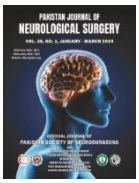

PAKISTAN JOURNAL OF NEUROLOGICAL SURGERY (QUARTERLY) – OFFICIAL JOURNAL OF PAKISTAN SOCIETY OF NEUROSURGEONS

Original Article

Chat GPT Develops Multiple Choice Questions (MCQs) for Postgraduate Specialty Assessment – A Reality or a Myth?

Faridah Amir Ali[1], Salman Sharif[2], Madiha Ata[1], Nirali Patel[3], Muhammad Rafay[2], Hasan Raza Syed[3], Saima Perwaiz Iqbal[4]

1Department of Family Medicine, Indus University of Health Sciences, Karachi
[2]Department of Neurosurgery, Liaquat National Hospital and Medical College, Karachi
[3]Department of Neurosurgery, Children's National Medical Center, Washington DC, USA
[4]Department of Family Medicine, Shifa College of Medicine, Islamabad, Pakistan

ABSTRACT

Objective: Multiple Choice Questions (MCQs) are a valuable assessment tool, but creating them to match learning goals needs experts. AI, like ChatGPT, might offer an alternative. A study showed MCQs made for medical programs by ChatGPT and the faculty. This study compares faculty-made MCQs to ChatGPT-made ones for a post-grad program.

Material & Methods: Specific learning objectives of a module from a medical and surgical program were extracted. One mid-level faculty and the AI software developed MCQ from each learning objective with a clinical scenario. Two subject and medical education experts from each specialty were blinded and given a standardized online tool to rate the technical and content quality of the MCQs in five domains; the item, vignette, question stem, response options, and overall quality.

Results: For the medicine and allied specialty, 23 MCQs in each set were assessed. There was no significant difference between each variable, the overall quality of MCQs, or the odds of the decision to accept the questionnaire. Two sets of 24 MCQs were assessed for the surgical and allied specialty. There was no difference between the domains for "Item" and "Vignette". For the domain "question stem", MCQs developed by faculty were more grammatically correct (p-value 0.02). There was no difference in the quality or odds of the decision to accept.

Conclusions: AI's impact on education is undeniable. Our findings indicate that in specific areas, faculty outperformed ChatGPT, though overall question quality was comparable. More research is necessary, but ChatGPT could potentially streamline assessment development, saving faculty substantial time.

Keywords: ChatGPT, AI, Artificial intelligence, Postgraduate Assessments, Specialty, MCQs.

Corresponding Author: Salman Sharif
Department of Neurosurgery
Liaquat National Hospital and Medical College
Karachi, Pakistan
E-mail: sharifsalman73@gmail.com

Date of Submission: 25-12-2023
Date of Revision: 01-01-2024
Date of Acceptance: 20-03-2024
Date of Online Publishing: 31-3-2024
Date of Print: 31-3-2024

DOI: 10.36552/pjns.v28i1.963

INTRODUCTION

The COVID-19 pandemic brought a paradigm shift in the teaching, learning, and assessment pedagogy, with e-learning becoming the "new normal".[1] Simultaneously, recent advances in Artificial Intelligence (AI) software have further revolutionized education throughout the world. It is imperative that Graduate Medical Education (GME) continues to evolve and keep pace with these advancements. AI-based tools are already being utilized in medicine for generating and checking tasks, creating clinical scenarios, formulating quizzes, and aiding in research.

As the teaching epistemology changes, so does the assessment methodology. In developing countries such as Pakistan, GME has recently been recognized as a critical institution in the education of post-graduate trainees. Although clinicians working in academic institutions offering post-graduate training are being encouraged to pursue a qualification in higher professional education (HPE), it is not a mandatory requirement yet.[2] Even in developed countries, due to excessive workload, engaging medicine and surgery consultants for post-graduate teaching and assessments is a challenge.[3] For the same reason, hospitals and universities offering post-graduate medical training struggle with enhancing the quality of teaching and assessment activities, as the educationists are not content experts and vice versa.

Multiple choice questions (MCQs) have long been used as an effective assessment tool for post-graduate medical education. Yet, developing MCQs aligned with specific learning outcomes requires content and process experts. AI could be a potential substitute for these experts. "ChatGPT" (Generative Pretrained Transformer), an Open AI software, was recently able to pass the USMLE which is developed by a "National Board of Examiners".[4] A recent study has also shown that MCQs developed for an undergraduate medical program by ChatGPT were comparable to those developed by faculty.[5] This makes academics question whether this software, which is capable of creating and passing a human-made assessment, can develop MCQs for GME programs. Consequently, this may reduce the workload on multi-tasking clinicians, who are all-in-one administrators, researchers, teachers, and clinicians. While the debate surrounding the ethical and moral implications of the use of AI in academia is ongoing, it is time to assess the quality of the work produced by it.

