Scientific Reports
https://doi.org/10.1038/s41598-026-44333-7
Article in Press

Comparative performance of recent and prior large language models and pediatric residents on pediatric in-training examination questions

Received: 1 September 2025

Accepted: 11 March 2026

Published online: 02 April 2026

Cite this article as: Kim M.J., Park J.S. & Kang S.H. Comparative performance of recent and prior large language models and pediatric residents on pediatric in-training examination questions. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-44333-7

Mi Jin Kim, Jun Sung Park & Sung Han Kang

Comparative performance of recent and prior large language models and pediatric residents on pediatric in-training examination questions

Mi Jin Kim[1], Jun Sung Park[2,*,†], Sung Han Kang[3,*,†]

[1]Division of Pediatric Cardiology, Department of Pediatrics, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea.

[2]Department of Pediatrics, Asan Medical Center Children's Hospital, University of Ulsan College of Medicine, Seoul, Korea

[3]Division of Pediatric Hematology/Oncology, Department of Pediatrics, Asan Medical Center Children's Hospital, University of Ulsan College of Medicine, Seoul, Korea

[†]Both authors contributed equally to the work

***Corresponding authors:**

Jun Sung Park, MD
Department of Pediatrics, Asan Medical Center,
University of Ulsan College of Medicine, Seoul, Republic of Korea
Tel: 82-2-3010-1497, Fax: 82-2-473-3725, E-mail: mrpiglet@naver.com

Sung Han Kang, MD
Division of Pediatric Hematology/Oncology, Department of Pediatrics, Asan Medical Center Children's Hospital, University of Ulsan College of Medicine, Seoul, Korea
Tel: 82-2-3010-1453, Fax: 82-2-473-3725, E-mail: sunghanie@gmail.com

Abstract

Recent vision-enabled multimodal large language models (LLMs) have achieved strong performance on high-stakes medical examinations, yet their capabilities in pediatrics, particularly for image-based questions, remain underexplored. We analyzed 498 unique questions with Korean–English terminologies, taken from 10 pediatric in-training examinations (ITEs) conducted by a single pediatric department between 2016 and 2023. Approximately 22% of items contained medical images. Three recent publicly accessible LLMs (GPT-4.1, Gemini-2.5-Pro, Claude-4.1-Opus) and three prior models (GPT-4o, Gemini-1.5-Pro, Claude-3.5-Sonnet) were tested. Recent LLMs significantly outperformed fourth-year residents (R4) (77.7–78.9% vs. 70.1%, all $P < 0.008$), while prior models showed comparable results. For text-only items, three recent LLMs achieved a superior proportion correct (PC) compared with R4 (80.1–81.0% vs. 69.6%). None of the evaluated LLMs surpassed R4 performance on image-included questions; both prior and recent models consistently exhibited inferior PC on image-included items than text-only questions. Outputs demonstrated high repeatability (intraclass correlation coefficient >0.98) across most models. In this study, multimodal LLMs achieved high performance on the Pediatric ITE, with further improvements observed over the past year and results exceeding those of senior residents. Nonetheless, performance on image-included questions was inferior to that of text-only questions and did not exceed that of senior residents.

Keywords: Natural Language Processing, Artificial Intelligence, Pediatrics

Introduction

Large language models (LLMs), a class of generative artificial intelligence (AI) systems, are increasingly being integrated into the medical landscape [1]. Recent studies have shown that these models can achieve accuracy on high-stakes examinations that meets or even exceeds expert-level passing thresholds [2]. In pediatrics, Generative Pre-trained Transformer-4 (GPT-4) scored 79.8% and 84.2% on the 2021 and 2022 American Academy of Pediatrics (AAP) self-assessments, exceeding the 75% average of pediatric trainees [3]. More recently, GPT-4 Omni achieved 90.4% accuracy on pediatric United States Medical Licensing Examination-style questions, outperforming GPT-3.5 and typical medical students [4].

Since early 2024, the widespread release of vision-enabled LLMs such as GPT-4, Claude, and Gemini has made large-scale multimodal, image-aware question solving feasible and further improved performance on medical examinations, sometimes matching or exceeding human test-takers [2, 4-8]. However, despite this rapid progress, several gaps persist in the pediatric field. First, few studies have compared multimodal LLMs with human clinicians on pediatric examinations that include image-based questions [5, 9]. Second, because multimodal capabilities are relatively recent, only a limited number of evaluations have tracked performance longitudinally across different LLM versions [10, 11]. Third, it is unclear whether findings based largely on English-language examinations generalize to non-English settings, where linguistic and contextual differences may affect both text processing and image–text alignment [12].

Therefore, to address these gaps, this study directly compared the proportion correct (PC) of state-of-the-art LLMs and pediatric residents on a single-institution in-training examination (ITE). We analyzed performance across both image-based and text-only items and quantified year-over-year performance changes by contrasting the latest models with their immediate predecessors using the same question bank.

Methods

This study was conducted in accordance with relevant guidelines and regulations. The study protocol was reviewed and approved, with a waiver of informed consent, by the Institutional Review Board of Asan Medical Center (IRB No. 2025-0722) because this study involved secondary analysis of educational assessment data and did not involve patient data or direct participant intervention. We included pediatric residents enrolled in a 4-year training program at Asan Medical Center. In this study, R1, R2, R3, and R4 refer to first-, second-, third-, and fourth-year pediatric residents, respectively. The pediatric ITE items and resident scores are proprietary to the Department of Pediatrics, Asan Medical Center, and their secondary use for this research and publication was approved by the department in accordance with institutional policies. This study was written according to the Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Healthcare [13].

Question preparation

Between 2016 and 2023, a total of 12 ITEs were conducted, excluding periods when examinations were not held due to the COVID-19 pandemic. Of these, one ITE was excluded because it was reserved for prompt optimization, and another was excluded due to the loss of answer sheets. Consequently, 546 questions from the remaining 10 ITEs were available for analysis. After removing 48 duplicated items, a total of 498 unique questions were included. The questions were written in Korean sentence structures; however, medical terminologies were predominantly presented in English. Approximately 20–30% of the items incorporated medical images. The original examination sheets were used without any modification to their content, preserving the format in which they had been distributed to the residents.

