Comparative Evaluation of ChatGPT and Microsoft Copilot in Solving Clinical Vignette-style multiple-choice questions (MCQs) in Physiology

Prabhu R[a], Prabhu G[a], Holla R[b]
[a]Faculty of Medicine, Manipal University College Malaysia (MUCM)
[b] Kasturba Medical College Mangalore, Manipal Academy of Higher Education, Manipal, India

Corresponding Author: Associate Professor Dr Rekha Prabhu
Department of Physiology, Faculty of Medicine, Manipal University College Malaysia (MUCM), Melaka, Malaysia.
Email Id: rekha.prabhu@manipal.edu.my

ABSTRACT

INTRODUCTION: Large language models (LLMs) are increasingly used by MBBS students as supplementary resources for exam preparation. The objective of this study was to evaluate the performance of ChatGPT and Microsoft Copilot in answering clinical vignette-style physiology MCQs from widely used resources for the United States Medical Licensing Examination (USMLE). **MATERIALS AND METHODS:** Fifty clinical vignette-style physiology multiple choice questions (MCQs) from the various USMLE question banks were submitted to ChatGPT and Microsoft Copilot to choose the correct option. The performance of ChatGPT and Microsoft Copilot was assessed using the provided answers in the question bank. Two experienced physiologists independently reviewed the explanations provided by ChatGPT and Microsoft Copilot for each MCQ. The explanations were rated between one to three points based on whether the answers were completely incorrect, partially correct with inaccurate information, or correct with adequate information. **RESULTS:** ChatGPT and Microsoft Copilot both correctly answered 48 and 47 out of 50 questions, reflecting a 96% and 94% accuracy rates respectively. One MCQ each on hypothyroidism and arrhythmia was incorrectly answered by both ChatGPT and Microsoft Copilot. For two MCQs, the explanations provided were inaccurate by ChatGPT and Microsoft Copilot provided inaccurate explanations for four of the MCQs. **CONCLUSION:** ChatGPT and Microsoft Copilot both demonstrated more than 90% accuracy in answering case-based MCQs from the USMLE Step 1 resources. Their incorrect option choices MCQs on hypothyroidism and inaccurate explanations for some MCQs highlight cautious use of AI by students.

KEYWORDS: ChatGPT; Large language models; Microsoft Copilot; United States Medical Licensing Examination; USMLE

INTRODUCTION

AI-powered large language models (LLMs), such as OpenAI's ChatGPT and GPT-4, Google's Gemini, and Microsoft's Copilot, have demonstrated remarkable potential in answering multiple choice questions (MCQs) of varied Bloom's taxonomy levels.[1] ChatGPT is a general LLM developed recently by OpenAI and has garnered significant attention due to its ability to perform diverse tasks. In early 2023, OpenAI released an updated LLM called GPT-4 which introduced multimodal capabilities, including the ability to process images.[2] Kung et al reported that ChatGPT performed near the passing threshold for United States Medical Licensing Examination (USMLE) (Steps 1, 2, and 3) without any specialised training or reinforcement.[3] The Copilot chatbot, earlier known as Bing Chat Enterprise (Microsoft, 2023) has also gained popularity in recent years. ChatGPT-3.5, Microsoft Bing, and Google Bard are some of the free AI-based LLMs whose applications are being explored extensively in medical education. The capabilities of ChatGPT-3.5 and Claude-2 in medical physiology education using curated sets of MCQs were evaluated in the competency-based medical curriculum (CBME) of the Indian National Medical Commission (NMC).[4]

MCQs cover a broader range of topics in the curriculum, and clinical vignette-style MCQs can assess higher-order learning,[5] hence MCQs are used for professional licensing examinations for doctors, lawyers, social workers, and others. Previous studies have highlighted significant performance differences among LLM chatbots on medical licensing examinations worldwide.[3] Medical licensing examinations, such as the USMLE, require candidates to apply their foundational knowledge in basic and clinical sciences to solve clinical vignette-style MCQs. The USMLE is a high-stakes, comprehensive three-step standardised testing programme covering all topics in basic science, clinical reasoning, medical management, and bioethics.[3] The USMLE Step 1 assesses students' readiness for clinical training by emphasizing the integration of physiology, pathology, and pharmacology. The Step 1 exam is taken by Bachelor of Medicine, Bachelor of Surgery (MBBS) students who have completed two years of learning that focuses on basic science, pharmacology, and pathophysiology.[6] Generative AI has demonstrated the ability to pass the USMLE which includes questions on communication skills, ethics, empathy, and professionalism.[7] The basic principles governing cardiac electrophysiology and myocardial contractility are applied in clinical scenarios, and fundamental knowledge of cardiovascular physiology is essential for managing patients during the perioperative period.[8] The cardiovascular system and endocrine system were selected for vignette-style assessments due to their emphasis on clinical application.