This study compares the quality and content of MCQs generated by medical and surgical faculty versus ChatGPT for a post-graduate program.

MATERIALS AND METHODS

Study Design & Setting

A cross-sectional analytical online study. The study was conducted at the Department of Neurosurgery, Liaquat National Hospital, Karachi with approval from the ethical committee of the hospital. The data was collected with permission from the participants.

Sampling Technique

Specific learning objectives of a module from a medical and surgical PGME program respectively were extracted by the principal investigators (PI) from Pakistan. One mid-level faculty with at least two years of post-graduate experience were identified from each specialty. The same instructions were given to the faculty and the AI software, to "develop one MCQ from each learning outcome with a clinical scenario and five plausible options with one best answer". The MCQs were randomized by PI and a unique ID was allotted to each.

Data Collection Tool

Two subject and medical education experts from both the medical and surgical specialties were given a standardized online tool to rate the technical and content quality of MCQs.[6] The MCQs were scored on five domains; the item, vignette, question stem, response options, and overall quality. The final assessment question was regarding the decision to accept, reject, or modify the MCQ.

Data Analysis

Data were analyzed on PRISM (version 9.2.0, GraphPad Software, San Diego CA). The statistician analyzing the data was blinded to which MCQs were written by faculty or AI. Responses to each item used to assess the MCQs as well as the five domains and the decision to accept, reject, or revise the MCQ were analyzed using Fischer's exact tests. Total scores for each set of MCQ for both the raters were compared through unpaired student's t-test. The odds ratio was calculated for the final question regarding the decision to accept, reject, or revise the MCQ. A p-value of <0.05 was considered as significant.

RESULTS

Table 1 shows the comparison of the difference between scores for various domains of MCQs developed by the faculty versus Chat GPT for medical and surgical postgraduate programs.

Table 1: Comparison of cumulative assessment of ChatGPT and faculty-developed MCQs by both raters from medicine and surgical specialty.

Specialty	Medicine (n=23)			Surgery (n=24)		
MCQs	**Chat GPT**	**Faculty**	**P-value**	**Chat GPT**	**Faculty**	**P-value**
ITEM			**0.0050**			**0.5795**
Assesses a Program/Course Learning Outcome	21	22	0.4889	24	22	0.4894
Uses Clear language	20	22	0.607	17	20	0.4936
Represents Current Knowledge / Best Practice	20	22	0.233	18	21	0.4614
Does **not** Measure opinion	20	22	0.233	17	17	>0.99
VIGNETTE						0.8593
Is necessary to respond to the question stem	0	0	-	17	13	0.3715
Presents clinical information in a logical sequence	0	0	-	20	17	0.4936
Does **not** contain red herrings (i.e., information meant to deceive)	0	0	-	9	15	0.3495
QUESTION STEM			0.1041			0.6958
Is a complete statement (Can stand-alone)	20	19	>0.99	21	19	0.70
Ends with a question mark	20	22	0.233	24	11	<0.0001*
Can be answered without viewing the response options	13	12	>0.99	18	17	>0.99
Is stated positively	21	22	0.488	21	23	0.608
Contains only necessary information	20	22	0.233	18	22	0.244
Is grammatically corrected	20	21	0.607	13	21	0.02*
Contains **no** vague or ambiguous terms	20	20	>0.99	13	19	0.124
RESPONSE OPTIONS			0.2287			0.2951
Are all plausible	18	22	0.049*	17	12	0.237
Are mutually exclusive (i.e., do not include overlapping content)	16	22	0.0092	13	15	0.77
Follow grammatically from the question stem	19	18	>0.99	19	19	>0.99
Are similar in length	18	15	>0.99	19	19	>0.99

Are similar in terminology	20	22	0.607	12	16	0.380
Are presented in a logical order (e.g., alphabetically, numerically)	3	8	0.09	20	13	0.059
Contains **no** extraneous trivia	20	22	0.233	13	15	0.77
Do not contain repeated elements that should appear in the stem	20	13	0.04*	11	16	0.244
Do not include all of the above	23	22	>0.99	12	18	0.135
Do not include any of the above	23	22	>0.99	12	17	0.23
The overall quality of MCQ	3.30 +/- 1.02	3.55 +/- 0.51	0.32	3.5 +/- 1.06	3.33 +/- 0.92	0.56
Decision to accept	12	13	0.76 1.32 (0.4 to 4.3)	12	19	0.068 3.8 (1.07 to 13.52)

