REVIEW ARTICLE

AI Integration in MCQ Development: Assessing Quality in Medical Education: A Systematic Review

Fizzah Ali[1*], Hajra Talat[2]

SUMMARY

This systematic review focuses on examining how artificial intelligence is included in multiple-choice questions and how this affects the efficacy and quality of assessments used in education. Several papers investigating the application of artificial intelligence in multiple-choice question creation have been found through a thorough literature analysis. The present study employed a systematic literature review to comprehensively analyze the existing literature and underscore the effects of incorporating artificial intelligence into creating multiple-choice questions on the standard and efficacy of assessments used in education. Between January 2019 and January 2024, we examined papers from credible publications, concentrating on sixteen chosen articles for in-depth examination. The results show how artificial intelligence can revolutionize traditional evaluation methods in education by improving the accuracy, efficiency, and diversity of multiple-choice questions. While artificial intelligence models like ChatGPT, Bard, and Bing have shown encouraging results in creating multiple-choice questions, issues with validity, complexity, and reasoning ability still need to be addressed. Notwithstanding its drawbacks, artificial intelligence-driven multiple-choice question holds great potential for enhancing evaluation processes and enhancing educational opportunities in a variety of subject areas. This Systematic review highlights the necessity of further research and advancement to fully utilize artificial intelligence in creating multiple-choice questions and its incorporation into frameworks for educational assessments.

Keywords: *Artificial Intelligence, Educational Measurement, Education, Medical Education, Questonnaires.*

How to cite this: *Ali F, Talat H. AI Integration in MCQ Development: Assessing Quality in Medical Education: A Systematic Review. Life and Science. 2024; 5(3): 413-426. doi: http://doi.org/10.37185/LnS.1.1.643*

Introduction

Education has always been a dynamic field that constantly changes to suit students' requirements. This trend has changed significantly in the twenty-first century with the introduction of artificial intelligence (AI) into educational evaluations. This revolutionary change is especially noticeable in the creation and caliber of multiple-choice questions (MCQs), which have long been a mainstay of evaluations used in education.[1]

MCQs have always been preferred due to their uniformity, impartiality, and effective scoring. They have, meanwhile, also come under fire for encouraging memorization and neglecting to gauge deeper comprehension. Because of this, there is now a need to move away from rote memorization and towards abilities like critical thinking, problem-solving, and teamwork. Richer, more interactive, and more adaptive question formats are now possible because of the growth of personalized learning and technology-enhanced assessments (TEAs), which are used more genuinely and thoroughly.[2]

Artificial intelligence is here to revolutionize the game in educational assessment. AI has special powers that have the potential to transform MCQ

[1]*Department of Medical Education*
University College of Medicine & Dentistry
University of Lahore, Pakistan
[2]*Department of Medical Education*
Fatima Memorial College of Medicine & Dentistry Lahore Pakistan

Correspondence:
Dr. Fizzah Ali
Assistant Professor, Medical Education
University College of Medicine & Dentistry
University of Lahore, Lahore, Pakistan
E-mail: fizzah.ali16@gmail.com

Received: Feb 24, 2024; Revised: May 15, 2024
Accepted: Jun 25, 2024

development and quality. In addition to saving teachers time and money, it may automate the creation of smarter multiple-choice questions (MCQs) based on learning objectives and domain expertise. This guarantees a more effective and dependable assessment procedure.[3] It may customize the question type and difficulty level to meet the needs of each learner, resulting in a more effective and interesting learning environment. It can give students thorough feedback on their answers, pointing out areas of knowledge that need improvement and encouraging introspection. It can even evaluate multiple-choice questions (MCQs) for possible language, substance, or difficulty level biases, facilitating more equitable evaluations.[4]

Many factors need to be considered when incorporating AI into the question development process. These include the general improvement of assessment quality, assessment fairness, and the matching of generated questions with learning objectives. To fully realize the benefits of this technological integration, it is essential to comprehend the complex interactions that exist between artificial intelligence and the qualitative components of evaluations. Comprehending these intricacies is essential to fully utilize artificial intelligence (AI) to improve assessment procedures and guarantee conformity with academic goals.[5]

This review conducts a thorough investigation to uncover the complex dynamics surrounding how the use of AI influences the creation and calibration of multiple-choice questions (MCQs) in modern medical educational assessments, as well as the potential of AI-based feedback in language learning, with a focus on student motivation and introspection.[6]