Microsoft Copilot has not been extensively studied in medical contexts, and its performance in answering basic science and clinical science MCQs remains underexplored, making this study a valuable contribution to evaluating its application in medical education. The objective of the study was to evaluate the performance of ChatGPT and Microsoft Copilot in answering clinical vignette-style physiology multiple-choice questions from widely used resources for board exam preparation in the United States.

Study tools

Our study used the GPT-4, free version of ChatGPT (developed by OpenAI, headquartered in San Francisco, California, USA, and is accessible at: https://chat.openai.com) and Microsoft Copilot (developed by Microsoft Corporation, headquartered in Redmond, Washington, USA, and is available via Microsoft 365 applications and at: https://copilot.microsoft.com). The answers provided by both AI platforms were assessed and compared.

Study design

The cross-sectional observational study was carried out in January 2025 by the Department of Physiology. Fifty case-based clinical vignette-style physiology MCQs (25 questions each from endocrine system and cardiovascular system) were extracted from various question banks (UWorld (n=20), AMBOSS (n=16), and True Learn (n=14)). The questions were vetted independently by two physiologists for relevance, cognitive level, and clarity. The questions were submitted to ChatGPT and Microsoft Copilot by two independent users simultaneously. Each MCQ was entered as a separate conversation and the answers obtained were converted to plain text format. The responses for each of the MCQs were independently reviewed by two experienced physiologists. The correct options provided by ChatGPT and Microsoft Copilot were assessed using the provided answers in the question bank. The first response given by each tool was considered final, and the 'regenerate response' was not utilised. A score of '0' indicated an incorrect answer, while a score of '1' indicated a correct response.

Accuracy assessment

Two experienced physiologists independently rated the answers based on responses ranging from 1 to 3 points (1 indicates that the answer is completely incorrect, 2 indicates that part of the answer is correct but contains inaccurate information, and 3 indicates a correct and adequate answer). A total score (TS) greater than 5 indicated 'good' responses, 3≤TS≤5 indicated 'moderate' responses, and TS less than 3 indicated 'poor' responses. In cases of discrepancy in scoring or interpretation between the two experienced physiologists, a consensus was reached through discussion.

Prompt construction

The prompt used was "*Choose the correct option for the MCQ and give an explanation for the right answer.*"

Data analysis

Microsoft Excel was used for initial input of data and subsequently statistical analysis was performed using IBM SPSS Statistics for Windows, Version 27. Pearson's correlation coefficient was used to evaluate the associations between variables, with $p<0.05$ considered statistically significant.

RESULTS

A total of 50 MCQs covering various clinical conditions in endocrine and cardiovascular physiology were assessed. ChatGPT provided the correct answers for 48 out of 50 MCQs, achieving a 96% accuracy while Microsoft Copilot provided the correct answers for 47 out of 50 MCQs, achieving a 94% accuracy. The difference in correct and incorrect responses between ChatGPT and Microsoft Copilot was not significant. There is a moderate positive correlation between the scores, with a Pearson correlation coefficient of $r=0.640$, which is statistically significant ($p<0.001$).

There is a moderate positive correlation between the scores of ChatGPT and Microsoft Copilot in the cardiovascular system ($r=0.482$, $p=0.015$), indicating a statistically significant similarity in performance. There is a strong positive correlation between the scores of ChatGPT and Microsoft Copilot in the endocrine system ($r=0.891$, $p<0.001$), suggesting a highly consistent pattern of responses between the two AI models.

Table I illustrates the various MCQs tested in this study. For MCQ 5, Microsoft Copilot chose the correct option but did not provide an accurate explanation. For MCQs 7 and 26, both ChatGPT and Microsoft Copilot chose the incorrect option and provided an inaccurate explanation, hence scoring 0. For MCQ 47, Microsoft Copilot chose the incorrect option and provided an inaccurate explanation.

