The Revival of Essay-Type Questions in Medical Education: Harnessing Artificial Intelligence and Machine Learning

Muhammad Shahid Shamim[1], Syed Jaffar Abbas Zaidi[2] and Abdur Rehman[3]

[1]*Directorate of Graduate Studies, The Aga Khan University, Karachi, Pakistan*
[2]*Department of Oral Biology Digital Learning Centre, Dow University of Health Sciences, Karachi, Pakistan*
[3]*Department of Science of Dental Materials, Hamdard College of Medicine & Dentistry, Hamdard University, Karachi, Pakistan*

ABSTRACT

Objective: To analyse and compare the assessment and grading of human-written and machine-written formative essays.

Study Design: Quasi-experimental, qualitative cross-sectional study.

Place and Duration of the Study: Department of Science of Dental Materials, Hamdard College of Medicine & Dentistry, Hamdard University, Karachi, from February to April 2023.

Methodology: Ten short formative essays of final-year dental students were manually assessed and graded. These essays were then graded using ChatGPT version 3.5. The chatbot responses and prompts were recorded and matched with manually graded essays. Qualitative analysis of the chatbot responses was then performed.

Results: Four different prompts were given to the artificial intelligence (AI) driven platform of ChatGPT to grade the summative essays. These were the chatbot's initial responses without grading, the chatbot's response to grading against criteria, the chatbot's response to criteria-wise grading, and the chatbot's response to questions for the difference in grading. Based on the results, four innovative ways of using AI and machine learning (ML) have been proposed for medical educators: Automated grading, content analysis, plagiarism detection, and formative assessment. ChatGPT provided a comprehensive report with feedback on writing skills, as opposed to manual grading of essays.

Conclusion: The chatbot's responses were fascinating and thought-provoking. AI and ML technologies can potentially supplement human grading in the assessment of essays. Medical educators need to embrace AI and ML technology to enhance the standards and quality of medical education, particularly when assessing long and short essay-type questions. Further empirical research and evaluation are needed to confirm their effectiveness.

Key Words: *Machine learning, Artificial intelligence, Essays, ChatGPT, Formative assessment.*

How to cite this article: Shamim MS, Zaidi SJA, Rehman A. The Revival of Essay-Type Questions in Medical Education: Harnessing Artificial Intelligence and Machine Learning. *J Coll Physicians Surg Pak* 2024; **34(05)**:595-599.

INTRODUCTION

A few decades ago, long or short essay-type questions were a part of traditional medical education to provide comprehensive assessments in medical education. They allowed learners to demonstrate their ability to think critically, analyse information, and communicate their understanding clearly and concisely.[1] They were regularly used to assess a wide range of knowledge and cognitive skills and were easily adapted to different subjects, levels, and learning outcomes. Furthermore, essay-type questions allow for detailed feedback on a learner's performance, highlighting strengths and areas for improvement.[2]

Correspondence to: Dr. Syed Jaffar Abbas Zaidi, Department of Oral Biology Digital Learning Centre, Dow University of Health Sciences, Karachi, Pakistan
E-mail: jaffar.zaidi@duhs.edu.pk

Received: April 22, 2023; Revised: December 07, 2023; Accepted: April 09, 2024
DOI: https://doi.org/10.29271/jcpsp.2024.05.595

However, the rise of multiple-choice questions (MCQs) and other objective formats has overshadowed the importance of essay-type questions in recent years.[3] MCQs were considered less time-consuming, especially for large classes and high-stakes exams. Unlike their predecessors, the MCQs were not beset with subjectivity and assessor bias issues.

Although the reasons are valid and have played a pivotal role in making MCQs the preferred method of assessment, the limitations of MCQs are also undeniable.[4] Unlike essay-type questions, MCQs allow cueing and guessing, limit the students' skills and ability to present ideas and thoughts in writing and are often unsuitable for testing higher-order thinking. MCQs can undoubtedly cover a broad breadth of content in the assessment. However, investigating the depth of knowledge through their use is unlikely.[5]

AI is the simulation of human intelligence processes by machines, such as computers. Machine learning (ML), a branch of artificial intelligence, focuses on the use of data to develop algorithms to solve problems. It allows systems to learn and

improve from experience without being explicitly programmed, gradually improving their accuracy.[6] These systems can enable humans with an unprecedented ability to analyse enormous data sets and discover complex relationships and patterns. Recent developments in AI technology have paved the ground for educators to think about reaping the benefits of AI in various educational endeavours.[7] Within the domain of assessment, AI and ML can provide a wide range of opportunities for educators, only limited by the boundaries of their imagination.

