ChatGPT as an Assessment Tool for Medical Physiology Exams: A Comparative Study on MCQ and SAQ Answering Efficiency

Rekha Prabhu[a], Sherly Deborah George[a*], M Ganesh Kamath[a], Sandheep Sugathan S[b]

[a]Department of Physiology, Faculty of Medicine, Manipal University College Malaysia (MUCM), Melaka, Malaysia
[b]Community Medicine, Faculty of Medicine, Quest International University (QIU), Ipoh, Perak, Malaysia

Corresponding Author: Dr. Sherly Deborah George,
Department of Physiology, Faculty of Medicine, Manipal University College Malaysia (MUCM), Melaka, Malaysia
Email: sherly.george@manipal.edu.my

ABSTRACT

INTRODUCTION: ChatGPT, a language model, is well-known for its capacity to generate human-like responses, but its use in medical education, particularly in assessment contexts, is underexplored. The aim of this study was to evaluate the efficiency of ChatGPT as an assessment tool in medical physiology examinations by comparing its performance in answering MCQs and SAQs. The findings of this study may impact the use of AI in medical education in a constantly digitised academic environment. **MATERIALS AND METHODS:** The study evaluated the performance of ChatGPT in answering 30 multiple-choice questions (MCQs) and 12 short-answer questions (SAQs) from each of the four physiology blocks. The questions were chosen from previous block exams to ascertain consistency. Two independent evaluators assessed the correctness and relevance of responses from ChatGPT using the answer key. The mean marks obtained by first-year medical students for 120 MCQs and 48 SAQs were compared with those of ChatGPT. **RESULTS:** ChatGPT performed better than first-year medical students in MCQs in all block exams and the difference in marks was statistically significant in blocks 1, 2, and 3. In SAQs, ChatGPT also performed better than the students in most questions. Students scored better in SAQ 11 in block 2, SAQ 12 in block 3 and SAQ 1, 2, 5 in block 3. **CONCLUSION:** ChatGPT is an effective AI tool for answering medical physiology questions. However, its performance varies across some MCQs and SAQs, indicating potential limitations in reasoning, contextual interpretation, and application-based problem-solving.

KEYWORDS: ChatGPT, medical physiology, block examination, multiple-choice questions (MCQ), short-answer questions (SAQ)

INTRODUCTION

Artificial intelligence (AI) has created instruments that can transform healthcare and medical education.[1] ChatGPT is a language model that helps in writing and producing sentences that seem human-like, hence making it the most sought-after AI platform of Open AI in recent years. ChatGPT has been useful for healthcare professionals and students.[2] Among the subject content in medical education, medical physiology plays a crucial role in understanding the function of human cells, organs, and homeostasis in both health and illness.[3] It is vital for all health professionals to learn physiology since it lays the groundwork for diagnosis and treatment of patients.[2] Assessment drives learning, and in the context of learning medical physiology, it is important that the assessment is accurate.[4] First-year medical students need to undertake theoretical exams which consist of multiple-choice questions MCQs and short answer questions SAQs.

It is important to have appropriate study material and tools in physiology to enhance preparation for examinations. The traditional study methods include textbooks, note-making, and group discussions. However, within the last five years, the rise of AI technology has been revolutionising the way students' study.[2] AI-assisted technologies such as ChatGPT generate mostly accurate, contextually relevant answers, as well as answer vast numbers of queries when specific questions or information is provided. ChatGPT has an enormous knowledge base which serves as a valuable source for students across the health sciences, including medical students preparing for their exams. By utilising ChatGPT's own language processing capabilities, medical students can employ interactive learning, clarify their doubts by asking questions, and obtain answers that would help bridge the knowledge gaps. ChatGPT can immediately generate answers which helps in answering MCQs and SAQs efficiently thereby gives students confidence and enhances exam strategy.