Table I: Comparison of ChatGPT and Microsoft Copilot in Answering MCQs on clinical conditions

MCQ No.	Clinical condition	ChatGPT		Microsoft Copilot	
		Answer option	Accuracy of answer explanation	Answer option	Accuracy of answer explanation
		Correct/ Incorrect	Accurate/ Inaccurate	Correct/Incorrect	Accurate/ Inaccurate
1	Brain tumour that impinges on the supraoptic nucleus	Correct	Accurate	Correct	Accurate
2	Tumours of the somatotrophs	Correct	Accurate	Correct	Accurate
3	Diabetes insipidus caused by a craniopharyngioma at the posterior pituitary stalk	Correct	Accurate	Correct	Accurate
4	SIADH	Correct	Accurate	Correct	Accurate
5	Conn syndrome	Correct	Accurate	Correct	Inaccurate
6	Primary hyperparathyroidism	Correct	Accurate	Correct	Accurate
7	Hypothyroidism	Incorrect	Inaccurate	Incorrect	Inaccurate
8	hypothyroidism	Correct	Accurate	Correct	Accurate
9	Short stature	Correct	Accurate	Correct	Accurate
10	Hypopituitarism	Correct	Accurate	Correct	Accurate
11	Stress	Correct	Accurate	Correct	Accurate
12	Steroid therapy	Correct	Accurate	Correct	Accurate
13	Acromegaly	Correct	Accurate	Correct	Accurate
14	Hypopituitarism	Correct	Accurate	Correct	Accurate
15	Phaeochromocytoma	Correct	Accurate	Correct	Accurate
16	Adenoma of the parathyroid gland	Correct	Accurate	Correct	Accurate
17	Hypovolemia	Correct	Accurate	Correct	Accurate
18	Cushing disease	Correct	Accurate	Correct	Accurate
19	Primary hypoparathyroidism	Correct	Accurate	Correct	Accurate
20	Graves' disease	Correct	Accurate	Correct	Accurate
21	Graves' disease	Correct	Accurate	Correct	Accurate
22	Addison disease	Correct	Accurate	Correct	Accurate
23	Primary hyperparathyroidism	Correct	Accurate	Correct	Accurate
24	Glucagonoma	Correct	Accurate	Correct	Accurate
25	Addison disease	Correct	Accurate	Correct	Accurate
26	Arrythmia	Incorrect	Inaccurate	Incorrect	Inaccurate
27	Heart failure	Correct	Accurate	Correct	Accurate
28	Supine hypotension	Correct	Accurate	Correct	Accurate
29	Coronary artery disease	Correct	Accurate	Correct	Accurate
30	Exercise treadmill test in diabetes mellitus	Correct	Accurate	Correct	Accurate
31	Chest discomfort	Correct	Accurate	Correct	Accurate
32	Dilated Cardiomyopathy	Correct	Accurate	Correct	Accurate
33	Paroxysmal supraventricular tachycardia (PSVT)	Correct	Accurate	Correct	Accurate
34	Atherosclerotic narrowing of the right renal artery	Correct	Accurate	Correct	Accurate
35	Carotid sinus hypersensitivity	Correct	Accurate	Correct	Accurate
36	Engorgement of the inferior vena cava	Correct	Accurate	Correct	Accurate
37	Heart catheterization	Correct	Accurate	Correct	Accurate
38	Atrial fibrillation (AF)	Correct	Accurate	Correct	Accurate
39	Chest pain	Correct	Accurate	Correct	Accurate
40	Upper GI Bleed	Correct	Accurate	Correct	Accurate
41	Exercise	Correct	Accurate	Correct	Accurate
42	Mitral Stenosis	Correct	Accurate	Correct	Accurate
43	Aortic regurgitation	Correct	Accurate	Correct	Accurate
44	Congestive heart failure	Correct	Accurate	Correct	Accurate
45	Complete Heart block	Correct	Accurate	Correct	Accurate

continue

continue

46	Aortic regurgitation	Correct	Accurate	Correct	Accurate
47	Cardiac transplantation for severe idiopathic cardiomyopathy	Correct	Accurate	Incorrect	Inaccurate
48	Lisinopril therapy	Correct	Accurate	Correct	Accurate
49	Abdominal aorta was constricted	Correct	Accurate	Correct	Accurate
50	ACE inhibitors	Correct	Accurate	Correct	Accurate

Table II illustrates the comparative analysis of the accuracy of the responses of ChatGPT and Microsoft Copilot to 50 MCQs. ChatGPT and Microsoft Copilot both achieved a "Good" accuracy score in 46/50 MCQs (92%).