Each test set (comprising 45–55 questions) was submitted as a whole sheet to the LLM, rather than processing each question as an individual application programming interface (API) call. The models were required to autonomously distinguish between context, choices, and images, and then generate answers. To maintain the independence of each question and to minimize recognition errors by the LLM, the following measures were implemented: 1) To clearly separate the context from the answer choices, the beginning and end of each context were explicitly marked with “##”. 2) To prevent image distortion or resolution degradation, the original image files were stored separately and made available for direct retrieval by the model. 3) To minimize recognition errors related to the use of Korean text, the answer and explanation for each question were accompanied by a complete English translation of the question. These translations were independently reviewed by pediatricians (BLINDED). All constructed-response items were short-answer rather than essay-type and were handled as follows. A single investigator (BLINDED) graded responses from both the LLMs and the residents. Answers were marked as correct if they matched the reference answer, an accepted synonym, or a commonly used abbreviation (e.g., “GAS,” “group A streptococcus,” and “Streptococcus pyogenes” were all considered correct); all other responses were classified as incorrect. No items were judged to be ambiguous

LLM Applications

To analyze temporal changes in LLM performance, we selected models based on their release dates relative to the study design time point (August 2025). Among the three major, publicly accessible LLM families (GPT, Gemini, and Claude), we included one “recent” model (released within the 12 months preceding August 2025) and one “prior” model (released 12–24 months before August 2025) from each family. We restricted our sample to general-purpose LLMs that were directly accessible to end users as web-based chatbots, without requiring API calls, local deployment, or fine-tuning, and that were publicly available at the time of the study. The final set of models consisted of GPT-4o, GPT-4.1, Gemini-1.5-Pro, Gemini-2.5-Pro, Claude-3.5-Sonnet, and Claude-4.1-Opus. All LLM model inferences, including prompt optimization and question solving, were performed programmatically via each company’s business API rather than through a web interface. All model inferences were performed between August 5 and August 15, 2025, using the versions of each LLM that were publicly available during this period. API calls, question set retrievals, and prompt submissions were performed using Python 3.9.13 (Python Software Foundation, USA). Test questions were retrieved from a secure repository and inserted into a standardized prompt template; no test content was entered into an interactive chatbot user interface. The API workspaces had enterprise data protection enabled by default (no training on business data, memory disabled, and no fine-tuning).

The prompt was constructed as follows. First, within each examination set, the context, choices, and images were extracted, and the question type was classified accordingly: multiple-choice questions with or without image, and constructed-response questions with image. Representative examples of multiple-choice and constructed-response questions, with and without images, are provided in Supplementary Table 1. Second, task-specific instructions were added to guide the reasoning process, including the expected knowledge level and the method of question solving according to the type. Finally, the prompt incorporated critical instructions for decision-making and output generation. Prompt refinement was conducted between August 5 and August 15, 2025. The finalized prompt is provided in Supplementary Table 2. Using this finalized prompt, the LLMs performed the assigned tasks between August 15 and August 20, 2025.

Evaluation

Given that all questions were medical decision–making tasks, the temperature parameter was set to 0 in order to minimize response variability and diverse output generation. Each examination set was processed five times to assess the robustness of outputs against randomness, and the first output generated was adopted as the final LLM response.

The primary outcome was the PC of LLMs on individual questions, which was compared with that of pediatric residents. To allow repeated participation without attribution errors, a unique identifier (year–exam–grade–sequence) was assigned to each participant, and multiple appearances of the same individual across examination sessions were treated as independent entries. In addition, we compared the performance between recent LLMs (released within the past year: GPT-4.1, Gemini-2.5-Pro, Claude-4.1-Opus) and prior LLMs (released more than 1 year earlier: GPT-4o, Gemini-1.5-Pro, Claude-3.5-Sonnet), as well as with the most senior resident group (fourth-year resident, R4).

Statistical analysis

PC and 95% confidence intervals were estimated using 10,000 nonparametric bootstrap replicates at the question level. Comparisons between residents and LLMs were conducted using a two-sided McNemar’s test, with resident group responses dichotomized at the item level by majority vote (≥50%), enabling direct binary comparisons with LLM predictions. Two-sided P values < 0.05 were considered statistically significant. To account for multiple comparisons, we controlled the family-wise error rate for two prespecified families of hypotheses using Bonferroni-adjusted thresholds, as described below. First, we tested six pairwise comparisons between R4 and each of the six LLMs, applying a Bonferroni-adjusted threshold of α = 0.008 (0.05/6). Second, we tested three comparisons involving R4, the prior LLMs, and the recent LLMs, using a Bonferroni-adjusted threshold of α = 0.017 (0.05/3). We did not perform formal pairwise hypothesis tests among LLMs themselves; instead, we report their relative performance descriptively.

To compare performance on image-included versus text-only questions for each LLM model, an independent-samples t-test was conducted on the mean PC values obtained through bootstrapping. Repeatability across five times of LLM responses was assessed using the ICC. Agreement levels were categorized as follows: poor (ICC < 0.20), fair (0.21–0.40), moderate (0.41–0.60), good (0.61–0.80), and excellent (> 0.80) [14]. Statistical analyses were performed using SPSS Statistics for Windows, version 21.0 (IBM Corp., New York, USA) and MedCalc Statistical Software version 23.3.5 (MedCalc Software Ltd, Ostend, Belgium).

Results

Characteristics of participating residents and examination questions

During the study period, a total of 308 residents participated in the ITEs, with 73–81 residents included per training year (Table 1). The distribution of participants by training level (R1 to R4) was even (range, 23.7%–26.3%). The median number of residents per training year per exam was 8.