*significant p-value (< 0.05). Odds ratios are calculated as the odds of accepting a question written by a faculty compared to that written by ChatGPT

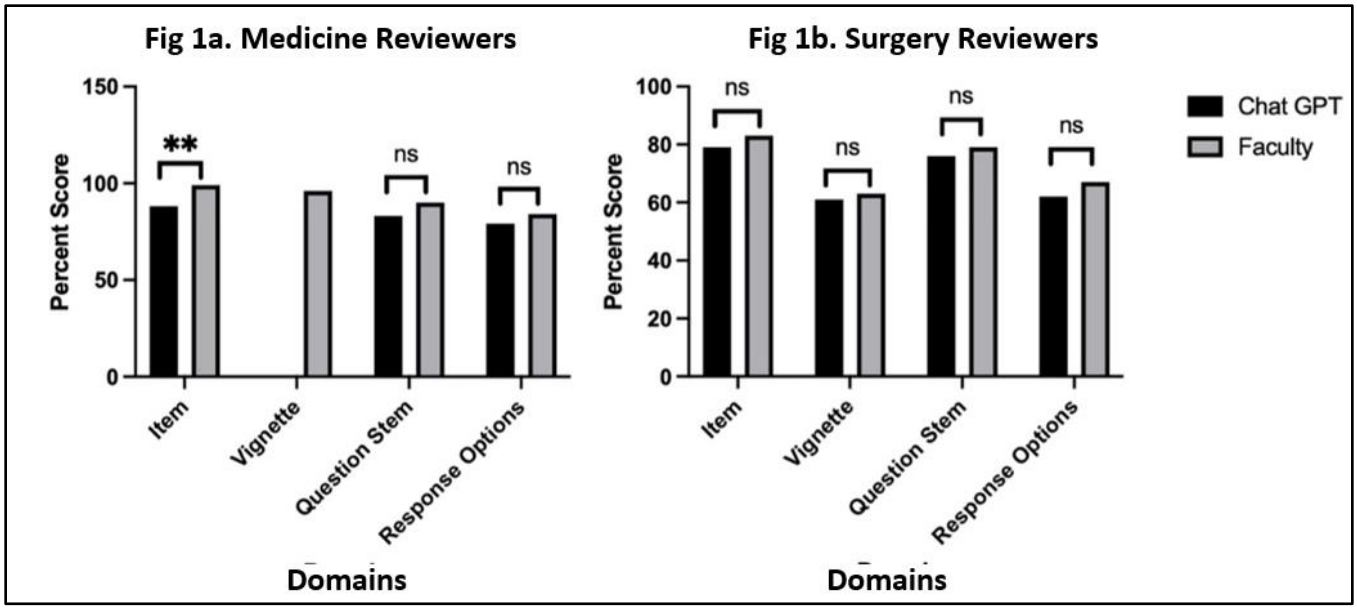

Figure 1: Comparison of various domains of MCQs developed by ChatGPT and faculty of Medical and Surgical Specialties. (ns-not significant, **p-value <0.005).

Medicine and Allied Specialty

23 MCQs in each set were assessed by two blinded raters. Table 1 and Figure 1 show that out of the four domains assessed, the scores were significantly different for "Item", where the overall score of MCQs developed by faculty was more than that of ChatGPT. However, there were no significant differences between each variable compared individually under the same domain.

For the "Vignette", the two MCQ sets were not comparable, as there was no vignette for those developed by ChatGPT. There was no significant difference between scores for variables in the "Question Stem" category among both MCQ sets. For the domain of "response options", there was no significant difference among overall scores. The faculty scored higher on a few variables, comparing if the MCQs are mutually exclusive (p-value 0.009) and plausible (p-value 0.049). Also, response options for MCQs developed by ChatGPT contained fewer repeated elements of the stem as compared to faculty (p-value 0.04).

There was no difference in the overall quality of MCQs (Figure 2) and the odds of the decision to accept among both groups were the same.