AI and ML techniques have been increasingly utilised in various aspects of medical education, ranging from personalised learning pathways[8] to automated grading.[9] They offer an array of possibilities for automating tasks, analysing vast datasets, and enhancing the learning experience. AI-driven adaptive systems can effectively tailor educational content to individual learning needs in medical settings.[10] AI enhances medical simulations by introducing more realistic, complex case scenarios, and enabling better decision-making skills among medical students.[11] ML models can grade medical essays with accuracy comparable to human graders, potentially alleviating workload for faculty.[12] NLP algorithms can generate meaningful feedback on essay content and structure, though they lack the ability to assess nuanced clinical reasoning.[13] Traditional essay grading is labor-intensive and time-consuming, which is especially problematic in the high-pressure, fast-paced medical educational environment. Even experienced graders can exhibit variance in scoring due to subjective interpretations of essay quality, clinical reasoning, and argumentative strength. This impacts fairness and objectivity in assessments.

By leveraging AI and ML capabilities, medical educators can streamline assessment processes and enhance the learning experience for students.[14] The growing interest in AI and ML presents an opportunity to reinvigorate the use of essay-type questions in medical education. This innovative study evaluated the potential of AI and ML in assessing human written assignments. This study was designed to explore the capabilities and limitations of AI and ML algorithms in the evaluation of human written academic assignments. Specifically, this study sought to determine the accuracy, efficiency, and reliability of AI and ML systems in grading human written work in comparison to traditional human-assessed methods. By focusing on these technological approaches, the study aimed to contribute to the broader conversation about the future of automated assessment tools in educational settings. By training the AI model in grading essays and providing written feedback, this study aimed to enhance the grading quality of summative essays.

METHODOLOGY

This research was approved by the Research and Ethics Committee of the Hamdard University Dental Hospital Ref. No. HCM&D/HUDH/423-23 dated 18-02-2023. This quasi-experimental, qualitative cross-sectional study was conducted at the Hamdard College of Medicine & Dentistry, Hamdard University, Karachi, from February to April 2023. Final-year BDS students of Hamdard College of Medicine & Dentistry, academic year 2022-2023, were included in the study after obtaining informed consent from the students. The authors conducted an informal experiment using an AI chatbot ChatGPT V.3.5 (software that simulates human-like conversations). Ten formative essays pertaining to final-year BDS subjects were selected and checked manually by faculty members of their respective clinical specialities as shown in Table I. These formative short essays are routinely given to students to assess their clinical reasoning skills in the clinical dental sciences. The authors selected essays from high, low, and average achievers to cover the range of performances. The scoring rubric is presented in Table II, with passing score from 60-100 points while a failing score below 60 points. High achievers were categorised as those scoring above 80 points, while low achievers scored below 60 points and average achievers scored between 60 to 80 points. These were then individually submitted to ChatGPT by another co-author. Four different prompts were given to the AI-driven platform of ChatGPT to grade the summative essays. The purpose of the prompts was to train the AI model to match the scores of the manually checked essays as the score points vary due to their subjective nature. The chatbot's responses were recorded. The third co-author compared the manually checked essays with the AI-checked ones and checked for the quality of the written feedback provided.

RESULTS

Four different prompts were provided to the AI-driven platform of ChatGPT to grade the summative essays and to train the AI model so that its scoring criteria matches that of the manually graded essays. These were the chatbot's initial responses without grading, the chatbot's response to grading against criteria, the chatbot's response to criteria-wise grading, and the chatbot's response to questions for the difference in grading. The AI chatbot was asked to assess and grade a student's essay on a scale of 1 to 10. The chatbot first declined to grade the essay. A second student's essay was then given to the chatbot to grade, compared to the first. The chatbot performed the grading as a 6 out of 10 for knowledge, coherence, and writing skills.

The author then asked for separate criteria-wise grading, which the chatbot did, but the grades differed from the previous grading. A detailed reasoning was provided with each rating.

