Open Access

ORIGINAL ARTICLE

ChatGPT-4 versus human generated multiple choice questions - A study from a medical college in Pakistan

Muhammad Ahsan Naseer[1], Yusra Nasir[2], Afifa Tabassum[3], Sobia Ali[4]

Department of Health Professions Education, Liaquat National Hospital & Medical College, Karachi, Pakistan[1,2,3,4]

ABSTRACT

Background: There has been a growing interest in using artificial intelligence (AI) generated multiple choice questions (MCQs) to supplement traditional assessments. While AI claims to generate higher-order questions, few studies focus on undergraduate medical education assessment in Pakistan.

Objective: To compare the quality of human-developed versus ChatGPT-4-generated MCQs for the final-year MBBS written MCQs examination

Methods: This observational study compared ChatGPT-4-generated and human-developed MCQs in four specialties: Pediatrics, Obstetrics and Gynecology (Ob/Gyn), Surgery, and Medicine. Based on the table of specifications, 204 MCQs were ChatGPT-4-generated and 196 MCQs were retrieved from the question bank of the medical college. ChatGPT-4-generated and human-generated MCQs were anonymized and MCQs quality was scored using a checklist based on the National Board of Medical Examiner criteria. Data was analyzed using SPSS version 23 and Mann-Whitney U and Chi square tests were applied.

Results: Out of 400 MCQs, 396 MCQs were included in the final review as four MCQs were not according to the table of specification. Total scores were not significantly different between human-generated and ChatGPT-4 generated MCQs (p=0.12). However, human-developed MCQs performed significantly better than ChatGPT-4-generated MCQ in Ob/Gyn (p=0.03). Human-developed MCQs scored better than ChatGPT-generated MCQs in the item checklist “stem includes necessary details for answering the question’’ in Ob/Gyn and Pediatrics ($p < 0.05$) as well as in "Is the item appropriate for cover the options rule"? in Surgery.

Conclusion: ChatGPT-4 has the potential to assist in medical examination MCQ development using a well-structured and specific prompting. However, ChatGPT-4 has limitations where in depth contextual item generation is required.

Key Words: Artificial intelligence, Multiple choice questions, Undergraduate medical examination, ChatGPT-4

Doi: https://doi.org/10.53685/jshmdc.v5i2.253

Corresponding Author:
Yusra Nasir
Lecturer
Department of Health Professions Education
Liaquat National Hospital & Medical College, Karachi, Pakistan
Email address: yusra.nasir@live.com
Received:13.09.2024, 1st Revision: 12.10.2024
2st Revision: 06.11.2024, Accepted: 10.12.2024

How to cite this article: Naseer MA, Nasir Y, Tabassum A, Ali S. ChatGPT-4 versus human generated final year MBBS multiple choice questions – A study from a medical college of Pakistan. J Shalamar Med Dent Coll. 2024; 5(2): 58-64. doi: https://doi.org/10.53685/jshmdc.v5i2.253

INTRODUCTION

In contemporary medical education, the quality of assessment tools, particularly multiple-choice questions (MCQs), plays a pivotal role in evaluating the knowledge and competency of medical students. MCQs evaluate understanding and develop critical thinking and decision-making skills crucial for real-world scenarios. Additionally, they offer efficient evaluation of a large volume of information quickly and provide objective grading criteria, ensuring fairness and consistency.[1]

It is essential for educators to create well-crafted MCQs that challenge students to think critically and apply their knowledge in real-world scenarios. These questions not only assess students' understanding of complex medical concepts and identify areas for improvement, but they also encourage deeper engagement by pushing students to synthesize knowledge and improve their critical thinking skills.[2]

With advancements in artificial intelligence (AI), particularly the emergence of powerful language

models like ChatGPT, there is growing interest in utilizing AI-generated MCQs to supplement the traditional question banks. AI-generated MCQs have the potential to provide personalized and adaptive learning experiences for students. This can be of help to enhance student engagement and to improve the retention of information.[3] Moreover; AI-generated MCQs can save teachers' time by automating the process of generating and grading assessments.[4]
Previous studies have shown that ChatGPT can solve higher-order questions, including those in the United States Medical Licensing Examination (USMLE). This indicates that the ChatGPT programming includes cognitive aspects such as logic and reasoning. However, evidence of AI's abilities to produce reasoning-based MCQs is still not available.[5]

Due to the significance of MCQs in medical education, it is vital to evaluate and compare the quality of AI-generated MCQs with those crafted by humans. Such evaluation would provide insights into the strengths and weaknesses of AI as an automated MCQ generator for undergraduate MBBS theory exams.

