MedCrave
Step into the World of Research

MOJ Anatomy & Physiology

Research Article

Open Access

CrossMark

Evaluating AI performance in answering questions related to thoracic anatomy

Abstract

Introduction: ChatGPT, an AI-enabled conversational language model, holds immense promise for applications across healthcare education, research, and clinical practice. A number of recent reports have highlighted promising outcomes from using ChatGPT for answering multiple-choice questions (MCQs), hinting at a potential reshaping of educational methodologies.

Aim: The objective of this particular study was to quantitatively assess the performance of ChatGPT-3.5 and ChatGPT-4 in the context of answering anatomical questions focusing on the Thorax in a Gross Anatomy course for medical students.

Methods: The research conducted for this study was focused on a comprehensive examination of ChatGPT-3.5 and ChatGPT-4 capabilities in answering 50 multiple choice questions MCQs designed in USMLE style. These questions were randomly selected from Gross Anatomy course exam database for medical students and reviewed by three independent experts. No questions with images were included in this study. The selected questions had different levels of difficulty.

Findings: The analysis revealed that ChatGPT-3.5 exhibited a remarkable ability to answer anatomical MCQs, achieving an average accuracy of 45.6±3.8%, significantly surpassing random guessing at 18.8±4.4%. The subsequent ChatGPT-4 demonstrated a 30% improvement, with an even higher accuracy of 75.6±1.7%.

Discussion: ChatGPT demonstrates considerable potential as an engaging educational tool for students immersed in the study of anatomy. Its unique strength lies in its ability to incite student engagement and arouse curiosity through an interactive, conversational mode of answer delivery.

Conclusion: It is vital to remember that ChatGPT should not be viewed as a replacement for the pivotal role teachers play in the educational journey. Instead, it should be visualized as an auxiliary instrument poised to enrich the educational experience. Future studies should aim to determine clear, holistic guidelines that illuminate the best ways to leverage and apply ChatGPT within the scope of anatomy instruction.

Keywords ChatGPT-3.5, ChatGPT-4, artificial intelligence, gross anatomy, medical education

Volume 10 Issue 1 - 2023

Volodymyr Mavrych, Olena Bolgova
Department of Anatomy and Genetics, Alfaisal University, Kingdom of Saudi Arabia

Correspondence: Volodymyr Mavrych, Professor of Neuroscience and Gross Anatomy, Department of Anatomy and Genetics, College of Medicine, Alfaisal University, Kingdom of Saudi Arabia, Tel +966 55 587 3958, Email vmavryc@alfaisal.edu

Received: December 01, 2023 | **Published:** December 14, 2023

Introduction

ChatGPT is an AI-powered conversational language model with potential applications in healthcare education, research, and practice. Many authors indicated different benefits, including improved scientific writing, enhanced research capabilities, efficiency in healthcare research, practice streamlining, cost savings, and personalized learning. However, some concerns were raised, including ethical, copyright, legal, plagiarism, inaccuracies, and cybersecurity issues. While ChatGPT holds promise, its use should be approached cautiously, and a code of ethics for its responsible application in healthcare and academia is needed. [1–3]

In medical education, the effectiveness and accuracy of Artificial Intelligence (AI) is currently under evaluation due to a need for more concrete data. However, researchers in this field are undertaking initiatives to address this notable gap. One such proposal includes establishing an international consortium database with robust, high-quality data. Many countries and regions worldwide are beginning to embrace AI as an integral part of their medical education curriculum. This strategic integration is aimed at counteracting the currently low usage rates of clinical AI and remedying the observed deficiency in AI awareness among practicing physicians. More importantly, it serves as an effective vehicle for driving the future advancement of AI in the healthcare sector.[4] However, the infusion of AI into medical education and realizing its potential benefits necessitates considerable policy support from local governments. Such backing is paramount in successfully incorporating AI into medical education. Despite the encouraging developments and proposed solutions, a solid foundation of high-quality research evidence chronicling AI's effectiveness, feasibility, and economic viability in medical education is required.[5]

