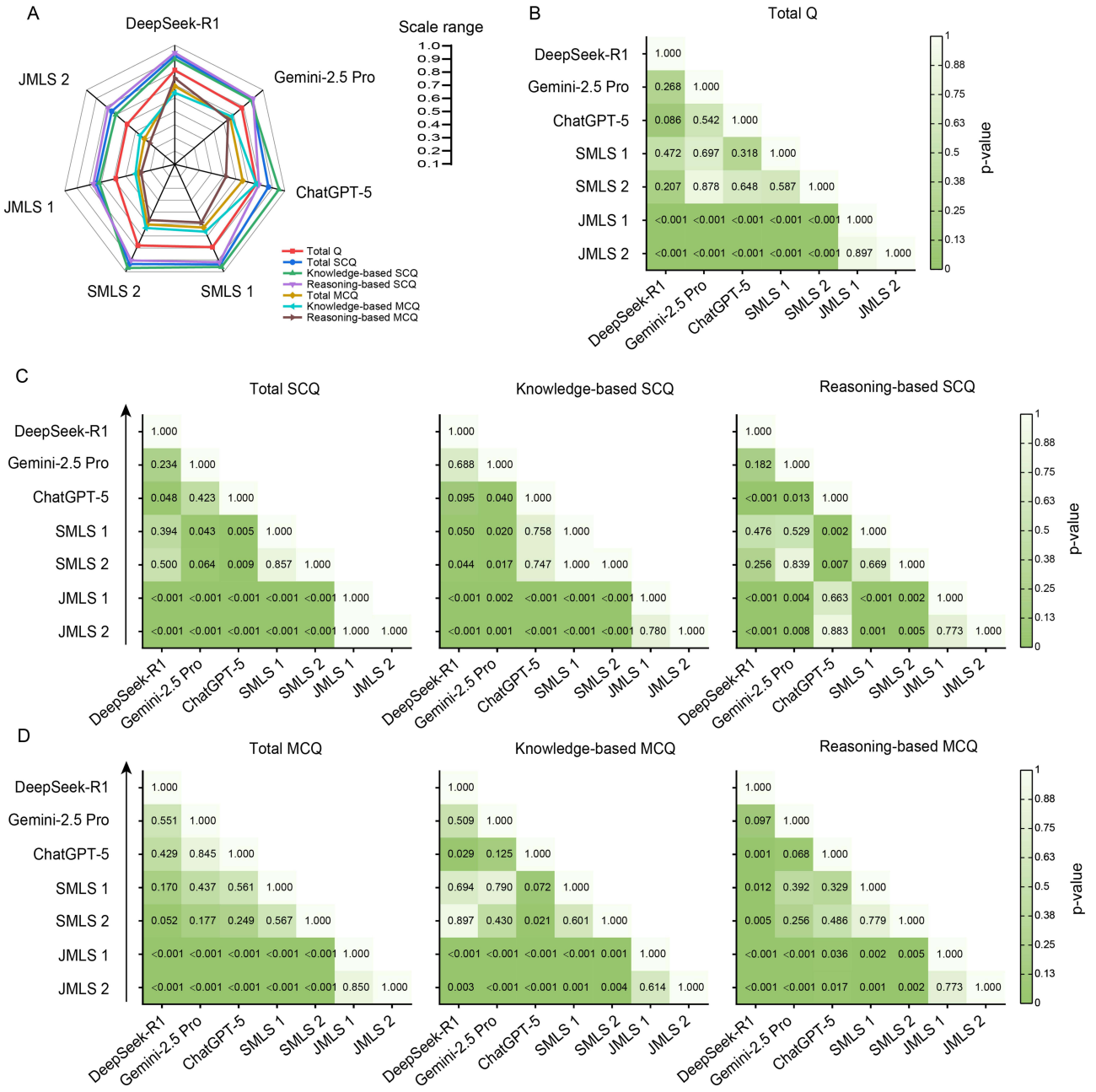

Figure 2 Performance accuracy and statistical significance of LLMs and human experts. Accuracy scores (**A**) and their associated p-values (**B**–**D**) for the three state-of-the-art LLMs (DeepSeek-R1, Gemini-2.5 Pro, and GPT-5) and four human experts (SMLS and JMLS) across different question categories are shown in a radar chart and heatmaps, respectively. The question types are divided into total questions (Total Q), single-choice questions (SCQ), and multiple-choice questions (MCQ), with further subdivisions into knowledge-based and reasoning-based subsets. P-values are provided to indicate the statistical significance of performance differences between readers. The significance between two readers was calculated by using Pearson's χ^2 test, or Continuous χ^2 Test where appropriate, respectively. Bonferroni correction was used to correct *P* values (< 0.007 was considered significant) for multiple comparisons.