Among the 498 included questions, 387 (77.7%) were text-only and 111 (22.3%) contained one or more images. The most common image modality was X-ray (42.3%), followed by clinical photographs (30.6%), magnetic resonance imaging (24.3%), ultrasound (20.7%), and computed tomography (18.0%). The majority of the questions were multiple-choice (492, 98.8%), while only 6 (1.2%) items were constructed-response questions. The number of questions by topic was relatively balanced, with 42–47 questions (8.4–9.4%) across most specialties, except for emergency medicine (30, 6%), medical genetics (28, 5.6%), endocrinology/metabolism (22, 4.4%), and critical care (13, 2.6%)

Comparison of LLM Performance Against R4 and Across Model Generations

Figure 1 shows the different LLMs and their performance. The PC scores of residents from R1 to R4 were 56.5%, 61.7%, 65.0%, and 70.1%, respectively (Fig. 1). Recent LLMs (GPT-4.1, Gemini 2.5-Pro, and Claude-4.1-Opus) significantly outperformed R4 (78.9%, 77.9%, and 77.7% vs. 70.1%, respectively; all $P < 0.008$). Among the LLMs, GPT-4.1 had the highest PC on the pediatric ITE (78.9%, Fig. 1). However, our statistical comparisons were prespecified to focus on R4 versus each LLM and on R4 versus prior versus recent LLM groups; we did not perform formal hypothesis testing for all pairwise differences among LLMs. When performance was analyzed by question type, all LLMs' performance was comparable to that of R4 on image-included questions. For text-only questions, recent LLMs (GPT-4.1, Gemini 2.5-Pro, and Claude-4.1-Opus) achieved significantly higher PC compared to R4 (80.1%, 81.0%, and 80.1% vs. 69.6%, respectively; all $P < 0.008$).

Compared with prior LLMs, recent models exhibited significantly higher PC across all questions, including image-included and text-only types (all $P < 0.017$, Fig. 2). Furthermore, compared with R4, recent LLMs showed significantly higher PC on total and text-only questions ($P = 0.01$ and 0.03, respectively) but demonstrated comparable performance on image-included questions.

For both Gemini 2.5-Pro and Claude 3.5-Sonnet, the PC was significantly lower on image-included questions compared to text-only questions (71.9% vs. 81.0%, $P = 0.02$ and 66.7% vs. 76.1%, $P = 0.024$, respectively; Fig. 3). This performance trend was consistent across both generational groups, with significantly lower PC on image-included questions than on text-only questions for both prior and recent LLMs (66.3% vs. 73.6%, $P = 0.038$ and 73.9% vs. 80.4%, $P = 0.032$, respectively).

PC comparisons between R4 and LLMs across topics and image subcategories are summarized in Table 2. Across all image types and topics, some pairwise comparisons showed significantly higher PC for LLMs than for R4, whereas none showed significantly higher PC for R4 than for any LLM (Bonferroni-adjusted $\alpha = 0.008$).

Repeatability of LLM outputs

Repeatability assessments are summarized in Table 3. With the exception of Gemini-2.5-Pro, all LLMs demonstrated excellent agreement (intraclass correlation coefficient [ICC] range, 0.98–1.00), while Gemini-2.5-Pro showed good agreement with an ICC of 0.73.

Discussion

In the study, we found that recent multimodal LLMs released within the past 2 years, both prior models and recent models, demonstrated performance that was either superior to or comparable to that of senior pediatric residents on the pediatric ITE. Prior models (released more than 1 year ago) showed performance comparable to senior residents, whereas recent models (released within the past year) exhibited improved accuracy compared to their predecessors and outperformed senior residents. These models also provided consistent responses across repeated tasks. Furthermore, the high performance of multimodal LLMs previously reported in English-based examinations was reproduced in a mixed Korean–English setting. However, both recent and prior LLMs exhibited lower performance on image-included questions compared to text-only questions, and none surpassed senior residents on image-included items.

Consistent with recent reports on the high performance of commercially accessible LLMs on ITEs [4-6, 9, 11, 15-17], our study confirms this trend and demonstrates even further improved performance. Although this study was conducted in ITE setting, the results suggest the potential for applying commercially accessible LLMs in clinical practice. Several studies have documented the real-world feasibility of LLMs as Clinical Decision Support Systems (CDSS) across various disease categories and clinical scenarios [18, 19]. Although most studies still caution against LLM-driven decisions, emphasizing their limited role as assistive tools [6, 20], our findings primarily show that multimodal LLMs can perform comparably to senior pediatric residents on ITE-style multiple-choice questions. However, recent work has highlighted important limitations of multiple-choice exam benchmarks for evaluating LLMs in healthcare and cautioned against interpreting such results as evidence of clinical readiness [21, 22]. Moreover, even if LLM performance on ITE-style questions is promising, multiple additional steps, including testing on real-world cases, and external validation, are still required. Accordingly, our results should not be taken as proof of readiness for autonomous pediatric CDSS, but rather as an initial signal that may justify carefully staged future evaluations of LLM-assisted decision support in pediatrics. Beyond clinical feasibility, the high accuracy of LLMs on the ITE directly underscores their educational utility [6, 23]. As adjunctive study tools, LLMs can provide trainees with rapid feedback, exposure to a wide range of clinical scenarios, and opportunities to practice critical appraisal of AI-generated answers [6]. While not a substitute for supervised clinical training, their consistent performance on standardized questions suggests a strong potential to complement traditional pediatric education.

Although multiple studies have investigated the performance of LLM on pediatric medical questions, the majority excluded image-based questions due to incomplete image recognition capabilities [4, 9, 15, 17]. Notably, Le et al. conducted the only study utilizing GPT-4.0 Vision to analyze image-included questions, but the model failed to interpret 15% of the items and achieved an accuracy of only 68% [3]. In our study, prior vision-enabled LLMs demonstrated lower accuracy (66.3%), whereas recent LLMs showed a meaningful improvement (73.9%), with the top-performing model, Claude-4.1-Opus, achieving an accuracy of 76.6%. Nevertheless, their performance remained inferior to that on text-only questions and did not exceed that of clinicians.

Several studies have similarly reported that LLM-based models exhibit weaker image analysis capabilities relative to their strong performance in text-based analysis [2, 24, 25]. This limitation is particularly pronounced in pediatrics, where studies have shown that LLMs demonstrate inadequate performance in image interpretation [26, 27]. Unlike real-world clinical practice, where imaging studies are variable and less standardized, ITE questions are carefully curated, which would be expected to favor higher AI performance. This insufficient performance raises concerns about the applicability of LLMs in pediatric settings, where diagnostic reasoning often relies on imaging studies due to the inherent limitations in history-taking and physical examination relative to adult populations. Consistent with prior studies, our findings underscore the limitations of LLMs in pediatric image interpretation and indicate the need for further improvements in commercial models or the development of dedicated fine-tuned systems [5, 6, 11, 28].

