Original article

Enhancing Short Essay Question Quality in MBBS Course: A Comparative Study of ChatGPT and Human Collaboration

Andleeb Kanwal[1], Tayyaba Azhar[2], Anam Zahra[3], Zahid Anwar[4], Hajra Talat[5]

[1]Assistant Professor, Obstetric and Gyne, Fatima Memorial Hospital Lahore, [2]Director Medical Education, Department of Medical Education, Fatima Memorial Hospital Lahore, [3]Senior Demonstrator, Department of Medical Education Fatima Memorial Hospital Lahore, [4]Associate Professor of Neonatology (Paediatrics), Department of Paediatrics, Fatima Memorial Hospital, Lahore, [5]Senior Demonstrator, Department Of Medical Education, Fatima Memorial Hospital Lahore

Correspondence to: Andleeb Kanwal, Email: sana755@hotmail.com

ABSTRACT

Background: The short essay questions quality plays an important role in assessing students' knowledge and understanding in educational settings. This study aims to enhance the quality of short essay questions by using ChatGPT and teacher collaboration. Objective of the study is to evaluate the short essay questions for a college-level MBBS course made by ChatGPT and human source.

Materials and methods: This qualitative exploratory study aimed to develop and evaluate essay questions for a college-level MBBS course at Fatima Memorial Hospital. Ethical approval was obtained, and four experienced subject specialists participated. The study involved selecting course learning objectives and developing questions with ChatGPT 3.5 and specialists. Both groups created 20 questions, which were reviewed by independent experts using a checklist with five components: clarity, problem inclusion, structure, English composition, and appropriate length. Questions were rated on a Likert scale from 1 to 5. SPSS version 25 was used for statistical analysis, including t-tests to compare ratings. The study found differences in quality and effectiveness between AI and human-generated questions.

Results: Data analysis was done which showed the mean scores given by human evaluators and AI. Human superseded AI in clarity with the mean score given by human evaluators was 3.51, while AI systems had 3.41. AI accomplished better in structuring the short essay question with the imply rating of 3.63. (p=0.557). AI was advanced in demonstrating the problem inclusion then the human with the score of 3.13 (p=0.774). SEQS made with the aid of the human had good English composition compared to AI (p=0.466). Appropriate length of question became the best factor in which the human and AI each completed same (p=0.917).

Conclusion: This study provides a comprehensive analysis by comparing human and ChatGPT in the quality of short essay questions. The results indicate that AI has the capability to replicate human judgment in certain aspects of question.

Keywords:

ChatGPT, short essay question, artificial intelligence, teacher collaboration

INTRODUCTION

Improving the quality of short essay questions is essential for educators to make sure that the questions accurately examine students' know-how and understanding of the concerned matter. The problem is construction of SEQ is usually its quality. The existing problem is that it requires a detailed knowledge of subject along with following the guidelines for writing good short essay questions as usage of clear and concise language, avoiding ambiguous or indistinct phraseology, and making sure that the questions are aligned with the learning objectives and outcomes. These may involve the use of rubrics for grading, peer evaluation, and automatic scoring structures. Automated scoring systems use algorithms to evaluate essay questions, reducing the workload for educators even as ensuring objectivity and consistency in grading. Overall, enhancing the quality of short essay questions is important for educators to correctly verify students' comprehension of subject and provide meaningful insight to the student knowledge. By following the guidelines and using specific tools, educators can be certain that their short essay questions are of best quality and offer meaningful insights into students' learning. The AI software ChatGPT has gained enormous interest and popularity within a quick span of months since its launch. In the field of education, it's far regarded as a game-changer.[1] The launch of ChatGPT, is predicted to have a much impact across various sectors of society. Nevertheless, the impact that this natural language processing model can also have on education is not yet completely understood. Given ChatGPT capabilities, its effect on education might be sizeable and might cause changes in gaining knowledge

Conflict of interest: The authors declared no conflict of interest exists.
Citation: Kanwal A, Azhar T, Zahra A, Anwar Z, Talat H. Enhancing short essay question quality in MBBS course: A comparative study of ChatGPT and human collaboration. J Fatima Jinnah Med Univ. 2024; 18(2):50-54.