Further affecting this investigation are the changing responsibilities that teachers and students play in the AI-driven educational environment. Teachers now must navigate an environment where working with AI systems is feasible, which may change their responsibilities in question design. Students then encounter exams that are impacted by artificial intelligence algorithms, raising concerns about how this would affect their educational experiences.[7]

This revolutionary change began with the shortcomings of traditional question development techniques, which were frequently laboring and time-consuming. With its advanced algorithms and machine learning powers, artificial intelligence (AI) holds the possibility of both automating and enhancing this process. Large-scale datasets can be analyzed by AI systems, which can also identify patterns and produce questions that are relevant to the given context and learning goals. This offers the promise of efficiency as well as opportunities to customize examinations to meet the various needs of students.[8] The convergence of artificial intelligence and educational assessments is fundamentally altering established models for creating and assessing questions. The application of AI technology in the development of MCQs stands out as a revolutionary force as educational institutions around the world struggle with the demand for improved efficiency and personalized learning experiences. As the educational landscape changes constantly, this evolution emphasizes how important it is for educators and stakeholders to adjust to new technological developments to keep assessment practices current and useful.[9]

Methods

Studies that examine the impact of AI integration on the creation and evaluation of multiple-choice questions (MCQs) in educational assessments are methodically included in this review. We evaluated the reliability of the papers we selected using a quality evaluation method. For our meta-a nalyses and systematic reviews, we used the PRISMA checklist. Studies using duplicates, editorials, conference reports, letters to editors, book chapters, conceptual papers, studies older than 5 years, studies other than in the English language, studies not relevant after reading abstracts, unrelated outcomes, and publication types such as abstracts and letters that only discuss AI methodology without providing pertinent instruction were eliminated. Complying with the 2020 Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) standards, as seen in Figure. 1, we conduct our analysis solely using data that has been taken from published studies; ethical approval is thus not required.[10]

Systematic Literature Search and Selected Studies:

We conducted a thorough search for relevant

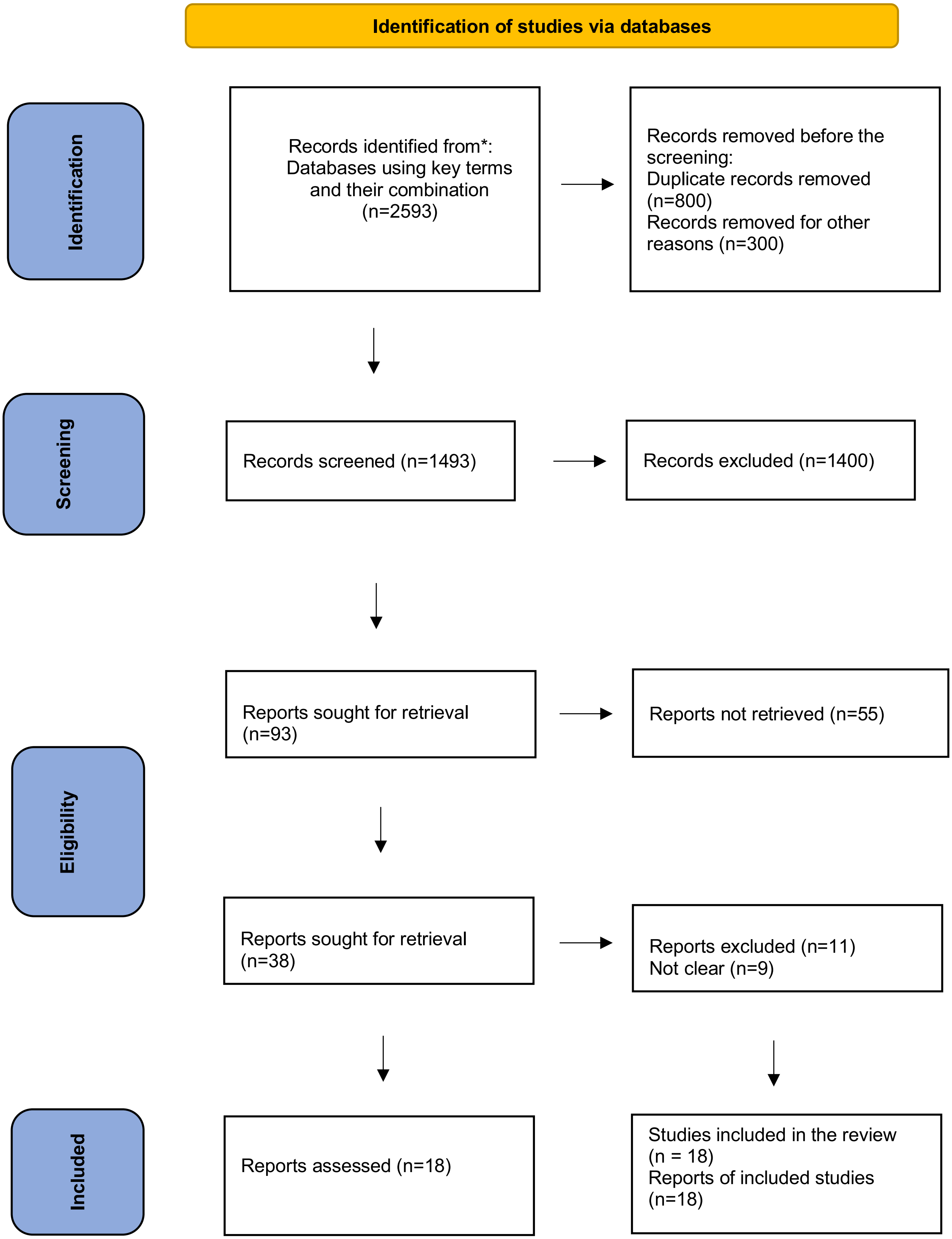
Identification of studies via databases
Identification
Records identified from*: Databases using key terms and their combination (n=2593)
Records removed before the screening: Duplicate records removed (n=800) Records removed for other reasons (n=300)
Screening
Records screened (n=1493)
Records excluded (n=1400)
Reports sought for retrieval (n=93)
Reports not retrieved (n=55)
Eligibility
Reports sought for retrieval (n=38)
Reports excluded (n=11) Not clear (n=9)
Included
Reports assessed (n=18)
Studies included in the review (n = 18) Reports of included studies (n=18)

