Discover Artificial Intelligence

Research

Evaluating ChatGPT-4 in medical education: an assessment of subject exam performance reveals limitations in clinical curriculum support for students

Brendan P. Mackey[1] · **Razmig Garabet**[1] · **Laura Maule**[1] · **Abay Tadesse**[1] · **James Cross**[1] · **Michael Weingarten**[1]

Received: 22 December 2023 / Accepted: 11 May 2024
Published online: 16 May 2024

Abstract

This study evaluates the proficiency of ChatGPT-4 across various medical specialties and assesses its potential as a study tool for medical students preparing for the United States Medical Licensing Examination (USMLE) Step 2 and related clinical subject exams. ChatGPT-4 answered board-level questions with 89% accuracy, but showcased significant discrepancies in performance across specialties. Although it excelled in psychiatry, neurology, and obstetrics and gynecology, it underperformed in pediatrics, emergency medicine, and family medicine. These variations may be potentially attributed to the depth and recency of training data as well as the scope of the specialties assessed. Specialties with significant interdisciplinary overlap had lower performance, suggesting complex clinical scenarios pose a challenge to the AI. In terms of the future, the overall efficacy of ChatGPT-4 indicates a promising supplemental role in medical education, but performance inconsistencies across specialties in the current version lead us to recommend that medical students use AI with caution.

Keywords Artificial intelligence · Machine learning · Medical education · Chatgpt-4 · Usmle step 2 · Nbme shelf

1 Introduction

Artificial intelligence (AI) has seen a substantial uptick in utilization for medical education since the release of ChatGPT and other language learning models (LLMs) [1–3]. As medical school curriculums continue to incorporate new innovations and changes to clinical recommendations, AI has helped meet a demand for more efficient and effective learning resources [4, 5]. In particular, medical students have been exploring the potential of LLMs to reinforce learned information, provide clarification on complex clinical topics, and aid in preparation for tests such as the United States Medical Licensing Examination (USMLE) series and related subject exams [1, 6]. Initial studies have demonstrated that ChatGPT-4 can process board-level exam questions and provide useful clinical insights [7–12]. No studies, however, have stratified these capabilities within the specialty-specific domains appearing on the USMLE Step 2 examination and associated clinical subject exams.

This study addresses this with an assessment of ChatGPT 4 performance on questions derived from the following specialties: Internal Medicine, Surgery, Family Medicine, Ambulatory Medicine, Emergency Medicine, Obstetrics and Gynecology, Pediatrics, Neurology, and Psychiatry. Primary objectives include an evaluation and comparison of accuracy across

✉ Brendan P. Mackey, brendanpatryck@gmail.com | [1]Drexel University College of Medicine, Philadelphia, PA, USA.

specialty domains. Secondary objectives include comparison to student performance and determination of whether the AI's understanding is authentic or a result of random correct answer selection. Furthermore, we evaluate whether ChatGPT-4 should be recommended as a study tool in its current version for medical students preparing for USMLE Step 2 and related subject examinations.

2 Materials and methods

2.1 Question acquisition

Authors were provided access to utilize questions from AMBOSS for the purpose of this study. AMBOSS is a comprehensive medical education platform that serves as a reference database for medical topics and offers question banks for various medical exams. The custom session interface within the question bank section allows users to extract practice questions from particular specialties and examinations.

For our analysis, questions were extracted from the "Clinical Shelf Exam" section. This section comprises nine specialties: Internal Medicine, Surgery, Pediatrics, Obstetrics and Gynecology, Neurology, Psychiatry, Family Medicine, Emergency Medicine, and Ambulatory Medicine. Each specialty was toggled "on" individually, and from each, 100 questions were randomly selected for a total of 900 questions.

2.2 Conversation input

Each question, along with its provided multiple choice answer options, were individually inputted into the ChatGPT4 interface. The output was assessed for accuracy by comparing to the correct answer choice provided by the AMBOSS question bank. Additional variables included the number of multiple-choice options provided by each question and the percentage of students that correctly answered the question. Figures 1 and 2 demonstrate an example of a ChatGPT input and output conversation. This sample question was obtained from NBME's free online USMLE Step 2 Orientation.