It is important to explore the ability of ChatGPT as an instrument to improve the answering ability of MCQs and SAQs in medical physiology.[5] In medical exams, especially in the Asian context MCQs and SAQs are formats that are commonly used for assessment of theoretical knowledge and its application in clinical scenarios.[6-8] MCQs are usually used to assess recall, interpretation, and application of knowledge.[8] SAQs require the medical students to communicate the responses clearly and show their ability to synthesise information and apply their knowledge in answering the physiological concepts or clinical cases.[9] Both types of assessments require the medical students to think critically and attain competency in their physiological concepts.[8,9] It is also important to investigate the benefits, potential challenges, and limitations of ChatGPT in the context of examinations in the subject of physiology conducted for medical students.[10] The ability of ChatGPT is impressive, however there are concerns about the accuracy of the topic especially as it deals with details about specialised medical content.[2] AI tools collate information from various online sources, but they lack the human judgement and clinical expertise to verify the medical concepts.[11] The information can be incomplete or misleading.[11] The aim of this study was to evaluate the efficiency of ChatGPT as an assessment tool in medical physiology examinations by comparing its performance in answering MCQs and SAQs.

By conducting this study, we can learn the potential benefits of ChatGPT, streamline its use in our institution for teaching-learning purposes, and assessment. As the field of AI continues to grow, the understanding of tools such as ChatGPT can be efficiently incorporated into medical education. This will maximise the benefits to the teachers who would act as facilitators to ensure that medical students are well equipped to face the challenges of contemporary healthcare.

MATERIALS AND METHODS

Study design

The current research is an analytical cross-sectional study conducted at a medical institution in Malaysia. The study was designed to compare the marks of first-year medical students and ChatGPT in answering MCQs and SAQs in medical physiology. The research focused on four distinct blocks of the traditional physiology curriculum commonly taught during the MBBS programme. Each of the blocks contains the following systems: i) Block 1:Basic concepts, Blood, Nerve, and Muscle, ii) Block 2: Cardiovascular system, Respiratory system and Gastrointestinal system, iii) Block 3: Endocrinology, Reproductive system, and Renal system, and iv) Block 4: Central nervous system (CNS) and Special senses.

Question selection

Previous block examination questions prepared according to the blueprint which included both recall and reasoning questions were selected. Approval from the departmental vetting committee was obtained to ensure content validity and relevance. The questions prepared for the study comprised of 30 MCQs and 12 SAQs for each block. The answer key for all the MCQs and SAQs was framed by subject matter experts (SMEs) using standard textbooks (Guyton and Hall Textbook of Medical Physiology, 14th edition; Ganong's Review of Medical Physiology, 26th edition; Physiology Linda S. Costanzo, 7th edition) and guidelines. The answer key was used as the gold standard for evaluation. Each MCQ was awarded one mark for one correct answer and zero mark for an incorrect answer. There was no negative marking. SAQs were awarded maximum of five marks each, based on predetermined criteria in the standard answer key.

ChatGPT was accessed through OpenAI in its default configuration. The ChatGPT version used was GPT-4-turbo (https://chatgpt.com/) developed by OpenAI (San Francisco, California, USA). Each question was entered into ChatGPT in its original form (with a prompt), without modification. The prompt to answer the MCQ by ChatGPT was *"Choose the appropriate answer for the given multiple-choice question"*. The prompt to answer the SAQ by ChatGPT was *"Answer the following five-mark short answer question"*. ChatGPT's first response was recorded verbatim.

Student selection

A cohort of n=75 first-year medical students served as the human benchmark for this study. The students answered the same set of questions as ChatGPT, following university examination protocols. Data was anonymized to ensure participant confidentiality.

Data analysis

Responses were evaluated for accuracy, relevance, and completeness. Two independent evaluators, each with over ten years of experience in teaching medical physiology and prior experience in evaluating student answers, were involved. They assessed the responses from both ChatGPT and the students. SAQ discrepancies of 2.5 marks for each question were resolved through consensus. ChatGPT marks were compared to those by the students during the block examination.