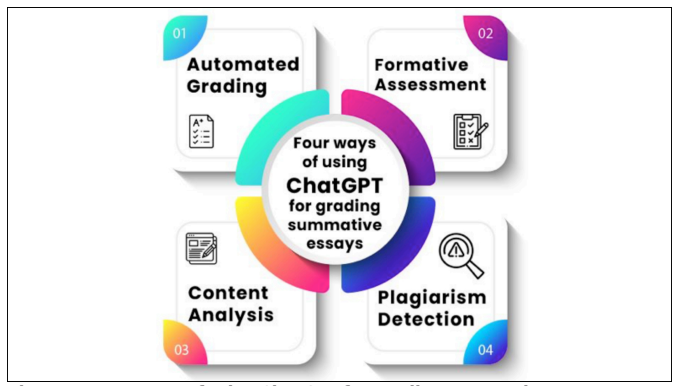

Figure 1: Four ways of using ChatGPT for grading summative essays.

Table I: Ten formative essay questions.

Speciality	Case Scenario
Prosthodontics	A 65-year-old patient presents with multiple missing teeth and complaints of difficulty in chewing and speaking. Describe the process for evaluating this patient for complete or partial dentures. What factors would you consider in deciding the type of prosthesis suitable for this patient?
Orthodontics	A 14-year patient presents with class II malocclusion and deep overbite. What diagnostic tests would you recommend? Discuss the treatment plan including the types of appliances that may be used.
Prosthodontics	You have a patient with a single missing anterior tooth. The patient wishes for a fixed solution. Describe the pros and cons of choosing a dental implant *versus* a fixed dental bridge. What are the considerations for long-term success?
Periodontics	A patient comes in with signs of advanced periodontitis including bleeding gums and mobile teeth. What is your initial diagnosis? What nonsurgical and surgical treatment options would you consider? How would you manage this patient in long-term?
Oral Medicine	A 50-year patient presents with recurring aphthous ulcers and complaints of extreme pain. Outline your differential diagnosis. What tests would you conduct? How would you manage the condition?
Oral Surgery	A 30-year patient presents with an impacted lower third molar causing recurrent pericoronitis. Discuss the surgical considerations and techniques for the removal of impacted molars.
Operative Dentistry	A patient with a history of poor oral hygiene presents with caries on multiple surfaces. Discuss your approach to diagnosis and treatment, including considerations for material selection in restorative procedures.
Pedodontics	A 7-year patient presents with severe early childhood caries. How would you diagnose and manage this case, keeping in mind the child's co-operation level and the potential for behaviour management issues?
Periodontics	A patient with diabetes presents with generalised gingival inflammation and bone loss around several teeth. How does diabetes complicate periodontal treatment and what are the considerations for managing this patient?
Oral Surgery	A patient comes in with a large cystic lesion in the maxillary sinus, which is discovered incidentally during a radiographic examination. Discuss your approach to diagnosis and treatment, considering the possible need for interdisciplinary collaboration.

Table II: Rubric for evaluating formative essay questions with passing score of 60 and above.

Categories	Criteria	Points	Pass	Fail
Clinical diagnosis (20 Points)	Identification of problem	10	Clearly identifies the presenting problem and related symptoms.	Fails to identify or incorrectly identifies the problem and symptoms.
	Diagnostic tests	10	Outlines appropriate diagnostic tests and justifies their necessity.	Fails to mention diagnostic tests or suggests irrelevant tests.
Treatment plan and management (30 Points)	Treatment options	10	Provides a comprehensive list of possible treatment options.	Misses key treatment options or suggests inappropriate treatments.
	Rationale	10	Clearly explains the rationale for choosing a particular treatment approach.	Does not provide or inadequately explain the rationale.
	Risks and benefits	10	Discusses the risks and benefits of the chosen treatment plan.	Omits or inadequately addresses risks and benefits.
Long-term management and follow-up (20 Points)	Follow-up protocols	10	Suggests realistic and evidence-based follow-up protocols.	Fails to suggest or inadequately detail follow-up protocols.
	Patient education	10	Discusses the importance of patient education and outlines topics to be covered.	Neglects to mention or inadequately discusses patient education.
Professionalism and carity of writing (30 Points)	Organisation	10	The essay is well-organised and flows logically from diagnosis to management.	The essay is disorganised or incoherent.
	Technical language	10	Uses appropriate medical terminology correctly.	Uses incorrect or inappropriate terminology.
	Grammar and syntax	10	Writes with correct grammar, punctuation, and sentence structure.	Contains multiple grammatical or syntactic errors.