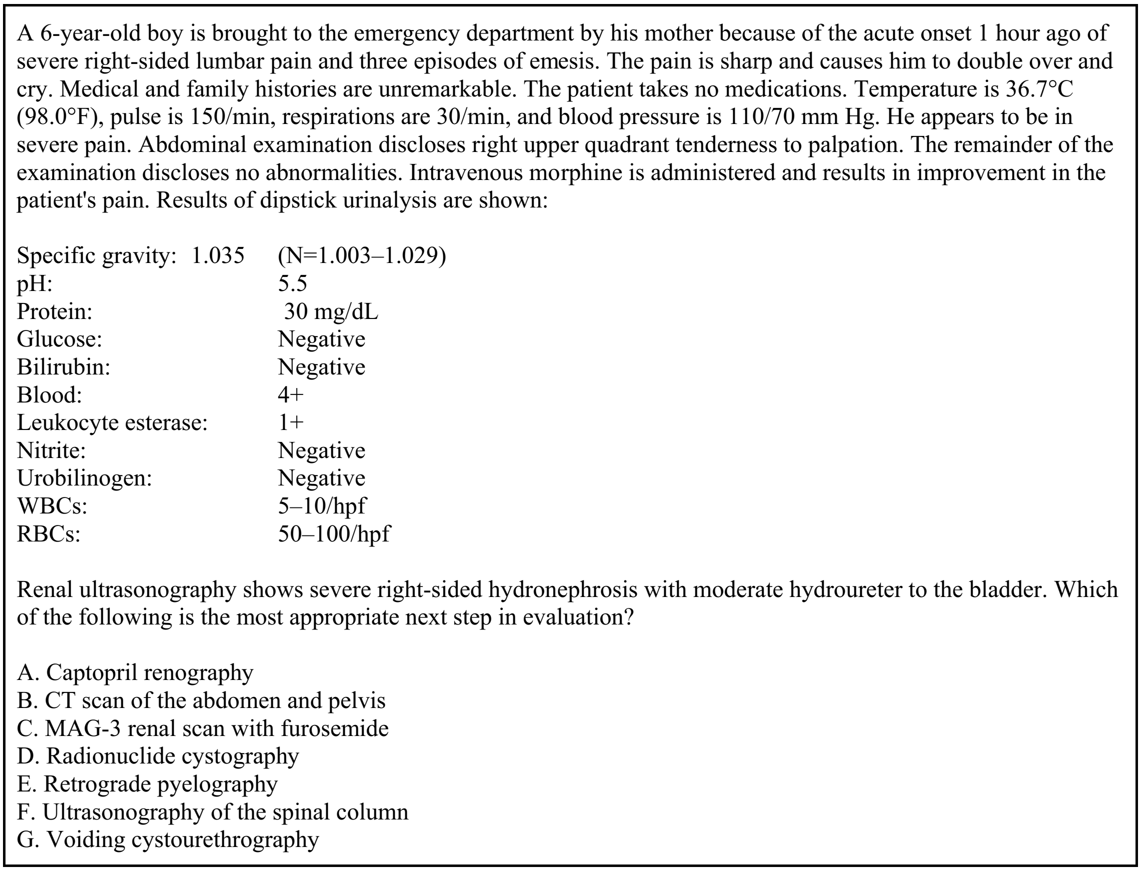
After each question, the ChatGPT input and output conversations were deleted, and a new conversation interface initiated. This was done to avoid the possibility of feedback knowledge from prior questions, which could potentially influence ChatGPT's algorithmic thought process in subsequent entries. Questions that included media interpretation, such as diagnostic imaging, electrocardiograms, or lesion appearances were excluded due to ChatGPT's limitation of image processing at the time of our analysis. Questions that incorporated all other modalities such as text or tables with lab results were included. In the event that ChatGPT did not provide a conclusive answer to a question or provided multiple answer choices, the question was omitted.

2.3 Statistical analysis

IBM SPSS 29 was used for all statistical analyses. Accuracy percentages for each specialty and for the total question set were determined. Any differences in accuracy across specialties were assessed by unpaired chi-square tests. Additional unpaired chi-square tests were then conducted to compare the accuracy of individual specialties to one another. The mean accuracy for student performance was determined for each specialty and for the total question set. Additionally, accuracy was compared to the amount of multiple choice questions available using unpaired chi-square tests.

