Advances in Medical Education and Practice

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

A Comparative Analysis of AI-Language Models' MCQ Performance versus Medical Students Across Different Pediatric Topics

Olena Bolgova[1], Volodymyr Mavrych[1], Eyad Almidani[2], Turki Alshareef[2], Sabri Kemahlı[3]

[1]Department of Anatomy, College of Medicine, Alfaisal University, Riyadh, Saudi Arabia; [2]Department of Pediatrics, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia; [3]Department of Pediatrics, College of Medicine, Alfaisal University, Riyadh, Saudi Arabia

Correspondence: Sabri Kemahlı, Email skemahli@alfaisal.edu

Background: Large Language Models (LLMs) are increasingly used in medical education, yet their performance in specialized fields like pediatrics remains understudied.

Objective: To evaluate and compare the performance of four leading LLMs on a standard set of pediatric multiple-choice questions (MCQs) and benchmark their results against those of medical students.

Methods: We assessed 4 leading LLMs: Copilot (Microsoft), Claude (Anthropic), ChatGPT (OpenAI), and Gemini (Google), on 120 MCQs from six pediatric topics: Pulmonology, Developmental Diseases, Infectious Diseases, General Pediatrics, Neonatology, and Nephrology. Each LLM attempted the questions three times to evaluate consistency. Medical student performance on the same questions served as a benchmark, with random responses providing baseline control.

Results: The LLMs demonstrated an average accuracy of 80.4±10.9%, outperforming medical students (72.1±13.1%) by 8.3% and exceeding random responses (20.5±6.2%) by nearly fourfold. Copilot (84.5±8.3%) and Claude (84.2±9.9%) achieved the highest accuracy, followed by GPT-4o (80.9±9.1%) and Gemini (72.0±18.9%). Among all LLM-to-student comparisons, only GPT-4o showed a statistically significant difference in accuracy (χ^2 = 4.277, p = 0.039), outperforming students by 8.8%. LLMs excelled in Pulmonology (96.3%), Developmental Diseases (86.3%), and Infectious Diseases (85%) while struggling with Nephrology (67.5%) and Neonatology (69.7%). Only 50% (60/120) of questions were correctly answered by all LLMs across all attempts, while 6.7% (8/120) were never answered correctly.

Conclusion: Modern LLMs demonstrate proficiency in pediatric knowledge that generally exceeds medical student performance, though with varying consistency across topics. These findings suggest LLMs may serve as valuable supplementary tools in medical education while highlighting the need for further improvements in specialized medical domains like nephrology and neonatology.

Keywords: artificial intelligence, medical education, pediatrics, large language models, multiple choice questions

Introduction

There has been growing interest in the role of artificial intelligence (AI) and large language models (LLMs) in medical education. AI-driven LLMs can mimic human cognition, including data analysis, pattern recognition, and decision-making.[1] These properties have paved the way for their use in medicine and medical education, including curriculum design, simulation of patient interactions, and generation of case reports and assessments.[2]

LLMs are particularly promising due to their ability to process and generate texts, which can support a wide range of applications from clinical documentation to patient communication, as well as generating and answering examination questions.[3–5] Several studies investigating the performance of LLMs in various high-stakes examinations have demonstrated good performance.[1,6,7] While some of this research addressed the performance of LLMs in general medical knowledge examinations, such as USMLE, some have assessed their performance in special fields of medical education, such as anatomy, histology, physiology, pathology, radiology, and head and neck surgery.[8–15]

https://doi.org/10.2147/AMEP.S601573
Received: 5 February 2026
Accepted: 28 April 2026
Published: 16 May 2026

The transition toward computer-based testing has created unprecedented demand for large numbers of high-quality assessment items that traditional item development approaches struggle to fulfill.[5] While efficient for assessment, multiple-choice questions (MCQs) primarily test recognition rather than knowledge recall, which is a significant limitation in evaluating future clinicians.[16] Short-answer questions address this limitation but impose a substantial grading burden on educators. The potential of LLMs to support both the creation and evaluation of various assessment formats represents a significant opportunity for medical education.[17,18]