Fig.1: PRISMA flow diagram illustrating the search strategy and study selection process for the systematic review PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

literature using PubMed, Google Scholar, Connected Paper, Semantic Scholar, and Research Rabbit. We searched for studies including research articles, systematic reviews, and meta analyses according to our inclusion criteria. We took studies from 2019-2024. (Table-1). We have set an inclusion/exclusion criterion that focuses on the use and impact of Artificial intelligence in developing MCQs for assessment.[11]

The search was conducted in different databases such as PubMed, Google Scholar, Connected Paper, and Semantic Scholar. The keywords used were artificial intelligence, multiple choice questions MCQ development, assessment, educational assessment, and medical education. The detailed search design is given in Table-2.

Table-1: The criteria followed during the literature search

S.No	Inclusion Criteria	Exclusion Criteria
1.	Publication date (2019-2024)	Publication Type
2.	Study design (full-text research articles, systematic review, meta-analysis)	Language
3.	Interventions	Irrelevant outcomes
4.	Outcomes related to MCQ development by using AI	Outdated technology

Table-2: Databases used and results

S.No	Database	Search	Results
1.	PubMed	Artificial intelligence and multiple-choice questions and assessment	36
	PubMed	Artificial intelligence and multiple-choice questions	64
	PubMed	Artificial intelligence and MCQs	11
2.	Google Scholar	Artificial intelligence, MCQs, assessment	2470
3.	Connected papers	Artificial intelligence, MCQ development	08
4.	Semantic Scholar	Artificial intelligence, MCQ development, assessment	01
	Semantic Scholar	Artificial intelligence, MCQ development	03

Results

After doing a thorough search across four carefully chosen databases: PubMed, Google Scholar, Connected Papers, and Semantic Scholar, 2593 articles were extracted. After giving each paper, a thorough evaluation and applying predetermined standards, we were able to exclude 1493 publications.[11] Of the articles left, we decided not to use them because of abstracts, duplicate papers, editorials, conference reports, letters to the editor, book chapters, conceptual papers, studies published more than five years ago, and studies not written in English. We carefully reviewed the remaining 93 publications, and 20 were eliminated because their material did not fit our inclusion criteria. Ultimately, we thoroughly examined the 18 papers that remained, all of which satisfied our requirements. Our final systematic review contains these sixteen articles. Table-3 includes a detailed representation of each.

Table-3: Summary of the results of the selected papers.

