Interdisciplinary Journal of Virtual Learning in Medical Sciences

Original Article

Assessment of the Capability of ChatGPT-3.5 in Medical Physiology Examination in an Indian Medical School

Himel Mondal[1], Anup Kumar Dhanvijay[1], Ayesha Juhi[1], Amita Singh[1], Mohammed Jaffer Pinjar[1], Anita Kumari[1], Swati Mittal[2], Amita Kumari[1], Shaikat Mondal[3]*

[1]*Department of Physiology, All India Institute of Medical Sciences, Deoghar, Jharkhand, India*
[2]*Department of Physiology, Kalyan Singh Government Medical College Bulandshahr, Uttar Pradesh, India*
[3]*Department of Physiology, Raiganj Government Medical College and Hospital, West Bengal, India*

ABSTRACT

Background: There has been increasing interest in exploring the capabilities of artificial intelligence (AI) in various fields, including education. Medical education is an area where AI can potentially have a significant impact, especially in helping students answer their customized questions. In this study, we aimed to investigate the capability of ChatGPT, a conversational AI model in generating answers to medical physiology exam questions in an Indian medical school.

Methods: This cross-sectional study was conducted in March 2023 in an Indian Medical School, Deoghar, Jharkhand, India. The first mid-semester physiology examination was taken as the reference examination. There were two long essays, five short essay questions (total mark 40), and 20 multiple-choice questions (MCQ) (total mark 10). We generated the response from ChatGPT (in March 13 version) for both essay and MCQ questions. The essay-type answer sheet was evaluated by five faculties, and the average was taken as the final score. The score of 125 students (all first-year medical students) in the examination was obtained from the departmental registery. The median score of the 125 students was compared with the score of ChatGPT using Mann-Whitney U test.

Results: The median score of 125 students in essay-type questions was 20.5 (Q1-Q3: 18-23.5) which corresponds to a median percentage of 51.25% (Q1-Q3: 45-58.75) ($P=0.147$). The answer generated by ChatGPT scored 21.5 (Q1-Q3: 21.5-22), which corresponds to 53.75% (Q1-Q3: 53.75-55) ($P=0.125$). Hence, ChatGPT scored like that of the students ($P=0.4$) in essay-type questions. In MCQ-type questions, ChatGPT answered 19 correctly in 20 questions (score=9.5), and this was higher than the median score of students (6) (Q1-Q3: 5-6.5) ($P<0.0001$).

Conclusion: ChatGPT has the potential to generate answers to medical physiology examination questions. It has a higher capability to solve MCQ questions than essay-type ones. Although ChatGPT was able to provide answers that had the quality to pass the examination, the capability of generating high-quality answers for educational purposes is yet to be achieved. Hence, its usage in medical education for teaching and learning purposes is yet to be explored.

Keywords: Distance, Education, Artificial intelligence, ChatGPT, Physiology, Examination, Students, Medical

**Corresponding author:*
Shaikat Mondal,
Department of Physiology,
Raiganj Government Medical College and Hospital, West Bengal, India
Email: drshaikat@gmail.com

Please cite this paper as:
Mondal H, Dhanvijay AK, Juhi A, Singh A, Pinjar MJ, Kumari A, Mittal S, Kumari A, Mondal S. Assessment of the Capability of ChatGPT-3.5 in Medical Physiology Examination in an Indian Medical School. Interdiscip J Virtual Learn Med Sci. 2023;14(4):311-317.doi:10.30476/IJVLMS.2023.98496.1221.

Received: 08-08-2023
Revised: 08-28-2023
Accepted: 09-13-2023

Introduction

In recent times, there has been a growing interest in the use of artificial intelligence (AI) in various fields, including education. One area where AI can potentially have a significant impact is in the field of medical education (1). With the advent of advanced AI Large Language Models (LLMs) such as ChatGPT, there has been an increasing interest in exploring the capabilities of such models in answering students' queries and respond to their customized questions (2).

The ability of AI models like ChatGPT to generate human-like text has been demonstrated in a variety of domains, including literature, news articles, and educational purposes. The discussion about ChatGPT is creating a ripple in the world with lots of discussion in social media (3). The AI-powered tool was found to be an effective instrument for writing various entrance exams with an acceptable level of accuracy (4, 5). However, a recent study showed that the capability of ChatGPT was not impressive to pass the cut-off marks of the Indian Union Public Service Commission examination that requires higher knowledge and analytic capability (6).