3 Results

ChatGPT-4 answered 89% of the total questions accurately. Table 1 demonstrates the AI and student accuracy for each specialty. There was a significant difference of accuracy among specialties [$\times 2(8) = 18.64$, $p = 0.017$]. Specifically, the accuracy of pediatrics questions were significantly less than psychiatry ($p = 0.004$), obstetrics and gynecology ($p = 0.009$), and neurology ($p = 0.019$). Emergency medicine was also significantly lower than psychiatry ($p = 0.007$), obstetrics and gynecology ($p = 0.015$), and neurology ($p = 0.030$). Family medicine was significantly lower than obstetrics and gynecology ($p = 0.038$), and psychiatry ($p = 0.018$). No notable difference in accuracy was observed between questions with variable numbers of available answer options [$\times 2(8) = 12.89$, $p = 0.116$].



A 6-year-old boy is brought to the emergency department by his mother because of the acute onset 1 hour ago of severe right-sided lumbar pain and three episodes of emesis. The pain is sharp and causes him to double over and cry. Medical and family histories are unremarkable. The patient takes no medications. Temperature is 36.7°C (98.0°F), pulse is 150/min, respirations are 30/min, and blood pressure is 110/70 mm Hg. He appears to be in severe pain. Abdominal examination discloses right upper quadrant tenderness to palpation. The remainder of the examination discloses no abnormalities. Intravenous morphine is administered and results in improvement in the patient's pain. Results of dipstick urinalysis are shown:

Specific gravity: 1.035	(N=1.003–1.029)
pH:	5.5
Protein:	30 mg/dL
Glucose:	Negative
Bilirubin:	Negative
Blood:	4+
Leukocyte esterase:	1+
Nitrite:	Negative
Urobilinogen:	Negative
WBCs:	5–10/hpf
RBCs:	50–100/hpf

Renal ultrasonography shows severe right-sided hydronephrosis with moderate hydroureter to the bladder. Which of the following is the most appropriate next step in evaluation?

A. Captopril renography
B. CT scan of the abdomen and pelvis
C. MAG-3 renal scan with furosemide
D. Radionuclide cystography
E. Retrograde pyelography
F. Ultrasonography of the spinal column
G. Voiding cystourethrography

Fig. 1 Conversation input within the ChatGPT-4 interface. Sample question obtained from USMLE Step 2 Orientation

