To Evaluate the Efficiency of ChatGPT in Medical Education: An Analysis of MCQ-Based Learning and Assessment

Syed Shirjeel Husain[1], Zaid Ansari[2], Azhar Hussain[3], Sahar Zubair Abbasi[4],

Tahera Ayoob[5], Rimsha Mujahid[6]

Abstract

Objective: This study aimed to evaluate the potential of ChatGPT to help students for their assessments via MCQ at different level of cognition by using different subjects of Internal medicine.

Methods: This cross-sectional study was conducted in the Department of Internal Medicine in collaboration with post graduate medical education department from June 2023 to August 2023. An MCQ bank was established from three books of MCQ's on subject of Internal Medicine. Total 1428 MCQ's were followed for scrutiny and 307 MCQ's were selected for the assigned task. The selected MCQ's were manually entered one by one in a fresh Chat GPT session. The response was noted against the replies given in respective MCQ's book and marked as correct, not correct or partially correct. MCQ's were categorized as per chapters in Internal medicine and as per cognition level of MCQ's i.e. C1, C2 and C3. Data was analyzed on SPSS version 21.00.

Results: Chat GPT replied with 199 correct replies while 98 were wrong and 10 were partially correct. Chat GPT scored 64% overall in all categories. At level of cognition, it solved C2 MCQ's by 80 % but scored 69% and 54% in C1 and C3 categories respectively. Chat GPT replied with 80% accuracy for C2 level MCQ's while results were low for C3 category at around 54%. C1 also had low percentage of correct answers standing close to 69.8%. Almost all subjects showed healthy responses around the mean except for endocrinology and hematology where responses are below 60% and 40% respectively.

Conclusion: This study findings suggest that ChatGPT is a useful tool for students and medical educationist with its current framework but a subtle approach should be inclined towards its role in future.

Key words; Artificial intelligence, educational assessment, Medical education

Introduction

The field of Science is witnessing a new era of artificial intelligence in form of Artificial intelligence tools (A.I) which world has never seen before reshaping the professional world and also the way healthcare professionals acquire and apply m-

edical knowledge. Among them, chatbots, powered by advanced artificial intelligence models like GPT-3 (Generative Pre-trained Transformer 3), have emerged as promising in medical education. San Francisco-based company Open AI Inc. developed and launched the revolutionary and enormous language model on November 30, 2022, which generates humanoid responses to varied and diverse natural queries. Within a week of its release, ChatGPT has crossed the over 1million user mark. This ChatGPT (Chat Generative Pre-trained Transformer) is a freely accessible conversational AI tool that was developed on the concept of reinforcement learning from human feedback[1,2].

[1-3,6] Department of Internal Medicine,
Liaquat College of Medicine and Dentistry
& Darul Sehat Hospital
[4] Directorate of Medical Education,
College of Physicians & Surgeons Pakistan
[5] Department of Oral Surgery, Liaquat College of Medicine and Dentistry & Qamar Dental Hospital

Correspondence: Syed Shirjeel Husain
Department of Internal Medicine,
Liaquat College of Medicine and Dentistry
& Darul Sehat Hospital
Email: dr_husain2003@yahoo.com
Date of Submission: 24th October 2023
Date of Acceptance: 30th November 2023

The ChatGPT model, which is based on deep learning, has demonstrated significant promise for a range of medical applications. These include medical record management, medical translation, and providing assistance to healthcare professionals in clinical, laboratory and radio-diagnostics. Its utility is expanding since its inception, all level of health care workers utilize the A.I efficiency in their workplace but with advent of time many institutes and organization are now planning to follow the process officially for better outcomes[3,4,5].