Author/Year	Country	Study Design	Database	Conclusion
Falcao et al./2020[12]	Portugal	Literature Review	PubMed	The study concludes that AIG is a promising solution to the increasing demand for new items in medical assessment. It offers efficiency, scalability, and diverse item formats through the application of artificial intelligence technologies and has the potential to revolutionize the item development process in medical education. AIG can improve productivity, and quality of teaching, and enhance

				assessment experiences for both students and educators.
Chahna Gosalves/2023[13]	London, UK	Qualitative analysis	Google Scholar	The study concludes that while ChatGPT has had an impact on summative MCQ assessments in unmonitored settings, MCQs are still useful for formative assessment, particularly when assessing higher-order cognitive skills that pose a challenge to AI models such as ChatGPT.
Sallam et al./2024[14]	Jordan	METRICS Checklist	Google Scholar	According to the study's findings, ChatGPT-4 significantly outperformed other AI models in managing challenging multiple-choice clinical chemistry problems. Some AI models were assessed as "Above average," however ChatGPT-4 excelled in completeness, accuracy/evidence, and appropriateness/relevance. It stated that existing methods of assessment in higher education should be reevaluated and support the incorporation of AI while maintaining the ability to think critically. The study suggested a well-rounded strategy that boosts pupils' intellectual growth and raises the bar for higher cognitive talent evaluation by fusing AI capabilities with human intelligence.
Marndi et al./2023[15]	West Bengal	Commands used for utilizing AI in the educational setting	Google Scholar	The study concludes that ChatGPT has the potential to be an effective teaching tool in a variety of subject areas, assisting both educators and learners in the process of teaching and learning. It can help with things like making test questions, preparing presentation slides, grading student responses,

				supporting lesson preparation, and producing information that is specifically tailored to each student's query.
Gardner et al./2020[16]	UK, Ireland, China	Literature Review	Google Scholar	The technological developments and the potential advantages of AI in improving assessment validity and efficacy—particularly in computerized adaptive exams and automated essay scoring systems—are emphasized in the conclusion. Machine learning and big data analysis play a major role in these developments.
Chaturvedi et al./2023[17]	Australia	Comparative Analysis	Google Scholar	The research found that frequent utilization of MCQs can result in the memorization of concepts. The study investigated whether creating MCQs encourages student engagement and learning, as evidenced by the number of likes received for each question from peers. Furthermore, the study explored various heuristics that could be transformed into MCQs to assist students in comprehending specific subjects.
Ali et al./2023[18]	Qatar	An exploratory approach to investigate the accuracy of ChatGPT	Google Scholar	The study's finding on ChatGPT's implications for healthcare education—specifically, about how dental students are assessed—highlights how revolutionary generative AI systems like ChatGPT have the potential to serve as virtual learning. To fully utilize the potential of AI-based technology, the study emphasizes the necessity for healthcare educators to modify their approaches to teaching and evaluation in the fields of medicine and dentistry. It implies

				that teachers ought to see these technology developments as chances to improve students' educational experiences.
Al Shuriaqi et al. 2023[19]	Oman	Narrative Review	Google Scholar	The vital role that case-based multiple-choice questions (MCQs) play in assessing clinical reasoning and decision-making abilities in medical education is highlighted by this study. It highlights how crucial it is to maintain the standards for reliability, content validity, and incorporating clinical reasoning into the creation of multiple-choice questions. With improvements in assessment quality brought about by technological innovations like virtual reality and artificial intelligence, the future of case-based multiple-choice questions seems bright.
Cheung et al. 2023[20]	(Multi prospective study) Hong Kong, UK, Ireland, Singapore	The study compared the quality of 50 multiple-choice questions (MCQs) generated by ChatGPT with 50 MCQs created by university professoriate staff, all based on standard medical textbooks. These MCQs were evaluated by five independent international assessors across five domains: appropriateness of the question, clarity and specificity, relevance,	Google Scholar	The conclusion of the study highlighted that ChatGPT, a large language model AI, demonstrated the ability to serve as an examiner for graduate medical exams with comparable performance to experienced human examiners. The research supports further exploration of how AI models can enhance efficiency in academia while maintaining high standards.