The use of AI-generated MCQs in undergraduate medical education in Pakistan is yet to be investigated. Farida et al. stated that faculty-created MCQs outperformed the ChatGPT-developed MCQs for a postgraduate program.[6] However, the details of the prompts were unclear, leaving a gap in understanding the methods used. The purpose of this study was to compare the quality of human-generated MCQs versus ChatGPT-4 generated MCQs for the final year MBBS written MCQs exam, based on National Board of Medical Examiners (NBME) guidelines. The MCQs were generated using a systematic approach, including contextual prompting and re-prompting.[7]

METHODS

The study used an observational design to compare GPT-4-generated MCQs to those crafted by human experts for final-year MBBS theory examinations based on National Board of Medical Examiner (NBME) recommendations for four specialties: Pediatrics, Obstetrics/Gynecology (Ob/Gyn), Surgery, and Medicine. The study was completed in three months (June – August, 2024) and was conducted at Liaquat National Hospital & Medical College, Karachi. To ensure a representative and unbiased selection of MCQs, each specialty used the Table of Specifications (TOS) established by concerned department faculty in consultation with the examination department. All MCQs were selected from mid-term and pre-professional final-year MBBS examination papers, administered during the previous three years (2021-2023), with the provision that the questions followed the TOS.

The sample size for MCQs was calculated using the OpenEpi calculator, based on 85% confidence interval and 80% statistical power.[8] The initial calculation resulted in a sample size of 362 MCQs. To ensure equal representation of the four specialties (Pediatrics, Ob/Gyn, Surgery, and Medicine), the sample size was adjusted to 400 MCQs, with 100 questions selected from each specialty. This adjustment allowed for a balanced comparison of ChatGPT-4 and human-generated MCQs in each field while maintaining statistical rigor.

Steps involved in data collection:

I. The examination department approved the selection and coding of 50 MCQs that aligned with the TOS for each specialty from the existing question bank. However, 50 MCQs from Ob/Gyn, 50 from Surgery, 47 from medicine, and 49 from Pediatrics were retrieved.

II. ChatGPT-4 was prompted to generate the remaining MCQs per specialty by attaching TOS. Number of questions generated were 50 from Ob/Gyn, 50 from surgery, 53 from Medicine and 51 from Pediatrics. This study used contextual prompting to guide the AI to generate MCQs.[9] Contextual prompt provides detailed guidance, and have been demonstrated to enhance task-specific outcomes.[7] In our case, we specified the subject (e.g., Surgery), TOS, level of the learner (final year MBBS), number of questions required, and cognitive level of the questions (application-based).

III. This structured approach for prompting ensured the generation of relevant and targeted MCQs for our study. For example, for Surgery, the initial prompt used was "Based on attached Surgery TOS for final year MBBS, generate (Number of questions) application based MCQs". As a result of this prompt, the created MCQs included 5 options. The prompt was further refined as "Based on attached Surgery TOS for final year MBBS, generate (Number of questions) knowledge application based MCQs. Give 4 options for each MCQ". Similar steps were used for the other 3 subjects i.e. Ob/Gyn, Pediatrics,

and Medicine. The ChatGPT-4 response and the sample of MCQ generated as a result of the prompt is presented in Figure 1.

IV. Both ChatGPT-4 and human-generated MCQs were anonymized and coded to maintain evaluator blindness during the faculty review process.

V. To assess the quality of ChatGPT-4 generated and human-constructed MCQs, a review checklist based on NBME guidelines was developed. The checklist consisted of 15 statements with a 'yes' or 'no' option for each statement. One mark was given for 'yes' while 'no' received zero. The maximum total score was fifteen (15). Three medical educationists examined and finalized the checklist to ensure its accuracy and relevance.

VI. A review team of faculty members (three subject specialists, from each discipline and one medical educationist) was then invited to participate in the evaluation process. Using the checklist each team evaluated the MCQs (ChatGPT-4-generated and Human-developed) of their own discipline. The overall process of data collection is shown in Figure 2.