LLMs such as ChatGPT, Claude, Copilot, and Gemini have demonstrated capabilities that extend beyond simple question-answering to include clinical reasoning, interpretation of medical knowledge, and application of concepts across various medical disciplines.[14,19] Their performance on standardized examinations like the USMLE, NBME, and various international medical licensing examinations has drawn significant attention from educators and researchers.[1,16] Their implementation on educational tasks in particular settings creates both excitement and concerns, as these models have been trained in a large volume of literature and can provide responses that are comparable to those of professionals in the medical field.[6,20] With these rapid shifts in the world of education, implementing LLMs brings unique opportunities and responsibilities that will require tremendous attention. Even though these technologies are expected to minimize the burden on faculty and augment support for students, many consider how to use them and how much human supervision is necessary to safeguard the quality and integrity of education.

Pediatrics represents a particularly relevant context for this evaluation, as it encompasses a broad and diverse range of clinical domains (from neonatology to developmental medicine and infectious diseases), each requiring distinct knowledge bases, making it a rigorous and meaningful testbed for LLM performance in undergraduate medical education.

Despite growing evidence of LLM competence in general medical knowledge assessments, no study to date has systematically compared multiple LLM platforms on undergraduate pediatric MCQs with topic-level analysis and repeated attempts to assess consistency. This research, therefore, aimed to evaluate the performance of four large language models (Copilot, Claude, GPT-4o, and Gemini) on multiple-choice questions from the pediatrics curriculum, compare their accuracy with medical student performance, and assess topic-specific variations and inter-platform differences. Assessing each LLM across three independent attempts was designed to evaluate response consistency - a critical dimension of reliability for any tool considered for use in educational assessment or self-directed learning.

Materials And Methods

This research evaluated the performance of four publicly available large language models (LLMs) on their proficiency in different pediatric topics: GPT-4o (OpenAI), Claude 3.5 Sonnet (Anthropic), Gemini 1.5 Flash (Google), and Copilot (Microsoft). The study assessed their ability to answer USMLE-style 120 multiple-choice questions (MCQs) across six pediatric domains: development, general pediatrics, infectious diseases, neonatology, nephrology, and pulmonology, with 20 questions per domain. The study excluded questions containing images and tables. Three independent experts (two pediatric clinicians and one medical educator) validated all questions drawn from a pediatric examination database, reviewing each item for scientific accuracy, clinical relevance, and appropriateness for the undergraduate level. Questions were excluded if consensus was not reached. The student benchmark data were drawn from Year 4 MBBS students at Alfaisal University College of Medicine who had completed the 9-week pediatrics clerkship, representing a cohort of clinical-phase undergraduate medical students.

The data collection took place in January 2025. Each model received the identical prompt: “Generate the list of correct answers for the following MCQs:” followed by three sets of 40 questions each, totaling 120 MCQs per attempt. The models tested were: GPT-4o (OpenAI, version gpt-4o), Claude 3.5 Sonnet (Anthropic), Gemini 1.5 Flash (Google), and Copilot (Microsoft, powered by GPT-4). All models were accessed via their standard publicly available web interfaces. Default generation settings were used for all models; no custom temperature or parameter adjustments were applied. Each attempt was conducted in a new independent session to prevent any carry-over of context between attempts. This process was repeated three times per model, yielding 1,440 total responses. The results of 3 successive attempts by each chatbot to answer this questionnaire set were recorded in an Excel spreadsheet (Microsoft®365) and evaluated based on accuracy. Further analysis examined data of consistency and accuracy in all successive attempts, tracking the percentage of repeated and correct answers between trials. A detailed item analysis was conducted for each

LLM across different questions and topics, comparing their results with the students' results for the same questionnaire set. Three random answer sets were generated using Microsoft Excel's RAND() function and compared with LLMs' results.

Basic data statistics was performed using Statistica 13.5.0.17 (TIBC® Statistica™), with the Pearson chi-squared test employed to compare performance between chatbots and students using a significance threshold of $p \leq 0.05$.

This study involved a retrospective analysis of an existing examination database without direct interaction with students or the collection of personal information. The statistical data received from the assessment office was extracted from previous exams, averaged, and fully anonymized. Based on our institutional guidelines (Alfaisal University), this study was classified as educational quality improvement research using de-identified retrospective academic performance data, which falls outside the scope requiring formal ethics review.