		discriminative power of alternatives, and suitability for medical graduate examination.		
Klang E et al./2023[21]	Israel	The study design involved utilizing GPT-4, an OpenAI application, to write a 210 multiple-choice questions (MCQs) examination based on an existing exam template. The output generated by GPT-4 was then thoroughly investigated by specialist physicians who were blinded to the source of the questions.	Google Scholar	According to the study's findings, it is possible to create medical multiple-choice questions (MCQs) using GPT-4 but doing so needs careful review by specialists in medicine. To guarantee the caliber and precision of tests, educators need to be aware of the shortcomings in GPT-4 and validate its results. Although GPT-4 can help students prepare for exams, teachers should be aware of its limitations and the requirement for expert validation to preserve exam quality.
Yunjiu et al./2022[22]	Bangladesh, South Korea, Pakistan, Russia, Thailand	Comparative Study	Google Scholar	The study concluded that vocabulary test items created by human experts and those generated by AI employ different processing methods. These variations are attributed to components like the difficulty of the item and the features of the item response format. According to the results, learners are more likely to employ lexical information techniques in more challenging items while context-based strategies are more commonly used in easy items.
Kiyak et al./2023[23]	Turkey	Methodological study	Connected papers	The study showcased the successful application of Automatic Item Generation (AIG) in Turkish to create case-based MCQs for evaluating clinical reasoning skills in medical education. Using a Python-based code, 1600 MCQs on

				hypertension were swiftly generated in 1.73 seconds. Evaluation by cardiologists confirmed the questions' suitability in assessing clinical reasoning skills. These results emphasize the potential of AI in producing MCQs that assess higher-order cognitive abilities in medical education, offering advantages such as cost efficiency and the development of a diverse question bank.
Roos et al./2023[24]	Germany	Quantitative research design	Semantic Scholar	The study revealed that large language models (LLMs) such as GPT-4 and Bing excelled in answering medical knowledge-based multiple-choice questions (MCQs). GPT-4 achieved the highest overall performance at 88.1%, closely followed by Bing at 86.0% and GPT-3.5-Turbo at 65.7%. When media-related questions were excluded, Bing performed the best at 90.7%, followed closely by GPT-4 at 90.4%. These findings suggest that LLMs, particularly GPT-4 and Bing, show great potential as effective tools in medical education and for testing examination questions. The superior performance of these AI models, even outperforming medical students, indicates promising opportunities for further advancement and integration of AI technology in both educational and clinical settings.
Huang et al./2023[25]	Canada	The study design included both quantitative and qualitative components to	Semantic Scholar	The study aimed to assess the utility of AI chatbots in medical education by comparing their performance with residents on a knowledge test, highlighting

		compare the performance of GPT-3.5 and GPT-4 AI chatbots with Family Medicine residents on a multiple-choice medical knowledge test.		their potential in medical education and MCQ assessments.
Kiyak/2023[26]	Turkey	The study design is quantitative, focusing on the use of ChatGPT for generating MCQs in medical education.	Semantic Scholar	The study investigates the use of ChatGPT, an AI model, for generating case-based MCQs in medical education. ChatGPT provides a fast and effective method for creating MCQs, generating interest in using AI for educational assessment. While template-based AIG has been successful, advancements in AI models like GPT-3.5 have enhanced question quality. The introduction of ChatGPT signifies a significant development in health professions education, leading researchers to differentiate between pre- and post-ChatGPT periods. It is essential to develop high-quality prompts to ensure AI-generated MCQs meet educational standards, underscoring the importance of human guidance in AI-generated content.
Meo et al./2023[27]	Saudi Arabia	Quantitative study design	Semantic Scholar	ChatGPT has shown proficiency in medical science exams, excelling in MCQs, and providing reasoning explanations, indicating its potential for medical education. ChatGPT offers benefits like assisting in drafting and generating assignments. MCQs are favored by students for assessing higher-order cognitive skills and are vital in medical education.

Discussion

The expansion of online educational platforms and evaluation mechanisms has markedly increased the necessity for premium multiple-choice questions. To fulfill this requirement, both researchers and educators are leveraging artificial intelligence to automate the creation of MCQs. These systems, grounded in AI, have demonstrated considerable potential in elevating both the precision and efficiency associated with the production of MCQs.[28]

According to Agarwal et al., AI will transform medical education by producing MCQs for lectures that are based on reasoning and will increase teacher-student engagement. For AI-generated instructional content to overcome issues with validity, difficulty, and reasoning capacity, more study and development are required.[29] Studies on whether AI models (ChatGPT, Bard, and Bing) are suitable for creating reasoning based multiple choice questions (MCQs) in the field of medical physiology have shown that these models require improvement to generate reasoning-based MCQs. Among the AI models, Bing offered the fewest valid multiple choice questions (MCQs), whereas ChatGPT yielded the most. In comparison to Bard and Bing, ChatGPT took longer to create MCQs. Bard employed certain inquiry formats, whereas Bing frequently utilized negative verb forms in his queries.[30]