The successful completion of an examination by ChatGPT would establish its immense value in medical education, particularly for self-learning purposes. With its comprehensive knowledge and ability to provide accurate explanations, ChatGPT could serve as an invaluable resource for medical students and professionals seeking additional information or clarification on various medical topics (7, 8). Acting as a virtual tutor or mentor, the availability and adaptability of ChatGPT would offer flexibility and convenience to individuals lacking immediate access to traditional educational settings. However, the training data limits its knowledge base, and current and reliable answers may not be obtained in many cases (9). Furthermore, its responses for different medical subjects may vary. Previous studies did not specifically investigate the capability of ChatGPT in a physiology examination.

In this study, we aimed to investigate the capability of ChatGPT in generating answers to medical physiology examination questions that are comparable in quality to those written by human examinee in an Indian medical school context. If ChatGPT proves capable of writing such examinations, it could serve as an additional learning resource that can provide personalized explanations and feedback to students' queries. Hence, it may be helpful for self-directed learning to supplement teaching and learning resources. The findings may also guide curriculum development and encourage further research on AI integration in medical education.

Methods

Study Design and Setting

This is a cross-sectional study conducted in an Indian Medical School, Deoghar, Jharkhand, India in March 2023. We used our personal computers and personal broadband Internet connections for generating the answers by ChatGPT. We used the ChatGPT-3.5 13 March 2023 version (free version).

Participants and Sampling

There were no human participants in this study. We collected the scores of the students' examination from the departmental registry. The scores of all medical students who wrote physiology semester examination in March 2023 as part of their first-year medical study were included. Students who did not answer more than two essay questions or 10% of multiple-choice questions (MCQ) were excluded from the research.

Data Collection Tool

The research tool consisted of an essay test (included 2 long essay questions and 5 short essay questions) and an MCQ test consisting of 20 four-choice multiple-choice questions with one correct answer. The scores for the long and short questions ranged from 0 to 40 and the range of the scores for MCQ questions was from 0 to 10.

The validity of the questions was

confirmed using the Content-Objective table based on the content presented during the course by 2 instructors of physiology course. The reliability of the tool was confirmed (>0.90) through the agreement of the scores of the examiners (5 examiners).

Data Collection

Questions: The question papers we used had two parts. The essay-type questions were further subdivided into long essay (two questions) and short essay types (five questions). The total score for the essay-type questions was 40. The MCQ questions consisted of a total of 20 questions, each having 0.5 point. All the questions had only one option to be correct from a set of four response options.

Answers: We collected the question-wise score (both essay-type and MCQ-type) of 125 students from the departmental data registry. The answers did not have a negative score. For getting answers of ChatGPT, each question was asked separately when we conversed with ChatGPT (March 13 free version). We converted the text to hand-written text for making it a human-written text. Then, the copies were distributed among five expert physiologists for checking and awarding the scores for each answer following the same answer key guidelines used for evaluating the students' papers. The scoring of the answers ranged from 0 to maximum number allotted for that answer. For an example, the answer of the following question - Describe skeletal muscle contraction under the following heads: a) excitation-contraction coupling b) sliding filament theory / modern theory of muscle contraction, a student or ChatGPT could get 0-3 for the part (a) of the question and 0-5 for part (b) of the question. We had detailed keywords and concepts for each question for semi-quantitative evaluation of the subjective essay-type answers. The same keywords or conceptual materials were searched when scoring the answers. We used five evaluators to reduce the bias.

Evaluation

In this study, we used the data available in the Department of Physiology and data generated from an online large language model – ChatGPT. The scores of students who wrote a semester examination were used for obtaining the result of the medical students' performance. Then, the answers of the ChatGPT were evaluated by five teachers, and the scores were used to obtain their performance in ChatGPT. A brief process of the study is presented in Figure 1. The details of the questions and answers are described in the next section.

Qualitative evaluation: We also asked the evaluators to provide their evaluation remarks in a short paragraph of the text at the end of the evaluations. These texts would be directly presented verbatim in the results section.