Incorporating AI into medical education can transform the approach to teaching biomedical sciences. Robust language models like ChatGPT can function as virtual teaching aides, offering students comprehensive and pertinent information and potentially evolving into interactive simulations. ChatGPT holds promise for heightening student engagement and improving learning outcomes, although further research is imperative to validate these benefits. Recognizing and tackling the issues and limitations that come with ChatGPT is of substantial importance. This involves taking into account ethical components, as well as the possibility of harmful repercussions. Such challenges may range from privacy concerns to bias in the AI's responses. As a result, there is a significant need for ongoing vigilance and moderation.[6] Medical educators, in particular, must stay constantly attuned to the swift technological changes. This rapid

evolution not only impacts how we transmit knowledge but, more importantly, it shapes the content and structure of the curriculum alongside assessment methodologies and pedagogical tactics.[7]

Integration of substantial volumes of data, complemented by images into AI networks, is also recommended. This would allow AI to have a large scale of information to learn and adapt from, consequently improving its performance in medical education. Moreover, standardized guidelines should be introduced instead of creating personalized guidelines that may vary from place to place. These general instructions would help foster uniform application and understanding of AI in the medical instruction context, thus addressing the existing issues effectively.[8]

Interesting research was done to evaluate the quality of multiple-choice questions generated by ChatGPT for medical graduate examinations compared to questions created by university professors. ChatGPT produced 50 MCQs in about 20 minutes, while human examiners took 211 minutes for the same number of questions.[9] The assessment by independent experts found that ChatGPT's questions were comparable in quality to those created by humans, except in the relevance domain, where it scored slightly lower. However, ChatGPT's questions showed a wider range of scores, while human-generated questions were more consistent. In conclusion, ChatGPT has the potential to generate high-quality MCQs for medical graduate examinations and solve current problems related to item development.[10]

Many studies explored ChatGPT's potential for guiding medical students in anatomy education and research. Questions were asked to ChatGPT to evaluate its accuracy, relevance, and comprehensiveness. ChatGPT provided accurate anatomical descriptions with clinical relevance and structure relationships. It also offered summaries and terminology assistance. However, its responses to anatomical variants must be improved with systematic classification.[11]

Some recent publications suggested good results of ChatGPT in answering multiple-choice questions, which can impact the educational system. Analysing the accuracy and consistency of responses from ChatGPT-3.5 and Google Bard when answering lung cancer prevention, screening, and radiology terminology questions, ChatGPT-3.5 provided 70.8% correct answers, while Google Bard answered 51.7% correctly.[12] ChatGPT generally provided relevant answers to typical patient questions about optic disc drusen and total hip arthroplasty. However, some answers needed to be more accurate, particularly regarding treatment and prognosis, which could potentially be harmful in some cases.[13] This highlights the need for caution when relying solely on ChatGPT for patient information.[14]

Even though ChatGPT was only recently introduced, there are many publications on this topic, but only very few have any statistical data included. The main objectives of our research were to develop an algorithm for the quantitative analysis of the Chatbot's ability to answer MCQ tests, specifically in material the thorax in Gross Anatomy course material course for medical students, and compare results for ChatGPT-3.5 and ChatGPT-4.

Materials and methods

The research conducted for this study was focused on a comprehensive examination of ChatGPT capabilities in answering a set of 50 Multiple-choice Choice Questions (MCQs) designed in USMLE style. They were randomly selected from the Gross Anatomy course exam database for medical students and reviewed by three independent experts. No questions with images were included in this study. The selected questions had different levels of difficulty. Since all questions were created in 2020, we avoided the lack of real-time information limitation for ChatGPT-3.5.