The machine learning and semantic techniques presented by Kumar et al. in their AI framework enable the creation of multiple-choice question stems that are automatically constructed by cognitive standards and learning goals. It also generates a variety of multiple-choice questions (MCQs) suited to the cognitive levels of Bloom's taxonomy. This discovery, which makes real-time automated multiple-choice question generation possible, has the potential to improve assessment processes in education, especially in technical subjects.[31]

Seetharaman et al. emphasized that ChatGPT is also used to enhance MCQ creation and evaluation procedures by giving feedback on students' answers and acting out patient interactions. It improves student's comprehension of medical concepts and their capacity for knowledge expression, which has an indirect impact on the caliber and effectiveness of MCQ-based assessments.[32]

Microsoft Bing and GPT-4 also outscored other bots and students in a different study assessing the effectiveness of AI chatbots in a multiple-choice medical licensing exam at the University of Antwerp. Even though the bots had to answer challenging questions, they performed better than people. Microsoft Bing demonstrated potential in identifying poor questions, indicating that artificial intelligence bots may be able to improve test quality. AI bots must always develop their algorithms to continue being useful for medical assessment and education.[33]

In another study, Hoch et al. assessed ChatGPT's accuracy in several subspecialties when responding to practice questions for the otolaryngology board certification. Its accuracy varied depending on the category; it was more accurate in allergology and less accurate in legal matters. These results highlight the necessity of continuous improvement and verification of AI models such as ChatGPT, especially when it comes to appropriately responding to multiple-choice questions in specialized medical fields.[34]

In another study, ontology-based methods are also used to create excellent multiple-choice questions (MCQs) in the field of medicine. It discusses the drawbacks of conventional MCQ creation procedures and emphasizes how effective automatic question generating (AQG) systems are at quickly and effectively creating a wide variety of MCQs. Ontologie's hierarchical form guarantees the correctness of the questions and helps to avoid typical item writing errors. All things considered, the combination of AI, AQG methods, and ontologies provides a viable way to raise the caliber and efficacy of multiple-choice questions in educational contexts.[35]

Cheung et al. also evaluated ChatGPT, against human examiners in terms of its capacity to produce exam questions for medical exams. In most evaluation fields, it generated questions with equivalent quality to those prepared by humans, but significantly faster. Some AI-generated questions were better than human generated ones. The potential of AI to boost efficiency in producing high quality multiple-choice questions (MCQs) for medical education by offering proof that it can help prepare exam content to a standard equivalent to that of skilled human examiners.[19]

The development of an algorithm that automatically creates valid and varied gap-fill multiple choice questions (MCQs) for assessment of knowledge in scientific areas by combining ontology-based design, text mining, and natural language processing. The program generated more than sixteen thousand multiple choice questions (MCQs) on software testing issues using 103 internet publications as inputs. The system showed excellent quality in choosing appropriate sentences for question phrases and producing powerful distractors. This highlights the algorithm's potential for automatically generating questions for online learning and knowledge assessment platforms.[36]

AI's capacity to generate a variety of item formats with effectiveness, versatility, and diversity is highlighted as a possible transformative tool for medical evaluation methods. Its usefulness also extends to formative situations, which prompts an evaluation of assessment techniques, and its remarkable ability to handle complex clinical/non-clinical MCQs. The potential of AI to improve evaluation procedures in higher education is highlighted by multiple-choice questions. It also acts as an adaptable teaching tool for a variety of academic subjects, highlighting how AI is revolutionizing healthcare education and how teachers must use it wisely.

Limitations

The integration of artificial intelligence (AI) in developing MCQs in medical education and assessment offers significant potential but also has some limitations. While ChatGPT and other AI systems are promising in producing multiple choice questions (MCQs), there are difficulties in producing questions that demand human-level subject matter knowledge. The necessity for thorough validation and the variation in question formats among AI models present substantial challenges. Further research is required to solve difficulties relating to question complexity, reasoning capacity, and content validity, even though AI can increase productivity and efficiency in generating multiple-choice questions. However, integrating AI offers the potential for transforming medical assessment and education, highlighting the significance of continuous improvement and prudent application in learning environments.