The formative essays graded by the faculty lacked explanations and feedback as opposed to those graded by ChatGPT. Manually-checked essays were graded on a scale of 1-10, with five being the passing grade. However, ChatGPT provided a comprehensive report with feedback on writing skills that manual scoring lacked.

DISCUSSION

The chatbot's responses during the experiment were fascinating and thought-provoking. They provide insight that AI and ML have a definite role in assessing essay-type questions in the future. Furthermore, it is now evident that there is a need to engage with AI and ML technology to ensure its timely and effective use in medical education. In this study, the AI model was initially trained by human-written essays so that its grading matches closely with that of human checkers. Subsequently, the AI model graded the essays while simultaneously providing individualised written feedback. Written feedback was absent from manually graded essays.

It is therefore imperative that educationists embrace AI's intelligence and learning capabilities with open arms to enhance the standards and quality of medical education in general, and assessment of long and short essay-type ques-

tions specifically. Medical educators should look forward to and work towards employing AI and ML technology for various assessment components, including automated essay scoring, feedback generation, plagiarism detection, and language translation of essay type questions, in less time and with minimal subjectivity.[15]

Here are some of the features that can be incorporated using AI and ML when assessing summative essays based on the author's experience, as shown in Figure 1.

Automated grading systems, powered by AI and ML algorithms, can provide an objective, consistent, and timely evaluation of essay-type questions. This can alleviate the workload of medical educators while ensuring fair grading practices. One such system, the Intelligent Essay Assessor (IEA), uses Latent Semantic Analysis to evaluate the content of essays, providing scores that correlate highly with human graders.[16] By utilising automated grading systems, medical educators can reintroduce essay-type questions without adding a significant grading burden.

ML algorithms can analyse students' essays to identify patterns, trends, and areas of weakness in their understanding. This information can help medical educators provide targeted feedback and create personalised learning experiences.[8] For example, a text mining study by Romero and Ventura demonstrated that ML techniques could be used to extract valuable insights from student writing to improve teaching and learning processes.[17]

AI-powered plagiarism detection systems can ensure academic integrity in essay-type questions. These systems can compare student submissions against a vast database of published literature and other student essays to identify potential instances of plagiarism.[18] This feature can help medical educators maintain high standards of academic integrity while fostering critical thinking and original thought.

AI and ML can be used in formative assessments, providing students with instant feedback on their essay submissions. By employing natural language processing (NLP) techniques, these systems can offer suggestions for improving writing quality, organisation, and content.[19] Immediate feedback can help students identify areas for improvement, promoting active learning and self-directed study.

While AI and ML have made significant progress, they are still not fully adept at understanding human language, including medical jargon and the subtleties of clinical reasoning, thereby limiting their efficacy. The intersection of AI and ML with medical education, particularly in essay-type assessments, offers promising avenues for innovation. However, inherent challenges in grading and technological limitations pose barriers that require further research and refinement. It should be noted that the AI used by the authors in their informal experiment was a free version with limited capabilities. A series of formal research on more advanced versions of AI is needed to establish meaningful outcomes in this regard. Moreover, medical educators should seek to collaborate with AI technology experts to explore the possibilities of using AI in educational endeavours.

CONCLUSION

AI and ML technologies are more likely to supplement human grading rather than replace them at this point unless the educator community empirically confirms their effectiveness through research and evaluations. By integrating AI and ML technologies, medical educators can foster their students' critical thinking, problem-solving, and communication skills, preparing them for the complex and evolving world of healthcare. However, by leveraging these technological advancements, the community should prepare to welcome essay-type questions in the assessment halls of medical education in the 21st century.

ETHICAL APPROVAL:
The study commenced after approval from the research and ethics committee of the Hamdard University Dental Hospital (Ref. No. HCM&D/HUDH/423-23, dated 18-02-2023).

STUDENTS' CONSENT:
Final-year BDS students of Hamdard College of Medicine & Dentistry academic year 2022-2023 were included in the study after obtaining informed consent from them.