The results of 5 successive attempts to answer this set of questions by ChatGPT-3.5 and ChatGPT-4 were evaluated based on accuracy, relevance, and comprehensiveness. Each ChatGPT attempt's data was recorded and compared with all previous attempts, finding the percentage of repeated and correct answers among them.

Seven sets of random answers were generated and analysed for the same MCQ sets utilizing the RAND () function in Microsoft Excel (Microsoft®365) to compare the results of ChatGPT performance with random guessing. Statistica 13.5.0.17 (TIBC® Statistica™) was used to analyse the data's basic statistics and compare ChatGPT-3.5 and ChatGPT-4 results.

Results

ChatGPT 3.5

According to our data, ChatGPT-3.5 provided accurate answers to 45.6±3.8% of the selected questions across five successive attempts, much superior to random guessing – 18.8±4.4%. The first attempt was the most successful, with 52% correct answers. The results of the following four attempts fluctuated in the 42% - 46% range. The coincidence of answers with the previous generations was 54% - 72%, and among them, the coincidence of correct answers was 34% - 42% (Table1).

Table 1 % of correct answers, coincidence with a previous attempt, and coincidence of correct answers with a previous attempt for 5 attempts from ChatGPT-3.5

Attempt	1	2	3	4	5
Correct answers	52	44	44	42	46
Coincidence with 1		68	58	66	70
Coincidence corrects with 1		36	34	40	42
Coincidence with 2			74	62	66
Coincidence corrects with 2			38	34	36
Coincidence with 3				54	62
Coincidence corrects with 3				32	34
Coincidence with 4					72
Coincidence corrects with 4					34

Fifteen questions (30%) were answered correctly across all five attempts and were considered a solid knowledge area for ChatGPT-3.5. The item analysis indicated that these MCQs were about anatomy, valves, blood supply, and heart embryology. They all were recall questions. ChatGPT did not show good results in answering more comprehensive questions about the pulmonary system and thoracic blood vessels.

ChatGPT 4

After five attempts, ChatGPT-4 generated 75.6±1.7% accurate answers for the set of questions, which is 30% superior to the results of ChatGPT-3.5 for the identical multiple-choice questions (Figure 1).

Citation: Mavrych V, Bolgova O. Evaluating AI performance in answering questions related to thoracic anatomy. *MOJ Anat Physiol.* 2023;10(1):55–59. DOI: 10.15406/mojap.2023.10.00339

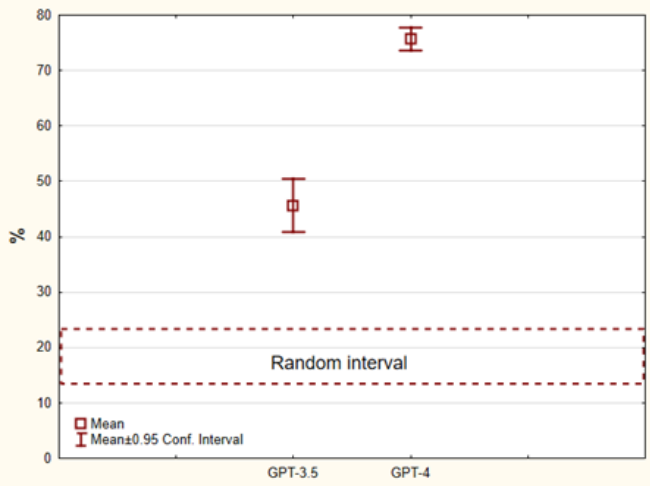

Figure 1 Percentile of correct answers of ChatGPT-3.5 and ChatGPT-4 on 50 MCQs for Thorax compared with random answers.

The initial attempt at ChatGPT-4 was also the most successful – 78% of the responses were correct. The following four tries yielded outcomes ranging from 74% to 76%. The answers coincided with those of the preceding attempts in a range of 88% to 90%; among them, the coincidence of correct answers was 68% to 74% (Table 2).