Conclusion

The endpoint emphasizes how AI has the potential to revolutionize several fields, including medical education and the development of multiple-choice questions (MCQs). Although AI approaches like ChatGPT and Bing demonstrate potential for automating multiple-choice question Preparation and improving productivity, however, there are significant constraints related to the validity, complexity, and reasoning ability of the questions. Even with these limitations, integrating AI into medical education has a lot of potential to improve student-teacher engagement and the caliber of multiple-choice questions (MCQs).

Acknowledgment: None

Conflict of Interest: The authors declare no conflict of interest

Grant Support and Financial Disclosure: None

REFERENCES

1. Felix J, Webb L. Use of artificial intelligence in education delivery and assessment. The Parliamentary Office of Science and Technology. 2024. doi: 10.58248/PN712
2. Smith VG, Szymanski A. Critical thinking: More than test scores. International Journal of Educational Leadership Preparation. 2013; 8: 16-25.
3. Owan VJ, Abang KB, Idika DO, Etta EO, Bassey BA. Exploring the potential of artificial intelligence tools in educational measurement and assessment. Eurasia Journal of Mathematics, Science and Technology Education. 2023; 19: em2307. doi: 10.29333/ejmste/13428
4. Andrew C, Raduescu C, Zeivots S, Elaine. Educator and student perspectives. Proceedings of the ACM on Human-Computer Interaction. 2024;8(CSCW1):1-21. doi: 10.1145/3573051.3596191
5. Nasution NEA. Using artificial intelligence to create biology multiple choice questions for higher education. Agricultural and Environmental Education. 2023; 2: em002. doi: 10.29333/agrenvedu/13071
6. De la Vall RR, Araya FG. Exploring the benefits and challenges of AI-language learning tools. International Journal of Social Sciences and Humanities Invention. 2023;

10: 7569-76. doi: 10.18535/ijsshi/v10i01.02.

7. M. Rizvi. Exploring the landscape of artificial intelligence in education: Challenges and opportunities, 2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications. 2023: 1-3. doi: 10.1109/HORA58378.2023.10156773
8. Izo F, Leão J, Pirovani JP, Oliveira E. Automatic Generation of Large-Scale Assessment Questions. InProceedings of the XVIII Brazilian Symposium on Information Systems. 2022: 1-8. doi: 10.1145/3535511.3535518
9. González-Calatayud V, Prendes-Espinosa P, Roig-Vila R. Artificial intelligence for student assessment: A systematic review. Applied sciences. 2021; 11: 5467. doi: 10.3390/app11125467
10. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Journal of Clinical Epidemiology. 2009; 62: e1-34. doi: 10.1016/j.jclinepi.2009.06.006
11. Brennan E. Guides: Systematic Reviews: Inclusion/Exclusion Criteria [Internet]. [cited 2024 Feb 19]. Available from: https://musc.libguides.com/ systematicreviews/eligibilitycriteria.
12. Falcão F, Patrício Costa, Pêgo JM: Feasibility assurance: A review of automatic item generation in medical assessment. Advances in Health Sciences Education. 2022; 27: 405-5. doi: 10.1007/s10459-022-10092-z
13. Gonsalves C. On ChatGPT: what promise remains for multiple choice assessment?. Journal of Learning Development in Higher Education. 2023; 27. doi: 10.47408/jldhe.vi27.1009
14. Sallam M, Al-Salahat K, Eid H, Egger J, Puladi B. Human versus artificial intelligence: ChatGPT-4 outperforming Bing, Bard, ChatGPT-3.5, and humans in clinical chemistry multiple-choice questions. medRxiv. 2024: 2024-01. doi: 10.1101/2024.01.08.24300995
15. Mondal H, Marndi G, Behera JK, Mondal S. ChatGPT for teachers: Practical examples for utilizing artificial intelligence for educational purposes. Indian Journal of Vascular and Endovascular Surgery. 2023; 10: 200-5. doi: 10.4103/ijves.ijves_37_23
16. Gardner J, O'Leary M, Yuan L. Artificial intelligence in educational assessment:' Breakthrough? Or buncombe and ballyhoo?'. Journal of Computer Assisted Learning. 2021; 37: 1207-16. doi: 10.1111/jcal.12577
17. Chaturvedi I, Cambria E, Welsch RE. Teaching simulations supported by artificial intelligence in the real world. Education Sciences. 2023; 13: 187. doi: 10.3390/educsci 13020187
18. Ali K, Barhom N, Tamimi F, Duggal M. ChatGPT—A double-edged sword for healthcare education? Implications for assessments of dental students. European Journal of Dental Education. 2024; 28: 206-11. doi: 10. 1111/eje.12937
19. Al Shuriaqi S, Aal Abdulsalam A, Masters K. Generation of Medical Case-Based Multiple-Choice Questions. International Medical Education. 2023; 3: 12-22. doi: 10. 3390/ime3010002
20. Cheung BH, Lau GK, Wong GT, Lee EY, 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: e0290691. doi: 10.1371/journal.pone.0290691
21. Klang E, Portugez S, Gross R, Brenner A, Gilboa M, Ortal T, et al. Advantages and pitfalls in utilizing artificial intelligence for crafting medical examinations: a medical education pilot study with GPT-4. BMC Medical Education. 2023; 23: 772. doi: 10.1186/s12909-023-04752-w
22. Yunjiu L, Wei W, Zheng Y. Artificial intelligence-generated and human expert-designed vocabulary tests: A comparative study. Sage Open. 2022; 12: 21582440221 082130. doi: 10.1177/215824402210821
23. Kıyak YS, Budakoğlu İİ, Coşkun Ö, Koyun E. The first automatic item generation in Turkish for assessment of clinical reasoning in medical education. Tıp Eğitimi Dünyası. 2023; 22: 72-90. doi: 10.25282/ted.1225814
24. Roos J, Kasapovic A, Jansen T, Kaczmarczyk R. Artificial intelligence in medical education: comparative analysis of ChatGPT, Bing, and medical students in Germany. JMIR Medical Education. 2023; 9: e46482. doi: 10.2196/46482
25. Huang RS, Lu KJ, Meaney C, Kemppainen J, Punnett A, Leung FH. Assessment of resident and AI chatbot performance on the University of Toronto family medicine residency progress test: comparative study. JMIR Medical Education. 2023; 9: e50514. doi: 10.2196/50514
26. Kıyak YS. A ChatGPT prompt for writing case-based multiple-choice questions. Revista Española de Educación Médica. 2023; 3: 98-103. doi: 10.6018/edumed.587451
27. Meo SA, Al-Masri AA, Alotaibi M, Meo MZ, Meo MO. ChatGPT knowledge evaluation in basic and clinical medical sciences: multiple choice question examination-based performance. Healthcare. 2023; 11: 2046. doi: 10.3390/

