Open Access Full Text Article

ORIGINAL RESEARCH

Comparing the Performance of ChatGPT-4 and Medical Students on MCQs at Varied Levels of Bloom's Taxonomy

Ambadasu Bharatha [1], Nkemcho Ojeh [1], Ahbab Mohammad Fazle Rabbi [2],
Michael H Campbell [1], Kandamaran Krishnamurthy [1], Rhaheem NA Layne-Yarde [1],
Alok Kumar [1], Dale CR Springer [1], Kenneth L Connell [1], Md Anwarul Azim Majumder [1]

[1]Faculty of Medical Sciences, The University of the West Indies, Bridgetown, Barbados; [2]Department of Population Sciences, University of Dhaka, Dhaka, Bangladesh

Correspondence: Md Anwarul Azim Majumder, Director of Medical Education, Faculty of Medical Sciences, The University of the West Indies, Cave Hill Campus, Barbados, Email azim.majumder@cavehill.uwi.edu; Ambadasu Bharatha, Lecturer in Pharmacology, Faculty of Medical Sciences, The University of the West Indies, Cave Hill Campus, Barbados, Email ambadasu.bharatha@cavehill.uwi.edu

Introduction: This research investigated the capabilities of ChatGPT-4 compared to medical students in answering MCQs using the revised Bloom's Taxonomy as a benchmark.

Methods: A cross-sectional study was conducted at The University of the West Indies, Barbados. ChatGPT-4 and medical students were assessed on MCQs from various medical courses using computer-based testing.

Results: The study included 304 MCQs. Students demonstrated good knowledge, with 78% correctly answering at least 90% of the questions. However, ChatGPT-4 achieved a higher overall score (73.7%) compared to students (66.7%). Course type significantly affected ChatGPT-4's performance, but revised Bloom's Taxonomy levels did not. A detailed association check between program levels and Bloom's taxonomy levels for correct answers by ChatGPT-4 showed a highly significant correlation (p<0.001), reflecting a concentration of "remember-level" questions in preclinical and "evaluate-level" questions in clinical courses.

Discussion: The study highlights ChatGPT-4's proficiency in standardized tests but indicates limitations in clinical reasoning and practical skills. This performance discrepancy suggests that the effectiveness of artificial intelligence (AI) varies based on course content.

Conclusion: While ChatGPT-4 shows promise as an educational tool, its role should be supplementary, with strategic integration into medical education to leverage its strengths and address limitations. Further research is needed to explore AI's impact on medical education and student performance across educational levels and courses.

Keywords: artificial intelligence, ChatGPT-4's, medical students, knowledge, interpretation abilities, multiple choice questions

Introduction

Academic interest in artificial intelligence (AI) is surging, but integration of technology and AI in pedagogical settings has been uneven.[1] ChatGPT (Chat Generative Pre-trained Transformer), a state-of-the-art language model, has emerged as a potential tool in medical education.[2] This tool operates primarily through prompt interpretation and is capable of producing reasoned responses that are difficult to distinguish from human-produced language.[3] Its intrinsic transformer architecture also enables ChatGPT to be proficient in understanding natural language. ChatGPT generates responses using models on a neural network that intelligently processes data and adapts to new information continuously.[4] Such novel technologies provide promising opportunities for improvements in teaching and learning.[4,5] AI may be especially useful to promote student engagement, as existing teaching and delivery methods are persistently met with challenges in this area.[5] Given the increasing reliance on AI in educational settings, there is a need to evaluate its performance against established educational benchmarks.