ChatGPT-4 Response
The provided Table of Specification (TOS) outlines topics and subtopics along with their classification into diagnosis, investigation, treatment, pathophysiology, and prevention for a General Surgery exam. Based on this information, I will generate knowledge application-based MCQs covering a broad range of topics from the table.
MCQ 6:
A 65-year-old woman presents with sudden onset of severe abdominal pain radiating to the back. She has a history of hypertension and smoking. On examination, her abdomen is tender with guarding and decreased bowel sounds. What is the most likely diagnosis?

A. Acute pancreatitis
B. Peptic ulcer perforation
C. Mesenteric ischemia
D. Aortic dissection
Correct Answer: C. Mesenteric ischemia

Figure 1: ChatGPT-4 response to the prompt and Sample MCQ generated

Ethical approval

Ethical approval for this study was obtained from the Liaquat National Hospital Ethical Review Board (1034-2024-LNH-ERC). Confidentiality and safe data storage were strictly observed throughout the process.

Statistical Analysis

Data analysis was performed using SPSS version 23. The Shapiro-Wilk test confirmed the data was not normally distributed, prompting the use of the Mann-Whitney U test to compare total scores of ChatGPT-4-generated and human-generated MCQs overall and by subject. Chi-square and Fisher's exact tests were employed to compare individual items across Surgery, Medicine, Ob/Gyn, and Pediatrics.

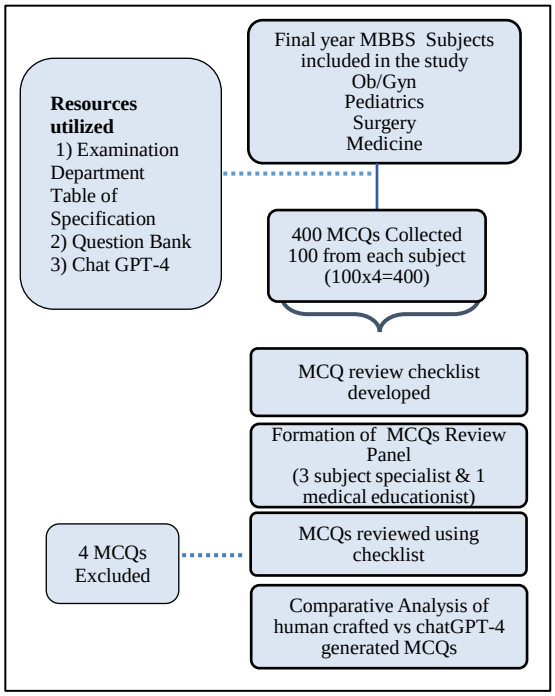

Figure 2: Sequential steps employed in the research process

RESULTS

A total of 400 MCQs were initially selected for analysis. However, four MCQs that did not align with the TOS were excluded leaving 396 MCQs across Surgery, Medicine, Ob/Gyn, and Pediatrics for the final analysis. The distribution of questions is shown in Table 1.

Table 1: Distribution of ChatGPT-4 and human-developed multiple choice questions (MCQs) across different subjects

Subjects	MCQs generated by		Total (n)
	ChatGPT-4 (n)	Human (n)	
Ob/Gyn	48	48	96
Surgery	50	50	100
Medicine	53	47	100
Pediatrics	51	49	100

Ob/Gyn=Obstetrics/ Gynecology

No significant difference was observed in the total scores for the ChatGPT-4-generated and human-

developed MCQs (p=0.12). However, in subject-specific comparisons, a significant difference was found in Ob/Gyn MCQs (p=0.03), while no significant differences were found in other subjects (Table 2). The difference between ChatGPT-4 and human-generate MCQs in the four subjects are shown in tables 3 and 4. For the item 'Does the stem include details necessary for answering the question?' significant differences were observed in the MCQs of Ob/Gyn and Pediatrics (p-values<0.05). Pediatric MCQs also had significant differences for the items 'Is the stem structured as a vignette' and 'Are they similar in length and parallel in structure?' (p-values <0.05). Human-developed MCQs performed better on the item 'Is the item appropriate for the "cover the options rule"?' for Surgery MCQs (p-value=0.04).

