PalStudent Journal
A Palestinian Scientific Journal for the Youth

ABSTRACTS: VOLUME 8, SPECIAL ISSUE {8th Undergraduate Conference}

CAN MACHINES WRITE YOUR EXAMS? A SYSTEMATIC REVIEW OF GENERATIVE ARTIFICIAL INTELLIGENCE FOR MULTIPLE-CHOICE QUESTION DEVELOPMENT IN MEDICAL EDUCATION

Ahmed M.J. Rabee[1], Abdulrahman Kharsa[1], Ammar Y.M. Saadawy[1], Mohammad Wasim Alnajjar[1], Ahmad M.A. Adi[1], Majid Ali[1]

Supervisor: Dr. Majid Ali[1]

[1] Department of Basic Medical Sciences, College of Medicine, Sulaiman AlRajhi University, AlBukayriyah, Qassim, Saudi Arabia

Abstract:

Background: Multiple-choice questions (MCQs) are essential for assessing medical education, whereas conventional item development is labour-intensive, inconsistent in quality, and often mismatched with evolving curricula. Generative artificial intelligence (AI), particularly large language models such as ChatGPT, has emerged as a promising tool for streamlining MCQ creation. However, evidence on the quality, psychometric properties, and educational effectiveness of AI-generated MCQs remains fragmented. Study Purpose/

Objectives: This review aimed to systematically assess the quality and effectiveness of generative AI in producing MCQs for medical education. Examining psychometric properties, content quality, cognitive level, resource efficiency, and user perceptions compared to human-authored questions.

Methods: This systematic review followed PRISMA 2020 guidelines and was registered on PROSPERO (CRD420251153109). Six databases (PubMed, Scopus, Embase, Web of Science, ERIC, and Europe PMC) were searched from January 2018 to November 2025. Two reviewers independently screened studies, extracted data, and assessed methodological quality using modified MERSQI and the Newcastle-Ottawa Scale.

Results: Of 701 records, 40 studies were included after removing duplicates (n=126) and screening. Most studies were published in 2024–2025 and focused on ChatGPT models (GPT-3.5 through GPT-4o). AI-generated MCQs showed comparable item difficulty to human-authored questions; however, discrimination indices were frequently lower. Items disproportionately assessed lower-order cognitive skills with limited depth in clinical reasoning. Factual inaccuracies were reported in up to 28% of items, with flawed distractors and answer key errors. Later-generation models (GPT-4 and above) consistently outperformed earlier versions. AI significantly reduced item development time, but expert review remained essential.

PalStudent Journal
Correspondence concerning this article should be addressed to the mentioned authors at the mentioned institutes.

E-mail: research@admin.alquds.edu
Palestine, Abu Dis, Al-Quds University

PalStudent Journal
A Palestinian Scientific Journal for the Youth

Conclusion: Generative AI is a promising adjunct to MCQ development in medical education, offering substantial efficiency gains while maintaining comparable item difficulty. Nevertheless, persistent limitations in discrimination, cognitive depth, and factual accuracy necessitate mandatory expert oversight. Current evidence supports AI as an assistive tool within a human-in-the-loop workflow rather than a standalone replacement for faculty-authored assessment items.

Key Words: Generative Artificial Intelligence; Multiple-Choice Questions; Medical Education; Psychometrics; Assessment.

Correspondence concerning this article should be addressed to the mentioned authors at the mentioned institutes.

E-mail: research@admin.alquds.edu
Palestine, Abu Dis, Al-Quds University