Abbreviations: SMLS, senior medical laboratory scientists; JMLS, junior medical laboratory scientists.

superior to that of JMLS readers (Figure 2C). However, ChatGPT-5 showed relatively poor performance, especially in reasoning-based SCQs which was comparably to JMLS readers (Figure 2C, right panel). In MCQs categories, DeepSeek-R1 and Gemini-2.5 Pro performed comparably to SMLS readers (all p > 0.007), but significantly outperformed JMLS readers (all p < 0.007) (Figure 2D). For reasoning-based MCQs, the accuracy of DeepSeek-R1 (75.0%, 78/104) and Gemini-2.5 Pro (64.4%, 67/104) was numerically higher than that of both SMLS readers (SMLS-1: 58.7%; SMLS-2: 56.7%) and ChatGPT-5 (51.9%). However, statistical comparisons showed that only a subset of these differences reached

https://doi.org/10.2147/AMEP.S576651

statistical significance (Figure 2D, right panel). DeepSeek-R1 demonstrated particular strength in reasoning-based questions. It significantly outperformed both ChatGPT-5 and JMLS readers in the reasoning subsets of SCQs and MCQs ($p < 0.007$), and its performance in reasoning-based MCQs even surpassed that of SMLS readers. This suggests that DeepSeek-R1's accuracy is on par with or superior to senior experts, especially in complex tasks like reasoning-based MCQs, while the other LLMs and junior experts generally underperformed.

We sought to further evaluate how well LLMs and human experts discriminate at the option level, focusing on their ability to identify correct options and exclude incorrect ones. To this end, we assessed their performance using Macro-F1 scores across multiple question categories. The radar charts revealed that DeepSeek-R1 and Gemini-2.5 Pro consistently exhibited high performance across all seven categories, approaching the level of the SMLS readers (Supplemental Figure 1A). For the total question set, the performance of the three LLMs was not statistically different from that of the two senior readers (SMLS) ($p > 0.05$). However, all three LLMs, along with the senior experts, collectively demonstrated significantly higher Macro-F1 scores than the junior readers (Supplemental Figure 1B). In the total SCQs, all three LLMs demonstrated comparable excellence. For knowledge-based SCQs, among the three LLMs, only ChatGPT-5 maintained a relatively poor performance; both DeepSeek-R1 and Gemini-2.5 Pro performed comparably well, approaching the level of the SMLS readers. For reasoning-based SCQs, DeepSeek-R1 and Gemini-2.5 Pro performed similarly, both outperforming ChatGPT-5, although the difference was not statistically significant (Supplemental Figure 1C). In MCQs, in contrast to ChatGPT-5, both DeepSeek-R1 and Gemini-2.5 Pro maintained high performance, demonstrating significant superiority over junior readers specifically in reasoning-based MCQs. Senior experts also consistently exceeded junior experts across all MCQ categories (Supplemental Figure 1D). These findings underscore the strong performance of DeepSeek-R1 and Gemini-2.5 Pro in the reasoning subsets of both SCQs and MCQs.

Efficiency in Processing Medical Laboratory Test Questions

The time efficiency of three LLMs and human experts (both senior and junior) was evaluated across various question types. DeepSeek-R1's mean reading time of 57.4 ± 51.2 seconds per report was significantly longer than that of ChatGPT-5 (11.6 ± 6.6 seconds, $p < 0.0001$, $|r| = 0.994$). Its performance was statistically comparable to the two JMLS readers (50.1 ± 30.0 and 45.4 ± 24.8 seconds, all $p > 0.05$) and Gemini-2.5 Pro (39.6 ± 10.0 seconds, $p > 0.05$). However, it remained significantly slower than both SMLS readers (SMLS-1: 36.1 ± 22.1 seconds, $p < 0.0001$; SMLS-2: 38.2 ± 24.0 seconds, $p = 0.0013$). In SCQ categories, DeepSeek-R1 was significantly faster than Gemini-2.5 Pro ($p < 0.0001$) but slower than ChatGPT-5 ($p < 0.0001$). Its performance was comparable to that of SMLS readers ($p > 0.05$), yet significantly faster than both JMLS-1 and JMLS-2 ($p < 0.01$). For MCQ categories, DeepSeek-R1 required substantially more time than all other readers ($p < 0.05$ for most comparisons), including both SMLS and JMLS evaluators. This pattern was consistent across both knowledge-based and reasoning-based MCQs, where DeepSeek-R1's processing times were the highest among all participants (Figure 3 and Supplemental Table 3). The results indicated that ChatGPT-5 was the fastest across all question types, with the smallest time variance. DeepSeek-R1 exhibited intermediate efficiency, often aligning with human experts in simpler question formats (such as SCQs) but requiring more time for complex tasks (such as MCQs). Human experts generally took longer for MCQs than SCQs, with JMLS readers being slower than SMLS readers in most categories.