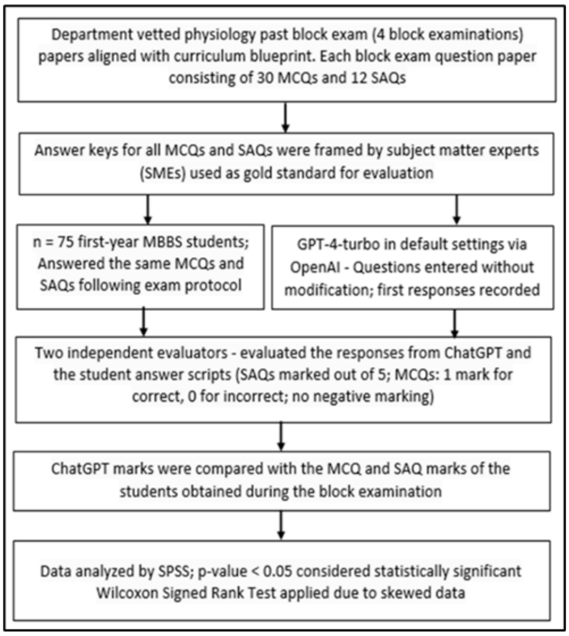
Department vetted physiology past block exam (4 block examinations) papers aligned with curriculum blueprint. Each block exam question paper consisting of 30 MCQs and 12 SAQs
Answer keys for all MCQs and SAQs were framed by subject matter experts (SMEs) used as gold standard for evaluation
n = 75 first-year MBBS students; Answered the same MCQs and SAQs following exam protocol
GPT-4-turbo in default settings via OpenAI - Questions entered without modification; first responses recorded
Two independent evaluators - evaluated the responses from ChatGPT and the student answer scripts (SAQs marked out of 5; MCQs: 1 mark for correct, 0 for incorrect; no negative marking)
ChatGPT marks were compared with the MCQ and SAQ marks of the students obtained during the block examination
Data analyzed by SPSS; p-value < 0.05 considered statistically significant Wilcoxon Signed Rank Test applied due to skewed data

Figure 1: Flow chart showing methodology

Declaration of Helsinki

Data was analysed using SPSS software. A p-value of <0.05 was considered statistically significant. The statistical significance was determined using the non-parametric Wilcoxon Signed Rank test. A non-parametric test was applied as the responses by students showed a skewed distribution. The summary of the methodology is presented in Figure 1.

RESULTS

Marks obtained by students and ChatGPT for 120 MCQs and 48 SAQs were compared. A total of 75 first-year medical students in a private university during the period between April 2023 and March 2024 were included in this research. The mean and median marks obtained by students and ChatGPT for both MCQs and SAQs for each block were compared.

Block 1

For the MCQ component with a maximum mark of 30, ChatGPT scored 26/30 marks while students scored mean 17.921 (5.821) and median 18 (9.75). The difference was found to be statistically significant (p<0.0005). For the SAQ component with a maximum mark of five for each question, ChatGPT scored significantly higher than the students in nine out of twelve9/12 SAQs. Table I tabulates and elaborates the comparison of marks between students and ChatGPT for Block 1 assessment.

Table I: Comparison of marks between students and ChatGPT for Block 1 assessment

Question	Students' Marks				ChatGPT's Marks	p-value	Better Score
	Mean	(SD)	Median	(IQR)			
MCQ	17.921	(5.821)	18	(9.75)	26	<0.0005*	ChatGPT
SAQ1	3.648	(1.408)	4	(2.5)	4	0.079	ND
SAQ2	3.497	(1.616)	4	(3)	3.5	0.908	ND
SAQ3	2.776	(1.687)	3.375	(2.94)	4.25	<0.0005*	ChatGPT
SAQ4	2.826	(1.845)	3.5	(3.88)	5	<0.0005*	ChatGPT
SAQ5	1.191	(1.481)	0.5	(2.5)	3.75	<0.0005*	ChatGPT
SAQ6	1.388	(1.784)	0	(3)	5	<0.0005*	ChatGPT
SAQ7	1.770	(2.012)	0	(4)	4.75	<0.0005*	ChatGPT
SAQ8	2.118	(1.831)	2	(4)	4.5	<0.0005*	ChatGPT
SAQ9	2.540	(1.570)	2.5	(2.88)	3	0.018*	ChatGPT
SAQ10	1.806	(1.788)	1	(3.38)	4.25	<0.0005*	ChatGPT
SAQ11	2.026	(1.219)	2	(1.75)	4	<0.0005*	ChatGPT
SAQ12	1.809	(1.183)	1.5	(2)	4.75	<0.0005*	ChatGPT