COMPETING INTEREST:
The authors declared that they have no conflict of interest.

AUTHORS' CONTRIBUTION:
SS: Drafted the work and analysed the essays through ChatGPT.
JZ: Drafted the work.
AR: Graded the essays manually using standard scales.
All authors approved the final version of the manuscript to be published.

REFERENCES

1. Hift RJ. Should essays and other "open-ended"-type questions retain a place in written summative assessment in clinical medicine? *BMC Med Educ* 2014; **14**:249. doi: 10. 1186/s12909-014-0249-2.
2. Javaeed A. Assessment of higher ordered thinking in medical education: Multiple choice questions and modified essay questions. *MedEdPublish (2016)* 2018; **7**:128. doi: 10.15694/mep.2018.0000128.1.
3. Epstein RM. Assessment in medical education. *N Engl J Med* 2007; **356(4)**:387-96. doi: 10.1056/NEJMra054784.
4. Boulet JR, Durning SJ. What we measure ... and what we should measure in medical education. *Med Educ* 2019; **53(1)**:86-94. doi: 10.1111/medu.13652.
5. Touissi Y, Hjiej G, Hajjioui A, Ibrahimi A, Fourtassi M. Does developing multiple-choice questions improve medical students' learning? A systematic review. *Med Educ Online* 2022; **27(1)**:2005505. doi: 10.1080/10872981.2021.2005505.

6. Lund BD, Wang T. Chatting about ChatGPT: How may AI and GPT impact academia and libraries? *Library Hi Tech News* 2023. doi: 10.2139/ssrn.4333415.
7. Rudolph J, Tan S, Tan S. ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *J Applied Learn Teach* 2023; **6(1)**. doi.org/10.37074/jalt. 2023.6.1.9.
8. Wartman SA, Combs CD. Reimagining medical education in the age of AI. *AMA J Ethics* 2019; **21(2)**:E146-52. doi: 10.1001/amajethics.2019.146.
9. Shermis MD, Hamner B. Contrasting state-of-the-art automated scoring of essays. *Handbook of automated essay evaluation*. *Routledge* 2013; 335-68.
10. Bakator M, Radosav D. Deep learning and medical diagnosis: A review of literature. *Multimodal Technol Interact* 2018; **2(3)**:47. doi: 10.3390/mti2030047.
11. Panayides AS, Amini A, Filipovic ND, Sharma A, Tsaftaris SA, Young A, *et al*. AI in medical imaging informatics: Current challenges and future directions. *IEEE J Biomed Health Inform* 2020; **24(7)**:1837-57. doi: 10.1109/JBHI.2020.2991043.
12. Gierl MJ, Latifi S, Lai H, Boulais AP, De Champlain A. Automated essay scoring and the future of educational assessment in medical education. *Med Educ* 2014; **48(10)**:950-62. doi: 10.1111/medu.12517.
13. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: Systematic review. *JMIR Med Inform* 2019; **7(2)**:e12239. doi: 10.2196/12239.
14. Kocon J, Cichecki I, Kaszyca O, Kochanek M, Szydlo D, Baran J, *et al.* ChatGPT: Jack of all trades, master of none. *Information Fusion* 2023; 101861. doi:10.1016/j.inffus.2023.101861.
15. Bang Y, Cahyawijaya S, Lee N, Dai W, Su D, Wilie B, *et al*. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *Assoc Comput Linguistics* 2023; **1:**675–718. doi:10.18653/v1/2023.ijcnlp--main.45.
16. Landauer TK. Automatic essay assessment. *Assessment Edu: Principles, Policy Practice* 2003; **10(3)**:295-308. doi:10.1080/0969594032000148154.
17. Romero C, Ventura S. Educational data mining: A review of the state of the art. *IEEE* 2010; **40(6)**:601-18. doi:10.1109/TSMCC.2010.2053532.
18. Bretag T. Academic integrity. *Oxford Res Encyclopedia Business Management* 2018. doi: 10.1093/acrefore/9780 190224851.013.147.
19. Graesser AC, McNamara DS, Kulikowich JM. Coh-Metrix: Providing multilevel analyses of text characteristics. *Edu Res* 2011; **40(5)**:223-34. doi: 10.3102/0013189X11413260.