Beyond insufficient image interpretation, several other barriers must be addressed before LLMs can be reliably applied in real-world pediatric practice. Clinical decision-making often requires nuanced contextual understanding, longitudinal patient information, and integration of multimodal data—areas in which current LLMs remain limited [29]. AI models also pose inherent challenges, including hallucinations, bias, performance drift, and the black-box issue, all of which require careful consideration [29, 30]. Given that errors may occur more frequently in pediatric scenarios than in adult scenarios, particular caution is warranted when applying LLMs in the pediatric field [11, 31, 32]. Given the potentially severe consequences of AI malfunctions in clinical practice, advancing toward a clinically applicable LLM-based CDSS will require rigorous validation alongside improvements in prompt engineering, fine-tuning, on-premise validation, and retrieval-augmented generation to ensure performance and reliability.

This study has several limitations. First, because it was conducted at a single center and included only residents who participated in one institutional ITE, the generalizability of the findings may be limited. Second, our evaluation reflects LLM performance during a specific time window. All experiments were conducted in 2025 using model versions that were available at that time; given the rapid pace of model updates, newer versions may already achieve higher accuracy than reported here. Accordingly, our results should be interpreted as a snapshot of contemporary performance rather than a definitive estimate of each model's long-term capabilities. Third, although two authors reviewed the English-translated questions, the examination consisted of mixed Korean–English items; therefore, the possibility that language recognition errors affected performance cannot be entirely excluded. Nevertheless, the findings of this study were consistent with the trends reported in prior publications and even reinforced some of the previously suggested conclusions, supporting a reasonable degree of reproducibility. Forth, we did not perform an in-depth analysis of the LLM error patterns, which could have clarified how their mistakes differed from those of residents and in what ways recent models may have overcome the shortcomings of prior versions. In addition, unlike most previous studies, our dataset included a small number of short-answer questions; however, this number was limited to six, which was insufficient for meaningful statistical analysis and prevented a direct comparison with multiple-choice questions. Fifth, we did not evaluate open-source LLMs because our study was designed around vendor-hosted models that are immediately deployable via managed APIs in real-world clinical and educational settings. Sixth, because we did not conduct a detailed review of the LLMs' generated reasoning, we were unable to evaluate reasoning transparency, hallucination risk, calibration, or error detectability. In addition, we did not analyze why the models failed to arrive at the correct answers on image-based items, nor did we assess the potential clinical severity or consequences of such errors. These aspects are critical for evaluating the feasibility and safety of clinical deployment of LLMs and should be addressed in future studies. Finally, it should be noted that our results should be interpreted as a comparison of LLMs and residents on a standardized written assessment, rather than as a direct comparison of their overall clinical performance, and that performance on ITE-style multiple-choice questions should not be overinterpreted as evidence of real-world clinical problem-solving ability. These limitations warrant further investigation in future studies.

Conclusion

Multimodal LLMs achieved high performance on the Pediatric ITE, with further improvements observed over the past year and results exceeding those of senior residents. Nonetheless, their performance on image-included questions remained inferior to text-based analysis, highlighting both the educational potential of LLMs and the persistent limitations of current models in pediatric image interpretation, which require further refinement before clinical application.

References

1. Clusmann, J., et al., The future landscape of large language models in medicine*. Commun Med (Lond)*, 2023. **3**(1): p. 141.
2. Singhal, K., et al., Toward expert-level medical question answering with large language models*. Nat Med*, 2025. **31**(3): p. 943-950.
3. Le, M. and M. Davis, ChatGPT Yields a Passing Score on a Pediatric Board Preparatory Exam but Raises Red Flags*. Glob Pediatr Health*, 2024. **11**: p. 2333794x241240327.
4. Bicknell, B.T., et al., ChatGPT-4 Omni Performance in USMLE Disciplines and Clinical Skills: Comparative Analysis*. JMIR Med Educ*, 2024. **10**: p. e63430.
5. Del Monte, F., et al., Diagnostic efficacy of large language models in the pediatric emergency department: a pilot study*. Front Digit Health*, 2025. **7**: p. 1624786.
6. Suresh, S. and S.M. Misra, Large Language Models in Pediatric Education: Current Uses and Future Potential*. Pediatrics*, 2024. **154**(3).
7. Yang, Z., et al., The dawn of lmms: Preliminary explorations with gpt-4v (ision)*. arXiv preprint arXiv:2309.17421*, 2023.
8. Jin, Q., et al., Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine*. NPJ Digital Medicine*, 2024. **7**(1): p. 190.
9. Katz, U., et al., GPT versus Resident Physicians — A Benchmark Based on Official Board Scores*. NEJM AI*, 2024. **1**(5): p. AIdbp2300192.
10. Park, J.S., et al., Accuracy of Large Language Models in Detecting Cases Requiring Immediate Reporting in Pediatric Radiology: A Feasibility Study Using Publicly Available Clinical Vignettes*. Korean J Radiol*, 2025. **26**(9): p. 855-866.
11. Mondillo, G., et al., ARE LLMS READY FOR PEDIATRICS? A COMPARATIVE EVALUATION OF MODEL ACCURACY ACROSS CLINICAL DOMAINS*. medRxiv*, 2025: p. 2025.04.25.25326437.
12. Yao, Z., et al., Performance of Large Language Models in the Non-English Context: Qualitative Study of Models Trained on Different Languages in Chinese Medical Examinations*. JMIR Med Inform*, 2025. **13**: p. e69485.
13. Park, S.H., et al., Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Healthcare (MI-CLEAR-LLM)*. Korean J Radiol*, 2024. **25**(10): p. 865-868.
14. Crewson, P.E., Reader agreement studies*. AJR Am J Roentgenol*, 2005. **184**(5): p. 1391-7.
15. Jaiswal, N., et al., PedMedQA: Comparing Large Language Model Accuracy in Pediatric and Adult Medicine*. Pediatrics Open Science*, 2025. **1**(2): p. 1-3.
16. Mondillo, G., et al., Large language models performance on pediatrics question: a new challenge*. Journal of Medical Artificial Intelligence*, 2024. **8**.
17. Abbas, A., M.S. Rehman, and S.S. Rehman, Comparing the Performance of Popular Large Language Models on the National Board of Medical Examiners Sample Questions*. Cureus*, 2024. **16**(3): p. e55991.
18. Gaber, F., et al., Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis*. NPJ Digit Med*, 2025. **8**(1): p. 263.
19. Busch, F., et al., Current applications and challenges in large language models for patient care: a systematic review*. Commun Med (Lond)*, 2025.