Text Reasoning Quality Assessed by ROUGE-L and METEOR

To comprehensively evaluate the NLG quality of LLMs, we employed two classic metrics, ROUGE-L and METEOR. These metrics were used to assess the similarity and fluency of explanations generated by several LLMs, including traditional ChatGPT models and three latest LLMs, against reference answers for medical laboratory test questions. The evaluation assessed model performance across the entire question set, including both SCQs and MCQs. DeepSeek-R1 consistently achieved the highest ROUGE-L scores in all categories (Figure 4A–C), demonstrating superior alignment with expert answers in terms of both content overlap and structural similarity. For the total question set, DeepSeek-R1's performance significantly surpassed that of all other models ($p < 0.0001$ for all pairwise comparisons) (Figure 4A). This leading trend was similarly evident in the SCQ and MCQ subsets, where its outputs showed greater semantic and syntactic coherence (Figure 4B and C). DeepSeek-R1 also excelled according to the METEOR metric, which places

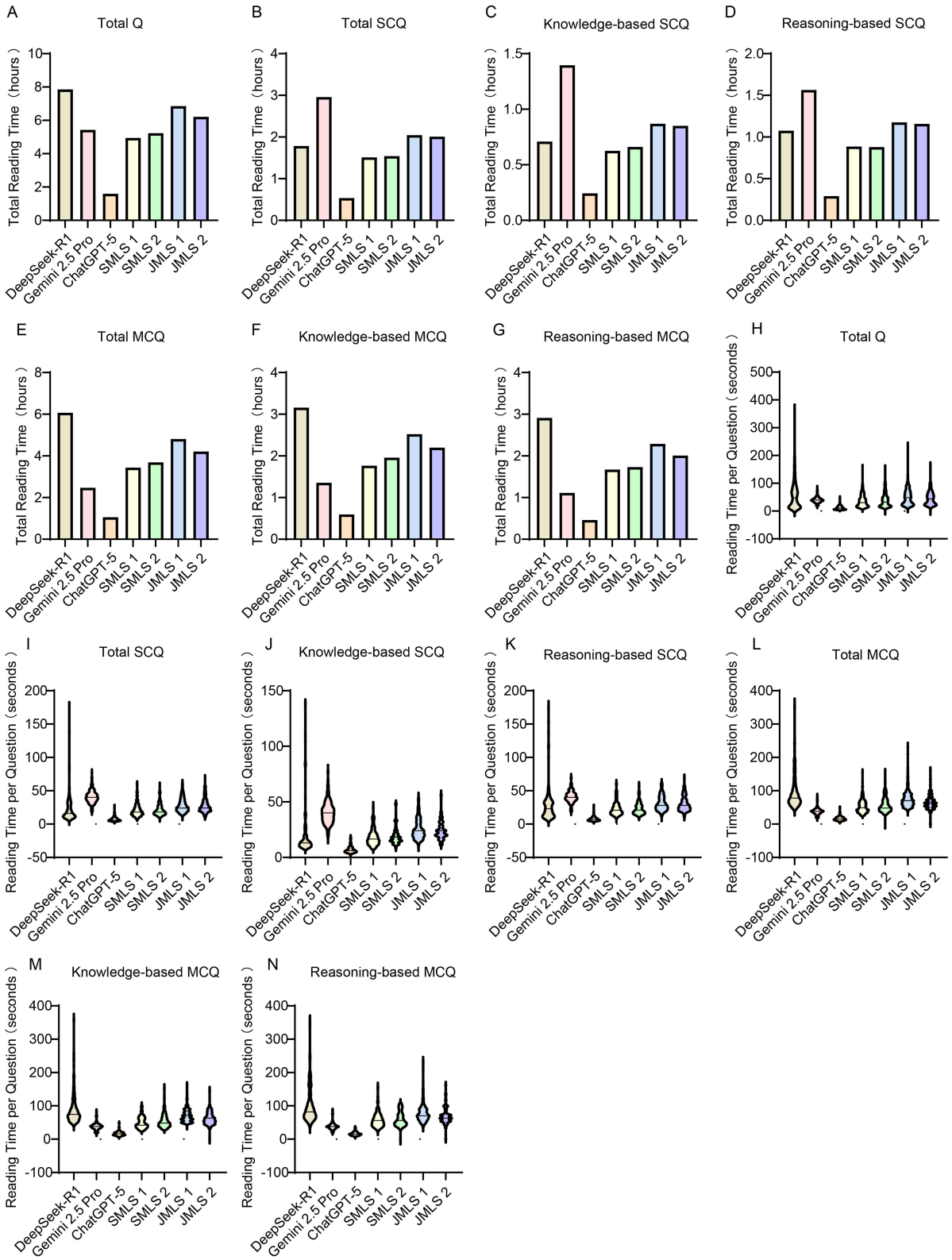

Figure 3 Comparison of reading times. The bar graph (**A**–**G**) and the violin plot (**H**–**N**) respectively illustrates the total reading time in hours and the reading times per question in seconds required by different groups. The groups include the three state-of-the-art LLMs (DeepSeek-R1, Gemini-2.5 Pro, and GPT-5) and four human experts (SMLS and JMLS).

Abbreviations: SMLS, senior medical laboratory scientists; JMLS, junior medical laboratory scientists.