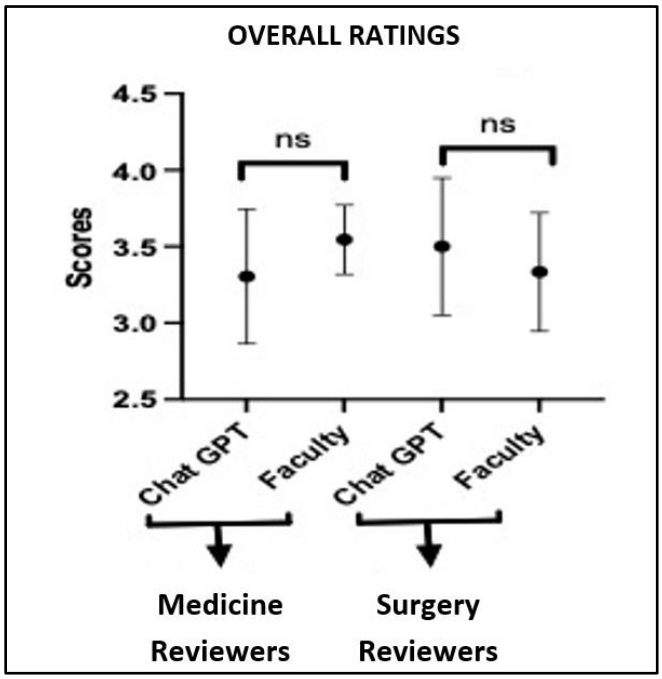

Figure 2: Comparison of overall quality of MCQs as rated by respective specialist raters. (ns-not significant).

Surgical and Allied Specialty

Table 1 and Figure 1 show that out of two sets of 24 MCQs assessed for Surgical and Allied specialty, there was no difference between the domains for "Item" and "Vignette". For the domain "question stem", the MCQs developed through ChatGPT were more likely to end with a question mark (p-value <0.0001) however, those of faculty were more likely to be grammatically correct (p-value 0.02).

For "response options" there was no difference between the scores of the 2 MCQ sets. Similarly, there was no significant difference in the overall quality. The odds of the decision to accept the MCQs were also not different among the two groups.

DISCUSSION

The key to maximizing the use of AI is to keenly identify the strengths and weaknesses of this technology and to use it to our advantage. Not every function can be performed as efficiently as the human brain by AI.[7] In this context, this study was designed to compare the quality of AI-generated MCQs and those made by faculty for GME programs. The results showed that in certain domains, the faculty performance was superior to that of ChatGPT.

For the MCQs developed for medicine specialty, faculty performed better in various domains as compared to ChatGPT. A recent study on biology MCQs created by AI found moderate reliability with a Cronbach's alpha of 0.623. The same study found all MCQs to be valid except one. However, there was no comparison of these MCQs with man-made ones.[8] Our study does not have any post-hoc analyses for comparison, as the questions have not been piloted on students yet.

In terms of distractors, those generated by ChatGPT for medicine were less plausible than those by the faculty. This limitation of automated MCQ generation using ontology has been recognized previously as well.[9]

Interestingly, the module objectives selected for the specialty of medicine dealt with ethical issues; hence this finding highlights the emotional and intellectual superiority of the human mind over machines; at least in the present time.[10] There is growing evidence that AI has inherent biases which are reflected in its responses when dealing with ethical scenarios.[11]

Two other variables regarding the quality of distractors were found to be significantly better in the faculty-made MCQs. These were the 'mutual exclusivity' of the distractors and the lack of 'repeated elements that should appear in the stem'. This reflects higher order thinking of the human brain, which makes a conscious effort to follow the 'best practices' of the item (MCQ) so that each

question is more palatable for the students without engaging them in frivolous details.

For the surgical MCQs, the lead-ins constructed by ChatGPT were better in terms of punctuation but those by the faculty were more grammatically sound. Given that ChatGPT is primarily a 'language software', this discovery comes as a surprise. Interestingly, the software itself admits that it is prone to making language errors when dealing with highly complex subject matter such as medicine, due to the limitation of technical knowledge. Likewise, it is consistently reported in the literature that the responses generated by ChatGPT have inconsistencies and inaccuracies. Even OpenAI, the developers of ChatGPT admit its limitations including the production of responses which are reasonable but wrong.[12]

LIMITATIONS & RECOMMENDATIONS

This study was conducted soon after Chat-GPT was released to the general public by open AI. The user giving the instructions to the AI model was still a novice in its use and lacked adequate training for prompt engineering to optimize the output.[13] So, it can be assumed that with more customized prompts, the quality of the MCQs generated by Chat-GPT may be different from the initial response. Moreover, only a small number of specialties from PGME were evaluated in the study which can lead to difficulty in the generalization of the findings.