Results

According to our data, the selected chatbots answered, on average, 80.4±10.9% of 120 MCQs across 6 topics in the pediatrics course. This result was 8.3% above the students' average (72.1±13.1%) and almost four times better than randomly generated responses (20.5±6.2%) for the same set of questions.

There was significant variation in the number of correct responses among the different LLMs. The best results were achieved by Copilot (84.5±8.3%) and Claude (84.2±9.9%), followed by GPT-4o (80.9±9.1%), all of which outperformed the students' performance. The lowest result was recorded for Gemini - 72.0±18.9% (Figure 1).

In the box plot analysis of AI system performance, Copilot (Z)'s perfect score of 100 (Figure 1) in the Pulmonology topic stands out as a statistical outlier, exceeding the upper bound of 92.5. This is particularly noteworthy because Copilot demonstrates quite consistent performance across topics, with most scores tightly clustered between 76.7–85.0% and a small interquartile range (IQR) of just 5.0. What makes this observation especially intriguing is that while the score of 100 is an outlier for Copilot, similar high scores in pulmonology are not statistical outliers for other AI systems. Claude, GPT-4, and Gemini all show much wider score distributions, with larger IQRs (18.4, 13.4, and 33.2, respectively). This suggests that the outlier status of Copilot's perfect score is not due to exceptional performance in

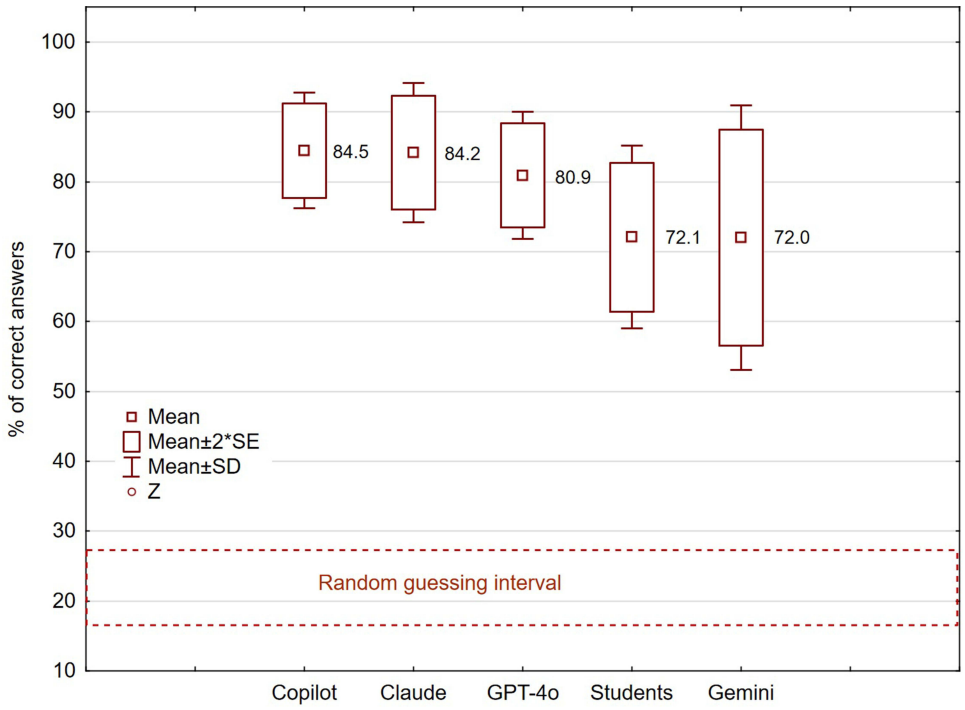

Figure 1 Percentile of correct answers from different chatbots and students on 120 MCQs from the pediatric course. Y-axis: % of correct answers; X-axis: different LLMs and Students' results.

https://doi.org/10.2147/AMEP.S601573

Topic	Students	Copilot	Claude	Copilot	GPT-4o	LLM Mean*
Development	79	85	93	87	80	86
General pediatrics	58	80	87	73	70	78
Infectious diseases	74	85	85	83	87	85
Neonatology	82	80	70	75	54	70
Nephrology	54	77	75	72	47	68
Pulmonology	85	100	95	95	95	96

** Mean across all 4 LLMs*

Figure 2 Heatmap of students' and LLMs' topic-wise performance (average % of correct answers) in pediatric courses.

pulmonology itself but rather reflects a departure from Copilot's characteristically consistent scoring pattern across other topics.