*SD: standard deviation; ND: no difference in scores; *statistically significant.*

Block 2

For the MCQ component with a maximum mark of 30, ChatGPT scored 23/30 marks, while students scored mean 14.868 (5.400) and a median 14 (8). ChatGPT performed significantly better than students in the MCQ component. For the SAQ component with a maximum mark of five for each question, ChatGPT performed significantly better than students in 11 out of 12 questions**.** Table II tabulates and elaborates the comparison of marks between students and ChatGPT for Block 2 assessment.

Block 3

ChatGPT performed significantly better than students in the MCQ component. In nine out of twelve SAQs, ChatGPT performed better than the students. Table III tabulates and elaborates the comparison of marks between students and ChatGPT for Block 3 assessment.

Table II: Comparison of marks between students and ChatGPT for Block 2 assessment

Question	Students' Marks				ChatGPT's Marks	p-value	Better Score
	Mean	(SD)	Median	(IQR)			
MCQ	14.868	(5.400)	14	(8)	23	<0.0005*	ChatGPT
SAQ1	3.428	(1.700)	4.25	(2)	3.25	0.139	ND
SAQ2	1.678	(1.724)	1	(3)	4.75	<0.0005*	ChatGPT
SAQ3	2.934	(1.742)	3.5	(3.5)	3.75	0.002*	ChatGPT
SAQ4	1.276	(1.556)	0.5	(2.5)	2.75	<0.0005*	ChatGPT
SAQ5	2.352	(1.267)	2.5	(1.88)	4	<0.0005*	ChatGPT
SAQ6	2.704	(1.565)	2.5	(2.5)	4.25	<0.0005*	ChatGPT
SAQ7	3.151	(1.934)	4	(3.88)	5	<0.0005*	ChatGPT
SAQ8	2.309	(1.864)	2	(3.88)	4	<0.0005*	ChatGPT
SAQ9	2.625	(1.528)	3	(3)	3.75	<0.0005*	ChatGPT
SAQ10	2.530	(1.348)	2	(1.5)	3.5	<0.0005*	ChatGPT
SAQ11	3.618	(1.487)	4	(2)	3	0.001*	Students
SAQ12	4.303	(1.334)	5	(1)	5	<0.0005*	ChatGPT

*SD: standard deviation; ND: No difference in scores; *statistically significant.*

Block 4

ChatGPT scored 22/30 marks for the MCQ component, while students scored a mean of 20.763 (5.506) and the difference was not statistically significant (p=0.140). In the SAQ component, students performed significantly better for SAQs 1, 2, and 5 (p<0.0005). Table IV tabulates and elaborates the comparison of marks between students and ChatGPT for the Block 4 assessment.

Table III: Comparison of marks between students and ChatGPT for Block 3 assessment

