

Exceptional Performance of DeepSeek on Pediatric Board Examination Preparation Questions

Masab Mansoor, Andrew Ibrahim, Ali Hamide

Submitted to: JMIR AI
on: April 15, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript.....	5
---------------------------------	----------

Preprint
JMIR Publications

Exceptional Performance of DeepSeek on Pediatric Board Examination Preparation Questions

Masab Mansoor¹ DBA, MBA, BS; Andrew Ibrahim² BS; Ali Hamide¹ MS, BS

¹ Louisiana Campus Edward Via College of Osteopathic Medicine monroe US

² School of Medicine Texas Tech University Health Sciences Center Lubbock US

Corresponding Author:

Masab Mansoor DBA, MBA, BS

Louisiana Campus
Edward Via College of Osteopathic Medicine
4408 bon aire dr
monroe
US

Abstract

Background: The integration of artificial intelligence in medical education raises questions about large language models' (LLMs) capabilities in specialized medical knowledge domains. Limited research exists evaluating AI performance on standardized pediatric assessments.

Objective: To evaluate and compare the performance of three leading LLMs on pediatric board examination preparation questions and contextualize their performance against human physician benchmarks.

Methods: We conducted a comparative analysis of DeepSeek 7B v2.5, ChatGPT-4, and ChatGPT-4.5 using 266 multiple-choice questions from the 2023 PREP® Self-Assessment (American Academy of Pediatrics). Each model was presented with identical questions covering the full spectrum of pediatric knowledge domains. Performance was measured by calculating the percentage of correct responses and compared to published first-time pass rates for the American Board of Pediatrics (ABP) examination.

Results: DeepSeek exhibited the highest accuracy at 98.12% (261/266 correct responses), exceeding typical human performance metrics. ChatGPT-4.5 achieved 96.6% accuracy (257/266), performing at the upper threshold of human performance. ChatGPT-4 demonstrated 82.7% accuracy (220/266), comparable to the lower range of human pass rates. Error pattern analysis revealed that AI models most commonly struggled with questions requiring integration of complex clinical presentations with rare disease knowledge.

Conclusions: Recent advancements in large language models have produced AI systems capable of performing at or above the level of board-certified pediatricians on standardized examination questions. These findings suggest potential applications in medical education, board examination preparation, and possibly clinical decision support. Further research should evaluate these AI systems on more complex clinical reasoning tasks and in simulated clinical scenarios.

(JMIR Preprints 15/04/2025:76056)

DOI: <https://doi.org/10.2196/preprints.76056>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <http://www.jmir.org>

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <https://www.jmir.org>



Original Manuscript

Exceptional Performance of DeepSeek on Pediatric Board Examination Preparation Questions

Masab Mansoor^a, DBA; Andrew F. Ibrahim^b, BS; Ali Hamide^a, MS

Affiliations: ^aEdward Via College of Osteopathic Medicine, Monroe, Louisiana; and ^bTexas Tech University Health Science Center School of Medicine, Lubbock, Texas

Address correspondence to: Masab Mansoor, Edward Via College of Osteopathic Medicine, Monroe, Louisiana [mmansoor@vcom.edu]

Short title: AI Model Performance on Pediatric Board Prep Questions

Conflict of Interest Disclosures (includes financial disclosures): The authors have no conflicts of interest to disclose.

Funding/Support: No funding was secured for this study.

Abbreviations: ABP, American Board of Pediatrics; LLM, Large Language Model; AI, Artificial Intelligence

Word Count: 904

Contributors Statement Page

Masab Mansoor, Andrew Ibrahim, and Ali Hamide conceptualized and designed the study, drafted the initial manuscript, critically reviewed and revised the manuscript, designed the data collection instruments, collected data, carried out the initial analyses, and critically reviewed and revised the manuscript.

All authors approved the final manuscript as submitted and agree to be accountable for all aspects of the work.

ABSTRACT

BACKGROUND

The integration of artificial intelligence in medical education raises questions about large language models' (LLMs) capabilities in specialized medical knowledge domains. Limited research exists evaluating AI performance on standardized pediatric assessments.

OBJECTIVE

To evaluate and compare the performance of three leading LLMs on pediatric board examination preparation questions and contextualize their performance against human physician benchmarks.