After that, a detailed evaluation of the results received from students and LLMs was performed (Figure 2).

Only 60 questions (50%) were answered correctly by all chatbots in all attempts. General item analysis revealed that pulmonary, Development, and infectious diseases were the three best categories, with the average results for all LLMs being 96.3%, 86.3%, and 85%, respectively. In contrast, the lowest results were recorded for nephrology and neonatology questions - 67.5% and 69.7%, respectively. Statistical analysis did not reveal significant differences between LLMs' performance in specific topics (all p-values > 0.05). The largest differences were observed in Nephrology: Copilot vs Gemini (χ^2 = 2.236, p = 0.1349) and Claude vs Gemini (χ^2 = 2.061, p = 0.1512). Pulmonology was performed consistently across all LLMs (lowest chi-square values).

Eight MCQs (6.7%) were never answered correctly by any chatbot. Item analysis revealed that 6 of them were high-level critical-thinking questions, and students struggled to answer them with index difficulty (ID) 0.02–0.44. The other 2 MCQs were recall questions from neonatology, which most students answered correctly, so LLMs were most likely not pre-trained on this specific material.

Copilot

Microsoft's Copilot can only accept up to 4000 characters in the prompt, so only 20 MCQs can be answered at a time, which makes it inconvenient to work with. However, the big advantage is that Copilot is now integrated into MS Windows and MS Office and is always available. On average, Copilot generated 84.5% accurate answers for 120 MCQs from the pediatrics course, showing the best result. It is 12.4% better than medical students, but the difference was statistically insignificant (χ^2 = 0.1671, p = 0.683). The last of three attempts was the most successful, with 85% correct answers; the two previous attempts had 84.2%. The coincidence generated by Copilot answers with the earlier attempts was 99.2% - 100%; among them, the coincidence of correct answers was 84.2%, so the standard deviation (SD) was only 8.3%.

Copilot answered 83.3% of MCQs (100/120) correctly in all three attempts consistently, reflecting high reliability. Conversely, 15% of questions (18/120) were not answered correctly in any of the three attempts, representing Copilot's knowledge gaps. These questions are high-order thinking questions, equally distributed through all topics. The item analysis revealed that Copilot performed well in all areas - 76.7% - 85%, with the best result in Pulmonology - 100% answers.

Claude

Claude, offered by Anthropic, provided 84.2% correct answers to the same 120 pediatric MCQs. Considering a statistical Copilot's outliers, the result was equally good, or probably even better.

It was 12.1% better than students, but the difference is still statistically insignificant (χ^2 = 0.137, p = 0.711). The first out of three attempts was the most successful, with 85% correct answers; the following were 83.3% and 84.2%. The coincidence generated by Claude's answers with the previous attempts was 96.7% - 98.3%, and among them, the coincidence of correct answers was 82.5% - 84.2%, with SD - 9.9%. Claude answered correctly 81.7% (98/120) across

all attempts and did not solve only 15% MCQs (18/120), almost identical to Copilot results. These were comprehensive questions from neonatology and nephrology topics. The item analysis suggested that Claude correctly answered 93.5% and 95% of development and pulmonology topics, 85% and 86.7% of infectious diseases and general pediatrics questions, respectively.

GPT-4o

The results of three successive ChatGPT-4 (Open AI) attempts to answer the 120 pediatrics MCQs showed 80.9 ±9.1% correct answers on average, which was 8.8% ($p<0.05$) better than students' results. Interestingly, that was the only statistically significant difference between students and LLM performance results ($\chi^2 = 4.2773$, p = 0.039). The last of three attempts was the most successful, with 84.2% correct answers; the previous two were 83.3% and 75%, and SD was 9.1%. The coincidence generated by GPT-4o's answers with the earlier attempts was 79.2% - 94.2%, and among them, the coincidence of correct answers was 70% - 80.8%, so the reliability was not as good as in the two previous chatbots. Only 68.3% (82/120) of questions were answered correctly across all attempts, indicating a solid knowledge area for GPT-4o. Most of these MCQs were recall questions, but some were complex and required critical thinking. GPT-4o did not answer only 10.8% MCQs (13/120) from the entire questionnaire set in any one out of 3 attempts, which is the best result in this regard. The item analysis indicated that pulmonology was the best topic, with 95% of responses correct. For the rest of the topics, the percentile of correct answers was in the interval 71.7% - 86.8%.