Table 2 % of correct answers, coincidence with a previous attempt, and coincidence of correct answers with a previous attempt for 5 attempts from ChatGPT-4

Attempt	1	2	3	4	5
Correct answers	78	74	74	76	76
Coincidence with 1		88	90	88	90
Coincidence corrects with 1		72	72	72	74
Coincidence with 2			86	84	90
Coincidence corrects with 2			68	68	70
Coincidence with 3				90	90
Coincidence corrects with 3				70	72
Coincidence with 4					88
Coincidence corrects with 4					72

Thirty-two questions (64%) were answered correctly across all ten attempts. The item analysis indicated that these 32 MCQs were not just simple recall questions but more comprehensive ones; the chatbot's replies were adequate and stayed consistent across the attempts.

Discussion

Artificial intelligence is a powerful driver that ceaselessly redefines and transforms different aspects of human life. Its role is even more noticeable and potent in the healthcare landscape, where it has become a significant pillar of evolution and advancement. Zooming into the intricate workings of AI, we find that this growth explosion has come on the back of sophisticated algorithms and machine learning techniques. These cutting-edge tools furnish AI with the capability to revolutionize various facets of healthcare.[15]

By leveraging the power of these advanced algorithms and machine learning strategies, there is an unprecedented possibility of elevating the standards and efficiency of healthcare processes. The potential of this symbiosis between AI and healthcare could lead to enhancements unheard of in traditional biological and medical environments, bringing forth unprecedented possibilities that could drastically reshape healthcare as we know it today. By assisting physicians in making more accurate diagnoses, identifying potential health risks, and devising personalized treatment plans, AI, such as the Chat Generative Pre-Trained Transformer (ChatGPT)-3.5, is playing a pivotal role in transforming the landscape of medical practices.[16–18]

ChatGPT-3.5, developed by OpenAI and made accessible to the general public on November 30, 2022, stands out as the first of its kind regarding broad availability. Ongoing investigations are delving into its healthcare applications, focusing on healthcare documentation, data interoperability, diagnostics, research, and education.[19] Evaluating the capabilities of such models involves tackling test problems and assessing performance metrics.[20]

Our study assessed the accuracy of ChatGPT versions GPT-3.5 and GPT-4, particularly in medical exams. We specifically targeted multiple-choice questions (MCQs) related to the challenging domain of Gross Anatomy, focusing on the Thorax. ChatGPT-3.5 exhibited a remarkable ability to answer medical students' MCQs, achieving an average accuracy of 45.6±3.8%, significantly surpassing random guessing at 18.8±4.4%. The subsequent ChatGPT-4 demonstrated a 30% improvement, with an even higher accuracy of 75.6±1.7%.

Our findings align with a parallel study assessing ChatGPT's performance in responding to the Japanese Nursing Examination (JNNE) conducted in February 2023. Despite random selection yielding accuracy rates of 20%-25%, GPT-3.5 and GPT-4 surpassed these figures with impressive performances at 59.9% and 80.2%, respectively. While GPT-3.5 fell short of meeting JNNE passing criteria, GPT-4 exceeded them, suggesting potential real-world applications in Japanese medical settings.[21]

Extending the scope to the US Medical Licensing Exams, ChatGPT-3.5 exhibited accuracy ranging from 42% to 64.4%, surpassing other models. However, performance declined with question difficulty, a trend also evident in our study.[22] It is also well correlated with our data, which indicated that only 30% of the questions were answered correctly across all five attempts. The item analysis indicated that all these 15 MCQs were simple recall questions. When it came to more comprehensive questions, the chatbot's replies could have been more adequate and varied across the attempts, unlike ChatGPT-4, which was able to answer the more comprehensive questions correctly.