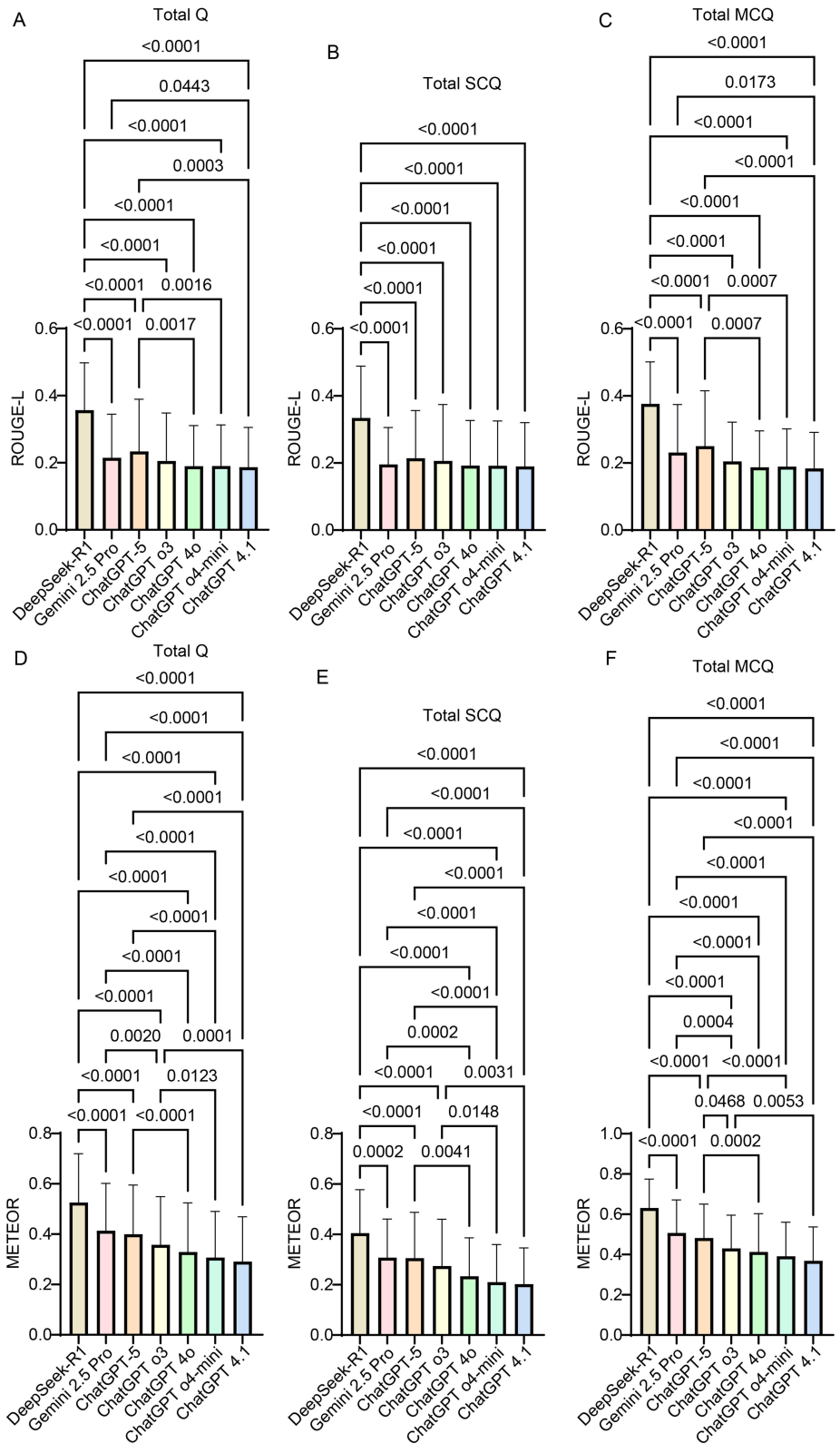

Figure 4 Comparative evaluation of textual reasoning capabilities across multiple LLMs. The performance of various LLMs on textual reasoning tasks was assessed using ROUGE-L (**A**–**C**) and METEOR (**D**–**F**) metrics. The models included three state-of-the-art LLMs (DeepSeek-R1, Gemini-2.5 Pro, and GPT-5) and multiple traditional ChatGPT variants (such as GPT-o3, GPT-4o, GPT-o4-mini, and GPT-4.1). Subplots correspond to different question categories: Total Questions (Total Q), Single-Choice Questions (SCQ), and Multiple-Choice Questions (MCQ). Kruskal–Wallis test followed by Dunn's test was used for multiple comparison.

https://doi.org/10.2147/AMEP.S576651

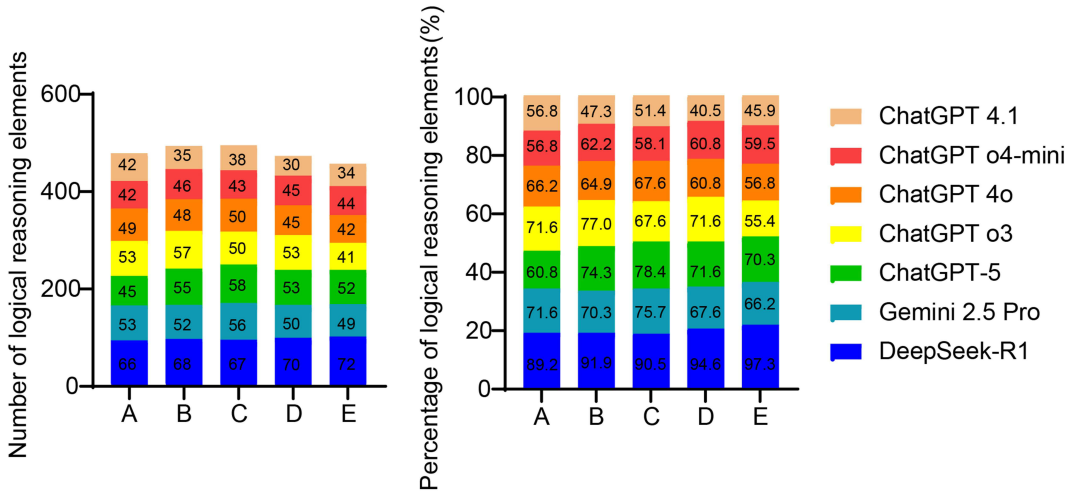

Figure 5 Analysis of logical reasoning performance among multiple LLMs. The logical reasoning capabilities of the three state-of-the-art LLMs (DeepSeek-R1, Gemini-2.5 Pro, and GPT-5) and multiple traditional ChatGPT variants (eg, GPT-o3, GPT-4o, GPT-o4-mini, and GPT-4.1) were evaluated. The left panel presents the absolute count of logical reasoning elements (and the exact score) for each chunk; the right panel displays their relative percentage distribution (and the percentage per chunk). Five critical elements for diagnostic reasoning were defined as follows: A: Essential positive history; B: Key positive findings; C: Essential negative history; D: Key negative findings; E: Diagnostic conclusions.

greater emphasis on fluency, synonym matching, and semantic similarity. It attained the highest METEOR score on the total question set ($p < 0.0001$ against all competitors) and maintained this superior performance in both SCQs and MCQs (Figure 4D–F). Notably, for the more complex MCQs, DeepSeek-R1's METEOR score was significantly higher than those of all ChatGPT variants and Gemini-2.5 Pro ($p < 0.0001$) (Figure 4F). This confirmed the robustness of DeepSeek-R1's advantage in generating high-quality, reasoning-based textual responses.