Gemini

Google's Gemini finished last with 72% correct answers to the same set of questions. This result is slightly below the students' performance, but the statistical difference is insignificant. The first out of three attempts was the most successful, with 75.8% correct answers; the following were 75% and 65%. The coincidence generated by Gemini's answers with the previous attempts was 73.3% - 99.2%, and among them, the coincidence of correct answers was 60% - 75%, with SD - 18.9%. Gemini correctly answered 52% (71/120) across all attempts and did not solve 20% MCQs (24/120), the worst result among tested chatbots. Item performance analysis revealed Gemini's notable strength in pulmonology with a 95% accuracy rate, in contrast to its considerably weaker performance in nephrology and neonatology, where it achieved only 46.7% and 53.3% correct answers, respectively.

Difference in LLMs Performance

Due to the binary nature of the data, we employed the Pearson Chi-square test to compare the performance of the different AI-driven chatbots (Table 1).

Table 1 Results of Pearson Chi-Square Test to Compare the Performance of Copilot, Claude, GPT-4o, and Gemini Against Each Other

LLMs	Chi-Square	df	p
Copilot vs. Claude	0.6098	1	0.435
Copilot vs GPT-4o	2.7972	1	0.094
Copilot vs Gemini	0.1671	1	0.683
Claude vs. GPT-4o	5.9469	1	0.015*
Claude vs Gemini	0.137	1	0.711
GPT-4o vs Gemini	4.2773	1	0.039*

Notes: * - Statistically significant difference.

https://doi.org/10.2147/AMEP.S601573

The results suggest that Claude significantly outperformed GPT-4o ($\chi^2 = 5.947$, $p = 0.015$), and GPT-4o significantly outperformed Gemini ($\chi^2 = 4.277$, $p = 0.039$). No other pairwise comparison between LLMs reached statistical significance.

Discussion

Alfaisal University College of Medicine has a 6-year MBBS program, the first 3 years as the pre-clerkship phase, years 4 and 5 as the clerkship phase, and year 6 as the 12-month rotating internship phase. Pediatrics is a 9-week clerkship in year 4, along with internal medicine, surgery, obstetrics, and gynecology. Each of these clinical clerkships is conducted at three hospitals.

King Faisal Specialist Hospital and Research Center (KFSHRC) is the main teaching hospital of Alfaisal University and is a tertiary care hospital. The students spend 4 weeks at KFSHRC and 4 weeks at one of the secondary care hospitals (King Saud Medical City or King Fahad Medical City). The assessment is made through MCQs and an objective structured clinical examination (OSCE) at the end of the clerkship, in addition to continuous assessment scores derived from clinical activities. Students' performance in MCQs may be affected by several factors, such as exam readiness and the quality and clarity of the questions. An exam committee reviews all questions for question quality, scientific content, and appropriateness for the students' level.

LLMs demonstrated good competence in pediatric knowledge assessment. On average, the four evaluated chatbots outperformed medical students by 8.3%. This suggests that modern LLMs possess substantial medical knowledge that may complement traditional educational approaches.

Performance varies across LLM platforms. Copilot and Claude achieved the highest accuracy rates, followed by GPT-4o and Gemini. While most differences between platforms were not statistically significant, Claude significantly outperformed GPT-4o ($p=0.015$), and GPT-4o significantly outperformed Gemini ($p=0.039$).

Topic-specific strengths and weaknesses are evident. All LLMs demonstrated exceptional performance in pulmonology, developmental pediatrics, and infectious diseases. However, they consistently struggled with nephrology and neonatology questions, suggesting gaps in specialized knowledge in these domains.