Table II: Comparison of accuracy assessment scores between ChatGPT and Microsoft Copilot

MCQ No	Chat GPT				Microsoft Copilot			
	Evaluator 1	Evaluator 2	Total Score (TS)	Category	Evaluator 1	Evaluator 2	Total Score (TS)	Category
	Accuracy assessment	Accuracy assessment			Accuracy assessment	Accuracy assessment		
1	3	3	6	Good	3	3	6	Good
2	3	3	6	Good	3	3	6	Good
3	3	3	6	Good	3	3	6	Good
4	3	3	6	Good	3	3	6	Good
5	3	3	6	Good	2	2	4	Moderate
6	3	3	6	Good	3	3	6	Good
7	1	1	2	Poor	1	1	2	Poor
8	3	3	6	Good	3	3	6	Good
9	3	3	6	Good	3	3	6	Good
10	3	3	6	Good	3	3	6	Good
11	3	3	6	Good	3	3	6	Good
12	3	3	6	Good	3	3	6	Good
13	3	3	6	Good	3	3	6	Good
14	3	3	6	Good	3	3	6	Good
15	3	3	6	Good	3	3	6	Good
16	3	3	6	Good	3	3	6	Good
17	3	3	6	Good	3	3	6	Good
18	3	3	6	Good	3	3	6	Good
19	3	3	6	Good	3	3	6	Good
20	3	3	6	Good	3	3	6	Good
21	3	3	6	Good	3	3	6	Good
22	3	3	6	Good	3	3	6	Good
23	3	3	6	Good	3	3	6	Good
24	3	3	6	Good	3	3	6	Good
25	3	3	6	Good	3	3	6	Good
26	1	1	2	Poor	1	1	2	Poor
27	3	3	6	Good	3	3	6	Good
28	3	3	6	Good	3	3	6	Good
29	3	3	6	Good	3	3	6	Good
30	3	3	6	Good	3	3	6	Good
31	3	3	6	Good	3	3	6	Good
32	3	3	6	Good	3	3	6	Good
33	3	3	6	Good	3	3	6	Good
34	3	3	6	Good	3	3	6	Good

continue

continue

35	3	3	6	Good	3	3	6	Good
36	3	3	6	Good	3	3	6	Good
37	3	3	6	Good	3	3	6	Good
38	3	3	6	Good	3	3	6	Good
39	3	3	6	Good	3	3	6	Good
40	2	2	4	Moderate	3	3	6	Good
41	3	3	6	Good	3	3	6	Good
42	2	2	4	Moderate	3	3	6	Good
43	3	3	6	Good	3	3	6	Good
44	3	3	6	Good	3	3	6	Good
45	3	3	6	Good	3	3	6	Good
46	3	3	6	Good	3	3	6	Good
47	3	3	6	Good	1	1	2	Poor
48	2	2	4	Moderate	3	3	6	Good
49	3	3	6	Good	3	3	6	Good
50	3	3	6	Good	3	3	6	Good

1- incorrect, 2- correct but contains inaccurate information, 3- correct and adequate answer.
A TS>5 - 'good' responses, 3≤ TS 5 - 'moderate' responses, and TS<3 - 'poor' responses.

DISCUSSION

The current study compared the accuracy responses of ChatGPT, and Microsoft Copilot, in answering the case-based clinical vignette-style physiology MCQs from widely used resources for the USMLE**.** To our knowledge, this study is the first to compare the accuracy responses of two AI models, namely ChatGPT and Microsoft Copilot, in answering the case based clinical vignette-style MCQs in endocrine and cardiac physiology**.** The number of correct answers obtained by ChatGPT, and Microsoft Copilot were calculated and the explanations provided were assessed. ChatGPT and Microsoft Copilot provided the correct answers for 48 and 47 out of 50 MCQs, achieving a 96% and 94% accuracy respectively. In a study conducted in the US on USMLE STEP 1 style questions ChatGPT, posted 86% accuracy across 1300 questions assessed.[9] Accuracy comparisons for AMBOSS questions between ChatGPT and student based on each system showed 82% accuracy by ChatGPT and 61% by students in answering endocrinology questions. The scores obtained by ChatGPT and Microsoft Copilot were significantly higher than the 60% required to pass the USMLE STEP 1 exam.[10] In another study, out of 100 MCQs in various disciplines of basic and clinical medical sciences, ChatGPT obtained 74% marks in basic medical sciences.[11] In answering MCQs about human genetics and its aspects, ChatGPT scored 68.2%.[12] In the endocrinology examination, the marks obtained by ChatGPT was 58%, and Bard obtained 58%.[13] Antaki et al reported that ChatGPT achieved 55.8% scores in basic science questions.[14] ChatGPT-3.5 and GPT-4 scored 73.4% and 83.4%, respectively in neurological surgery (ABNS) Self-Assessment Examination 1 and demonstrated examination scores of 80.2% with ChatGPT-3.5 and 91% with GPT-4 with text-only questions.[15] GPT-4 outperformed with an 87.5% correct response rate, while Bard and Copilot had 57.5% and 62.5% accuracy, respectively when tested on a 40-question Advanced Life Support MCQ test by the European Resuscitation Council.[16] In a neuroscience USMLE format 200 MCQs, Claude outperformed other chatbots with 83% correct answers, GPT-4 PRO scored 81.7%, and Copilot scored 59.5%.[17] More than 90% accuracy in this study suggests that ChatGPT-4 and Microsoft Copilot can be a useful tool for MBBS students preparing for USMLE. Over the years, AI performance has improved as large language models (LLM) are continuously updated and are able to provide more accurate responses. A moderate positive correlation of total scores between ChatGPT and Microsoft