DOI: http://doi.org/10.37018/JFJMU/5054

of objectives, teaching methodologies, and other different elements.[2]

It is much important to apprehend that language model like ChatGPT do not possess the cognitive skills of human beings to manage semantic content efficaciously. Despite this, it has become more and more effective due to the exponential growth, quicker processing speeds, and better algorithms. Language models are able to carry out tasks statistically that human beings can do semantically.[3] The collaboration between human intelligence (HI) and artificial intelligence (AI) may be effective, provided that AI can deliver correct and dependable outputs. The capacity advantages of AI in healthcare were previously mentioned, along with its applications in customized fields of medicine. Furthermore, exploring the usage of AI chatbots in healthcare education gives an interesting road, considering the significant quantity of statistics and complicated standards that healthcare settings may require. Despite this, their effectiveness in medical education remains unexplored.[4] Due to the well-defined and systematic regulations of coding, we anticipated that AI could excel people in the obligations described.

ChatGPT has raised issues among educators approximately the capacity for AI in training.[5] Therefore, the purpose of this study was to investigate the future use of ChatGPT in healthcare education, primarily based on existing proof. ChatGPT was used in making of short essay question for final year MBBS students and then the quality was assessed by subject experts. It aimed to identify potential limitations and concerns that could arise in the application of ChatGPT in the aforementioned contexts.

MATERIALS AND METHODS

This study aimed to develop and evaluate essay questions in a college-level MBBS course. The qualitative exploratory study was conducted in the Medical Education Department of Fatima Memorial Hospital. The study received ethical approval from the institutional review board (IRB). Participants were subject specialists who had a fellowship degree with eight years of post-fellowship experience, total of four specialists participated who provided informed consent before participation. The duration of the study was one month following IRB approval.

Ethical considerations included obtaining informed consent from all subject specialists. Confidentiality and anonymity were maintained throughout the study. For the use of ChatGPT, ethical guidelines were followed to ensure the responsible use of AI in educational research. The commands given to ChatGPT did not involve any sensitive or personal data.

The development of the essay questions involved several steps as first step was selection of learning objectives of the MBBS course which were identified according to study guidelines. Key concepts to be covered in the essay questions were outlined, second step was question development. The subject specialists and ChatGPT developed short essay questions of moderate difficulty aimed at final-year MBBS students based on the identified concepts. ChatGPT 3.5 was used for this task. Subject specialists and ChatGPT each developed 20 questions. The commands given to ChatGPT included specific prompts such as "Develop short essay questions of moderate difficulty for final-year MBBS students based on the following key concepts in Gynaecology (list of concepts). The commands to ChatGPT were given by the researcher who had expertise in medical education.

A team of subject experts, who possessed higher degrees in the subject as well as a Master's in Medical Education, reviewed and evaluated the questions. Both sets of SEQs (developed by specialists and ChatGPT) were evaluated by subject experts who were not involved in the initial question development. They ensured the questions were clear, relevant, and aligned with the course learning objectives.

An SEQ (Short Essay Question) checklist was created, encompassing five components for each question; Clarity of the essay question, Inclusion of a problem, Structure of the question, English composition, Appropriate length of the question. A Likert scale was used to rate each component, with scores ranging from 1 to 5 (1 indicating poor, 2 fair, 3 average, 4 good, and 5 excellent). The scale was applied to individual components. Both sets of SEQs (developed by specialists and ChatGPT) were evaluated by a different team of medical educationist subject experts who were not involved in the initial question development. The minimum number of SEQs selected for significant comparison with confidence was 30. Descriptive statistical analysis was performed using SPSS version 25. The statistical tests used included t-tests to compare the ratings of SEQs developed by specialists and ChatGPT.

The SEQs developed by specialists and ChatGPT were compared based on the ratings provided by the subject experts. The analysis focused on the average scores for each component and overall score to

determine the quality and effectiveness of the questions. Descriptive statistical analysis was done.

RESULTS

The data provides a comparative analysis of the mean scores given by human evaluators and AI systems for each component of SEQ. On analysing the data, it was found that human superseded ChatGPT in clarity. The mean score given by human evaluators was 3.51, while AI systems received a slightly lower score of 3.41 (0.42). This suggests that human evaluators and ChatGPT exhibited a similar understanding and assessment of clarity in the evaluated content. AI accomplished better in structuring the short essay questions with the imply rating of 3.63+SD (p=0.557). AI was advanced in demonstrating the problem inclusion of SEQ's compared to the human with the score of 3.13+SD (p=0.774). SEQ's made with the aid of the human had better English composition compared to AI (p=0.466). Appropriate length of question became the best factor in which the human and AI were equally comparable (p=0.917) (Table 1). The graphical illustration of comparison between mean scores of human and AI evaluation of SEQ's is given in Figure 1.

Correlation among the ratings of human and ChatGPT assessment of SEQ's was assessed. A robust statistically significant effective correlation among the scoring of human and AI evaluation of SEQ for all the components was observed (Table 2). Table 2 gives precious insights into the correlation among human assessment scores and the corresponding assessments conducted by AI algorithms. Each evaluation component demonstrated a robust effective correlation, indicating a regular alignment among human judgment and ChatGPT generated rankings.