LLMs show varying levels of consistency. Copilot demonstrated remarkable consistency across attempts, while Gemini showed greater variability. This suggests that reliability differs significantly between models, with potential implications for their educational use.

Critical thinking questions remain challenging. Among the 8 MCQs (6.7%) that no chatbot answered correctly, 6 were identified as high-level critical-thinking questions. This indicates that, despite their impressive performance, LLMs still struggle with complex clinical reasoning tasks requiring nuanced judgment.

Our findings regarding the performance of different LLMs on pediatric MCQs align with several previous studies and also reveal important nuances. The average accuracy of 80.4±10.9% achieved by the four evaluated chatbots compares favorably with a comprehensive analysis of popular LLMs on NBME clinical subject exam questions, in which GPT-4 achieved 100% accuracy, Claude scored 84.7%, and Bard (predecessor to Gemini) reached 75.5%.[20] Similarly, a systematic review and meta-analysis found that GPT-4 achieved an overall accuracy rate of 81% across various medical licensing examinations worldwide, while GPT-3.5 reached 58%.[6] The performance hierarchy observed in our study - with Copilot and Claude leading, followed by GPT-4o and Gemini - differs somewhat from these previous studies, particularly in Copilot's strong performance, suggesting possible improvements in Microsoft's platform or specific strengths in pediatric knowledge.

The topic-specific performance variations identified in our research echo findings from a study comparing GPT-4.1, Claude 3.7 Sonnet, Gemini 2.0 Flash, Copilot, and DeepSeek R1 on medical histology topics, which found that Histological Methods, Blood and Hemopoiesis, and Circulatory System achieved complete accuracy, while Muscle tissue and Lymphoid System presented the greatest challenges.[9] Our observation that LLMs performed best in pediatric pulmonology, development, and infectious diseases, while struggling in nephrology and neonatology, reinforces the conclusion that LLMs demonstrate uneven mastery across medical specialties.

This pattern is further supported by an evaluation of LLMs in radiology board exams, which found significant performance variations across different radiological subspecialties,[14] and by an assessment of multiple LLMs in gross

anatomy, which revealed stronger performance in MCQs of Head & Neck and Abdomen material compared to questions of Upper Limb, where only 29.5% were answered correctly by all models.[8] Similarly, a comparative study in medical biochemistry demonstrated that LLMs achieved the highest accuracy for eicosanoids, bioenergetics, and the electron transport chain, and ketone bodies, with notable performance differences across platforms.[19]

The consistency analysis in our study revealed important differences between LLM platforms, with Copilot demonstrating remarkable consistency across attempts, whereas Gemini showed greater variability. This parallels findings in pathology, where ChatGPT showed 80–85% consistency across three tests while Bard exhibited lower consistency rates of 54–61%.[11] The challenge of critical thinking questions identified in our study - with 6 of the 8 MCQs that no chatbot answered correctly being high-level critical-thinking questions - aligns with an assessment of ChatGPT in solving questions based on core concepts in physiology, which found significant performance differences among various core concepts that required different levels of reasoning.[10] Similarly, an evaluation of ChatGPT in head and neck surgery found that while the AI provided fully or nearly fully correct diagnoses in 81.7% of clinical scenarios, its proposed diagnostic or therapeutic procedures were judged complete in only 56.7% of cases, indicating challenges with complex clinical reasoning.[15]

Our finding that LLMs generally outperformed medical students contributes to an ongoing discussion about the comparative performance of AI systems and medical students. An evaluation of ChatGPT on the USMLE found the model performed at or near the passing threshold without specialized training,[1] while a study of GPT-3.5 and GPT-4 on the Polish Medical Final Examination found that while GPT-4 passed all exam versions with a mean accuracy of 79.7%, its score was generally lower than that of the average medical student.[7] This contrasts with an assessment of ChatGPT and GPT-4 on USMLE soft-skill questions, in which GPT-4 outperformed historical user data from the AMBOSS question bank.[4] In the specialized domain of ophthalmology, a study found that ChatGPT achieved significantly lower diagnostic accuracy than residents and attendings, highlighting the current limitations of LLMs in specialized clinical reasoning.[12] These varied results suggest that LLM performance relative to human medical trainees may depend on the specific domain, question type, and level of reasoning required.