In the context of medical education, where precision and depth of understanding can have crucial implications, evaluating AI's efficacy and relationship to relevant outcomes is essential.[6,7] Moreover, comparing the performance of AI models like ChatGPT to that of medical students may provide insights into areas where AI can complement human learning and, importantly, where it might fall short[8,9] Several studies have investigated the capabilities of AI in medical education, exploring both its potential and limitations.[10–13] Analyzing MCQs in medical education is crucial as it allows educators to assess the effectiveness of questions in testing higher-order thinking and clinical reasoning skills, ensuring assessments accurately reflect the skills required for medical practice.[14–16] While AI's capabilities in answering queries and simulating scenarios are noteworthy, the depth and breadth of its understanding, especially concerning multiple-choice questions (MCQs) of medical exams, still need to be thoroughly evaluated.[17] Several studies have demonstrated that ChatGPT outperforms medical students on MCQ items in board and licensing exams.[18–20] It is important to highlight the significance of MCQs in medical licensing exams, as they are extensively utilized in crucial assessments such as the United States Medical Licensing Examination (USMLE), Medical Council of Canada Qualifying Examination (MCCQE), United Kingdom Medical Licensing Assessment (UKMLA), and Australian Medical Council (AMC) Exam.[9,20,21] This widespread use is attributed to MCQs' effectiveness in evaluating higher-order skills through complex clinical scenarios, analyzing, and problem-solving. These questions assess students' ability to integrate information, reflecting real-world challenges, and shaping competent professionals.[14–16]

A systematic analysis using Bloom's Taxonomy as a benchmark can offer a more structured and comprehensive understanding of AI's role in this domain.[22] Bloom's Taxonomy, a foundational educational framework, categorizes cognitive learning objectives into hierarchical levels, ranging from basic knowledge recall to complex evaluation and synthesis.[23] Bloom's taxonomy serves as a guideline for educators in curriculum development and also provides a structured approach to assess the cognitive depth of questions and responses.[24]

Against this background, we proposed to identify gaps by analyzing and comparing exam scores in medical sciences, within the framework of the revised Bloom's Taxonomy,[10] to evaluate the performance of ChatGPT-4 compared to that of medical students. The main aim of this research was to perform a comparative evaluation of ChatGPT-4's and medical students' knowledge and interpretation skills, employing MCQs from exam papers of basic science courses and clinical clerkships courses. In particular, the following were investigated:

- scores of ChatGPT-4 compared to those of medical students;
- correct answer rate of ChatGPT-4 according to item knowledge level, assessing how well AI can adapt to various cognitive demands; and
- acceptability of ChatGPT-4's explanations to reflect current basic and clinical medical knowledge, to understand the practicality and reliability of AI-generated content in medical education.

Materials and Methods

This cross-sectional study was conducted at the Faculty of Medical Sciences (FMS), The University of the West Indies, Cave Hill Campus, Barbados.

We treated ChatGPT-4 uniquely as a single examinee for this study. We incorporated the exam grades for Year 1 (n=51) and Year 3 (n=46) MBBS students. The following courses were incorporated:

1. Year 1 MBBS - Basic Medical Sciences:
   a. Cell Biology (MDSC1201)
   b. Introduction to Molecular Medicine (MDSC1104)
   c. Respiratory System (MDSC1205)
2. Year 3 MBBS - Clinical Medical Sciences
   a. Junior Medicine Clerkship (MDSC3201)
   b. Junior Surgery Clerkship (MDSC3202)
   c. Aspects of Family Medicine (MDSC3203)

The MCQ midterm papers (for basic medical sciences) and end-of-clerkship exams (for clinical sciences) were administered to ChatGPT-4 using computer-based testing with identical items to those that were administered to first- and

third-year medical students, respectively. To ensure an unbiased comparison, the exam items for ChatGPT-4 were kept identical to those given to the medical students.

AB and NO classified the questions using the updated Bloom's taxonomy. To maintain uniformity and avoid inconsistencies, AM re-checked the classification process. Next, AB and NO presented the MCQs to ChatGPT and recorded their answers. A thorough analysis was then performed by AB, NO, and AM, to analyze the ChatGPT-4's responses and compare with those provided by medical students.

The included courses commenced on September 5, 2022, and concluded on April 14, 2023. During this period, the curriculum included the following number of lecture and laboratory practice hours, respectively: MDSC1201 – 38 and 4; MDSC1104 – 30 and 4; and MDSC1205 – 28 and 12.