5(1): p. 26.
20. Thirunavukarasu, A.J., et al., Large language models in medicine*. Nature Medicine*, 2023. **29**(8): p. 1930-1940.
21. Awasthi, R., et al., Human evaluation of large language models in healthcare: gaps, challenges, and the need for standardization*. npj Health Systems*, 2025. **2**(1): p. 40.
22. Raji, I.D., R. Daneshjou, and E. Alsentzer, It's Time to Bench the Medical Exam Benchmark*. NEJM AI*, 2025. **2**(2): p. AIe2401235.
23. Zhang, J. and S.H. Fenton, Preparing healthcare education for an AI-augmented future*. Npj Health Syst*, 2024. **1**(1): p. 4.
24. Kim, H., et al., ChatGPT Vision for Radiological Interpretation: An Investigation Using Medical School Radiology Examinations*. Korean J Radiol*, 2024. **25**(4): p. 403-406.
25. Liu, Z., et al., Holistic evaluation of gpt-4v for biomedical imaging*. arXiv preprint arXiv:2312.05256*, 2023.
26. Reith, T.P., D.M. D'Alessandro, and M.P. D'Alessandro, Capability of multimodal large language models to interpret pediatric radiological images*. Pediatr Radiol*, 2024. **54**(10): p. 1729-1737.
27. Suh, P.S., et al., Comparing Diagnostic Accuracy of Radiologists versus GPT-4V and Gemini Pro Vision Using Image Inputs from Diagnosis Please Cases*. Radiology*, 2024. **312**(1): p. e240273.
28. Hager, P., et al., Evaluation and mitigation of the limitations of large language models in clinical decision-making*. Nat Med*, 2024. **30**(9): p. 2613-2622.
29. Templin, T., et al., Addressing 6 challenges in generative AI for digital health: A scoping review*. PLOS Digit Health*, 2024. **3**(5): p. e0000503.
30. Jung, K.H., Large Language Models in Medicine: Clinical Applications, Technical Challenges, and Ethical Considerations*. Healthc Inform Res*, 2025. **31**(2): p. 114-124.
31. Barile, J., et al., Diagnostic Accuracy of a Large Language Model in Pediatric Case Studies*. JAMA Pediatr*, 2024. **178**(3): p. 313-315.
32. Muralidharan, V., et al., Recommendations for the use of pediatric data in artificial intelligence and machine learning ACCEPT-AI*. NPJ Digit Med*, 2023. **6**(1): p. 166.

Acknowledgment: Not applicable

Funding statement: Not applicable

Author contributions: JSP and SHK conceived and designed the study. MJK performed data analysis and interpretation. MJK, JSP and SHK contributed to data acquisition and clinical review. MJK drafted the manuscript. All authors critically revised the manuscript for important intellectual content and approved the final version for submission.

Data availability statement: The datasets generated during and/or analyzed during the current study are not publicly available due to restrictions from the institutional review board of the Asan Medical Center, Seoul, Korea (IRB no.2025-0722), which prohibits data sharing with out-of-hospital facilities for ethical reasons. However, data are available from the corresponding author upon reasonable request.

Competing interest: The authors declare no competing interests.

Figure Legends

Figure 1. Comparison of performance between pediatric residents and LLMs

*P<0.008 compared to R4

LLM, large language model; GPT, Generative Pre-trained Transformer

Point estimates indicate the mean bootstrap PC based on 10,000 bootstrap resamples.

Figure 2. Comparison of the performance between recent LLMs (GPT-4.1, Gemini-2.5-pro, and Claude-4.1-opus) and older LLMs (GPT-4o, Gemini-1.5-pro, and Claude-3.5-Sonnet)

*P<0.017

LLM, large language model; GPT, Generative Pre-trained Transformer

Point estimates indicate the mean bootstrap PC based on 10,000 bootstrap resamples.

Figure 3. Performance of each LLM model on image-included versus text-only questions

LLM, large language model; GPT, Generative Pre-trained Transformer

*P<0.05

Point estimates indicate the mean bootstrap PC based on 10,000 bootstrap resamples.

Tables

Table 1. Distribution of Participating Residents and Composition of Questions

Distribution of Residents by Training Year	308
1st grade	80 (26)
2nd grade	74 (24)
3rd grade	81 (26.3)
4th grade	73 (23.7)
Composition of Questions	498
Text-only	387 (77.7)
Image-included	111 (22.3)
X-ray	47 (42.3)
Clinical photograph	34 (30.6)
MRI	27 (24.3)
US except echocardiography	23 (20.7)
CT	20 (18)
Microscopy	19 (17.1)
Illustration	14 (12.6)
ECG	12 (10.8)
EEG	9 (8.1)
Endoscopy	6 (5.4)
Echocardiography	4 (3.6)
Nuclear scan	3 (2.7)
Fluoroscopy	1 (0.9)
Fundoscopy	1 (0.9)
Multiple-choice questions	492 (98.8)
Constructed-response questions	6 (1.2)
Neonatology	47 (9.4)
Hemato-Oncology	47 (9.4)
Nephrology	45 (9)
Cardiology	44 (8.8)
Gastroenterology	44 (8.8)
Infectious disease	44 (8.8)
Neurology	43 (8.6)
Pulmonology	42 (8.4)
Emergency medicine	30 (6)
Medical genetics	28 (5.6)
Endocrinology/Metabolism	22 (4.4)
Critical care medicine	13 (2.6)