This deep learning model of ChatGPT has dynamic utilization in the field of medical education too. Among many of the frontiers which it may support includes Question-Answer Sessions, Case Studies, Clinical Decision Support, Anatomy and Physiology explanation, study and research assistance, generating research ideas, designing assessment tool, Continuing Medical Education (CME) and support in Communication Skills. The Dynamicity of this tool is not limited only to saving of precious time but it also produces and creates ideas collected from the World Wide Web in a single note.[6,7,8]

The classical teaching in medical education have witnessed didactic lectures, textbooks, and paper based assessments. Though time tested, but often these methods fell short in providing dynamic and interactive experience that today's digital-native medical students and professional demands. Along with ever expanding volume of medical knowledge which demands innovative approaches to facilitate efficient and continuous learning, the recent COVID era raised the need of new innovations which exploded the scene of medical education with lots of new gaps in learning and education.[9]

In just one year, the incorporation of A.I (artificial intelligence) and specially chatbots powered by GPT-3 in medical education is enormous and beyond words. Every technology when introduced usually faces resistance at various levels but in this case due to its interactive and conversational nature have emerged as a core player in field of medical education along with many other fields. As

the technology is new and is yet to be fully explored; its application in many dimension at different level of teacher, students, educationist and even patients need careful follow up & instructions. Currently; it employeS at level of personal learning, teaching module, assessment like MCQ generation and solving, reducing administrative burden, integration with learning management system, to act as virtual teacher and many more.[10,11]

ChatGPT is enriched with advanced technology and techniques to respond to users requirements. ChatGPT can swiftly obtain, interpret and provide the required information on a topic. It can compose the information with specified content for a particular person. ChatGPT is a highly capable and effective tool for language translation, article summarization and draft generation, which can be used for various scholarly performances.[12]

Like any other technology, ChatGPT may develop the option of cheating on online exams and minimize critical thinking skills. With all its expertise, it is always wiser to go slow and blend its usage with most suitable outcome for the medical community.[13]

Their ability to engage learners in dynamic, personalized and interactive ways has gained significant attention. To what extent can ChatGPT assist students in successfully answering MCQs of different cognitive levels (C1, C2, and C3) across various topics within Internal Medicine, and what are the potential implications of ChatGPT's performance on medical education is topic of our interest.

Therefore, this research explored the application of ChatGPT in medical education and its potential to solve Multiple Choice questions of varying cognitive level (C1, C2 & C3) in backdrop of different chapters of Internal medicine at graduate and post-graduate level.

Methodology
This cross-sectional study was conducted in the Department of Internal Medicine in collaboration with post graduate medical education from June 2023 to August 2023. The research team comp-

rised of clinician, surgeons, educationist, post graduate resident and a house officer. The inclusion criteria was commonly readable internal medicine books of MCQs selected through convenience sampling. Exclusion criteria was MCQs rejected by subject experts of research team.

The research team utilized three MCQ's book as given in table 1. An MCQ bank of 1428 MCQ's was established from three books of MCQ's on subject of Internal Medicine used widely by undergraduate and post graduate medical students. MCQ's were only selected from different subjects of Internal Medicine. As the MCQ's books don't mention the cognitive level of assessment, the senior team members with experience in medical education mark individual MCQ with category of cognition level as C1, C2 and C3 respectively. The research team members carefully reviewed the MCQs' contents and assured that the MCQs were relevant and appropriately challenging. Each question was scenario-based with four sub-stems or had a single correct answer. MCQs were evaluated for quality, and research team concurred with the final answer. Finally, a total of 307 MCQ's were finally selected for the task by the research team.

The investigation team members also proofread the MCQs for any errors, typos, confusing or misleading statements or inconsistencies. It was also checked that the options were well constructed and that there were no clear hints or clues within the questions.

The selection of MCQ's was based on two factors, subject in internal medicine and cognitive level as apprehended by the research team. All MCQ's selected were text based and any other format like pictures, graphs, flowcharts were excluded from the list.

ChatGPT is currently a free, open-source online tool that is accessible to users with a registration on the website, and all information that was collected was executed from its most recent version (version 3.5 as of July 2023). Selected MCQ's were copied to ChatGPT version 3.5 (free ware) and replies obtained were recorded against the items given in respective MCQ source. Replies were recorded as correct, incorrect or partially correct in a separate file. The MCQs were entered manually one by one, and a fresh ChatGPT session was started for each entry to avoid memory retention bias by the A.I tool. The first response that was obtained was taken as the final response, and we did not use the choice of "regenerate response".