Data Analysis

We compared student and ChatGPT-4 performance using the average percent answered correctly by course type and student education level (year in MBBS program). Further, each course was assessed considering its purpose according to revised Bloom's Taxonomy using the following 6 levels: remembering, understanding, applying, analyzing, evaluating, and creating.[23] Univariate analysis (frequency distribution) was conducted to measure each of the study variables. We conducted bivariate analyses to examine the association between selected background variables and correct responses. The data analyses were performed using SPSS software (version 26).

Ethical Approval

We applied for ethical approval from The University of the West Indies-Cave Hill/Barbados Ministry of Health Research Ethics Committee/Institutional Review Board. The study was exempted under "Category 2" by the IRB, which includes surveys, interviews, educational tests, and public observation studies (Ref: CREC-CH.00173/03/2023).

Results

Of the 332 questions, 304 were analyzed. Twenty-eight questions containing images were excluded because they were not supported by ChatGPT-4 (in April 2023). The excluded questions were from the Junior Medicine Clerkship (MDSC3201). Seventy-eight percent of the students answered 90% of the problems correctly. However, the overall performance of students was lower (66.7%) than that of ChatGPT (73.7%). The performance of ChatGPT-4 is summarized in Table 1.

According to the revised Bloom's Taxonomy[23] levels, 109 of the items were for "remember" (35.9%), 39 were for "understanding" (12.8%), 44 were for "apply" (14.5%), 27 were for "analysis" (8.9%), and the remaining 85 were for "evaluation" (28%). There were no questions at the "create" level.

We performed bivariate analysis to find the association of ChatGPT-4 performance with course type and Bloom's Taxonomy level. The results are summarized in Table 2. We found that course type was significantly associated with ChatGPT-4 performance (p = 0.011). However, the association between ChatGPT-4 performance and revised Bloom's Taxonomy levels was found to be insignificant (p=0.577).

Next, we examined the association between program levels and revised Bloom's Taxonomy only for the MCQs ChatGPT-4 answered correctly. The findings are given in Table 3. For correct answers given by ChatGPT-4, a highly significant association between Bloom's Taxonomy and course level was found (p = 0.000).

Table 1 Overall Performance Metrics of ChatGPT-4

Performance	Frequency	Percentage (%)
Incorrect	80	26.3
Correct	224	73.7
Total	304	100.0

Table 2 Relationship of Program Levels, Courses, Bloom's Taxonomy Levels and Performance of ChatGPT-4

Background information	Response of ChatGPT		χ^2
	Correct	**Incorrect**	
Program levels			
Preclinical	101 (77.7%)	29 (22.3%)	χ^2 (1, n=304) = 1.882, p=0.107
Clinical	123 (70.7%)	51 (29.3%)	
Course titles			
Cell Biology	41 (93.2%)	3 (6.8%)	χ^2 (5, n=304) = 14.865, p=0.011
Introduction to Molecular Medicine	36 (78.3%)	10 (21.7%)	
Respiratory System	24 (60.0%)	16 (40.0%)	
Junior Medicine Clerkship	16 (66.7%)	8 (33.3%)	
Junior Surgery Clerkship	69 (69.0%)	31 (31.0%)	
Aspects of Family Medicine	38 (76.0%)	12 (24.0%)	
Bloom's Taxonomy Levels			
Remember	82 (75.2%)	27 (24.8%)	χ^2 (4, n=304) = 2.889, p=0.577
Understand	29 (74.4%)	10 (25.6%)	
Apply	29 (65.9%)	15 (34.1%)	
Analyse	18 (66.7%)	9 (33.3%)	
Evaluate	66 (77.6%)	19 (22.4%)	
Create	0	0	

Table 3 Association Between Bloom's Taxonomy and Program Levels for the Correct Answers from ChatGPT-4

Bloom's Taxonomy Levels	Correct answers		χ^2
	Preclinical	**Clinical**	
Remember	61 (74.4%)	21 (25.6%)	χ^2 (4, n=224) = 91.978, p<0.001
Understand	21 (72.4%)	8 (27.6%)	
Apply	11 (37.9%)	18 (62.1%)	
Analyse	8 (44.4%)	10 (55.6%)	
Evaluate	0	66 (100%)	
Create	0	0	

The majority of correct remember-level MCQs (74.4%) were found in preclinical courses. All correct evaluation-level MCQs (100%) were found in clinical courses.