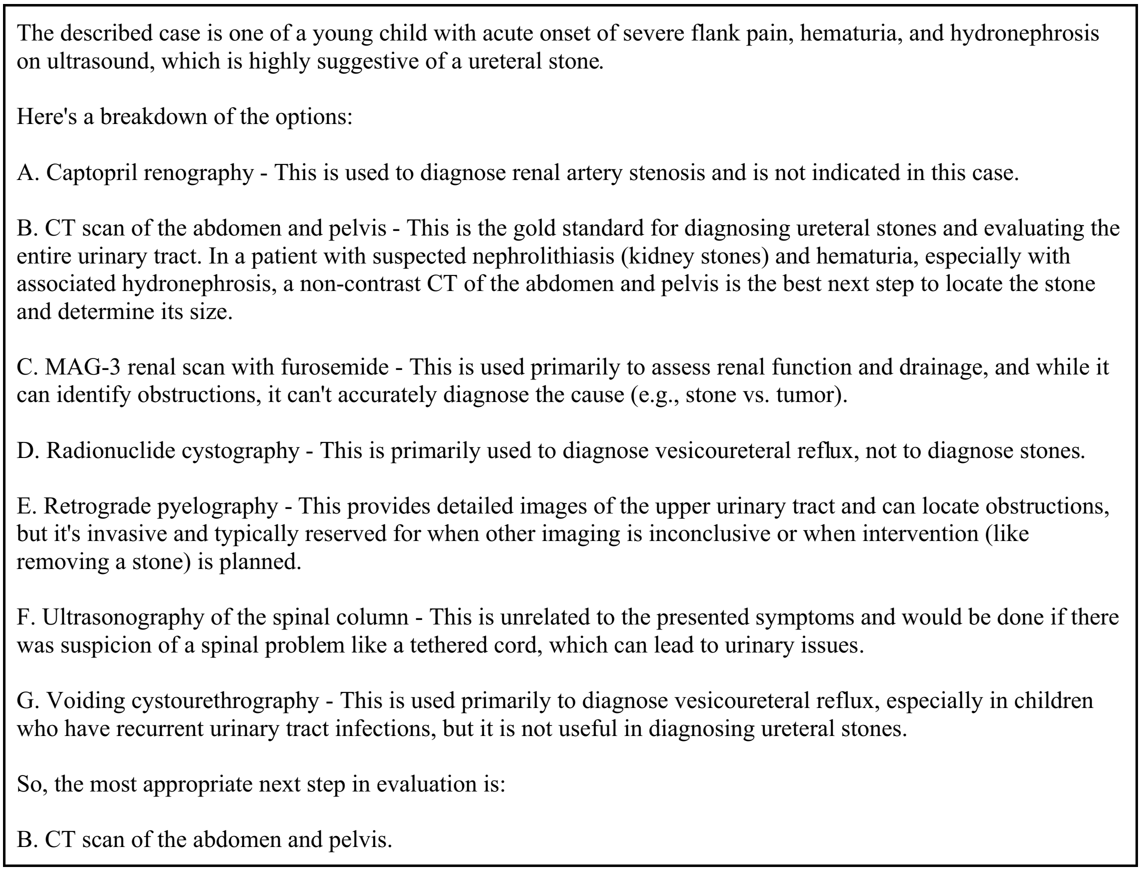
4 Discussion

ChatGPT-4 was found to be proficient in answering clinical subject exam questions with an overall accuracy of 89%. Further investigation, though, revealed specialty-specific performance discrepancies. This is of particular importance to medical students considering the use of AI-based tools to enhance their preparation for both shelf subject and clinical knowledge board examinations.

ChatGPT-4 had an impressive performance in areas like psychiatry, neurology, obstetrics and gynecology, but its accuracy was notably lower in pediatrics, emergency medicine, and family medicine. There are several factors that should be taken into consideration to explain this performance variation. Inherently, performance is influenced by the comprehensiveness and timeframe of the data used to train the model. For example, responses for specialties that are more represented in the data would be expected to be more accurate than those that are not. Furthermore, because the training data for ChatGPT-4 was extracted in 2021, specialties that have changed their recommendations for medical conditions over the past 2 years would be expected to result in outdated AI responses [13]. Although the depth of the training data remains a point of discussion, it was encouraging to note no variations in accuracy based on the number of multiple-choice options provided. This suggests that the AI's responses are authentic and not random answer selections.

Building upon this principle, we observed that the specialties with the lowest AI performance were those with significant interdisciplinary overlap in their questions. For instance, specialties like family medicine, emergency medicine, and pediatrics, which frequently feature complex, multifaceted clinical scenarios that require intricate clinical reasoning skills, displayed the lowest performance. Conversely, specialties such as psychiatry, obstetrics and gynecology, and neurology, have the narrowest scope and showcased the best AI performance among the fields we assessed.

Given these findings, one can infer the broader implications for the application of AI in medical education. Although the current version of ChatGPT displays variable performance across specialties, its overall efficacy suggests immense potential as an adjunct in medical school curricula. However, due to these observed inconsistencies, it becomes crucial for medical students, especially those in the clinical phase of their studies, to exercise caution when incorporating AI into their studies.



The described case is one of a young child with acute onset of severe flank pain, hematuria, and hydronephrosis on ultrasound, which is highly suggestive of a ureteral stone.

Here's a breakdown of the options:

A. Captopril renography - This is used to diagnose renal artery stenosis and is not indicated in this case.

B. CT scan of the abdomen and pelvis - This is the gold standard for diagnosing ureteral stones and evaluating the entire urinary tract. In a patient with suspected nephrolithiasis (kidney stones) and hematuria, especially with associated hydronephrosis, a non-contrast CT of the abdomen and pelvis is the best next step to locate the stone and determine its size.

C. MAG-3 renal scan with furosemide - This is used primarily to assess renal function and drainage, and while it can identify obstructions, it can't accurately diagnose the cause (e.g., stone vs. tumor).

D. Radionuclide cystography - This is primarily used to diagnose vesicoureteral reflux, not to diagnose stones.

E. Retrograde pyelography - This provides detailed images of the upper urinary tract and can locate obstructions, but it's invasive and typically reserved for when other imaging is inconclusive or when intervention (like removing a stone) is planned.

F. Ultrasonography of the spinal column - This is unrelated to the presented symptoms and would be done if there was suspicion of a spinal problem like a tethered cord, which can lead to urinary issues.