The authors recommend that the use of AI should not be admonished when it can be used to increase the productivity of busy clinicians. However, the onus of the quality of the work produced shall lie with the 'human being' giving the prompts. Only a subject expert confident in vetting the content of the output should opt for using generative AI tools like Chat-GPT. Educational institutes and governments should also hasten in making policies for the use of AI in health professions education including PGME to ensure its ethical use.

CONCLUSION

The reality of AI in education is irrefutable. The key is not to 'beat' it but to 'join' it. While medical educators are finding ways to identify content developed by AI in academia, more research needs to be done to synergize its use to reduce workload. The time saved from the ethical and efficient use of language models such as ChatGPT can be utilized in completing more creative and intellectual tasks by clinical faculties. Though appealing, the limitations of ChatGPT in the field of medicine warrant its judicious use as the quality may not be optimal. Thus, it is recommended that AI be used as an assisting tool, but the final output should be reviewed and refined by content experts.

ACKNOWLEDGEMENT

We are thankful to Dr. Shaista Saghir for facilitating the study.

REFERENCES

1. Shima Tabatabai, P. D. COVID-19 impact and virtual medical education. J Adv Med Educ Prof. 2020;8(3):140. Doi: 10.30476/jamp.2020.86070.1213
2. Latif MZ, Wajid G. Reforming medical education in Pakistan through strengthening departments of medical education. Pak J Med Sci. 2018;34(6):1439. Doi 10.12669/pjms.346.15942
3. Nassar AK, Waheed A, Tuma F. Academic clinicians' workload challenges and burnout analysis. Cureus. 2019;11(11). Doi: 10.7759/cureus.6108
4. 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 Digital Health. 2023;2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198
5. Cheung BHH, Lau GKK, Wong GTC, Lee EYP, Kulkarni D, Seow CS, et al. 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;18(8):e0290691.

https://doi.org/10.1371/journal.pone.0290691
6. Reniers J. Writing Fair and Effective MCQs - Checklist. Office of Teaching and Learning - University of Guelph. https://otl.uoguelph.ca/system/files/Writing%20Fair%20and%20Effective%20MCQs%20-%20Checklist.pdf. Updated: February 2020. Accessed August 2023.
7. Grace K, Salvatier J, Dafoe A, Zhang B, Evans O. When will AI exceed human performance? Evidence from AI experts. J Artif Intell Res. 2018;62:729-754. Doi: https://doi.org/10.1613/jair.1.11222
8. Nasution NEA. Using artificial intelligence to create biology multiple choice questions for higher education. Agric Environ Educ. 2023;2(1). https://doi.org/10.29333/agrenvedu/13071
9. Vinu EV. A novel approach to generate MCQs from domain ontology: Considering DL semantics and open-world assumption. J Web Semantics. 2015;34:40-54. https://doi.org/10.1016/j.websem.2015.05.005
10. Kambur E. Emotional intelligence or artificial intelligence? Emotional artificial intelligence. Florya Chronicles of Political Economy. 2021;7(2):147-168. DOI: 10.17932/IAU.FCPE.2015.010/fcpe_v07i2004
11. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. JAMA. 2019;322(24):2377-2378. Doi: 10.1001/jama.2019.18058
12. Fitria TN. Artificial intelligence (AI) technology in OpenAI ChatGPT application: A review of ChatGPT in writing English essay. ELT Forum: J English Lang Teach. 2023;12(1):44-58. Doi: https://doi.org/10.15294/elt.v12i1.64069
13. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. Journal of Medical Internet Research. 2023;25:e50638. Doi: 10.2196/50638

Additional Information

Disclosures: Authors report no conflict of interest, and all data can be provided if needed.
Ethical Review Board Approval: The study conformed to the ethical review board requirements. Approval was taken from the ethical board of the hospital.
Human Subjects: Consent was obtained by all patients/participants in this study.
Conflicts of Interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following:
Financial Relationships: All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work.
Other Relationships: All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.
Financial Support: No financial support has been taken from any funding agency or organization.
Data Availability Statement: The data record could be available at the request of the corresponding author.

AUTHORS CONTRIBUTIONS

Sr.#	Author's Full Name	Intellectual Contribution to Paper in Terms of:
1.	Salman Sharif	1. Study design and methodology.
2.	Madiha Ata, Faridah Amir Ali & Nirali Patel	2. Paper writing.
3.	Muhammad Rafay & Saima Perwaiz Iqbal	3. Data collection and calculations.
4.	Nirali Patel	4. Analysis of data and interpretation of results.
5.	Faridah Amir Ali	5. Literature review and referencing.
6.	Hasan Raza Syed	6. Editing and quality insurer.