Firstly, the thing of clarity demonstrates a correlation coefficient of 0.909, indicating a strong relationship between human and AI evaluations in terms of the way a piece of writing conveys its message. This finding shows that AI systems can efficaciously figure and evaluate the readability of written content. The assessment of structure of SEQ's shows a correlation coefficient of 0.926, proving that both human evaluators and ChatGPT algorithms understand the importance of well-organized and logically written portions. The high correlation means that ChatGPT algorithms can efficiently assess the structural integrity of numerous texts. The factor of problem inclusion displays a high correlation coefficient of 0.951, indicating a good alignment among human and

Table 1: Comparison between mean scores of human and AI evaluation of SEQ

Components	Human	ChatGPT	p-value[1]
Clarity	3.51	3.41	0.420
Structure	3.56	3.63	0.557
Problem inclusion	3.09	3.13	0.774
Good English composition	3.74	3.67	0.466
Appropriate length	3.37	3.36	0.917

[1]Paired sample t test

Table 2: Correlation between the scores of human and AI evaluation

Component	Correlation[1]	p-value
Clarity	0.909	0.005**
Structure	0.926	0.003**
Problem inclusion	0.951	0.001**
Good English composition	0.970	0.0001**
Appropriate length	0.973	0.0001**

** Statistically significant (p≤0.01)

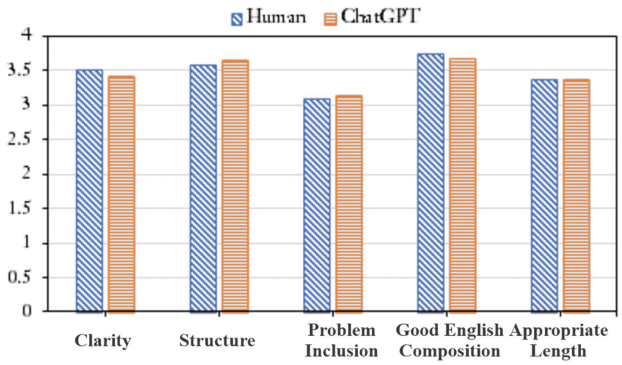

Figure 1: Comparison between mean scores of Human and ChatGPT evaluation of SEQ

ChatGPT evaluations regarding the inclusion and coverage of applicable issues within written content material. The AI algorithms appear to correctly understand and evaluate the presence of trouble-cantered factors. Furthermore, the assessment of good English composition demonstrates a correlation coefficient of 0.970. This indicates a sturdy agreement among human evaluators and AI systems in assessing the fineness of English composition, which includes grammar, syntax, and ordinary linguistic proficiency. AI algorithms show their capability in gauging the linguistic factors of written content. Lastly, the factor of appropriate length exhibits a correlation coefficient of 0.973. This shows a noteworthy concurrence among human and AI reviews in determining the most fulfilling length of written content. The capability of AI algorithms to evaluate the appropriateness of the textual content's period can be useful in various contexts, along with summarization or content material technology. All the correlations noted above are statistically big, as denoted through the associated p-values (p≤0.01). This similarly strengthens the reliability and validity of the

found relationships between human and AI evaluation ratings.