Logical Reasoning Capability in Diagnostic Questions

To further evaluate the logical reasoning capabilities of LLMs, a subset of 74 diagnostic questions (61 single-choice and 13 multiple-choice) was selected from a total pool of 493. The performance of various LLMs was assessed by analyzing the logical structure of their responses to these diagnostic prompts. The assessment framework was based on five critical elements essential for logical reasoning. As shown in Figure 5, DeepSeek-R1 significantly outperformed all ChatGPT variants (including ChatGPT-4.1, o4-mini, 4o, o3, and ChatGPT-5) and Gemini-2.5 Pro in the comprehensiveness of its reasoning. While most models adequately incorporated positive findings (Elements A and B), only DeepSeek-R1 consistently integrated critical negative elements (C and D), which are vital for a thorough differential diagnosis. Furthermore, DeepSeek-R1 achieved a substantially higher score in Element E (diagnostic conclusions), demonstrating superior synthetic and diagnostic reasoning capabilities.

Discussion

LLMs hold the potential to enhance students' learning experiences in medical education, particularly through providing personalized learning and replicating human-level performance in medical knowledge.[16] Based on a comprehensive evaluation across multiple dimensions of medical laboratory science expertise, DeepSeek-R1 demonstrates excellent capabilities, establishing its potential as a valuable tool in specialized medical education and assessment. The model's performance reveals several key insights regarding the current state of large language models in specialized medical domains.

The most striking finding is DeepSeek-R1's exceptional accuracy in tasks, particularly multiple-select questions, achieving significantly higher accuracy than both senior experts and other state-of-the-art models. This superiority in complex reasoning scenarios suggests that DeepSeek-R1's architectural optimization for reasoning tasks translates effectively to the medical laboratory domain. The model's ability to handle the cognitive complexity inherent in MCQs, which require simultaneous consideration of multiple variables and exclusion of distractors, indicates

a sophisticated understanding of diagnostic relationships that extends beyond simple pattern recognition. This aligns with previous research showing specialized LLMs' potential in clinical reasoning tasks,[7,17–19] but extends it by demonstrating this capability specifically in laboratory medicine contexts. Additionally, different prompt strategies, especially Chain-of-Thought, can enhance AI reliability in complex scenarios.[20] However, this study utilized a direct prompting strategy and did not explore methods such as Chain-of-Thought or self-reflection. In future work, we will continue to investigate how LLMs respond to different prompt strategies for specific problems.

Our multi-faceted evaluation approach reveals that DeepSeek-R1's advantages extend beyond mere accuracy. The model demonstrated superior performance on established NLG metrics, specifically ROUGE-L and METEOR. These metrics were selected due to their strong correlation with human judgments of explanation quality in prior medical-AI studies. This superior performance indicates the model's capacity to produce explanations that closely mirror expert reasoning in both content and structure. More importantly, the logical reasoning analysis revealed a key finding. DeepSeek-R1 consistently incorporated the highest level of negative reasoning elements among all models. Notably, another article indicated that DeepSeek can approach expert-level performance in Top-3 and Top-5 diagnoses, demonstrating clinical auxiliary value, although its Top-1 accuracy still lags significantly.[21] In this study, we evaluated the logical reasoning processes of seven LLM models, including scoring based on five key elements involved in reasoning. We did not separately rank diagnostic accuracy, such as Top-1 or Top-2 etc. Future research will focus on this aspect and explore the accuracy of LLMs in Top-5 differential diagnosis ranking.

The efficiency results present a more nuanced picture. While DeepSeek-R1 required substantially more processing time than ChatGPT-5, particularly for complex MCQs, its performance aligned with human experts on simpler SCQs. This time-accuracy trade-off suggests that DeepSeek-R1 may be engaging in more deliberate, analytical processing rather than relying on heuristic responses. The fact that its processing patterns mirrored human experts' increased time investment for more complex tasks further supports the notion that it may be employing reasoning strategies analogous to human experts.