Question	Students' Marks				ChatGPT's Marks	p-value	Better Score
	Mean	(SD)	Median	(IQR)			
MCQ	21.053	(5.650)	22.5	(7)	23	0.010*	ChatGPT
SAQ1	4.375	(1.184)	5	(0.88)	5	<0.0005*	ChatGPT
SAQ2	3.211	(1.367)	3	(2)	4.75	<0.0005*	ChatGPT
SAQ3	2.895	(1.206)	3	(2)	4	<0.0005*	ChatGPT
SAQ4	3.618	(1.516)	4	(2)	5	<0.0005*	ChatGPT
SAQ5	2.918	(1.636)	3.375	(3)	4.75	<0.0005*	ChatGPT
SAQ6	2.526	(1.655)	3	(3)	4.75	<0.0005*	ChatGPT
SAQ7	3.993	(1.539)	5	(1.38)	3.25	<0.0005*	Students
SAQ8	2.691	(1.705)	3	(2.5)	3.5	<0.0005*	ChatGPT
SAQ9	3.434	(1.372)	4	(1.5)	5	<0.0005*	ChatGPT
SAQ10	3.401	(1.873)	4	(3.75)	3.75	0.799	ND
SAQ11	3.901	(1.740)	5	(1.38)	2.75	<0.0005*	Students
SAQ12	3.474	(1.677)	4	(1.88)	0.5	<0.0005*	Students

*SD: standard deviation; ND: No difference in scores; *statistically significant.*

Table IV: Comparison of marks between students and ChatGPT for Block 4 assessment

Question	Students' Marks				ChatGPT's Marks	p-value	Better Score
	Mean	(SD)	Median	(IQR)			
MCQ	20.763	(5.506)	21	(8)	22	0.140	ND
SAQ1	4.658	(1.123)	5	(0)	1	<0.0005*	Students
SAQ2	3.605	(1.563)	4	(1.5)	0	<0.0005*	Students
SAQ3	2.033	(1.698)	2.5	(3)	2.75	0.001*	ChatGPT
SAQ4	3.447	(2.062)	5	(4)	5	<0.0005*	ChatGPT
SAQ5	2.901	(1.997)	3.5	(4)	1.5	<0.0005	Students
SAQ6	3.066	(2.336)	4.75	(5)	4.75	0.001*	ChatGPT
SAQ7	3.901	(1.201)	4.25	(2)	4.50	<0.0005*	ChatGPT
SAQ8	2.586	(1.710)	3	(3)	4.50	<0.0005*	ChatGPT
SAQ9	3.382	(1.233)	4	(1)	4	<0.0005*	ChatGPT
SAQ10	4.441	(1.092)	5	(0.88)	4.75	0.480	ND
SAQ11	2.487	(2.214)	3	(5)	3.25	0.003*	ChatGPT
SAQ12	2.750	(1.854)	2.5	(4)	3.50	0.002*	ChatGPT

*SD: standard deviation; ND: No difference in scores; *statistically significant.*

DISCUSSION

Previous studies in medical education demonstrated the remarkable performance of large language model (LLM) based chatbots on many high-stakes medical examinations.[12] In this study, we compared the performance of ChatGPT and first-year medical students in answering both MCQs and SAQs in medical physiology.

The performance of ChatGPT was better than that of the first-year medical students in answering MCQs in all four block exams. The mean mark of students in MCQs was considerably lower than the ChatGPT mark. Previous studies have shown that ChatGPT versions 3.5 and 4 outperformed most medical students in the American Board of Neurological Surgery exam[13] and the German state licensing exam for medicine.[14] In addition, ChatGPT scored more than 75% in a physiology exam in a university in India[3] and 74% in answering basic medical science MCQs.[15]

Out of four blocks, MCQ marks obtained by ChatGPT were highest in Block 1 (26/30), followed by Blocks 2 and 3 (23/30 each), and lowest in Block 4 (22/30). In Block 1, ChatGPT surpassed the mean mark of students in answering MCQs in physiology. Basic concepts, muscle physiology, and blood were the topics taught in Block 1, which covered the basic physiological principles. The better marks obtained by ChatGPT may be attributed to its ability to recall information effectively and apply fundamental physiological concepts. This is consistent with earlier studies, which showed that AI models scored better on questions that test basic medical knowledge.[16] This is attributed to ChatGPT-4's exposure to large training datasets and its ability to identify repetitive patterns in MCQs.[17] Students' lower MCQ marks in Block 1 may be attributed to their transition from high school to a professional course, change in study strategies, and differences in question patterns.