DISCUSSION

The use of ChatGPT in education is still being evaluated as in terms of its benefits and capabilities. This study highlights the use of ChatGPT for developing and assessing SEQs. The use of AI models in various fields, including education and healthcare, has become increasingly prevalent. AI can increase efficiency and reduce the workload of healthcare professionals, enabling them to focus on more complex tasks that require their expertise and clinical judgement. However, it is essential to acknowledge that the ultimate source of knowledge and decision-making ability comes from the human mind. Therefore, AI models, like ChatGPT, should be viewed as a supporting tool for healthcare professionals rather than a replacement for their decision-making processes.[6] The release of ChatGPT, is anticipated to have a far-reaching impact across various sectors of society. Nevertheless, the impact that this natural language processing tool may have on education is not yet fully understood. Given ChatGPT's capabilities, its impact on education could be significant and may lead to changes in learning objectives, teaching methodologies, and other aspects of education.[7] This research evaluated the opportunities and challenges of using ChatGPT for assessments in medical education, including the risks and benefits of these tools. It also addresses the challenges in appropriate use of AI. It is obvious from results of this study that although it is an advancement in the capabilities of AI, the technology relies on a combination of supervised and reinforcement learning methods and employs human trainers to facilitate both approaches.[8] When it comes to development of short essay question, a task mainly linked to cognitive function, AI performed well in the domain of structure of question and problem inclusion. This specifically needs a higher level of training be it a human or AI. AI models have the capability to automate certain tasks that are traditionally performed by humans. However, it is essential to acknowledge that the ultimate source of knowledge and decision-making ability comes from the human mind. The other aim of this study was to investigate the pros and cons of ChatGPT use for assessment in medical education. Based on review of the existing literature along with results of present study, ChatGPT benefits included the possibility of improving personalized learning, clinical reasoning and understanding of complex medical concepts.[9] Hence, there exists a potential for the AI language models in overcoming language barriers, it is vital to establish ethical standards and best practices to mitigate these risks. Although these models can support better, it is still necessary to acknowledge their use appropriately. With the continuous improvement of AI technology, one can anticipate that AI will be employed in more advanced ways to aid in the identification and creation of novel treatments and therapies.[10] After ChatGPT is created and implemented, it is essential to continuously monitor and evaluate its performance and effects and make necessary changes to maintain its effectiveness and ethics over time. Hence it can generate conversational responses to user prompts and has potential applications in education.[11] ChatGPT can access large amounts of medical data and provide accurate and timely answers to clinical questions, thus improving access to up to date information.[12]

CONCLUSION

The study presents a comparison of ratings among human evaluators and AI structures for numerous components of SEQ. The effects imply that, usual, there wasn't any statistically big differences between the reviews of human judges and AI algorithms. This indicates that AI structures can efficiently emulate human judgment in assessing structure, problem inclusion, and appropriate length in written content. However, it's important to note that further research and validation are required to ascertain the robustness and generalizability of these findings.

It also gives a complete analysis of the correlation between human and AI evaluation scores throughout diverse additives. The constant and statistically large, tremendous correlations indicate a robust alignment between human judgment and AI-generated assessments. These findings spotlight the potential of AI algorithms in appropriately comparing written content in terms of clarity, structure, problem inclusion, good English composition, and appropriate length. As AI keeps to adapt, it holds great promise in enhancing the efficiency and objectivity of content and material evaluation techniques. This study evaluates the opportunities of use of ChatGPT in assessment, together with the shortcoming. It additionally addresses the problems in its appropriate use. The research paper concludes that AI's utilization in short essay question formation requires better training which will give possibility of its use to enhance the quality of question and challenges that require cautious attention.

The limitation of study was that the essay questions were evaluated by subject experts only, they can be administered to students enrolled in the course, who can provide feedback on their experience. Additionally, a team of subject matter experts can be increased who has reviewed the questions and provided feedback on their quality.

REFERENCES

1. Tong Y, Zhang L. Discovering the next decade. Synthetic and Systems Biotechnology. 2023;8(2):220-3.
2. Zhai X. ChatGPT User experience: Implications for education. SSRN Electronic Journal. 2022. 4312418
3. Floridi L. AI as agency without Intelligence: On ChatGPT, large language models, and other generative models. Philos. Technol. 36, 15 (2023). https://doi.org/10.1007/s13347-023-00621-y
4. Merow C, Serra-Diaz JM, Enquist BJ, Wilson AM. AI chatbots can boost scientific coding. Nat Ecol Evol. 2023 Jul;7(7):960-962. doi: 10.1038/s41559-023-02063-3.
5. Gussow L. Technology & Inventions. Emergency Medicine News. 2023;45(3):19.
6. Xue VW, Lei P, Cho WC. The potential impact of ChatGPT in clinical and translational medicine. Clinical and Translational Medicine. 2023;13(3). e1216.
7. Zhai X. ChatGPT user experience: implications for education. SSRN Electronic Journal. 2022. Available at SSRN 4312418
8. Macdonald C, Adeloye D, Sheikh A, Rudan I. Can ChatGPT draft a research article? An example of population-level vaccine effectiveness analysis. Journal of Global Health. 2023;13. doi: 10.7189/jogh.13.01003
9. Sallam M, Salim N, Barakat M, Al-Tammemi A. ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations. Narra J. 2023;3(1).e103. doi: 10.52225/narra.v3i1.103
10. King MR. The Future of AI in medicine: A perspective from a Chatbot. Annals of Biomedical Engineering. 2022;51(2):291-5
11. Fergus S, Botha M, Ostovar M. Evaluating academic answers generated using ChatGPT. J Chem Educ. 2023;100(4):1672-5.
12. Baumgartner C. The opportunities and pitfalls of ChatGPT in clinical and translational medicine. Clinical and Translational Medicine. 2023;13(3).