Values are presented as number (%) or median (interquartile range)

The percentage of image subcategories was calculated based on the total number of image-included questions

MRI, magnetic resonance image; US, ultrasound; CT, computed tomography; ECG, electrocardiography; EEG, electroencephalography

Table 2. Accuracy by Image Modality and Topic

[	R4	GPT 4o		GPT 4.1		Gemini 1.5 pro		Gemini 2.5 pro		Claude 3.5 sonnet		Claude 4.1 opus	
[	PC	PC	P value	PC	P value	PC	P value	PC	P value	PC	P value	PC	P value
Topic													
Neonatology	46.2	55.8	0.332	**67.3**	**0.007**	59.6	0.167	**73.1**	**0.003**	59.6	0.118	65.4	0.013
Hemato-Oncology	67.3	73.1	0.607	86.5	0.021	73.1	0.581	84.6	0.035	76.9	0.302	86.5	0.013
Nephrology	80.0	68.0	0.146	74.0	0.607	70.0	0.227	78.0	1.000	78.0	1.000	76.0	0.774
Cardiology	77.6	69.4	0.454	77.6	1.000	63.3	0.167	87.8	0.267	77.6	1.000	81.6	0.791
Gastroenterology	69.4	69.4	1.000	77.6	0.388	57.1	0.263	63.3	0.607	61.2	0.454	81.6	0.210
Infectious disease	54.2	64.6	0.424	58.3	0.839	52.1	1.000	64.6	0.383	60.4	0.629	54.2	1.000
Neurology	75.0	81.3	0.581	93.8	0.012	81.3	0.607	79.2	0.774	81.3	0.581	87.5	0.109
Pulmonology	80.9	78.7	1.000	83.0	1.000	68.1	0.180	78.7	1.000	74.5	0.581	72.3	0.424
Emergency medicine	84.0	96.0	0.250	96.0	0.375	84.0	1.000	88.0	1.000	84.0	1.000	88.0	1.000
Medical genetics	66.7	66.7	1.000	63.3	1.000	63.3	1.000	76.7	0.549	56.7	0.581	70.0	1.000
Endocrinology/Metabolism	97.1	91.4	0.625	94.3	1.000	77.1	0.016	94.3	1.000	94.3	1.000	94.3	1.000
[Critical care medicine	92.3	92.3	1.000	92.3	1.000	69.2	0.250	69.2	0.375	84.6	1.000	92.3	1.000
Image Modality													
X-ray	74.5	68.1	0.508	72.3	1.000	57.4	0.115	63.8	0.227	63.8	0.227	76.6	1.000
Clinical photograph	64.7	52.9	0.424	64.7	1.000	61.8	1.000	67.6	1.000	55.9	0.581	61.8	1.000
MRI	63.0	74.1	0.453	85.2	0.109	81.5	0.125	70.4	0.688	81.5	0.125	85.2	0.070
US except echocardiography	60.9	52.2	0.754	78.3	0.289	52.2	0.754	69.6	0.754	56.5	1.000	60.9	1.000
CT	70.0	70.0	1.000	80.0	0.625	70.0	1.000	80.0	0.688	70.0	1.000	75.0	1.000
Microscopy	73.7	78.9	1.000	78.9	1.000	84.2	0.625	84.2	0.625	73.7	1.000	68.4	1.000
Illustration	64.3	57.1	1.000	71.4	1.000	50.0	0.688	85.7	0.375	57.1	1.000	57.1	1.000
ECG	100	83.3	0.500	83.3	0.500	66.7	0.125	91.7	1.000	75.0	0.250	83.3	0.500
EEG	77.8	88.9	1.000	100	0.500	88.9	1.000	88.9	1.000	88.9	1.000	100	0.500
Endoscopy	83.3	66.7	1.000	66.7	1.000	83.3	1.000	66.7	1.000	50.0	0.500	83.3	1.000
Echocardiography†	25.0	0.0	NA	25.0	NA	25.0	NA	25.0	NA	0.0	NA	25.0	NA
Nuclear scan†	100	66.7	NA	100	NA	66.7	NA	33.3	NA	33.3	NA	66.7	NA
Fluoroscopy†	100	100	NA	100	NA	0.0	NA	100	NA	100	NA	100	NA
[Fundoscopy†	100	100	NA	100	NA	0.0	NA	100	NA	0.0	NA	100	NA

Abbreviations: R4, 4th grade resident; GPT, Generative Pre-trained Transformer; MRI, magnetic resonance image; US, ultrasound; CT, computed tomography; ECG, electrocardiography; EEG, electroencephalography, NA, not applicable.

P values are calculated using R4 as the reference.

†Statistical analysis was not performed due to extremely low number of questions.