Implications of Findings

Our study evaluating the performance of LLMs on pediatric MCQ tests reveals both promising capabilities and important limitations. These AI systems demonstrated strong overall performance, generally exceeding medical student averages, with notable strengths in areas such as pulmonology and development, while struggling with topics such as nephrology and neonatology. Different platforms showed varying levels of consistency and reliability, with Copilot and Claude achieving the highest accuracy rates. Across all models, there was a lack of success with high-level critical-thinking questions, indicating that while LLMs have considerable medical knowledge, they struggle with more sophisticated clinical reasoning. These results indicate that current LLMs can serve as valuable supplementary tools in medical education but require thoughtful implementation with appropriate human oversight, particularly for advanced reasoning tasks and specialized knowledge domains. The evolution of these technologies will certainly expand their role in medical education. Careful consideration would have to be given to the best way to integrate educational tools and their limitations within educational frameworks.

Strengths and Limitations

This study assesses the performance of leading LLMs across a range of pediatric topics to evaluate their proficiency with specialized medical knowledge. The inclusion of repeated attempts for each LLM establishes reliability metrics and consistency patterns that would not be apparent from single attempt testing. Benchmarking against both medical student performance and random responses creates a meaningful context for interpreting the results. The topic-specific performance focus reveals strong and weak areas across a multitude of pediatric domains that educators and developers can target to address knowledge gaps. Another important dimension that this study highlights is the detailed analysis of questions which no one LLM was able to answer meaningfully. With this, the boundaries of LLMs' capability for high-level critical thinking are examined, providing guidance for future LLM development and highlighting the gaps that need to be bridged.

Several limitations should be taken into consideration. The study employed a pre-set bank of 120 MCQs, which may not capture the full depth of pediatric knowledge. Additionally, this study used a single fixed prompt for all LLMs and did not account for known LLM sensitivities to prompt variation; future studies should explore how prompt design influences accuracy and consistency. The potential for LLM hallucination was also not assessed and remains an important area for future investigation. The evaluation was limited to verifying the answers and did not consider how well the LLMs could explain them, which can be useful in teaching situations. There are also some operational limitations, for example, Copilot's limit on the number of characters that can be used per question, which might have affected the way the questions were asked to the different LLMs. The Pearson chi-squared test was employed to compare binary (correct/incorrect) response rates between chatbots and students, using a significance threshold of $p \leq 0.05$. Given the number of pairwise comparisons performed, results should be interpreted with appropriate caution regarding Type I error. Lastly, these findings are the result of a cross-sectional study and therefore do not reflect the performance of newer model versions or of models post-release.

Conclusions

This study demonstrates that large language models show promising capabilities in pediatric education, with the evaluated chatbots outperforming medical students on average across MCQs. While Copilot and Claude achieved thAlme highest accuracy rates with remarkable consistency across attempts, all platforms exhibited clear topic-specific strengths and weaknesses. The LLMs excelled in pulmonology, development, and infectious diseases, yet consistently struggled with questions in nephrology and neonatology. The differences in performance across various platforms and topics revealed by this analysis suggest that careful selection of models and consideration of particular areas of strength and weakness, or the harnessing of these technologies in medical education, are essential. Based on MCQ accuracy and response consistency alone, current LLMs show promise as supplementary study tools in undergraduate pediatric education. However, this study did not assess clinical reasoning ability or performance on applied clinical tasks, so conclusions about broader educational integration should therefore be drawn cautiously and await further evidence.

Clinical Trial Number

The clinical trial number is not pertinent to this study as it does not involve medicinal products or therapeutic interventions.

Data Sharing Statement

The data supporting this study's findings are available on request from the corresponding author.

Ethics Declarations

Ethics approval, Consent to Participate, and Consent to Publish declarations: not applicable.

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 authors received no funding for this study.

Disclosure

The authors declare no conflicts of interest, financial or otherwise.