Table 2: Comparison of the total scores of item review checklist between using ChatGPT-4 and human-developed multiple choice questions (MCQs)

Subjects	MCQs Generated By			
	ChatGPT-4 Median (IQR)	Human Median (IQR)	U value	p-value
Ob/Gyn	13 (11-15)	14 (13-15)	869.5	0.032*
Surgery	12 (11-13)	12(11-15)	1245	0.972
Medicine	13 (11-15)	12 (10-14)	1089	0.273
Pediatrics	12 (10-14)	14 (11-15)	1005	0.086
Combined score	13 (11-14)	13 (11-15)	17848	0.120

*Mann-Whitney U was applied. *p-value <0.05 is statistically significant. Ob/Gyn=Obstetrics/ Gynecology*

Table 3: Comparison of individual item scores between AI-generated using ChatGPT-4 and human-developed multiple choice questions (MCQs) in Obstetric-Gynecology and Surgery

	Obstetrics and Gynecology				Surgery			
	MCQs by				MCQs by			
	No/ Yes	AI (n=48) n(%)	Human (n=48) n(%)	p-value	No/ Yes	AI (n=48) n(%)	Human (n=48) n(%)	p-value
Is the stem structured as a vignette	No	16 (33)	11 (23)		No	14 (28)	18 (36)	
	Yes	32 (67)	37 (77)	0.364	Yes	36 (72)	32 (64)	0.521
Does the stem include details	No	24 (50)	11 (23)		No	26 (52)	19 (38)	
necessary for answering the question?	Yes	24 (50)	37 (77)	0.010*	Yes	24 (48)	31 (62)	0.228
Is the scenario written in simple and	No	7 (15)	4 (8)		No	1 (2)	3 (6)	
clear language	Yes	41 (85)	44 (92)	0.523[a]	Yes	49 (98)	47 (94)	0.617[a]
Is the lead-in given as a question or	No	3 (6)	1 (2)		No	1 (2)	11 (22)	
completion statement format?	Yes	45 (95)	47 (98)	0.617[a]	Yes	49 (98)	39 (78)	0.004*[a]
Is the lead-in aligned with the	No	11 (23)	7 (15)		No	3 (6)	10 (20)	
scenario and options?	Yes	37 (77)	41 (85)	0.433	Yes	47 (94)	40 (80)	0.071[a]
Are they homogeneous in content and	No	3 (6)	2 (4)		No	7 (14)	13 (26)	
phrasing?	Yes	45 (94)	46 (96)	1[a]	Yes	43 (86)	37 (74)	0.211
Are they similar in length and parallel	No	2 (4)	2 (4)		No	7 (14)	8 (16)	
in structure?	Yes	46 (96)	46 (96)	1[a]	Yes	43 (86)	42 (84)	1
Does each follow the lead-in both	No	0	1 (2)		No	11 (22)	12 (24)	
grammatically and logically	Yes	48 (100)	47 (98)	1[a]	Yes	39 (78)	38 (76)	1
Are they plausible?	No	9 (19)	5 (10)		No	8 (16)	8 (16)	
	Yes	39 (81)	43 (90)	0.386	Yes	42 (84)	42 (84)	1
Is the mentioned key correct?	No	10 (21)	6 (13)		No	23 (46)	16 (32)	
	Yes	38 (79)	42 (88)	0.412	Yes	27 (54)	34 (68)	0.218
Does it test the application of	No	17 (35)	11 (23)		No	11 (22)	19 (38)	
knowledge rather than recall of isolated facts	Yes	31 (65)	37 (77)	0.261	Yes	39 (78)	31 (62)	0.126
Is the item appropriate for the "cover	No	4 (8)	3 (6)		No	28 (56)	17 (34)	
the options rule"?	Yes	44 (92)	45 (94)	1[a]	Yes	22 (44)	33 (66)	0.044*
Is the item positively framed and	No	0	0		No	0	3 (6)	
avoid words like not, except?	Yes	48 (100)	48 (100)	NA	Yes	50 (100)	47 (94)	0.242[a]
Is the item free of words like usually,	No	0	1 (2)		No	0	2 (4)	
always, never, rarely, all of the above, none of the above?	Yes	48 (100)	47 (98)	1[a]	Yes	50 (100)	48 (96)	0.495[a]
Is the item free of grammatical errors	No	0	0		No	1 (2)	6 (12)	
	Yes	48 (100)	48 (100)	NA	Yes	49 (98)	44 (88)	0.112[a]

*Chi-square test was applied. [a]Fisher's Exact test applied where cell count is less than 5. *p-value <0.05 is statistically significant. AI=artificial intelligence, NA=not applicable*

Table 4: Comparison of individual item scores between AI-generated using ChatGPT-4 and human-developed multiple choice questions (MCQs) in Medicine and Pediatrics