Data Analysis

First, we tested the data for normality using the Shapiro Wilk test and found that it did not follow a normal data distribution. Hence, we decided to express the descriptive data in median and quartiles and used non-parametric inferential statistical tests. The median score of 125 students was compared with hypothetical value of 50% using Wilcoxon signed rank test to check if their score significantly differed from the pass grade. Similarly, the score obtained by ChatGPT was also compared using Wilcoxon signed rank test with hypothetical 50% (as 50% is the pass score in the examination). Here, a non-significant P value or a significant P value where the median score is more than 50% indicates passing the median score. The median score of individual answers by 125 students was compared to that by five evaluators using the Mann-Whitney U test. For all tests, a P<0.05 was considered significant; we used GraphPad Prism 7 for conducting the tests.

Results

The median score of 125 students in essay-type questions was 20.5 (Q1-Q3: 18-23.5) which corresponds to a median percentage of 51.25% (Q1-Q3: 45-58.75). In the examination, 50% is considered the minimum score for

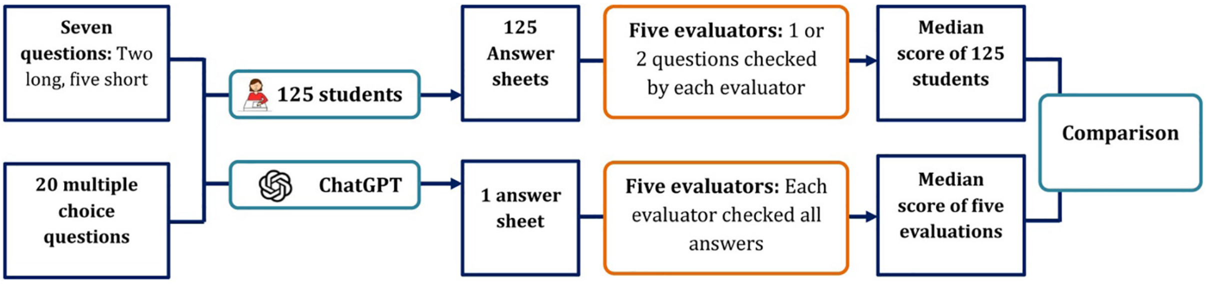
Seven questions: Two long, five short
20 multiple choice questions
125 students
ChatGPT
125 Answer sheets
1 answer sheet
Five evaluators: 1 or 2 questions checked by each evaluator
Five evaluators: Each evaluator checked all answers
Median score of 125 students
Median score of five evaluations
Comparison

Figure 1: Brief data collection method of the study

Table 1: Question-wise and overall score of the students and ChatGPT in the examination

Question type	Question number (full marks)	Students (n=125)	ChatGPT (rater=5)	P
		Median (Q1-Q3)		
Long essay	1 (8)	6 (5.5-6.5)	4.5 (4-5)	0.001*
	2 (7)	3 (2.5-4)	4 (4-5)	0.03*
Short essay	3 (5)	3 (1.5-3)	2.5 (2-3)	0.85
	4 (5)	2.5 (1.5-3)	2.5 (2-3)	0.49
	5 (5)	2.5 (2-3)	2.5 (2.5-3)	0.46
	6 (5)	2.5 (2-3)	3 (3-3)	0.19
	7 (5)	2 (0.5-3)	2.5 (2.5-3)	0.11
Total	1-7 (40)	20.5 (18-23.5)	21.5 (21.5-22)	0.4

*Statistically significant P value of Mann Whitney U test

Table 2: Comments of the evaluators on the answer sheet written by ChatGPT

Evaluator 1	Many of the answers are written as general rather than a professional examination. Although the answers are written with logical flow, the content could be improved.
Evaluator 2	Information is too basic and more explanation is required. In majority of the answers, there were missing essential information in the answers.
Evaluator 3	The answers could be presented as flow charts in some of the answers but it is written in a paragraph.
Evaluator 4	More detailed answers are required. The explanations are of high literary value than the scientific value.
Evaluator 5	Average paper, incomplete answers, improvement required.

passing the examination. There was no difference between the students' score and pass score (P=0.147).

The answer generated by ChatGPT scored a median of 21.5 (Q1-Q3: 21.5-22) in five evaluations which corresponds to 53.75% (Q1-Q3: 53.75-55). As 50% was the cut-off point, the ChatGPT passed the examination. There was no difference between the score of ChatGPT and the pass score (P=0.125).