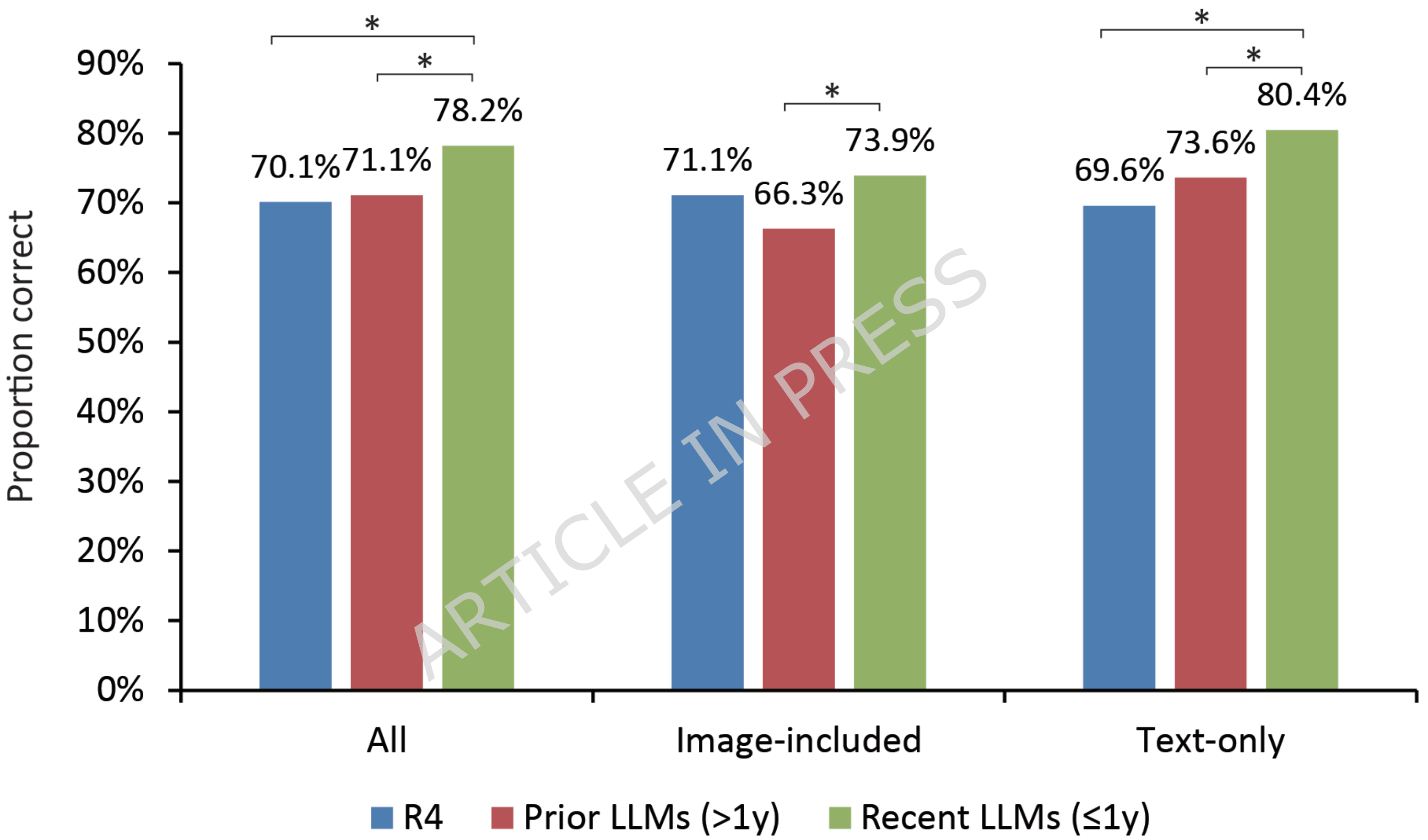
Bold indicates statistical significance at the Bonferroni-adjusted threshold (α = 0.05/6 = 0.008) for the six R4-vs-model comparisons.

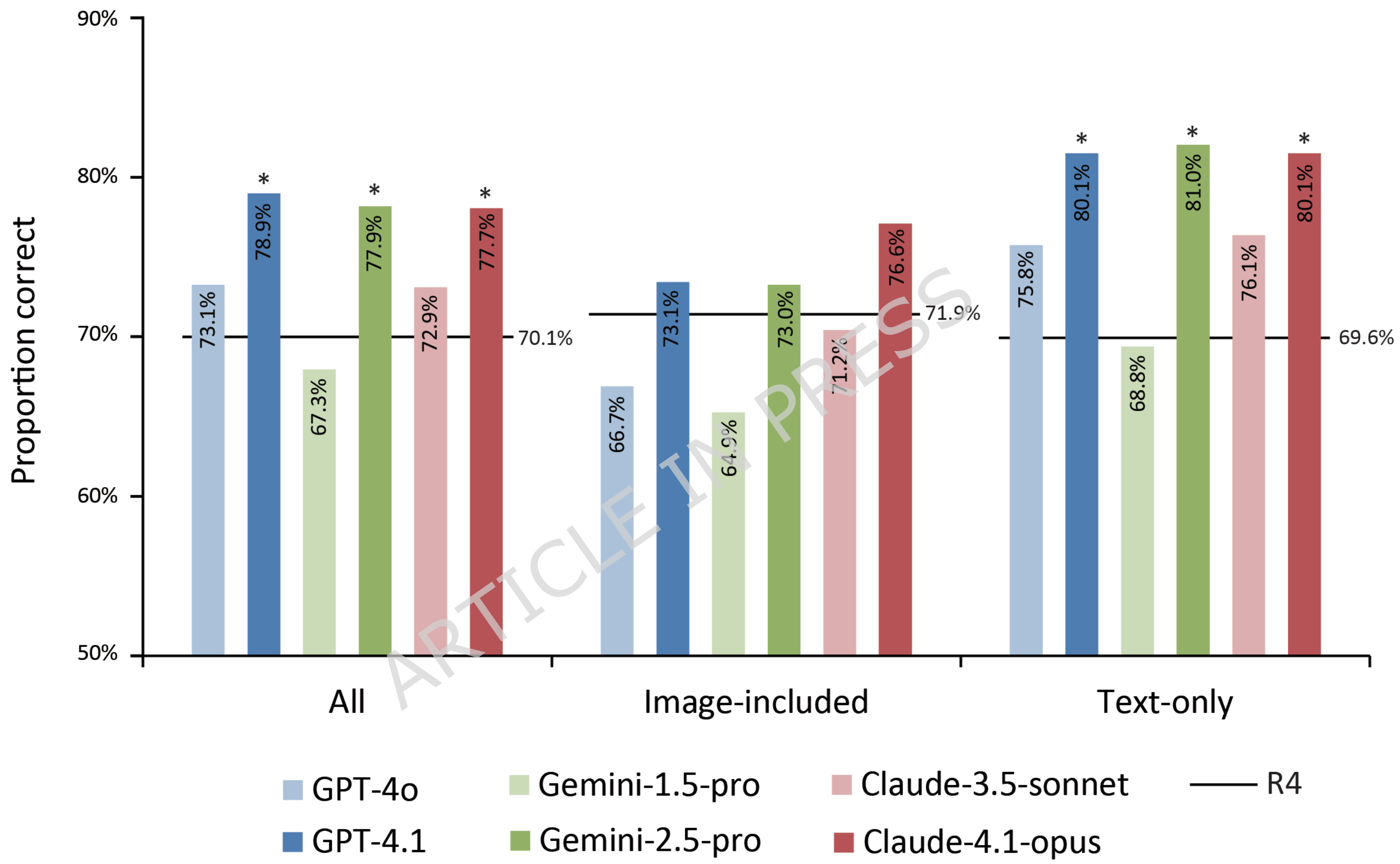
PC is the mean from 10,000 bootstrap resamples, expressed as a %.