Results

The research team followed 1428 MCQ's for the three book and a total of 307 MCQ's from source books were enquired via ChatGPT ver3.5. ChatGPT replied with 199 correct replies while 98 were wrong and 10 were partially correct. (Table 2)

A further subdivision was created to categorize the question based on cognition; C1 being a simple recall question, C2 comprising simple question scenarios and C3 being a vignette style question with more than one probable answer. Due to length of some of the C3 case scenarios containing more than one sub questions, each sub-question was treated as an individual question. The A.I. tool reply to each answer is being marked separately.

With level of cognition, ChatGPT performed as per table 3. ChatGPT replied with 80% accuracy for C2 level MCQ's while the results were low for C3 category at around 54%. Surprisingly C1 which were only recall questions also had low percentage of correct answers standing close to 69.8%. (Table 3)

The replies obtained were further sequestered as per subjects in internal medicine as shown in table 4. Almost all subjects showed healthy responses around the mean except for endocrinology and hematology where responses are below 60% and 40% respectively. (Table 4)

Table 1. MCQ's books utilized to gather MCQ pool by the research team.

Book name	Author	Edition/publisher
1000 Questions and Answers from Kumar & Clark's Clinical Medicine	Parveen Kumar Micheal Clarke	Second edition SaundersElsevier
Rapid review of ClinicalMedicinefor MRCP Part 2	Sanjay Sharma Rashmi Kaushal	Second Edition Manson publishing
Master the IMM (Medicine)	Usman Muzaffar Ali	4th Edition Nishtar Publications

Table 2. Marks obtained by ChatGPT (freeware; ver 3.5) after solving selected MCQ's pool.

Total MCQ's	307	Percentage (%)
Correct replies	199	64.82
In correct replies	98	31.92
Partially correct	10	3.25

Table 3. ChatGPT performance as per cognition level of respective MCQ's

Level of Cognition	Correct (n & %)	In correct (n & %)	Partially correct (n & %)	Total n=307
C1	95 (69.8)	40 (29.41)	1 (0.73)	136
C2	33 (80.48)	8 (19.51)	0	41
C3	71 (54.61)	50 (38.46)	9 (6.9)	130

Table 4. ChatGPT replies to MCQ's as per chapters of Internal Medicine MCQ Book

Chapters in Internal Medicine Medicine	Correct (n)	In Correct (n)	Partially correct (n)	Correct score percentage (%)
CVS	32	13	1	68
Rheumatology	18	6	0	75
Endocrine	26	17	1	56.5
GIT	21	8	1	67.7
Pulmonology	27	13	0	67
Hematology	6	8	2	37.5
Infectious Disease	13	8	0	61.9
Nephrology	22	9	2	66.6
Neurology	23	10	1	65.7
Pharmacology	1	0	0	100
Genetics	3	2	0	50
Toxicology	5	1	2	55.5
A&E	0	1	0	100
Immunology	0	2	0	0
Psychiatry	1	0	0	100
Maternal medicine	1	0	0	100

Discussion

Assessment of students is an integral part of medical education and while assessing knowledge in various disciplines. Acquisition of knowledge is essential in medical education and can be achieved by capability of candidate to interoperate and apply that knowledge in real life scenario with effectiveness. These skills are assesses through medical case presentation, essays and MCQs.Therefore MCQs must be well designed to assess deep learning and making them the most renowned tool of medical assessment throughout the world.[14]

MCQ's are worldwide accepted tool in medical education. Internationally, medical schools and medical licensing examination bodies are utilizing MCQ-based examination during various stages of undergraduate and post graduate assessment. As students face highly variable challenges during their course preparation days, they often look for numerous ways to consolidate their knowledge and approaches that are helpful for their exam needs.[15,16, 17]

ChatGPT has gathered great attention from the public, students, academicians, researchers and science communities in a very short span of time. Response of ChatGPT is highly swift, it articulates different dimensions of subject and moreover is able to generate multiple times with variations to meet the demand of query. Its utility is universal and has shown great enhancement in medicine and medical education too at all levels.[18]

ChatGPT is designed to generate human-like responses and engage the users in conversational interactions and rapid responses within seconds. ChatGPT is guided by a wide range of internet text data, which allows it to understand and produce text in a variety of contexts. It can answer questions, provide explanations, offer suggestions, create conversational dialogues and assist with multiple tasks.