References

1. 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. doi:10.1371/journal.pdig.0000198
2. 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
3. Artsi Y, Sorin V, Konen E, Glicksberg BS, Nadkarni G, Klang E. Large language models for generating medical examinations: systematic review. *BMC Med Educ*. 2024;24(1):354. doi:10.1186/s12909-024-05239-y
4. Brin D, Sorin V, Vaid A, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Sci Rep*. 2023;13(1):1–5. doi:10.1038/s41598-023-43436-9
5. Gierl MJ, Lai H, Turner SR. Using automatic item generation to create multiple-choice test items. *Med Educ*. 2012;46(8):757–765. doi:10.1111/j.1365-2923.2012.04289.x
6. Liu M, Okuhara T, Chang X, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. *J Med Internet Res*. 2024;26:e60807. doi:10.2196/60807
7. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the polish medical final examination. *Sci Rep*. 2023;13(1):1–13. doi:10.1038/s41598-023-46995-z
8. Bolgova O, Mavrych V. Evolution of AI in anatomy education study based on comparison of current large language models against historical ChatGPT performance. *Sci Rep*. 2025;15(1):37545. PMID: 41152541; PMCID: PMC12569271. doi:10.1038/s41598-025-22437-w
9. Mavrych V, Yousef EM, Yaqinuddin A, Bolgova O. Large language models in medical education: a comparative cross-platform evaluation in answering histological questions. *Med Educ Online*. 2025;30(1):2534065. PMID: 40651009; PMCID: PMC12258195. doi:10.1080/10872981.2025.2534065
10. Banerjee A, Ahmad A, Bhalla P, Goyal K. Assessing the efficacy of ChatGPT in solving questions based on the core concepts in physiology. *Cureus*. 2023;15(8):e43314. PMID: 37700949; PMCID: PMC10492920. doi:10.7759/cureus.43314
11. Du W, Jin X, Harris JC, et al. Large language models in pathology: a comparative study of ChatGPT and Bard with pathology trainees on multiple-choice questions. *Ann Diagn Pathol*. 2024;73:152392. doi:10.1016/j.anndiagpath.2024.152392
12. Shemer A, Cohen M, Altarescu A, et al. Diagnostic capabilities of ChatGPT in ophthalmology. *Graefes Arch Clin Exp Ophthalmol*. 2024;262(7):2345–2352. doi:10.1007/s00417-023-06363-z
13. Totlis T, Natsis K, Filos D, et al. The potential role of ChatGPT and artificial intelligence in anatomy education: a conversation with ChatGPT. *Surg Radiol Anat*. 2023;45(10):1321–1329. doi:10.1007/s00276-023-03229-1
14. Wei B. Performance evaluation and implications of large language models in radiology board exams: prospective comparative analysis. *JMIR Med Educ*. 2025;11:e64284. doi:10.2196/64284
15. Vaira LA, Lechien JR, Abbate V, et al. Accuracy of ChatGPT-generated information on head and neck and oromaxillofacial surgery: a multicenter collaborative analysis. *Otolaryngol Head Neck Surg*. 2024;170(6):1492–1503. doi:10.1002/ohn.489
16. Grévisse C. LLM-based automatic short answer grading in undergraduate medical education. *BMC Med Educ*. 2024;24(1:1060. doi:10.1186/s12909-024-06026-5
17. Kıyak YS, Budakoğlu Iİ, Coşkun Ö, Koyun E. The first automatic item generation in Turkish for assessment of clinical reasoning in medical education. *Tip Egitimi Dünyasi*. 2023;22(66):72–90. doi:10.25282/ted.1225814
18. Kıyak YS, Coşkun Ö, Budakoğlu Iİ, Uluoğlu C. Psychometric Analysis of the First Turkish Multiple-Choice Questions Generated Using Automatic Item Generation Method in Medical Education. *Tip Egitimi Dünyasi*. 2023;22(68):154–161. doi:10.25282/ted.1376840
19. Bolgova O, Shypilova I, Mavrych V. Large language models in biochemistry education: comparative evaluation of performance. *JMIR Med Educ*. 2025;11:e67244. PMID: 40209205; PMCID: PMC12005600. doi:10.2196/67244
20. Abbas A, Rehman MS, Rehman SS. Comparing the Performance of popular large language models on the national board of medical examiners sample questions. *Cureus*. 2024;16(3):e55991. doi:10.7759/cureus.55991