Furthermore, ChatGPT scored the same as the students' score(P=0.4). Question-wise and overall scores in essay-type answers are shown in Table 1. There was a significantly lower performance of ChatGPT in the first long essay but better in the second long essay answer. The rest of the scores including the overall score were equal to those of the students.

In MCQ-type questions, ChatGPT answered 19 correctly in 20 questions (score=9.5), and this was also significantly higher than the students' median score (6 (Q1-Q3: 5-6.5)) (P<0.0001). Comments by the evaluations are shown in Table 2. These comments were directly quoted from the answer sheets.

From the comments, we found that the answers were too general and lacked essential information; also, some of the answers were incomplete, and that more detailed explanations were required. The evaluator suggested that further explanation was required to provide a scientific value to the answers. However, the content was interesting or engaging but lacked the necessary depth and detail required for a professional examination.

Discussion

The results of this study demonstrate the potential of natural language processing models to be performed in both essay-type and MCQ-type questions in the physiology exam. Interestingly, the ChatGPT model showed lower performance in the first long-essay question compared to the students but performed better in the second long essay answer. The overall performance of ChatGPT was equal to that of the students, in essay-type questions.

In MCQ-type questions, ChatGPT answered 19 out of 20 questions correctly, and it was a significantly higher performance than the students. Hence, if students used the model for their learning, they would easily get their MCQ questions solved. However, ChatGPT is a machine learning model, and its performance is dependent on the quality and quantity of data it is trained on. Hence, the answers should always be checked from other credible sources as well. Few studies were conducted to ascertain the capability of ChatGPT in solving MCQ questions, and the pooled percentage of correct answers was 53.1% (10). We found a higher accuracy in MCQ type answer in our study. This finding is in the same line with those of the study carried out by Huynh et al. who found that ChatGPT performed better in MCQ than open-ended question (11). The underlying reason may be the difference in the level of difficulty. Overall, the higher performance of ChatGPT in answering MCQs can be attributed to two training data and ease of retrieval. ChatGPT extensive training on a great number of text data allows to the therecognition of patterns commonly found in MCQs. It can identify the keywords and understand the context of the question, enabling it to generate accurate responses (12).

ChatGPT can assist students in their learning in medical schools in several ways. It can be used to personalize learning by generating content tailored to the individual learning needs of students. The major advantage is the time it requires for generating the answers. ChatGPT can provide immediate feedback to students based on their responses to generated questions. It is relatively difficult for faculties to provide the students with such services. ChatGPT can be used to provide continued learning opportunities outside the classroom, enabling students to learn at their own pace and on their own schedule (13). This would be beneficial for flipped classroom and self-directed learning (14).

There are several disadvantages of ChatGPT for its use in medical education. ChatGPT is limited by the quality and quantity of the training data it receives, which can affect the quality of the content it generates (15). It does not provide the recent data immediately. Free access to full version of ChatGPT is not possible, and particularly those in resource-limited settings would face difficulty even in accessing the free version due to no or poor Internet connectivity.

Limitation and Suggestions

This study has several limitations. It was done on a single pre-clinical subject. Although the answer sheet was written by human and presented to the evaluators and they were instructed to evaluate it like they did for students, still some bias might be present that was beyond our control. The results presented in the study may not be generalizable to all types of essay or multiple-choice questions, so further research is needed to explore the capabilities of natural language processing models in educational assessment.

Conclusion

ChatGPT has the potential to generate answers to medical physiology exam questions. However, its capability was found to be higher for MCQ-type questions compared to essay-type questions. The model was able to provide answers that had the quality to pass the examination. However, its ability to generate high-quality answers for educational purposes is yet to be fully achieved. The answers may be logically and grammatically correct, but they may not necessarily provide the depth of understanding or conceptual clarity that is required in educational contexts. Therefore, it is crucial to continue evaluating and improving the performance of natural language processing models like ChatGPT to ensure that they can provide high-quality answers that meet the medical educational standards.

Acknowledgement

We would like to thank Sarika Mondal and Ahana Aarshi for sacrificing their time during preparation of this manuscript. We also thank Ruchika Kumari and Sabnam Sadhna of the Department of Physiology, AIIMS Deoghar for their help in tabulation of the data. Some parts of the text was edited for grammar and presentation by ChatGPT free research version (May 24 2023 version) and it improved the quality.