Table 3. Summary of agreement between five repeated sessions in each LLM

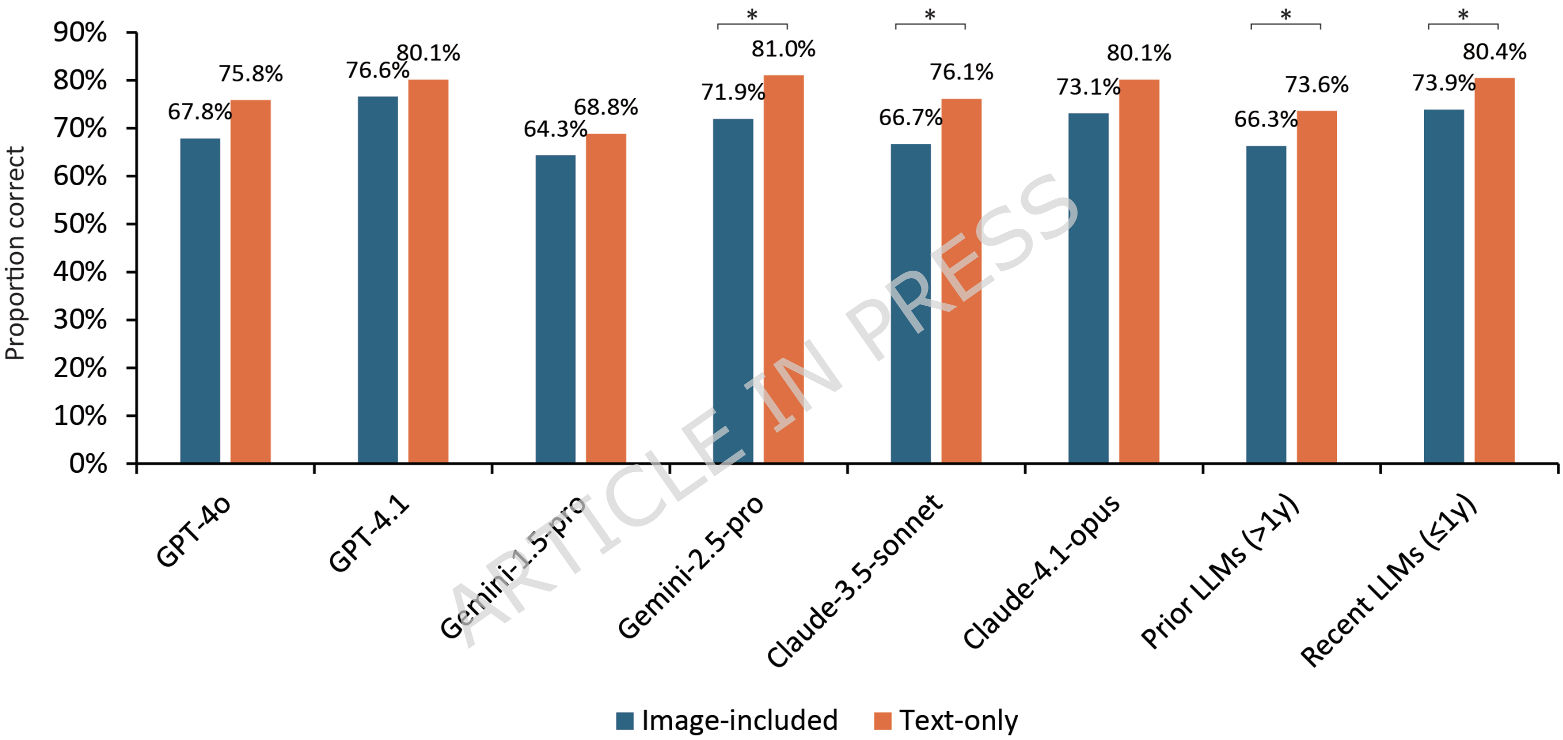
	ICC	95% CI
GPT-4o	0.98	0.98–0.99
GPT-4.1	0.99	0.99–0.99
Gemini-1.5-pro	1.00	1.00–1.00
Gemini-2.5-pro	0.73	0.68–0.76
Claude-3.5-sonnet	1.00	1.00–1.00
Claude-4.1-opus	1.00	1.00–1.00

LLM, large language model; GPT, Generative Pre-trained Transformer; ICC, Intraclass correlation coefficient; CI, confidence interval


*
*
78.2%
70.1%
71.1%
*
73.9%
71.1%
66.3%
*
*
80.4%
73.6%
69.6%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Proportion correct
All
Image-included
Text-only
R4
Prior LLMs (>1y)
Recent LLMs (≤1y)



90%
80%
70%
60%
50%
Proportion correct
73.1%
78.9%
67.3%
77.9%
72.9%
77.7%
70.1%
66.7%
73.1%
64.9%
73.0%
71.2%
76.6%
71.9%
75.8%
80.1%
68.8%
81.0%
76.1%
80.1%
69.6%
*
All
Image-included
Text-only
GPT-4o
GPT-4.1
Gemini-1.5-pro
Gemini-2.5-pro
Claude-3.5-sonnet
Claude-4.1-opus
R4



Proportion correct
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
67.8%
75.8%
76.6%
80.1%
64.3%
68.8%
71.9%
81.0%
66.7%
76.1%
73.1%
80.1%
66.3%
73.6%
73.9%
80.4%
*
*
*
*
GPT-4o
GPT-4.1
Gemini-1.5-pro
Gemini-2.5-pro
Claude-3.5-sonnet
Claude-4.1-opus
Prior LLMs (>1y)
Recent LLMs (≤1y)
Image-included
Text-only