		Medicine MCQs by				Pediatrics MCQs by		
	No/ Yes	AI (n=48) n(%)	Human (n=48) n(%)	p-value	No/ Yes	AI (n=48) n(%)	Human (n=48) n(%)	p-value
Is the stem structured as a vignette	No	5 (9)	10 (21)		No	18 (35)	8 (16)	
	Yes	48 (91)	37 (79)	0.159	Yes	33 (65)	41 (84)	0.04*
Does the stem include details necessary for answering the question?	No	23 (43)	11 (23)		No	29 (57)	14 (29)	
	Yes	30 (57)	36 (77)	0.056	Yes	22 (43)	35 (71)	0.005*
Is the scenario written in simple and clear language	No	2 (4)	14 (30)		No	9 (18)	5 (10)	
	Yes	51 (96)	33 (70)	0.001 *[a]	Yes	42 (82)	44 (90)	0.390
Is the lead-in given as a question or completion statement format?	No	2 (4)	6 (13)		No	2 (4)	2 (4)	
	Yes	51 (96)	41 (87)	0.143[a]	Yes	49 (96)	47 (96)	1 [a]
Is the lead-in aligned with the scenario and options?	No	11 (21)	9 (19)		No	5 (10)	3 (6)	
	Yes	42 (79)	38 (81)	1	Yes	46 (90)	46 (94)	0.715 [a]
Are they homogeneous in content and phrasing?	No	14 (26)	11 (23)		No	10 (20)	8 (16)	
	Yes	39 (74)	36 (77)	0.819	Yes	41 (80)	41 (84)	0.796
Are they similar in length and parallel in structure?	No	6 (11)	7 (15)		No	0	6 (12)	
	Yes	47 (89)	40 (85)	0.767	Yes	51 (100)	43 (88)	0.012* [a]
Does each follow the lead-in both grammatically and logically	No	6 (11)	2 (4)		No	3 (6)	1 (2)	
	Yes	47 (89)	45 (96)	0.276 [a]	Yes	48 (94)	48 (98)	0.618 [a]
Are they plausible?	No	16 (30)	14 (30)		No	11 (22)	8 (16)	
	Yes	37 (70)	33 (70)	1	Yes	40 (78)	41 (84)	0.613
Is the mentioned key correct?	No	9 (17)	11 (23)		No	18 (35)	11 (22)	
	Yes	44 (83)	36 (77)	0.461	Yes	33 (67)	38 (78)	0.189
Does it test the application of knowledge rather than recall of isolated facts	No	14 (26)	17 (36)		No	28 (55)	22 (45)	
	Yes	39 (74)	30 (64)	0.387	Yes	23 (45)	27 (55)	0.424
Is the item appropriate for the "cover the options rule"?	No	23 (43)	25 (53)		No	21 (41)	11 (22)	
	Yes	30 (57)	22 (47)	0.423	Yes	30 (59)	38 (78)	0.055
Is the item positively framed and avoid words like not, except?	No	0	1 (2)		No	0	0	
	Yes	53 (100)	46 (98)	0.470 [a]	Yes	51 (100)	49 (100)	NA
Is the item free of words like usually, always, never, rarely, all of the above, none of the above?	No	1 (2)	0		No	0	0	
	Yes	52 (98)	47 (100)	1 [a]	Yes	51 (100)	49 (100)	NA
Is the item free of grammatical errors	No	0	7 (15)		No	0	0	
	Yes	53 (100)	40 (85)	0.004* [a]	Yes	51 (100)	49 (100)	NA

*Chi-square test was applied. [a]Fisher's Exact test applied where cell count is less than 5. *p-value <0.05 is statistically significant. AI=artificial intelligence, NA=not applicable*

DISCUSSION

This study compared the quality of MCQs generated by ChatGPT-4 and those developed by humans. The human-developed MCQs performed significantly better in Ob/Gyn and provided better stem details across Ob/Gyn and Pediatrics. The findings from this study highlight the strengths and limitations of both ChatGPT-4 and human-generated MCQs in medical education, particularly in the context of final-year MBBS written MCQs examination. Our study demonstrates mixed results as regards to the ability of AI tools, such as ChatGPT-4, to generate MCQs that are comparable to human-developed questions.