As per literature search and capability of A.I tools, they can help teachers and medical educationist in exam preparation or formation of assessment tools like MCQ's or any other format. The comparison of ChatGPT vs Human teachers is beyond any match as it can generate MCQ's ten times faster than a whole group of teachers. With such abilities it was very well expected that when the technology will be used for solving the available MCQ's, results would be very helpful and encouraging for students to prepare for their examinations and assessment.[19]

As the technology is new and huge area of research is pending to explore its dimension, validate its functioning and explain its utility even as a teacher in a class room. This study aims to validate ChatGPT ability to help students who seek help while solving MCQ's during their exam preparation tenure.[20]

With the aura of Artificial intelligence, it was expected that ChatGPT would be a master trainer for students preparing for their exam with MCQ's. The present study results reveal that ChatGPT answered and obtained good grades in internal medicine subjects at an average of 64.82%. This figure is quite less than expected by the team of researchers engaged,as ChatGPT is designed to generate human-like responses and engage the users in conversational interactions; its responses were similar but with limitations. This study is part of the project as the management of LCMD is thinking and supporting the idea of utilizing A.I tools for students; and multiple studies are also undergoing for exploration of its potential usage.

Another aspect is the availability of higher version of ChatGPT which is paid version and was made available on March 14, 2023, on a subscription basis thus limiting its usage at student level especially in third world countries; however its potential usage in medical education still holds a very strong question for teacher and students when compared to free one.

In literature search, a research review provides valuable information regarding not only for ChatGPT potential for MCQ solving but also the comparison of the two available version. Among selected studies in the review, 114 data sheets exhibit variable results ranging from 36.5 to 80% capacity of ChatGPT freeware to reply MCQ's correctly. Though not all datasheets compared both versions, but it was significant that ChatGPT paid version performed better than the free ware and in an instance reaching to 100% accuracy.[21]

While comparing the two version of ChatGPT, similar result were also observed in a study that recorded the score of answering multiple-choice questions by 2 available versions. ChatGPT-3.5 achieved an overall score of 63.1%,and ChatGPT-4 scored 90.5%, significantly higher than the free ware. This favors the ability of ChatGPT paid version to be more dependable in terms of MCQ's solving capacity.[22]

Though with higher expectations initially by the research team, ChatGPT performed as per available literature. In our study where only MCQ's from domain of internal medicine were used for evaluation, ChatGPT scored 64% accuracy with 3.2% of partially correct replies. ChatGPT scored remarkably well in C2 category of MCQ's i.e. up to 80 % but scored much lower scores in C1 and C3 with 69% and 54% respectively. As C1 are simple recalls, ChatGPT was expected to score well with available resources on internet. A potential drawback of the model in addressing complex medical queries can be seen in the fact that more challenging questions (C3) appeared to have potentially marginally less accurate ratings than easier ones. C3 level MCQ's; as it is more complex and demands apprehension is the area of betterment for any A.I tool in the future.

For different subjects in medicine, average score remains close to mean score of 64% with exception of hematology scoring very low with 16 queries of various cognitive strength.

In our study, we only used freeware of ChatGPT therefore a comparison with paid version is not possible with current statistics. The number of questions selected was from a pool of MCQ's; where the number of MCQ's from a subject of internal medicine are not equal in number; therefore statistics in percentage may not represent the real strength of the study.

Our findings appear to be in line with other studies that reach the conclusion with an average ChatGPT's accuracy in solving MCQ's is around 65% grossly with highest efficiency in C2 type of MCQ's.

As previously indicated, ChatGPT's propensity to elicit hallucinatory reactions is also a problem.