Discussion

Emerging educational scholarship has highlighted the potential value of AI for teaching, learning, and assessment in medical education.[17,25,26] However, AI has important limitations. The ability of AI, as demonstrated by ChatGPT, to

excel in standardized tests does not necessarily translate to clinical expertise. While AI can assist in data analysis, pattern recognition, and even suggesting diagnoses or treatments based on large data sets, it currently lacks the human elements of empathy, ethical judgment, and the ability to understand the subtleties of patient communication and cultural context.[9] Further, research and clinical medicine are dynamic, requiring continuous learning and adaptation that goes beyond the static knowledge base of an AI trained on past data.[12]

The current study suggests that the performance of ChatGPT-4 on MCQ items exceeds that of medical students, a finding consistent with other recent research.[21,27–29] For example, ChatGPT versions 3.5 and 4 scored higher than the students on the American Board of Neurological Surgery exam.[18] ChatGPT also passed the German state licensing exam for medicine, outperforming most medical students.[19] ChatGPT performed well (76.3%) on the UKMLA.[20] In contrast, the performance of ChatGPT vs dental students on a medical microbiology MCQ exam found that ChatGPT 3.5 correctly answered 64 out of 80 MCQs (80%), scoring 80.5 out of 100 which was below the student average of 86.21 out of 100.[29]

A recent study examining ChatGPT responses to the USMLE-type questions demonstrated that each response from ChatGPT, whether correct or incorrect, demonstrated a degree of reasoning.[21] Another study tasked ChatGPT to respond to complex pathology questions necessitating advanced reasoning. The AI achieved an impressive 80% success rate across a range of pathology subjects, showcasing its capabilities in critical thinking.[30] Both studies were conducted using an earlier version of ChatGPT, and more recent iterations are likely to exhibit enhanced performance.[27,31] A scoping review by Newton and Xiromeriti[32] revealed that ChatGPT's performance varied across different evaluation methods and subjects, with ChatGPT 3 passing 20.3% and ChatGPT-4 passing 92.9% of exams. Notably, ChatGPT 3 outperformed human students in 10.9% of exams, while ChatGPT 4 did so in 35%, indicating significant performance improvement in the more advanced version. Brin and colleagues[9] observed that GPT-4 outperformed ChatGPT in answering USMLE MCQs related to soft skills like empathy, ethics, and judgment, with a 90% correct response rate compared to ChatGPT's 62.5%. Agarwal et al[5] also found that Claude-2, an alternative Generative AI system, outperformed ChatGPT-3.5 by correctly answering 40 medical physiology MCQs compared to 26 by ChatGPT-3.5. It also received significantly higher ratings for its explanations.

ChatGPT was previously reported to perform well on both basic and clinical medical science examinations.[2] A recent study by Sallam et al noted that the performance of ChatGPT varied across different cognitive domains, with the best performance in the "Remember" domain and the weakest in the "Evaluate" domain,[29] which was the case for preclinical courses (74.4% of the correct answers) in our study, but not for clinical courses (25.6% of the correct answers) in remember-level MCQs. The evaluate-level MCQs for clinical exams in our study showed the best performance (100% of the correct answers – see Table 3). In a study of psychosomatic medicine exam items, Herrmann-Werner et al[33] found that ChatGPT-4 made the most mistakes for the 'remember' (42.6%) and 'understand' (33.8%) categories; however, the error rates for both levels were higher than found in the current study (25.5% and 27.6%, respectively). Sallam et al[29] reported varied ChatGPT 3.5 performance across cognitive domains – best in the "Remember" (88.5% correct answers) and in "Understand" (82.4%) domains but decreased performance in higher-order domains. In sum, existing findings regarding ChatGPT performance are mixed, with some emerging patterns of differences associated with cognitive and content domains. Hence, as ChatGPT and other AI technologies are rapidly developing, continuous evaluation of AI models is needed.[1,12]