Our findings revealed no significant difference in total scores between ChatGPT-4 and human-generated MCQs across three out of four specialties (Surgery, Medicine, and Pediatrics), implying that the ChatGPT-4-generated questions are more broadly comparable to those created by human experts in terms of overall item quality. A study by Ahmed et al. also concluded that ChatGPT can create good-quality examination questions if criteria is defined and instructions given to ChatGPT are clear.[9] This could be attributed to the use of a rigorous strategy of prompting and re-prompting to develop the AI-generated MCQs in our methodology.[10] In a study by Rezigalla, when the prompt was simple, the questions were direct and none of the stems had scenarios or vignettes.[11]

In the Ob/Gyn subject, human-generated MCQs had a higher total score on the review checklist than ChatGPT-4-generated MCQs with a p-value of 0.03.

This highlights the importance of domain expertise in humans, particularly in terms of their experience and deep cognitive skills for developing high-quality MCQs.[12]

In Ob/Gyn and Pediatrics MCQs, ChatGPT-4 could not meet the quality requirement of "Does the stem include details necessary for answering the question?" These results suggest that ChatGPT-4, while proficient in generating technically correct content, often lacks the depth required to formulate questions. This aligns with prior research highlighting AI's limitations in generating content that requires an in-depth understanding. A study on generating MCQs of medical physiology, all the AIs models used were unable to generate a considerable number of MCQs that assessed reasoning ability.[13]

Human-generated MCQs were better in the score for ''Is the item appropriate for the ''cover the options rule''?' on our checklist in the subject of surgery. This rule means that stem and lead-in contain sufficient information for content experts to answer questions without relying on the options. Human experts, with their contextual understanding, are better at crafting stems that meet this requirement, which aligns with previous research emphasizing the importance of well-structured stems for assessing higher-order cognitive skills.[14] In contrast, AI, while proficient in generating structurally correct questions, often lacks the depth needed to ensure well-developed stems, which is consistent with studies showing AI's limitations in creating complex, context-driven questions.[15] A study by Cheung et al. found significant differences between ChatGPT-4 and human-developed MCQs in the relevance domain and highlighted ChatGPT's limitation in generating relevant clinical scenarios.[8]

The performance of ChatGPT-4-generated MCQs was better than human-developed questions in the items: "Are the options similar in length and parallel in structure?", "Is the lead-in given as a question or completion statement format?" and "Is the scenario written in simple and clear language" reflects the potential strengths of ChatGPT-4 in specific aspects of question design that require mechanical consistency rather than detailed content understanding. The potential of AI to handle such aspects of MCQ generation with greater consistency has been supported by studies showing that AI systems can reduce human error in repetitive tasks.[16,17] In a similar study on generating programming education MCQs, authors concluded that the AI-generated MCQs provided sufficient information in clear language.[18]

The strength of this study is its sample size across four specialties being assessed in final year MBBS examination.

Limitations of the study

The study has several limitations that must be considered. It focused on final-year MCQs only, limiting the generalizability of the findings to other phases of medical schools. Moreover, although our checklist was based on NBME item writing flaws; while analyzing, we felt that it should also have more items relevant to the overall quality of MCQs in terms of the complexity of the scenario, alignment with the objectives, and difficulty level targeting specific learners. Lastly, our focus was mainly on assessing quality of the MCQs with no evidence from item responses on the actual test. Psychometric analysis of test items will further enhance the evidence of the utility of ChatGPT-4 in MCQ development.

CONCLUSION

ChatGPT-4 has the potential to assist medical examination multiple choice questions development to ensure high quality assessments, when carefully used. ChatGPT-4 can generate large volumes of questions quickly, using well-structured and specific prompting for effective item generation. Its limitations in generating content that requires an in-depth understanding is further highlighted by this study. Additional researches are required to explore additional applications and other limitations of the flourishing artificial intelligence platforms.

REFERENCES

1. Tangianu F, Mazzone A, Berti F, Pinna G, Bortolotti I, Colombo F, et. al. Are multiple-choice questions a good tool for the assessment of clinical competence in Internal Medicine? Ital J Med. 2018; 12(2): 88-96. doi: 10.4081/itjm.2018.980
2. Towns MH. Guide to developing high-quality, reliable, and valid multiple-choice assessments. J Chem Educ. 2014; 91(9) : 1426-1431. doi: 10.1021/ed500076x
3. Diwan C, Srinivasa S, Suri G, Agarwal S, Ram P. AI-based learning content generation and learning pathway augmentation to increase learner engagement. Comput Educ: Artif Intell. 2023; 4: 100110. doi: 10.1016/j.caeai.2022.100 110
4. Owan VJ, Abang KB, Idika DO, Etta EO, Bassey BA. Exploring the potential of artificial intelligence tools in educational measurement and assessment. EURASIA J Math Sci Tech Ed. 2023; 19(8): em2307. doi: 10.29333/ejmste/13428