healthcare11142046

28. Teaching Commons. AI tools in teaching and learning [Internet]. [cited 2024 Jul 24]. Available from: https://teachingcommons.stanford.edu/news/ai-tools-teaching-and-learning.

29. 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: e40977. doi: 10.7759/cureus.40977

30. Kumar AP, Nayak A, K MS, Chaitanya, Ghosh K. A novel framework for the generation of multiple choice question stems using semantic and machine-learning techniques. International Journal of Artificial Intelligence in Education. 2023; 34: 332-375. doi: 10.1007/s40593-023-00333-6

31. Seetharaman R. Revolutionizing medical education: Can ChatGPT boost subjective learning and expression?. Journal of Medical Systems. 2023; 47: 61. doi: 10.1007/s10916-023-01957-w

32. Morreel S, Verhoeven V, Mathysen D. Bing outperforms five other generative artificial intelligence chatbots in the Antwerp University multiple choice medical license exam. medRxiv. 2024; 3; er000349. doi: 10.1371/ journal.pdig.0000349

33. Hoch CC, Wollenberg B, Lüers JC, Knoedler S, Knoedler L, Frank K, et al. ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification preparation questions. European Archives of Oto-Rhino-Laryngology. 2023; 280: 4271-8. doi: 10.1007/s00405-023-08051-4

34. Leo J, Kurdi G, Matentzoglu N, Parsia B, Sattler U, Forge S, et al. Ontology-based generation of medical, multi-term MCQs. International Journal of Artificial Intelligence in Education. 2019; 29: 145-88. doi: 10.1007/s40593-018-00172-w

35. Sirithumgul P, Prasertsilp P, Olfman L. An Algorithm for Generating Gap-Fill Multiple Choice Questions of an Expert System. arXiv preprint arXiv: 2109.11421. 2021. doi: 10.48550/arXiv.2109.11421

36. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Journal of Clinical Epidemiology. 2009; 62: e1-34. doi: 10.1016/j.jclinepi.2009.06.006.

Authors Contribution

FI: Idea conception, study designing, data collection, data analysis, results and interpretation, manuscript writing and proofreading

RQ: Data collection, data analysis, results and interpretation, manuscript writing and proofreading