Beyond the challenges of Gross Anatomy, ChatGPT showed proficiency in other medical disciplines, excelling in physiology, head and neck surgery, and biochemistry, particularly when confronted with non-MCQ formats. Research by Banerjee A et al. highlighted ChatGPT's effectiveness in tackling reasoning questions across diverse physiology modules, achieving an impressive 74% correctness.[23] In head and neck surgery, it responded correctly to 84.7% of closed-ended questions. It provided accurate diagnoses in 81.7% of clinical scenarios, with room for improvement in procedural details and bibliographic references.[24]

OpenAI subsequently launched GPT-4 on March 14, 2023. This latest iteration, which powers ChatGPT, was found to have an 82% decline in processing unauthorized content requests and a 40% boost in generating fact-based responses compared to GPT-3.5. Its enhancements also extended to dealing with images, not just text. The bot even passed the United States Bar legal exam with a score that far outstripped its predecessor.[25]

Citation: Mavrych V, Bolgova O. Evaluating AI performance in answering questions related to thoracic anatomy. *MOJ Anat Physiol.* 2023;10(1):55–59. DOI: 10.15406/mojap.2023.10.00339

Nonetheless, it is critical to approach the use of ChatGPT with discernment. Its accuracy is not entirely reliable, and there have been instances of it giving "hallucinated," or erroneous, responses.[26] It also occasionally falls short in more specialized domains.[27] Unquestioned acceptance of all generated content could lead to inaccurate healthcare advice. Therefore, critical appraisal alongside rigorous and targeted training should be pursued to enhance ChatGPT's medical field performance further. Moreover, because of its wide accessibility, there is a conceivable risk of disseminating incorrect health information, emphasizing the need for vigilance.

Conclusion

The growing utilization of cutting-edge technologies, such as artificial intelligence (AI), in education, makes conducting comprehensive and rigorous analyses imperative. It is necessary to verify and validate these AI-based resources for their seamlessly effective integration into the fundamental structure of medical education. Such practices form an essential part of the process to ensure that the deployment of these tools fits perfectly with the ongoing teaching and learning dynamics.

While introducing these innovative tools certainly hints at an evolved learning landscape, it does not necessarily ensure an immediate elevation in the quality of the learning experience. Hence, educators must continually engage in assessing the true potential of AI. Additionally, it is crucial to identify its limitations and potential risks. This continuous mapping and evaluation will help ensure that AI integration adds substantial value to the education process instead of simply leading to superficial enhancements.

Despite the growing prominence of AI tools like ChatGPT, it is crucial to remember that they are meant to supplement teachers' invaluable role in the educational process. These tools should be seen as additional aids designed to enrich and expand the scope of learning. Medical educators should make it a priority to maintain a proactive approach in devising a balanced and ethically rooted strategy toward the integration of this technology.

The importance of persistent research and re-evaluation cannot be overstated, especially concerning critical areas such as curriculum development, teaching techniques, and evaluation methods. These domains directly affect the overall effectiveness of the education system and its capacity to adapt to technological advancements.

A detailed comparison of AI versions, such as identifying situations where GPT-3.5 made errors that GPT-4 managed to correct and evaluating differences in their accuracy based on specific domains, could be wildly illuminating. Such trials, mainly conducted on larger datasets, can point out critical areas needing improvement or further development. As one looks forward, these elements should serve as focal points for subsequent studies, driving the future of AI integration in medical education.

Acknowledgments

The authors thank the Dean of the College of Medicine at Alfaisal University, Prof. Khaled Al-Kattan, and the Head of the Anatomy and Genetic Department, College of Medicine at Alfaisal University, Prof. P. Ganguly, for their support in this research.