This study has several limitations. First, the ecological validity of the choice-based evaluation is limited, as such tasks do not adequately reflect real-world clinical decision-making processes. Second, there is a risk of data contamination, as large language models may have been exposed to similar questions during their pretraining phase. Third, the expert sample size was small. While it was sufficient to establish a clear senior-junior performance gradient for benchmarking the LLMs, it may not fully capture the performance variability within a broader expert population. This restricts the generalizability of the results. Forth, the study focused on models at a specific stage of development, while the rapid evolution of large language model technology necessitates ongoing assessment. Furthermore, while the DeepSeek model achieved performance comparable to that of human experts on multiple-choice benchmarks, the absolute accuracy (0.6–0.8) observed across the three datasets may not yet meet the robustness requirements for unattended real-world applications. Therefore, future work should focus not only on achieving parity with humans but also on enhancing the absolute reliability of AI systems in complex, high-stakes reasoning tasks.

This study found that DeepSeek-R1 demonstrates exceptional complex reasoning and logical completeness in structured testing, laying a foundation for its future application in clinical decision support systems. Although the ecological validity of the current multiple-choice-based evaluation is limited, the model's capabilities in systematic analysis, negative reasoning, and generating coherent explanations represent core elements for building trustworthy AI-assisted diagnostic tools. In the future, if such reasoning abilities-validated in closed tasks-can be transferred to open clinical scenarios (such as differential diagnosis ranking based on complete medical records, interpretation of test results, and recommendations), it will be possible to develop assistive decision-making systems capable of logical dialogue with physicians and providing transparent reasoning processes, thereby enhancing diagnostic accuracy and efficiency.

The implications for medical laboratory education are substantial. DeepSeek-R1 holds significant application value in this field. Specifically, its capabilities can support intelligent tutoring systems that guide students through diagnostic reasoning processes. It can also serve as a scalable, standardized assessment tool for large-scale educational evaluation. Furthermore, its logical reasoning abilities make it well-suited for clinical reasoning training and the development of diagnostic thinking. Future research should explore its performance in open diagnostic scenarios. It should also examine how effectively it integrates into authentic learning environments. Additionally, researchers should develop hybrid solutions that leverage the strengths of complementary models.

Conclusion

Our study conducted a multidimensional evaluation to systematically compare the performance of advanced large language models on laboratory medicine test questions. In terms of accuracy, DeepSeek-R1 closely approached, and in certain tasks even surpassed, the performance of senior experts. In processing efficiency, DeepSeek-R1 required time similar to human experts on single-choice questions but significantly more time on multiple-select questions. This pattern suggests that the model engages in deeper analytical processing rather than relying on heuristic responses for complex tasks. In textual reasoning quality, DeepSeek-R1 significantly outperformed all other models on both the ROUGE-L and METEOR metrics, indicating that its generated explanations align more closely with expert reasoning in content, structure, and semantics. Concerning logical reasoning completeness, DeepSeek-R1 consistently incorporated the highest level of negative reasoning elements among all models, reflecting a structured reasoning capability that better approximates clinical diagnostic thinking.

In summary, DeepSeek-R1 achieves expert-level accuracy in laboratory medicine education tasks and even surpasses it in certain scenarios. It also exhibits strong capabilities in complex reasoning and generating explanations. While its processing efficiency, particularly for complex questions, could be improved, its robust analytical and logical reasoning skills underscore its potential as an educational and assessment tool.

Data Sharing Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Consent for Publication

The present study analyzed exam items sourced from the Medical Laboratory Test Bank within the internal resources of Wannan Medical College's official website. As the study did not involve any human or animal subjects, it was exempt from ethical approval. All data were anonymized and handled in accordance with relevant data privacy regulations.

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.

Funding

The present study was supported by the Natural Science Foundation of Universities in Anhui Province (grant no. 2023AH051742, and 2024AH051936), the Health Research Foundation of Anhui Province (grant no. AHWJ2023A20546), Natural Science Foundation of China (grant no. 82201195).

Disclosure

The authors declare no competing interests in this work.