Clinical questions are known to require the application of knowledge relevant to the unique clinical situation of the patient in question. In the present study, the lack of a significant association with Bloom's Taxonomy levels suggests that ChatGPT-4's performance does not markedly differ across cognitive domains from recall to evaluation. This finding is congruent with a previous study in which ChatGPT's performance also did not correlate with the knowledge level of MCQs.[8] These findings suggest that technical progress in ChatGPT development has increased utility for application to medicine. Moreover, the consistent proficiency observed across all levels of the cognitive hierarchy can be attributed to the increased reasoning capabilities of the language model.

Our study found a significant association between ChatGPT-4 performance and course level (year of medical school). This implies that AI's effectiveness in providing correct answers is influenced by the course contents. ChatGPT-4's enhanced performance in the Cell Biology course (MDSC1201) can be attributed to the structured and factual nature of the material, aligning with AI's strengths. In contrast, courses where AI showed lower performance likely involve more complex clinical reasoning or practical skills, which are more challenging for AI to replicate.

The high association between Bloom's Taxonomy and course level for correct answers in our study indicates that the course level might mediate the relationship between the type of cognitive skill assessed and the likelihood of a correct response by ChatGPT-4. Despite its promise, concerns about the reliability of ChatGPT and other AI tools, especially for high-stakes medical decision-making, are important to address empirically as AI technology continues to develop. Establishing lower limits of accuracy required for appropriate clinical use is crucial. Some researchers have maintained that unless ChatGPT consistently achieves 95% accuracy or higher in medical tests, it should not be used without supervision. For example, in a recent study, ChatGPT scored 76% on medical pharmacology MCQs.[34] The total number of MCQ options associated with the accuracy of responses. This illustrates the non-trivial current limitations of ChatGPT and the importance of defining parameters for appropriate future use, especially in clinical settings. Indeed, the authors concluded that, unless ChatGPT consistently achieves 95% accuracy or higher in medical tests, it should not be used without supervision.

The present study adds to the growing literature on the implications of AI for medical education[1,35] and positions AI as a valuable educational tool. As AI is rapidly evolving, responsive research and professional development for medical educators is needed to maximize the benefits (and minimize risks) of AI in learning and teaching. This study contributes to understanding of the application of AI in medical education, balancing its strengths and limitations. We intend to utilize ChatGPT-4 to evaluate MCQs to further explore its potential application in student assessment. This includes using it to provide automated feedback on student responses, generate customized practice questions, and possibly assist in grading student assessments. Leveraging ChatGPT-4 or similar tools in this manner hold promise to improve the efficiency and effectiveness of student assessment processes. Regarding the breakdown of items according to the revised Bloom's Taxonomy, we found that 109 items were categorized under "remember" (35.9%), while 39 were classified under "understanding" (12.8%). Based on this, we plan to revise our MCQs to include higher-order items to align with the revised Bloom's Taxonomy levels. The strategic use of AI can augment traditional learning, making education more efficient and comprehensive. Future research efforts are needed to broaden the technical, pedagogical, and ethical application of AI in medical education.[12,17,26]

The study has several limitations, including the use of ChatGPT-4, which may not have capabilities found in newer AI models. The focus on MCQs might bias results in favour of known strengths of ChatGPT-4. The cross-sectional design does not permit causal attributions or longitudinal analysis of learning trajectories. Sampling was limited to a single institution and may not be generalizable to other medical education settings. Additionally, potential biases in MCQ selection and evaluation criteria could influence outcomes. These factors suggest that findings should be cautiously interpreted and call for further research in diverse educational settings. Finally, the study's narrow focus on MCQs may not fully capture the diverse range of learning experiences in medical training, such as clinical skills, patient interactions, and critical thinking exercises, thereby limiting its relevance and broad applicability to medical education.