Authors' Contribution

HM and SM conceived the study. AKD, AJ, MJP, AK, SMi, AS, and AmK collected the data. HM and SM analyzed the data. HM, SM, AKD, AJ, MJP, AK, SMi, AS, and AmK interpreted the data. HM and SM wrote the manuscript. AKD, AJ, MJP, AK, SMi, AS, and AmK revised the manuscript and improved it critically. HM revised the manuscript after peer review. All the authors approved the final version of the manuscript and take responsibility for the contents of the manuscript.

Ethical Considerations and Participants Consent

This study did not involve any human and animal participants and only used departmental data registry and online (on public domain) data audit. Hence, as per prevailing law in the country, this study does not require ethical approval.

Conflict of Interest: None declared.

Funding: None to declare.

References

1 Civaner MM, Uncu Y, Bulut F, Chalil EG, Tatli A. Artificial intelligence in medical education: a cross-sectional needs assessment. BMC Med Educ. 2022;22:772. PMID: 36352431 PMCID: PMC9646274. doi: 10.1186/s12909-022-03852-3.

2 Tlili A, Shehata B, Adarkwah MA, Bozkurt A, Hickey DT, Huang R, et al. What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. Smart Learn Environ. 2023;10:15. doi: 10.1186/s40561-023-00237-x.

3 Taecharungroj V. "What Can ChatGPT Do?" Analyzing Early Reactions to the Innovative AI Chatbot on Twitter. Big Data and Cognitive Computing. 2023; 7(1):35. doi: 10.3390/bdcc7010035

4 Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, 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;9:e45312. PMID: 36753318 PMCID: PMC9947764 doi: 10.2196/45312.

5 Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2:e0000198. PMID: 36812645 PMCID: PMC9931230 doi: 10.1371/journal.pdig.0000198.

6 ChatGPT fails to clear the prestigious Civil Service Examination. IndiaAI. Available from: https://indiaai.gov.in/news/chatgpt-fails-to-clear-the-prestigious-civil-service-examination (Last accessed on 22 March 2023).

7 Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT - Reshaping medical education and clinical management. Pak J Med Sci. 2023;39:605-7. PMID: 36950398

PMCID: PMC10025693 .doi: 10.12669/pjms.39.2.7653.

8 Sinha RK, Deb Roy A, Kumar N, Mondal H. Applicability of ChatGPT in Assisting to Solve Higher Order Problems in Pathology. Cureus. 2023;15:e35237. PMID: 36968864 PMCID: PMC10033699 .doi: 10.7759/cureus.35237.

9 Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595. PMID: 37215063 PMCID: PMC10192861 doi: 10.3389/frai.2023.1169595

10 Newton PM. ChatGPT performance on MCQ-based exams 2023. Preprint. doi:10.35542/osf.io/sytu3.

11 Huynh LM, Bonebrake BT, Schultis K, Quach A, Deibert CM. New Artificial Intelligence ChatGPT Performs Poorly on the 2022 Self-assessment Study Program for Urology. Urol Pract. 2023;10:409-15. PMID: 37276372 .doi: 10.1097/UPJ.0000000000000406.

12 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). Adv Physiol Educ. 2010;34:86-9. PMID: 20522902. doi: 10.1152/advan.00087.2009.

13 Arif TB, Munaf U, Ul-Haque I. The future of medical education and research: Is ChatGPT a blessing or blight in disguise? Med Educ Online. 2023;28:2181052. PMID: 36809073 PMCID: PMC9946299 doi: 10.1080/10872981.2023.2181052.

14 Das D, Kumar N, Longjam L, Sinha R, Deb Roy A, Mondal H, et al. Assessing the Capability of ChatGPT in Answering First- and Second-Order Knowledge Questions on Microbiology as per Competency-Based Medical Education Curriculum. Cureus 2023;15:e36034. PMID: 37056538 PMCID: PMC10086829. doi: 10.7759/cureus.36034.

15 Quintans-Júnior LJ, Gurgel RQ, Araújo AAS, Correia D, Martins-Filho PR. ChatGPT: the new panacea of the academic world. Rev Soc Bras Med Trop. 2023;56:e0060. PMID: 36888781 PMCID: PMC9991106 doi: 10.1590/0037-8682-0060-2023.