References

1. Jator EK, Phillips HL, Latchem SR, Catalano TA. Establishing the need for standardized clinical educator training programs for medical laboratory professionals. *Lab Med*. 2023;54(2):e63–e69. doi:10.1093/labmed/lmac108
2. Lubin IM, Astles JR, Shahangian S, et al. Bringing the clinical laboratory into the strategy to advance diagnostic excellence. *Diagnosis*. 2021;8(3):281–294. doi:10.1515/dx-2020-0119
3. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digit Med*. 2024;7(1):20. doi:10.1038/s41746-024-01010-1
4. Srinivasan S, Ai X, Zou M, et al. Ophthalmological question answering and reasoning using OpenAI o1 vs other large language models. *JAMA Ophthalmol*. 2025;143(9):740–748. doi:10.1001/jamaophthalmol.2025.2413
5. Yang X, Li T, Su Q, et al. Application of large language models in disease diagnosis and treatment. *Chin Med J*. 2025;138(2):130–142. doi:10.1097/CM9.0000000000003456

6. Kang D, Wu H, Yuan L, et al. Evaluating the efficacy of large language models in guiding treatment decisions for pediatric refractive error. *Ophthalmol Ther*. 2025;14(4):705–716. doi:10.1007/s40123-025-01105-2
7. Xu P, Wu Y, Jin K, Chen X, He M, Shi D. DeepSeek-R1 outperforms gemini 2.0 pro, OpenAI o1, and o3-mini in bilingual complex ophthalmology reasoning. *Adv Ophthalmol Pract Res*. 2025;5(3):189–195. doi:10.1016/j.aopr.2025.05.001
8. Han W, Wan C, Shan R, et al. Evaluation of error detection and treatment recommendations in nucleic acid test reports using ChatGPT models. *Clin Chem Lab Med*. 2025;63(9):1698–1708. doi:10.1515/cclm-2025-0089
9. Jin K, Yu T, Grzybowski A. Multimodal artificial intelligence in ophthalmology: applications, challenges, and future directions. *Surv Ophthalmol*. 2025;(25):S0039–6257. doi:10.1016/j.survophthal.2025.07.003
10. Yu E, Chu X, Zhang W, et al. Large language models in medicine: applications, challenges, and future directions. *Int J Med Sci*. 2025;22 (11):2792–2801. doi:10.7150/ijms.111780
11. Yang Y, Jin Q, Zhu Q, et al. Beyond Multiple-Choice accuracy: Real-World challenges of implementing large language models in healthcare. *Annu Rev Biomed Data Sci*. 2025;8(1):305–316. doi:10.1146/annurev-biodatasci-103123-094851
12. Yang H, Hu M, Most A, et al. Evaluating accuracy and reproducibility of large language model performance on critical care assessments in pharmacy education. *Front Artif Intell*. 2025;7:1514896. doi:10.3389/frai.2024.1514896
13. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172–180. doi:10.1038/s41586-023-06291-2
14. Banerjee S, Lavie A. Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments. *Proceedings of ACL-WMT*. 2007; 2004:65–72.
15. Ji J, Hou Y, Chen X, Pan Y, Xiang Y. Vision-Language model for generating textual descriptions from clinical images: model development and validation study. *JMIR Form Res*. 2024;8e32690. doi:10.2196/32690
16. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. *Med Educ*. 2024;58(11):1276–1285. doi:10.1111/medu.15402
17. 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
18. Mikhail D, Farah A, Milad J, et al. DeepSeek-R1 vs OpenAI o1 for ophthalmic diagnoses and management plans. *JAMA Ophthalmol*. 2025; e252918. doi:10.1001/jamaophthalmol.2025.2918
19. Hassanein FEA, El Barbary A, Hussein RR, et al. Diagnostic performance of ChatGPT-4o and DeepSeek-3 differential diagnosis of complex oral lesions: a multimodal imaging and case difficulty analysis. *Oral Dis*. 2025. doi:10.1111/odi.70007
20. Hassanein FEA, Ahmed Y, Maher S, Barbary AE, Abou-Bakr A. Prompt-dependent performance of multimodal AI model in oral diagnosis: a comprehensive analysis of accuracy, narrative quality, calibration, and latency versus human experts. *Sci Rep*. 2025;15(1):37932. doi:10.1038/s41598-025-22979-z
21. Abou-Bakr A, El Barbary A, Hassanein FEA. ChatGPT-5 vs oral medicine experts for rank-based differential diagnosis of oral lesions: a prospective, biopsy-validated comparison. *Odontology*. 2025. doi:10.1007/s10266-025-01242-x