5. Mihalache A, Huang RS, Popovic MM, Muni RH. ChatGPT-4: an assessment of an upgraded artificial intelligence chatbot in the United States Medical Licensing Examination. Med Teach. 2024; 46(3): 366-372. doi: 10.1080/0142159X. 2023.2249588
6. Ali FA, Sharif S, Ata M, Patel N, Muhammad Rafay M, Syed HR, et. al. The Chat GPT develops multiple choice questions (MCQs) for postgraduate specialty assessment–A reality or a myth? Pak J Neruol Surg. 2024; 28(1): 142-149. doi: 10.36552/pjns.v28i1.963
7. Giray L. Prompt engineering with ChatGPT: A guide for academic writers. Ann Biomed Eng. 2023; 51(12): 2629–2633. doi: 10.1007/s10439-023-03272-4
8. Cheung BHH, Lau GKK, Wong GTC, Lee EYP, Kulkarni D, Seow CS, et.al. ChatGPT versus human in generating medical graduate examination multiple choice questions-A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). PLoS One. 2023; 18(8): e0290691. doi: 10.1371/journal.pone.0290691
9. Ahmed A, Jamil E, Abubakar M, Batool A, Masoom Akhtar M, Iqbal Nasiri M, et al. Harnessing the power of ChatGPT to develop effective MCQ-based clinical pharmacy examinations. J Res Technol Edu. 2024; 2: 1-1. doi: 10.1080/15391523.2024.2425435
10. Laverghetta AJ, Licato J. Generating better items for cognitive assessments using large language models BEA. Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications. 2023; 414-428. doi: 10.18653/v1/2023.bea-1.34
11. Rezigalla AA. AI in medical education: uses of AI in construction type A MCQs. BMC Med Educ. 2024; 24(1): 247. doi: 10.1186/s12909-024-05250-3
12. Haataja ES, Tolvanen A, Vilppu H, Kallio M, Peltonen J, Metsäpelto RL. Measuring higher-order cognitive skills with multiple choice questions–potentials and pitfalls of Finnish teacher education entrance. Teach Educ. 2023; 122: 103943. doi: 10.1016/j.tate.2022.103943
13. Agarwal M, Sharma P, Goswami A. Analysing the applicability of ChatGPT, Bard, and Bing to generate reasoning-based multiple-choice questions in medical physiology. Cureus. 2023; 15(6): e40977. doi: 10.7759/cureus.40977
14. Morrison S, Free KW. Writing multiple-choice test items that promote and measure critical thinking. J Nurs Educ. 2001; 40(1): 17-24. doi: 10.3928/0148-4834-20010101-06
15. Liu J, Zheng J, Cai X, Wu D, Yin C. A descriptive study based on the comparison of ChatGPT and evidence-based neurosurgeons. iScience. 2023; 26(9). doi: 10.1016/j.isci.2023.107590
16. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ. 2019; 5(2): e16048. doi: 10.2196/16048
17. Adiguzel T, Kaya MH & Cansu FK. Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. Contemp Educ Technol. 2023; 15(3): ep429. doi: 10.30935/cedtech/13152
18. Doughty J, Wan Z, Bompelli A, Qayum J, Wang T, Zhang J, et.al. A comparative study of AI-generated (GPT-4) and human-crafted MCQs in programming education. In Proceedings of the 26th ACEC. 2024; 114-123. doi: 10.1145/3636243.363625

AUTHORS' CONTRIBUTION:

MAN: Conception of the study, design of work, data acquisition, analysis, manuscript drafting, approval of final version to be published

YN: Data collection, analysis & interpretation, manuscript drafting, approval of final version to be published

AT: Data collection, data analysis, manuscript drafting, approval of final version to be published

SA: Data collection & interpretation, critical review, approval of final version to be published

All Authors agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved

CONFLICT OF INTEREST:

All authors declared no conflict of interest.

GRANT SUPPORT AND FINANCIAL DISCLOSURE:

No specific grant was taken for this research from any funding agency in the public, commercial or not-for-profit sectors.

DATA SHARING STATEMENT:

The data are available from the corresponding author upon reasonable request.

……