Conclusion

The study demonstrates that ChatGPT-4 exhibits high accuracy in answering MCQs compared to the performance of medical students. AI should be viewed as a promising supplementary tool to medical education as ChatGPT and similar technologies develop and are continuously evaluated. The findings suggest that AI's effectiveness varies with the course content and levels. ChatGPT-4 demonstrated proficiency across different cognitive levels but struggled with complex clinical reasoning and practical skills. This study underscores the need for strategic integration of AI in medical training, highlighting its strengths in knowledge dissemination and limitations in clinical judgment. Further exploration is needed to understand AI's nuanced impact on medical education and student performance, particularly in areas where AI cannot replicate human expertise.

Disclosure

Dr. Md Anwarul Azim Majumder is the Editor-in-Chief of *Advances in Medical Education and Practice*. The other authors report no conflicts of interest in this work.

References

1. Chen X, Xie H, Zou D, Hwang G-J. Application and theory gaps during the rise of artificial intelligence in education. *Artl Intel*. 2020;1:100002.
2. Meo SA, Al-Masri AA, Alotaibi M, Meo MZS, Meo MOS. ChatGPT knowledge evaluation in basic and clinical medical sciences: Multiple choice question examination-based performance. *Health care*. 2023;11(14). doi:10.3390/healthcare11142046

3. Ignjatović A, Stevanović L. Efficacy and limitations of ChatGPT as a biostatistical problem-solving tool in medical education in Serbia: A descriptive study. *J Educ Eval Health Prof*. 2023;20:28. doi:10.3352/jeehp.2023.20.28

4. Roos J, Kasapovic A, Jansen T, Kaczmarczyk R. Artificial Intelligence in medical education: Comparative analysis of ChatGPT, Bing, and medical students in Germany. *JMIR Med Educ*. 2023;9(e46482):e46482. doi:10.2196/46482

5. Agarwal M, Goswami A, Sharma P. Evaluating ChatGPT-3.5 and Claude-2 in answering and explaining conceptual medical physiology multiple-choice questions. *Cureus*. 2023;15(9):e46222.

6. Khorshidi H, Mohammadi A, Yousem DM, et al. Application of ChatGPT in multilingual medical education: how does ChatGPT fare in 2023's Iranian residency entrance examination. *Inf Med Unlocked*. 2023;41:101314.

7. Cheung BHH, Lau GKK, Wong GTC, et al. ChatGPT versus human in generating medical graduate exam multiple choice questions-A multinational prospective study (Hong Kong S.A.R. Singapore, Ireland, and the United Kingdom). *PLoS One*. 2023;18(8): e0290691. doi:10.1371/journal.pone.0290691

8. Anderson, LW Krathwohl, DR. A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives New York: Longman, 2021.

9. Brin D, Sorin V, Vaid A, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Sci Rep*. 2023;13(1):16492. doi:10.1038/s41598-023-43436-9

10. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. *J Educ Eval Health Prof*. 2023;20:1. doi:10.3352/jeehp.2023.20.1

11. Buabbas AJ, Miskin B, Alnaqi AA, et al. Investigating students' perceptions towards artificial intelligence in medical education. *Healthcare*. 2023;11(9):1298. doi:10.3390/healthcare11091298

12. Mir MM, Mir GM, Raina NT, et al. Application of artificial intelligence in medical education: Current scenario and future perspectives. *J Adv Med Educ Prof*. 2023;11(3):133–140. doi:10.30476/JAMP.2023.98655.1803

13. Li Q, Qin Y. AI in medical education: medical student perception, curriculum recommendations and design suggestions. *BMC Medical Education*. 2023;23(1):852. doi:10.1186/s12909-023-04700-8

14. Kumar A, George C, Harry Campbell M, et al. Item analysis of multiple choice and extended matching questions in the final MBBS medicine and therapeutics examination. *J Med Edu*. 2022;21(1):e129450. doi:10.5812/jme-129450

15. Epstein RM, Cox M, Irby DM. Assessment in medical education. *N Engl J Med*. 2007;356(4):387–396. doi:10.1056/NEJMra054784

16. van der Vleuten C. Validity of final examinations in undergraduate medical training. *BMJ*. 2000;321(7270):1217–1219. doi:10.1136/bmj.321.7270.1217

17. Sallam M, Salim N, Barakat M, Al-Tammemi A. ChatGPT applications in medical, dental, pharmacy, and public health education: a descriptive study highlighting the advantages and limitations. *Narra J*. 2023;3:e103.

18. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. *Neurosurgery*. 2023;93 (6):1353–1365. doi:10.1227/neu.0000000000002632

19. Friederichs H, Friederichs WJ, März M. ChatGPT in medical school: how successful is AI in progress testing? *Med Educ Online*. 2023;28 (1):2220920. doi:10.1080/10872981.2023.2220920

20. Lai UH, Wu KS, Hsu TY, Kan JKC. Evaluating the performance of ChatGPT-4 on the United Kingdom medical licensing assessment. *Front Med Lausanne*. 2023;10:1240915. doi:10.3389/fmed.2023.1240915

21. 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

22. Park SH, Do KH, Kim S, Park JH, Lim YS. What should medical students know about artificial intelligence in medicine? *J Educ Eval Health Prof*. 2019;16:18. doi:10.3352/jeehp.2019.16.18

23. Lw A, Dr K, Pw A, et al. A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives;2001.

24. Leupen SM, Kephart KL, Hodges LC, Knight J. factors influencing quality of team discussion: discourse analysis in an undergraduate team-based learning biology course. *CBE Life Sci Educ*. 2020;19(1):ar7. doi:10.1187/cbe.19-06-0112

25. Varma JR, Fernando S, Ting BY, Aamir S, Sivaprakasam R. The global use of artificial intelligence in the undergraduate medical curriculum: a systematic review. *Cureus*. 2023;15(5):e39701. doi:10.7759/cureus.39701

26. Ibrahim H, Liu F, Asim R, et al. Perception, performance, and detectability of conversational artificial intelligence across 32 university courses. *Sci Rep*. 2023;13(1):12187. doi:10.1038/s41598-023-38964-3

27. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: comparison Study. *JMIR Med Educ*. 2023;9(e48002):e48002. doi:10.2196/48002

28. Johnson D, Goodman R, Patrinely J, et al. assessing the accuracy and reliability of AI-Generated medical responses: An evaluation of the Chat-GPT model. *Res Sq*. 2023.

29. Sallam M, Al-Salahat K. Below average ChatGPT performance in medical microbiology exam compared to university students. *Front Educ*. 2023;8.

30. Sinha RK, Deb Roy A, Kumar N, Mondal H. Applicability of ChatGPT in assisting to solve higher order problems in pathology. *Cureus*. 2023;15(2).

31. Yanagita Y, Yokokawa D, Uchida S, Tawara J, Ikusaka M. Accuracy of ChatGPT on medical questions in the national medical licensing examination in japan: Evaluation study. *JMIR Form Res*. 2023;7:e48023.

32. Newton P, Xiromeriti M. ChatGPT performance on multiple choice question examinations in higher education. A pragmatic scoping review. *Assess Eval High Educ*. 2024;1–18. doi:10.1080/02602938.2023.2299059

33. Herrmann-Werner A, Festl-Wietek T, Holderried F, et al. Assessing ChatGPT's Mastery of Bloom's Taxonomy Using Psychosomatic Medicine Exam Questions: Mixed-Methods Study. *J Med Internet Res*. 2024;26:e52113.

34. Choi W. Assessment of the capacity of ChatGPT as a self-learning tool in medical pharmacology: a study using MCQs. *BMC Med Educ*. 2023;23 (1):864. doi:10.1186/s12909-023-04832-x

35. Sallam M, Al-Salahat K, Al-Ajlouni E. ChatGPT performance in diagnostic clinical microbiology laboratory-oriented case scenarios. *Cureus*. 2023;15(12).