The Chat bot also modifies its responses when a question's phrase or tone is slightly changed, producing several responses emphasizing a new domain of prompt writing for A.I tools in order to avoid deception or misleading replies.

The study has several limitations worth noting. Firstly, it is primarily focused on Multiple Choice Questions (MCQs) within the field of Internal Medicine, limiting its representation of other medical specialties and assessment formats. The study was conducted in a specific educational setting (Liaquat College of Medicine and Dentistry, Karachi, Pakistan). The results may not be fully generalizable to other medical institutions or regions with different curricula and student populations. Additionally, the study relied on the free version of ChatGPT, potentially missing out on the enhanced capabilities of the paid version. In order to alleviate these limitations, it would be advisable to conduct similar studies with more subject options, using paid version of ChatGPT with more diverse collection of MCQ's available for students.

Students, teachers and medical educationist should follow a subtle route and allow a healthy amalgam of the new technology with conventional teaching, assessment and skills. All assessment or exam preparation with help of MCQ's should be conducted under secure conditions and referral to source books or database should always be consult for accuracy. . Its strengths in natural language understanding and quick responses streamline information acquisition. Yet, users must remain cautious, understanding the limitations that arise from the system's lack of clinical judgment and potential inaccuracies.

Conclusion

In conclusion, ChatGPT's application in medical question solving presents a dynamic tool for initial inquiries, offering accessibility and convenience A balanced approach, combining the system's efficiency with expert medical consultation, is crucial for optimal outcomes.

References

1. Rahaman MS, Ahsan MM, Anjum N, Rahman MM, Rahman MN. The AI race is on! Google's Bard and OpenAI's ChatGPT head to head: an opinion article. Mizanur and Rahman, Md Nafizur, The AI Race is on. 2023. Available on https://papers.ssrn.com/sol3/pap ers. cfm? abstract_id= 4351785 .Accessed on 29th November 2023.

2. Thorp HH. ChatGPT is fun, but not an author. Science. 2023 Jan 27;379(6630):313-. Available on https://www.science.org/doi/10.1126/science.adg7879 . Accessed on 29th November 2023.

3. Biswas SS. Role of chat gpt in public health. Annals of biomedical engineering. 2023;51 (5):868-9.Available on https://pubmed. ncbi. nlm. nih.gov/36920578/.Accessed on 29th nove mber.

4. Sedaghat S. Early applications of ChatGPT in medical practice, education and research. Clinical Medicine. 2023: 23(3):278-9. Available on https://pubmed.ncbi.nlm.nih.gov/37085182/ . Accessed on 29th November 2023

5. Javaid M, Haleem A, Singh RP. ChatGPT for healthcare services: An emerging stage for an innovative perspective. BenchCouncil Transactions on Benchmarks, Standards and Evaluations. 2023 Feb 1;3(1):100105.Available on https://www. science direct.com/science/article/pii/S2772 4859230 00224 .Accessed on 29th November 2023.

6. Lee H. The rise of ChatGPT: Exploring its potential in medical education. Anatomical Sciences Education [Internet]. 2023 Mar 28 ; Available on https://anatomypubs.onlinelibrary.wiley.com/doi/10.1002/ase.2270 . Accessed on 29th november 2023.

7. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. JMIR Medical Education. 2023 8;9(1):e45312.Available on https://pubmed. ncbi. nlm.nih.gov/36753318/ .Accessed on 29th November 2023

8. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In Healthcare 2023 Mar 19 (Vol. 11, No. 6, p. 887). MDPI.Available on https://www.mdpi.com/2227-9032/11/6/887. Accessed on 29th November 2023.

9. Ahuja V, Nair LV. Artificial Intelligence and technology in COVID Era: A narrative review. Journal of anaesthesiology, clinical pharmacology. 2021 Jan;37(1):28.Available on https://pubmed. ncbi. nlm.nih.gov/34103818/ . Accessed on 29th November 2023.

10. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in

medical education: a conversation with ChatGPT and a call for papers. JMIR Medical Education. 2023 Mar 6;9(1):e46885. Available on https://mededu.jmir.org/2023/1/e46885/ .Accessed on 29th November 2023.

11. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT- Reshaping medical education and clinical management. Pakistan Journal of Medical Sciences. 2023 Mar;39(2):605. Available on https://www.pjms.org.pk/index.php/pjms/article/view/7653. Assessed on 29th November 2023.

12. Salvagno M, Taccone FS, Gerli AG. Can artificial intelligence help for scientific writing?. Critical care. 2023 Dec;27(1):1-5. Available on https://ccforum.biomedcentral.com/articles/10.1186/s13054-023-04380-2 .Accessed on 29th November 2023.

13. Rahman MM, Watanobe Y. ChatGPT for education and research: Opportunities, threats, and strategies. Applied Sciences. 2023 May 8;13(9):5783. Available on https://www.mdpi.com/2076-3417/13/9/5783 . Accessed on 29th November 2023.

14. Vegi VA, Sudhakar PVC, Bhimarasetty DM, Pamarthi KEdara L, Kutikuppala LS et al. Multiple-choice questions in assessment: Perceptions of medical students from low-resource setting. Journal of Education and Health Promotion. 2022;11. Available on https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9093664/ . Accessed on 29th November 2023.

15. Ali R, Sultan AS, Zahid N. Evaluating the effectiveness of' MCQ development workshop using cognitive model framework: A pre-post study. JPMA. The Journal of the Pakistan Medical Association. 2021;71(1 (A)):119. Available on https://pubmed.ncbi.nlm.nih.gov/33484534/ . Accessed on 29th November 2023.

16. Grainger R, Dai W, Osborne E, Kenwright D. Medical students create multiple-choice questions for learning in pathology education: a pilot study. BMC medical education. 2018 Dec;18:1-8. Available on https://pubmed.ncbi.nlm.nih.gov/33484534/. Accessed on 29th November 2023.

17. Kenwright D, Dai W, Osborne E, Gladman T, Gallagher P, Grainger R. Just tell me what I need to know to pass the exam!" can active flipped learning overcome passivity. TAPS. 2017 Jan 1;2(1):1-6. Available on https://www. researchgate.net/publication/315096327_Just_tell_me_what_I_need_to_know_to_pass_the_ exam_Can_active_ flipped_ lea rning_ overcome_ passivity . Accessed on 29th November 2023.

18. 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. Available on https://pubmed.ncbi.nlm.nih.gov/37510487/ . Accessed on 29th November 2023.

19. Cheung BH, Lau GK, Wong GT, Lee EY, Kulkarni D, Seow CS, etal ChatGPT versus human in generating medical graduate exam multiple choice questions—A multinational prospective study (Hong Kong SAR, Singapore, Ireland, and the United Kingdom). PloS one. 2023 Aug 29;18 (8):e0290691. Available on https://journals. plos.org/plosone/article?id=10.1371/journal.pone.0290691 . Accessed on 29th November 2023.

20. Roumeliotis KI, Tselikas ND. ChatGPT and Open-AI Models: A Preliminary Review. Future Internet. 2023 May 26;15(6):192. Available on https://www.mdpi.com/1999-5903/15/6/192 Accessed on 29th November 2023.

21. Meo SA, Al-Khlaiwi T, AbuKhalaf AA, Meo AS, Klonoff DC. The Scientific Knowledge of Bard and ChatGPT in Endocrinology, Diabetes, and Diabetes Technology: Multiple-Choice Questions Examination-Based Performance. Journal of Diabetes Science and Technology. 2023 Oct 5:19322968231203987. Available on https://www.mdpi.com/1999-5903/15/6/192 .Accessed on 29th November 2023.

22. Passby L, Jenko N, Wernham A. Performance of ChatGPT on Specialty Certificate Examination in Dermatology multiple-choice questions. ClinExp Dermatol [Internet]. 2023 . Available on https://pubmed.ncbi.nlm.nih.gov/37264670/ .Accessed on 29th November 2023.