The comparison of MCQ marks obtained between students and ChatGPT in blocks 2 and 3 were statistically significant. Our findings align with previous study, in which ChatGPT scores were highest in cardiovascular physiology module, followed by neurophysiology module, and endocrine physiology. For our institution, cardiovascular physiology is taught in Block 2, while endocrine physiology is taught in Block 3. The difference in the scores may be attributed to the limitations in training the AI application.[18]

ChatGPT also performed better than first-year medical students in Block 4, which included CNS, and special senses topics, but the results were not statistically significant. The remedial teaching sessions provided to the underperforming medical students in Block 4 may have contributed to the non-significant statistical finding between MCQ marks of students and ChatGPT. A previous study comparing underperforming medical students with controls showed a significant improvement in assessment marks due to post-remedial teaching sessions.[19] Many of these topics required higher-order cognitive skills, concept integration, and correlation with clinical disorders. This finding is consistent with earlier studies, which showed that AI models have difficulty answering clinical case-based questions and higher-order conceptual-type questions.[20] A recent study found that the performance of OpenAI ChatGPT-4 and Google Gemini in answering higher-order virology MCQs was poor.[21] For the internal medicine exam questions provided by the Taiwan Internal Medicine Society, GPT-4o underperformed in specialties like neurology and endocrinology.[22] The topics taught in CNS required an understanding of several complex concepts and relate them to real-life cases. Out of all the systems taught in the preclinical stages of the MBBS program, CNS is often considered the most difficult for medical students.[23]

The ChatGPT model showed lower performance in a few of the physiology SAQs (7/48) compared with students but performed better in most SAQs. For most SAQs, the responses provided by ChatGPT contained acceptable explanation, although some responses were found to be inappropriate. ChatGPT scored 50.75/60 marks in Block 1 and only 39.5/60 in Block 4. Block 1 included basic physiological concepts for which large training dataset enables comprehensive coverage and supports correct answers in ChatGPT. The high score in Block 1 demonstrated ChatGPT's aptitude for answering questions on fundamental concepts, definitions, and simple applications. These findings are consistent with a study showing that ChatGPT-4 performs well on questions requiring basic recall and structured knowledge.[16] Block 4 covered topics such as CNS and special senses, which require integration of anatomical knowledge, physiological mechanisms, and clinical correlations. Limited integration of information from multiple disciplines by ChatGPT-4 may have contributed to the lower SAQ scores in Block 4.

Out of the total 48 SAQs, ChatGPT scored less than 2.5 marks in four SAQs, three of which were from Block 4 (SAQs 1, 2, and 5) and one from Block 3 (SAQ 12). In Block 3, for SAQ 12 was a case-based question on homeostatic control of extracellular fluid (ECF) osmolarity with two sub-questions. The mean student mark was 3.474 (1.677) out of 5, whereas ChatGPT scored only 0.5 out of 5. Based on the case and data provided, the question on regulation of ECF osmolarity was incorrectly answered as regulation of ECF volume by ChatGPT. The difference in marks demonstrates how effectively students used physiological concepts to construct a flowchart describing the homeostatic control of ECF osmolarity.

A similar study showed that ChatGPT was able to provide appropriate diagnoses and answers for sub-questions in a case-based essay question on myocardial infarction.[3] In Block 4, SAQ 1 was a recall question for which ChatGPT scored 1/5 whereas students scored 4.65/5. The SAQ 1 recall question was; *"Explain the circulation and drainage of CSF"*. ChatGPT stated the sites of production and absorption of CSF but did not describe drainage across the ventricles or the foramina involved. Students' familiarity with textbook definitions and structured learning during face-to-face teaching sessions may have contributed to their better performance in SAQ 1. In contrast answers generated by ChatGPT are derived from large and diverse datasets that may not align with standard physiology textbooks.[24] For case-based SAQs on the pain pathway (SAQ 2) and spinal cord lesion (SAQ 5), ChatGPT scored 0/5 and 1.5/5 respectively, while students scored higher. This decrease in marks in Block 4 highlights ChatGPT's difficulty in answering questions requiring critical thinking and problem-solving. These findings contrast with another study showing that ChatGPT-4 achieved correct diagnoses for complex clinical vignettes compared to physicians.[25]