G. Voiding cystourethrography - This is used primarily to diagnose vesicoureteral reflux, especially in children who have recurrent urinary tract infections, but it is not useful in diagnosing ureteral stones.

So, the most appropriate next step in evaluation is:

B. CT scan of the abdomen and pelvis.

Fig. 2 Conversation output within the ChatGPT-4 interface

Table 1 Accuracy percentages for ChatGPT-4 and medical students on specialty-specific subject examination questions

Specialty	AI accuracy (%)	Student accuracy (%)
Psychiatry	95	71
Obstetrics and gynecology	94	61
Neurology	93	66
Ambulatory medicine	90	59
Internal medicine	90	59
Surgery	90	62
Family medicine	85	64
Emergency medicine	83	62
Pediatrics	82	63
Overall	89	63

As we make this recommendation, it remains pivotal to consider the limitations of our study methods. A notable constraint was in our question selection, such that not all of the topics within specialties were examined and image-based questions were not assessed. Furthermore, our study did not assess the specific characteristics of questions where ChatGPT4 faltered. Future studies should take these considerations into account in order to further enhance recommendations for medical students who intend to use AI to support their education.

5 Conclusion

LLMs like ChatGPT-4 provide promise for the future landscape of medical education. Nonetheless, our findings emphasize the need for caution. At this time, medical students should be aware of the model's specialty-specific strengths and weaknesses and use a well-rounded approach to their clinical curriculum.

Author contributions Conceptualization: BM, JC, RG. Methodology: BM, JC, RG. Validation: BM, RG. Formal analysis: RG. Data Curation: AT, BM, JC, LM, RG. Investigation: AT, LM. Resources: BM, RG. Writing—original draft: BM, RG. Writing—Review and editing: BM, RG, JC. Visualization: BM. Supervision: MW. Project Administration: BM, RG.

Funding No funding was received for this study. The authors have no financial or non-financial interests to disclose.

Data availability The dataset obtained and analyzed during the current study is available from the corresponding author upon reasonable request.

Declarations

Competing interests The authors declare no competing interests.

References

1. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT—reshaping medical education and clinical management. Pak J Med Sci. 2023;39:605–7. https://doi.org/10.12669/pjms.39.2.7653.
2. Ray PP. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Phys Syst. 2023. https://doi.org/10.1016/j.iotcps.2023.04.003.
3. Jeyaraman M, Jeyaraman N, Nallakumarasamy A, Yadav S, Bondili SK. ChatGPT in medical education and research: a boon or a bane? Cureus. 2023;29:44316–10. https://doi.org/10.7759/cureus.44316.
4. Grabb D. ChatGPT in medical education: a paradigm shift or a dangerous tool? Acad Psychiatry. 2023. https://doi.org/10.1007/s40596-023-01791-9.
5. Lee H. The rise of ChatGPT: exploring its potential in medical education. Anat Sci Educ. 2023. https://doi.org/10.1002/ase.2270.
6. Feng S, Shen Y. ChatGPT and the future of medical education. Acad Med. 2023;98(8):867–8. https://doi.org/10.1097/ACM.0000000000005242.
7. Kung TH, Cheatham M, Medenilla A, Sillos C, Leon LD, Elepaño C, Madriaga M, Aggabao R, DiazCandido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2:0000198. https://doi.org/10.1371/journal.pdig.0000198.
8. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. 2023;9:45312. https://doi.org/10.2196/45312.
9. Mbakwe AB, Lourentzou I, Celi LA, Mechanic OJ, Dagan A. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. PLOS Dig Health. 2023;9:0000205. https://doi.org/10.1371/journal.pdig.0000205.
10. Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, McCoy AB, Sittig DF, Wright A. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc. 2023;30:1237–45. https://doi.org/10.1093/jamia/ocad072.
11. Pugliese G, Maccari A, Felisati E, Felisati G, Giudici L, Rapolla C, Pisani A, Saibene AM. Are artificial intelligence large language models a reliable tool for difficult differential diagnosis? An a posteriori analysis of a peculiar case of necrotizing otitis externa. Clin Case Rep. 2023;11:7933. https://doi.org/10.1002/ccr3.7933.
12. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res. 2023;28:48568. https://doi.org/10.2196/48568.
13. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, Moy L. ChatGPT and other large language models are double-edged swords. Radiology. 2023. https://doi.org/10.1148/radiol.230163.