METHODS

We conducted a comparative analysis of DeepSeek 7B v2.5, ChatGPT-4, and ChatGPT-4.5 using 266 multiple-choice questions from the 2023 PREP® Self-Assessment (American Academy of Pediatrics). Each model was presented with identical questions covering the full spectrum of pediatric knowledge domains. Performance was measured by calculating the percentage of correct responses and compared to published first-time pass rates for the American Board of Pediatrics (ABP) examination.

RESULTS

DeepSeek exhibited the highest accuracy at 98.12% (261/266 correct responses), exceeding typical human performance metrics. ChatGPT-4.5 achieved 96.6% accuracy (257/266), performing at the upper threshold of human performance. ChatGPT-4 demonstrated 82.7% accuracy (220/266), comparable to the lower range of human pass rates. Error pattern analysis revealed that AI models most commonly struggled with questions requiring integration of complex clinical presentations with

rare disease knowledge.

CONCLUSION

Recent advancements in large language models have produced AI systems capable of performing at or above the level of board-certified pediatricians on standardized examination questions. These findings suggest potential applications in medical education, board examination preparation, and possibly clinical decision support. Further research should evaluate these AI systems on more complex clinical reasoning tasks and in simulated clinical scenarios.

KEYWORDS

Artificial intelligence; large language models; medical education; pediatrics; board examination; DeepSeek; ChatGPT

INTRODUCTION

The integration of artificial intelligence in medical education and assessment raises important questions about the capabilities of large language models (LLMs) in understanding and applying pediatric knowledge. Recent advancements in AI have produced models with increasingly advanced medical reasoning capabilities^{1,2}, but limited research exists evaluating AI performance on standardized medical assessments. This study evaluates the performance of three leading LLMs (DeepSeek 7B v2.5 [DeepSeek-AI, 2024], ChatGPT-4 [OpenAI, 2023], and ChatGPT-4.5 [OpenAI, 2024]) on a set of 2023 pediatric board examination preparation questions (2023 PREP® Self-Assessment, American Academy of Pediatrics), a comprehensive resource containing case-based multiple-choice questions designed to simulate actual board examinations³. We hypothesized that newer AI models would demonstrate improved accuracy on pediatric knowledge assessment, potentially approaching the performance levels of board-certified pediatricians taking certification examinations.

METHODS

We conducted a comparative analysis of three advanced large language models (DeepSeek, ChatGPT-4, and ChatGPT-4.5) using a set of 266 questions from the 2023 PREP® Self-Assessment, American Academy of Pediatrics (AAP). In compliance with fair use copyright law and with methods deemed exempt by the Healthy Steps Pediatrics Ethics Committee, we entered 266 questions and answer choices from the Pediatrics 2023 PREP® Self-Assessment into the three LLM platforms. DeepSeek 7B v2.5 [DeepSeek-AI, 2024], ChatGPT-4 [OpenAI, 2023], and ChatGPT-4.5 [OpenAI, 2024] were accessed through their respective web interfaces in February 2025. The 2023 version of PREP® Self-Assessment was chosen as it represents the most current board preparation material available at the time of the study and contains questions designed to reflect the content and format of the actual American Board of Pediatrics (ABP) examination. These questions covered the

full spectrum of pediatric knowledge domains including general pediatrics, neonatology, critical care, infectious diseases, and all pediatric subspecialties. Each AI model was presented with identical questions in their original multiple-choice format. Responses were collected and evaluated against the established correct answers. Performance was measured by calculating the percentage of correct responses for each model.

To contextualize these findings, we compared the AI models' performance to published data on first-time pass rates for board-certified pediatricians taking the American Board of Pediatrics (ABP) examination. This comparison provides a benchmark for evaluating the clinical relevance of AI performance in pediatric knowledge assessment. The Healthy Steps Pediatrics Ethics Committee is an institutional committee that evaluates research proposals within our affiliated private practice network. This committee consists of three board-certified pediatricians who review research for ethical considerations. The committee determined this study was exempt from formal IRB approval as it involved publicly available AI tools and did not include human subjects or protected health information.