Strength of the study

This study assessed the knowledge level of ChatGPT in various physiology topics and compared its performance with the mean marks of first-year medical students. To the best of our knowledge, this study is among the first to assess ChatGPT's abilities in physiology examinations using both MCQs and SAQs across various blocks. The questions were selected from the university question bank and reviewed by two faculty members to ensure quality. ChatGPT may be used by students in preparing for formative assessments, potentially improving summative assessment performance. It may also assist teachers in preparing physiology practice questions. AI tools can serve as learning support tools to enhance understanding of physiological concepts.

Limitation of study

The study included only 120 MCQs and 48 SAQs. Different prompts could have been used; however, we chose to use the exact questions verbatim as prompts. The study was conducted in a single pre-clinical discipline. Evaluator bias may be present. The findings from this study may not apply to all SAQ or MCQ formats.

CONCLUSION

The findings of this study confirmed that ChatGPT is an effective AI tool for answering medical physiology exam questions, with particular strength in MCQs and SAQs. ChatGPT performed significantly better than students in many questions across most blocks. However, its performance was inconsistent in specific SAQs, with students outperforming ChatGPT in some cases. These inconsistencies suggest limitations in ChatGPT's ability to interpret contexts, integrate complex medical concepts, and provide detailed explanations required for certain SAQs. Despite these challenges, the results highlight the potential of AI models as effective adjunct learning tools in medical education. Future research should focus on enhancing AI models to improve accuracy in complex medical reasoning and application-based problem-solving. Exploring the use of AI in adaptive learning and interactive tutoring systems may further optimise its educational benefits.

DATA AVAILABILITY

The datasets used and analysed in this study, including MCQs, SAQs, and ChatGPT responses, are not publicly available due to institutional confidentiality regulations. However, they may be provided upon reasonable requests to the corresponding author.

CONFLICT OF INTEREST

There is no conflict of interest.

REFERENCES

1. Paranjape K, Schinkel M, Panday RN, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. *JMIR medical education*. 2019;5(2):e16048. doi:10.2196/16048
2. Lee H. The rise of ChatGPT: Exploring its potential in medical education. Anatomical sciences education. 2024 Jul;17(5):926-31. 3. https://doi.org/10.1002/ase.2270
3. Subramani M, Jaleel I, Krishna Mohan S. Evaluating the performance of ChatGPT in medical physiology university examination of phase I MBBS. Advances in physiology education. 2023 Jun 1;47(2):270-1. https://doi.org/10.1152/advan.00036.2023
4. Rae MG, Abdulla MH. An investigation of preclinical medical students' preference for summative or formative assessment for physiology learning. Advances in Physiology Education. 2023 Sep 1;47(3):383-92. https://doi.org/10.1152/advan.00013.2023
5. Kıyak YS, Emekli E. ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: a literature review. Postgraduate Medical Journal. 2024 Nov;100(1189):858-65. https://doi.org/10.1093/postmj/qgae065
6. Othman AS, Idris S, Latiff AA, Manan NA, Omar NS. Enhancing Assessment towards Flexible Medical Education: A Deep Dive into Preclinical Short Answer Questions through Item Analysis. Education in Medicine Journal. 2024 Jun 2;16. https://doi.org/10.21315/eimj2024.16.s1.3
7. Tjandra S, Tsurayya G, Manik AV, et al. Asian Medical Students' Perspectives on Medical Education Curricula Standards: A qualitative research. Journal of Asian Medical Students' Association. 2024 Nov 22;11(1). https://doi.org/10.52629/jamsa.v11i1.747
8. Bajpai R, Gavali Y, Rukadikar C. Assessment by modified essay type questions and multiple-choice questions in medical undergraduate students: Which are better in addressing higher cognitive skills?. National Journal of Physiology, Pharmacy and Pharmacology. 2024 Sep 30;14(10):2041-. DOI: 10.5455/njppp.2024.14.05217202419052024
9. Saeed GT, Hassan AE, Al Omary HL, Alawad ZM. Multiple choice questions and essay questions in assessment of success rate in medical physiology. Journal of the Faculty of Medicine Baghdad. 2017;59(4):333-5. https://doi.org/10.32007/jfacmedbagdad.59481
10. 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). DOI: 10.7759/cureus.46222
11. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC medical education. 2023 Sep 22;23(1):689. https://doi.org/10.1186/s12909-023-04698-z
12. OpenAI. Gpt4. https://openai.com/research/gpt-4, 2023. Accessed: 2023-07-01.
13. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. Neurosurgery. 2023 Dec 1;93(6):1353-65. *DOI:* 10.1227/neu.0000000000002632
14. Friederichs H, Friederichs WJ, März M. ChatGPT in medical school: how successful is AI in progress testing?. Medical Education Online. 2023 Dec 31;28(1):2220920. https://doi.org/10.1080/10872981.2023.2220920
15. Meo SA, Al-Masri AA, Alotaibi M, Meo MZ, Meo MO. ChatGPT knowledge evaluation in basic and clinical medical sciences: multiple choice question examination-based performance. InHealthcare 2023 Jul 17 (Vol. 11, No. 14, p. 2046). MDPI. https://doi.org/10.3390/healthcare11142046

16. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLoS digital health. 2023 Feb 9;2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198
17. Pepple DJ, Young LE, Carroll RG. A comparison of student performance in multiple-choice and long essay questions in the MBBS stage I physiology examination at the University of the West Indies (Mona Campus). Advances in physiology education. 2010 Jun;34(2):86-9. https://doi.org/10.1152/advan.00087.2009
18. Banerjee A, Ahmad A, Bhalla P, Goyal K. Assessing the efficacy of ChatGPT in solving questions based on the core concepts in physiology. Cureus. 2023 Aug 10;15(8). DOI: 10.7759/cureus.43314
19. Kumari S, Panda TK, Pradhan T, Subba SH. Modified formative assessment and its impact on undergraduate medical learning. Int J Health Sci Res. 2017;7(7):86-91.
20. Bharatha A, Ojeh N, Fazle Rabbi AM, et al. Comparing the performance of ChatGPT-4 and medical students on MCQs at varied levels of Bloom’s taxonomy. Advances in Medical Education and Practice. 2024 Dec 31:393-400. https://doi.org/10.2147/AMEP.S457408
21. Sallam M, Al-Mahzoum K, Almutawaa RA, et al. The performance of OpenAI ChatGPT-4 and Google Gemini in virology multiple-choice questions: a comparative analysis of English and Arabic responses. BMC Research Notes. 2024 Sep 3;17(1):247. https://doi.org/10.1186/s13104-024-06920-7
22. Lin SY, Hsu YY, Ju SW, et al. Assessing AI efficacy in medical knowledge tests: A study using Taiwan's internal medicine exam questions from 2020 to 2023. Digital Health. 2024 Oct;10:20552076241291404. https://doi.org/10.1177/20552076241291404
23. Lieu RM, Gutierrez A, Shaffer JF. Student Perceived Difficulties in Learning Organ Systems in an Undergraduate Human Anatomy Course. HAPS Educator. 2018 Apr;22(1):84-92. doi: 10.21692/haps.2018.011
24. Alawida M, Mejri S, Mehmood A, Chikhaoui B, Isaac Abiodun O. A comprehensive study of ChatGPT: Advancements, limitations, and ethical considerations in natural language processing and cybersecurity. Information. 2023 Aug 16;14(8):462. https://doi.org/10.3390/info14080462
25. Hirosawa T, Kawamura R, Harada Y, et al. ChatGPT-generated differential diagnosis lists for complex case-derived clinical vignettes: Diagnostic accuracy evaluation. JMIR Medical Informatics. 2023 Oct 9;11:e48808. doi: 10.2196/48808