Copilot indicates that as the scores from one AI model increases, the scores from the other tends to increase as well. These results suggest a consistent scoring pattern between the two AI platforms used in the study. For a clinical vignette MCQ in endocrine physiology with clinical features suggestive of hypothyroidism, the question asked about the laboratory value that would be expected to decrease after treatment. Both ChatGPT and Microsoft Copilot chose the incorrect option (TSH) instead of the correct option (cholesterol). The AI models focussed on the feedback axis (TSH) rather than on systemic metabolic effects. This reflects the tendency of AI models to prioritise direct endocrine relationships rather than secondary physiological consequences. Previous studies have found that LLMs could correctly identify isolated facts but struggled to prioritise the most relevant answer in clinical case scenarios.[18]

In our study, correct answers provided by ChatGPT and Microsoft Copilot for the majority of the questions were accompanied by accurate explanations, while incorrect answers were associated with inaccurate explanations. ChatGPT and Microsoft Copilot were able to provide highly readable responses. However, the explanations for two MCQs by ChatGPT and four MCQs by Microsoft Copilot were inaccurate. For two MCQs, incorrect options were selected, and the explanations were inaccurate. There were no significant differences between ChatGPT and Microsoft Copilot regarding the overall accuracy of their explanations, with both being potentially useful in educational settings. This result is in line with a recent study that found ChatGPT-4 performed better than ChatGPT-3.5 on MCQs in cardiovascular physiology.[8] In a study conducted in Korea that tested the knowledge and interpretative ability of ChatGPT in a parasitology examination, the relationship between ChatGPT's explanations and the correctness of the answers was statistically significant.[19] In a study conducted by Rahsepar et al., ChatGPT achieved a superior performance to Google Bard in answering questions related to lung cancer screening and prevention.[20]

ChatGPT and Microsoft Copilot both achieved a "Good" accuracy score in 46/50 MCQs (92%). ChatGPT appeared to be more reliable in avoiding "Poor" responses compared to Microsoft Copilot. For all MCQs, the accuracy scores provided by both evaluators were identical, indicating a strong inter-rater reliability. Consistency in scoring is crucial to ensure reproducibility and objectivity.[21] In MCQ 5, Microsoft Copilot scored 2 from each evaluator (TS=4; Moderate), whereas ChatGPT scored a perfect score (TS=6; Good).

Implications for medical education

ChatGPT and Microsoft Copilot can be integrated into self-directed learning and formative assessments to help medical students understand the clinical application of physiological concepts. They are also beneficial in flipped classrooms and case-based learning sessions, as they provide immediate feedback and explanations.

Limitations

The limitations of our study include a small sample size of questions. The study focused solely on the USMLE resources which may limit the generalisability of the findings to other medical examinations or licensing examinations. Other question formats, such as 'true or false' questions, recall-based questions, higher-order cognitive questions, image-based questions, or open-ended questions were not evaluated. Our study did not categorize MCQs according to Bloom's taxonomy levels. Future studies should classify items by cognitive domain.

Data availability

The datasets used and/or analysed during the current study are available from the corresponding author on a reasonable request.

CONCLUSION

The high accuracy in answering MCQs suggests that both ChatGPT and Copilot could be used as learning tools for medical students. Both models demonstrate potential as supplementary tools in medical education. However, to ensure the accuracy of explanations provided, medical students and medical educators should critically evaluate the performance of ChatGPT and Microsoft Copilot on different questions formats, such as image-based questions and data interpretation tasks, to better understand their utility in diverse educational settings.

FUNDING

There is no external or internal research grant support for executing this work.