RESULTS

Performance across the three AI models demonstrated marked differences when tested on 266 pediatric board examination preparation questions. DeepSeek exhibited the highest accuracy at 98.12% (261/266 correct responses), outperforming both ChatGPT models (Table 1). ChatGPT-4 achieved an accuracy of 82.7% (220/266 correct responses), while ChatGPT-4.5 showed improvement over its predecessor with approximately 96.6% accuracy (missing only 9 questions).

Error pattern analysis revealed that AI models most commonly struggled with questions requiring integration of complex clinical presentations with rare disease knowledge. For example, DeepSeek's five incorrect answers primarily involved metabolic disorders and rare genetic syndromes, particularly questions requiring correlation between subtle biochemical abnormalities and uncommon clinical presentations. ChatGPT models additionally struggled with complex medication dosing calculations and interpretation of pediatric growth parameters in the context of genetic

disorders. Notably, there was minimal overlap in the specific questions missed by each model, suggesting that different LLMs have distinct knowledge gaps despite similar training paradigms.

These results were compared to the published first-time pass rates for board-certified pediatricians taking the ABP examination, which typically range from 80-89% (80% in 2022 and 89% in 2024 for general pediatrics)⁴. As illustrated in Figure 1, DeepSeek's performance exceeded the typical range for human pediatricians on first-attempt board examinations, while ChatGPT-4.5 also performed at the upper threshold of human performance. ChatGPT-4's performance was comparable to the lower range of human pass rates.

DISCUSSION

Our findings demonstrate that recent advancements in large language models have produced AI systems capable of performing at or above the level of board-certified pediatricians on standardized examination questions. DeepSeek's exceptional performance (98.12% accuracy) represents a significant milestone in AI medical knowledge representation, exceeding typical human performance metrics on similar assessments. The substantial performance gap between AI models highlights the rapid evolution of these technologies, with newer iterations showing marked improvements.

These results have important implications for medical education, board examination preparation, and potentially clinical decision support. AI models could serve as supplementary educational tools for pediatric trainees, offering accurate content knowledge while human educators focus on clinical reasoning, ethics, and patient communication skills that remain challenging for AI systems^{5,6}.

Limitations of this study include the use of multiple-choice questions rather than freeresponse clinical scenarios and the focus on knowledge recall rather than practical clinical decision-making. We cannot determine whether the AI models' performance reflects true clinical reasoning or pattern

recognition based on similar questions in their training data. Additionally, while PREP® Self-Assessment questions are designed to simulate board examinations, they may differ in difficulty and content distribution from actual ABP examinations, complicating direct comparisons with human pass rates. Furthermore, these models have not been tested on their ability to gather historical information, perform physical examinations, or develop appropriate management plans in real clinical settings ⁷. Future research should evaluate these AI systems on more complex clinical reasoning tasks and directly compare their performance to practicing pediatricians in simulated clinical scenarios.

References

1. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. *Nature*. 2023;616(7956):259-265. doi:10.1038/s41586-023-058814
2. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med*. 2022;28(1):31-38. doi:10.1038/s41591-021-01614-0
3. 2023 PREP Self-Assessment | AAP. American Academy of Pediatrics Accessed April 3, 2025. <https://www.aap.org/en/catalog/categories/maintenance-of-certification/2023-prep-self-assessment/>
4. Exam Pass Rates | The American Board of Pediatrics. Accessed April 3, 2025. <https://www.abp.org/content/exam-pass-rates>
5. Ramgopal S, Sanchez-Pinto LN, Horvat CM, Carroll MS, Luo Y, Florin TA. Artificial intelligence-based clinical decision support in pediatrics. *Pediatr Res*. 2023;93(2):334341. doi:10.1038/s41390-022-02226-1
6. Shah N, Arshad A, Mazer MB, Carroll CL, Shein SL, Remy KE. The use of machine learning and artificial intelligence within pediatric critical care. *Pediatr Res*. 2023;93(2):405-412. doi:10.1038/s41390-022-02380-6
7. Booven DV, Cheng-Bang C, Meenakshy M. Chapter 8 - Limitations of artificial intelligence in healthcare. In: Arora H, ed. *Artificial Intelligence in Urologic Malignancies*. Academic Press; 2025:231-246. doi:10.1016/B978-0-443-15504-8.00008-9

