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

Syed Shirjeel Husain1, Zaid Ansari2, Azhar Hussain3, Sahar Zubair Abbasi4, Tahera Ayoob5, Rimsha Mujahid6

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


Citation: Husain SS, Ansari Z, Hussain A, Abbasi SZ, Ayoob T, Mujahid R. To Evaluate the Efficiency of ChatGPT in Medical Education: An Analysis of MCQ-Based Learning and Assessment [Online]. Annals ASH & KMDC. (ASH & KMDC 26(4):194-200;2023)

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-
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 outcomes3,4,5.

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 Dynam-icity 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 under graduate 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 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.

<table>
<thead>
<tr>
<th>Book name</th>
<th>Author</th>
<th>Edition/publisher</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 Questions and Answers from Kumar &amp; Clark’s Clinical Medicine</td>
<td>Parveen Kumar, Michael Clarke</td>
<td>Second edition Saunders Elsevier</td>
</tr>
<tr>
<td>Rapid review of Clinical Medicine for MRCP Part 2</td>
<td>Sanjay Sharma, Rashmi Kaushal</td>
<td>Second Edition Manson publishing</td>
</tr>
</tbody>
</table>

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

| Total MCQ’s | 307  |
| 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

<table>
<thead>
<tr>
<th>Level of Cognition</th>
<th>Correct (n &amp; %)</th>
<th>In correct (n &amp; %)</th>
<th>Partially correct (n &amp; %)</th>
<th>Total n=307</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>95 (69.8)</td>
<td>40 (29.41)</td>
<td>1 (0.73)</td>
<td>136</td>
</tr>
<tr>
<td>C2</td>
<td>33 (80.48)</td>
<td>18 (41.951)</td>
<td>0</td>
<td>41</td>
</tr>
<tr>
<td>C3</td>
<td>71 (54.61)</td>
<td>50 (38.46)</td>
<td>9 (6.9)</td>
<td>130</td>
</tr>
</tbody>
</table>

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

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

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 assessed 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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{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


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