Conflicts of interest

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

References

1. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel).* 2023;11(6):887.
2. Duong MT, Rauschecker AM, Rudie JD, et al. Artificial intelligence for precision education in radiology. *Br J Radiol.* 2019;92(1103):20190389.
3. Khan RA, Jawaid M, Khan AR, et al. ChatGPT - Reshaping medical education and clinical management. *Pak J Med Sci.* 2023;39(2):605–607.
4. Sun L, Yin C, Xu Q, et al. Artificial intelligence for healthcare and medical education: a systematic review. *Am J Transl Res.* 2023;15(7):4820–4828.
5. Li HH, Chen B, Li JP, et al. Status, problems and countermeasures of artificial intelligence application in medical education. *Chin J Evid Based Med*. 2020;20:1092–1097.
6. Lee H. The rise of ChatGPT: Exploring its potential in medical education. *Anat Sci Educ*. 2023.
7. Jamal A, Solaiman M, Alhasan K, et al. Integrating ChatGPT in medical education: adapting curricula to cultivate competent physicians for the AI era. *Cureus.* 2023;15(8):e43036.
8. Sapci AH, Sapci HA. Artificial intelligence education and tools for medical and health informatics students: systematic review. *JMIR Med Educ*. 2020;6:e19285.
9. 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.
10. Falcão F, Costa P, Pêgo JM. Feasibility assurance: a review of automatic item generation in medical assessment. *Adv Health Sci Educ Theory Pract.* 2022;27(2):405–425.
11. Totlis T, Natsis K, Filos D, et al. The potential role of ChatGPT and artificial intelligence in anatomy education: a conversation with ChatGPT. *Surg Radiol Anat*. 2023;45(10):1321–1329.
12. 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.
13. Potapenko I, Malmqvist L, Subhi Y, et al. Artificial intelligence-based ChatGPT responses for patient questions on optic disc drusen. *Ophthalmol Ther*. 2023;12(6):3109–3119.
14. Mika AP, Martin JR, Engstrom SM, et al. Assessing ChatGPT responses to common patient questions regarding total hip arthroplasty. *J Bone Joint Surg Am*. 2023;105(19):1519–1526.
15. Akshaya AVR, Kumar C. Artificial Intelligence is changing health and eHealth care. *EAI Endorsed Trans Smart Cities*. 2022;6: e3.
16. Lin SY, Mahoney MR, Sinsky CA. Ten ways artificial intelligence will transform primary care. *J Gen InternMed*. 2019;34:1626–1630.
17. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng.* 2018;2:719–31.
18. Nagpal K, Foote D, Liu Y, et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *NPJ Digit Med*. 2019;2:48.
19. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med*. 2023;388(13):1233–1239.
20. 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.
21. Kaneda Y, Takahashi R, Kaneda U, et al. Assessing the performance of GPT-3.5 and GPT-4 on the 2023 Japanese nursing examination. *Cureus.* 2023;15(8):e42924.

22. Gilson A, Safranek CW, Huang T, et al. 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;8;9:e45312.

23. Banerjee A, Ahmad A, Bhalla P, et al. Assessing the efficacy of ChatGPT in solving questions based on the core concepts in physiology. *Cureus.* 2023;10;15(8):e43314.

24. Vaira LA, Lechien JR, Abbate V, et al. Accuracy of ChatGPT-generated information on head and neck and oromaxillofacial surgery: a multicenter collaborative analysis. *Otolaryngol Head Neck Surg*. 2023;18.

25. GPT-4 is OpenAI's most advanced system, producing safer and more useful responses. 2023.

26. Anderson N, Belavy DL, Perle SM, et al. AI did not write this manuscript, or did it? Can we trick the AI text detector into generated texts? The potential future of ChatGPT and AI in sports & exercise medicine manuscript generation. *BMJ Open Sport Exerc Med.* 2023;9(1):e001568.

27. Kaneda Y, Tsubokura M, Ozaki A, et al. Are the issues pointed out by ChatGPT can be applied to Japan? Examining the reasons behind high COVID-19 excess deaths in Japan. *New Microbes New Infect.* 2023;53:101116.

Citation: Mavrych V, Bolgova O. Evaluating AI performance in answering questions related to thoracic anatomy. *MOJ Anat Physiol.* 2023;10(1):55–59. DOI: 10.15406/mojap.2023.10.00339