Table 1: Performance of AI Models on Pediatric Board Examination Questions

AI Model	Correct Answers	Accuracy (%)	Comparison to ABP Pass Rates*
Deepseek R1	261	98.12	Exceeds typical human performance
ChatGPT-4.5	257	96.6	Upper threshold of human performance
ChatGPT-4	220	82.7	Comparable to lower range of human performance

Table 1: Performance of Large Language Models on 2023 Pediatric Board Examination Preparation Questions. Each model was tested on 266 multiple-choice questions from the American Academy of Pediatrics 2023 PREP® Self-Assessment. Accuracy was calculated as the percentage of correct responses. Performance is contextualized relative to the typical first-time pass rates (80–89%) for board-certified pediatricians on the American Board of Pediatrics (ABP) examination.

*ABP first-time pass rates for board-certified pediatricians typically range from 80-89% (80% in 2022 and 89% in 2024 for general pediatrics)⁵

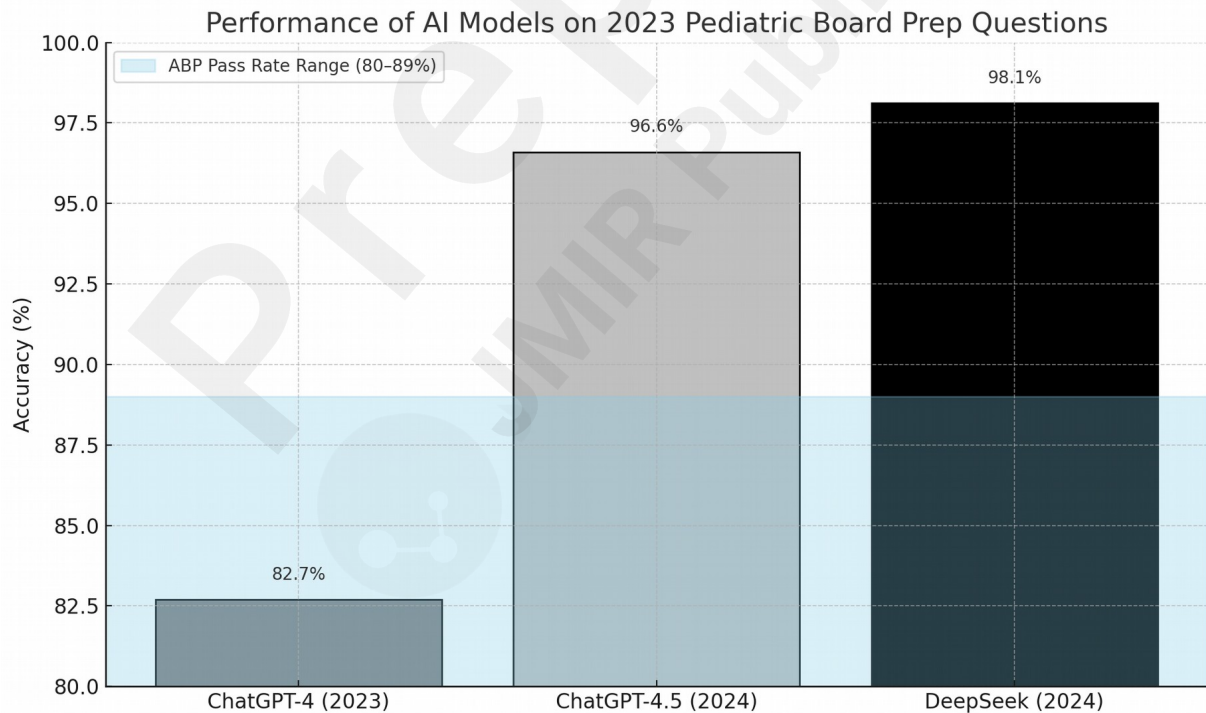


Figure 1: Accuracy of large language models (LLMs) on pediatric board examination preparation questions from the 2023 PREP® Self-Assessment. ChatGPT-4, ChatGPT-4.5, and DeepSeek were each tested on 266 multiple-choice questions. The shaded area represents the typical first-time pass rate range (80–89%) for board-certified pediatricians on the American Board of Pediatrics (ABP) examination from 2022 to 2024. DeepSeek achieved the highest performance at 98.1%, exceeding typical human pass rates.