CONFLICT OF INTEREST

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

INSTITUTIONAL REVIEW BOARD (ETHICS COMMITTEE)

Ethical assessment by the institutional review board was not necessary because no human or animal research subjects were involved in the study.

ACKNOWLEDGEMENTS

Nil

REFERENCES

1. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023 Mar 30;388(13):1233-9. DOI: 10.1056/NEJMsr2214184
2. OpenAI. GPT-4 Technical Report. 2023; https://cdn.openai.com/papers/gpt-4.pdf.
3. 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. https://doi.org/10.1371/journal.pdig.0000198
4. 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 Sep 29;15(9): e46222. DOI: 10.7759/cureus.46222
5. Newton, P, Xiromeriti M. ChatGPT performance on multiple choice question examinations in higher education. A pragmatic scoping review. Assessment & Evaluation in Higher Education 2024; 49(6), 781–98. DOI:10.1080/02602938.2023.2299059
6. Burk-Rafel J, Santen SA, Purkiss J. Study behaviors and USMLE Step 1 performance: Implications of a Student Self-Directed Parallel Curriculum. Acad Med. 2017;92: S67–74. doi: 10.1097/ACM.0000000000001916.

7. Brin D, Sorin V, Vaid A, et al. Comparing ChatGPT and GPT-4 Performance in USMLE Soft Skill Assessments. Scientific Reports 2023; 13 (1):16492. https://doi.org/10.1038/s41598-023-43436-9
8. Banerjee A, Chatterjee M, Goyal K, et al. Performance of ChatGPT-3.5 and ChatGPT-4 in Solving Questions Based on Core Concepts in Cardiovascular Physiology. Cureus 2025; 17(5): e83552. doi: 10.7759/cureus.83552
9. Garabet R, Mackey BP, Cross J, et al. ChatGPT-4 Performance on USMLE Step 1 Style Questions and Its Implications for Medical Education: A Comparative Study Across Systems and Disciplines. Medical Science Educator 2024; 34:145–52 doi: 10.1007/s40670-023-01956-z
10. United States Medical Licensing Examination, Examination Results and Scoring. https:// www. usmle. org/ bulletin- information/ scoring- and- score- reporting. Accessed 29 Jul 2023.
11. Meo SA, Al-Masri AA, Alotaibi M, et al. ChatGPT Knowledge Evaluation in Basic and Clinical Medical Sciences: Multiple Choice Question Examination-Based Performance. Healthcare 2023; 11: 2046. doi: 10.3390/healthcare11142046
12. Duong D, Solomon BD. Analysis of large-language model versus human performance for genetics questions. Eur. J. Hum. Genet. 2023: 1-3. doi: 10.1038/s41431-023-01396-8
13. Meo SA, Al-Khlaiwi T, Abu Khalaf AA, et al. The Scientific Knowledge of Bard and ChatGPT in Endocrinology, Diabetes, and Diabetes Technology: Multiple-Choice Questions Examination-Based Performance. J Diabetes Sci Technol. 2023 Oct 5:19322968231203987. doi: 10.1177/19322968231203987.
14. Antaki F, Touma S, Milad D, et al. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. Ophthalmol Sci. 2023;3(4):100324. doi: 10.1016/j.xops.2023.100324.
15. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank. Neurosurgery 2023 Jun 12. doi: 10.1227/neu.0000000000002551
16. Semeraro, Carmona F, Monsieurs F, et al. Clinical questions on advanced life support answered by artificial intelligence. A comparison between ChatGPT, Google Bard and Microsoft Copilot. Resuscitation 2024; 195: 110114. doi: 10.1016/j.resuscitation.2024.110114.
17. Mavrych V, Yaqinuddin A, Bolgova O. Claude, ChatGPT, Copilot, and Gemini performance versus students in different topics of neuroscience. Adv Physiol Educ. 2025 Jun 1;49(2):430-7. doi: 10.1152/advan.00093.2024.
18. Wang D, Zhang S. Large language models in medical and healthcare fields: applications, advances, and challenges. Artif Intell Rev 2024; 57, 299. doi: 10.3390/bioengineering12060631.
19. Sun Huh. 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.
20. Rahsepar AA, Tavakoli N, Kim GHJ, et al. How AI responds to common lung cancer questions: ChatGPT vs Google Bard. Radiology. 2023;307(5): e230922. doi: 10.1148/radiol.230922. PMID: 37310252.
21. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. The American Journal of Medicine, 2006; 119(2), 166.e7–166.e16. doi: 10.1016/j.amjmed.2005.10.036