

RESEARCH

Open Access



Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations

Bernadette Quah^{1,2}, Lei Zheng^{1,2†}, Timothy Jie Han Sng^{1,2†}, Chee Weng Yong^{1,2†} and Intekhab Islam^{1,2*} 

Abstract

Background This study aimed to answer the research question: How reliable is ChatGPT in automated essay scoring (AES) for oral and maxillofacial surgery (OMS) examinations for dental undergraduate students compared to human assessors?

Methods Sixty-nine undergraduate dental students participated in a closed-book examination comprising two essays at the National University of Singapore. Using pre-created assessment rubrics, three assessors independently performed manual essay scoring, while one separate assessor performed AES using ChatGPT (GPT-4). Data analyses were performed using the intraclass correlation coefficient and Cronbach's α to evaluate the reliability and inter-rater agreement of the test scores among all assessors. The mean scores of manual versus automated scoring were evaluated for similarity and correlations.

Results A strong correlation was observed for Question 1 ($r=0.752-0.848$, $p<0.001$) and a moderate correlation was observed between AES and all manual scorers for Question 2 ($r=0.527-0.571$, $p<0.001$). Intraclass correlation coefficients of 0.794–0.858 indicated excellent inter-rater agreement, and Cronbach's α of 0.881–0.932 indicated high reliability. For Question 1, the mean AES scores were similar to those for manual scoring ($p>0.05$), and there was a strong correlation between AES and manual scores ($r=0.829$, $p<0.001$). For Question 2, AES scores were significantly lower than manual scores ($p<0.001$), and there was a moderate correlation between AES and manual scores ($r=0.599$, $p<0.001$).

Conclusion This study shows the potential of ChatGPT for essay marking. However, an appropriate rubric design is essential for optimal reliability. With further validation, the ChatGPT has the potential to aid students in self-assessment or large-scale marking automated processes.

Keywords Artificial intelligence, Education, Dental, Academic performance, Models, Educational, Mentoring, Educational needs assessment

Background

Large Language Models (LLMs), such as OpenAI's GPT-4, LLaMA by META, and Google's LaMDA (Language Models for Dialogue Applications), have demonstrated tremendous potential in generating outputs based on user-specified instructions or prompts. These models are trained using large amounts of data and are capable of natural language processing tasks. Owing to their ability to comprehend, interpret, and generate natural language text, LLMs allow human-like conversations with coherent contextual responses to prompts. The capability

[†]Lei Zheng, Timothy Jie Han Sng and Chee Weng Yong contributed equally to this work.

*Correspondence:
Intekhab Islam
denii@nus.edu.sg

¹ Faculty of Dentistry, National University of Singapore, Singapore, Singapore

² Discipline of Oral and Maxillofacial Surgery, National University Centre for Oral Health, 9 Lower Kent Ridge Road, Singapore, Singapore



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

of LLMs to summarize and generate texts that resemble human language allows the creation of task-focused systems that can ease the demands of human labor and improve efficiency.

OpenAI uses a closed application programming interface (API) to process data. Chat Generative Pre-trained Transformer (OpenAI Inc., California, USA, <https://chat.openai.com/>) was introduced globally in 2020 as ChatGPT3, a generative language model with 175 billion parameters [1]. It is based on a generative AI model that can generate new content based on the data on which they have been trained. The latest version, ChatGPT-4, was introduced in 2023 and has demonstrated improved creativity, reasoning, and the ability to process even more complicated tasks [2].

Since its release in the public domain, ChatGPT has been actively explored by both healthcare professionals and educators in an effort to attain human-like performance in the form of clinical reasoning, image recognition, diagnosis, and learning from medical databases. ChatGPT has proven to be a powerful tool with immense potential to provide students with an interactive platform to deepen their understanding of any given topic [3]. In addition, it is also capable of aiding in both lesson planning and student assessments [4, 5].

The potential of ChatGPT for assessments

Automated Essay Scoring (AES) is not a new concept, and interest in AES has been increasing since the advent of AI. Three main categories of AES programs have been described, utilizing regression, classification, or neural network models [6]. A known problem of current AES systems is their unreliability in evaluating the content relevance and coherence of essays [6]. Newer language models such as ChatGPT, however, are potential game changers; they are simpler to learn than current deep learning programs and can therefore improve the accessibility of AES to educators. Mizumoto and Eguchi recently pioneered the potential use of ChatGPT (GPT-3.5 and 4) for AES in the field of linguistics and reported an accuracy level sufficient for use as a supportive tool even when fine-tuning of the model was not performed [7].

The use of these AI-powered tools may potentially ease the burden on educators in marking large numbers of essay scripts, while providing personalized feedback to students [8, 9]. This is especially crucial with larger class sizes and increasing student-to-teacher ratios, where it can be more difficult for educators to actively engage individual students. Additionally, manual scoring by humans can be subjective and susceptible to fatigue, which may put the scoring at risk of being unreliable [7, 10]. The use of AI for essay scoring may thus help reduce

intra- and inter-rater variability associated with manual scoring by providing a more standardized and reliable scoring process that eases the time- and labor-intensive scoring workload of human assessors [10, 11].

The current role of AI in healthcare education

Generative AI has permeated the healthcare industry and provided a diverse range of health enhancements. An example is how AI facilitates radiographic evaluation and clinical diagnosis to improve the quality of patient care [12, 13]. In medical and dental education, virtual or augmented reality and haptic simulations are some of the exciting technological tools already implemented to improve student competence and confidence in patient assessment and execution of procedures [14–16]. The incorporation of ChatGPT into the dental curriculum would thus be the next step in enhancing student learning. The performance of ChatGPT in the United States Medical Licensing Examination (USMLE) was recently validated, with ChatGPT achieving a score equivalent to that of a third-year medical student [17]. However, no data are available on the performance of ChatGPT in the field of dentistry or oral and maxillofacial surgery (OMS). Furthermore, the reliability of AI-powered language models for the grading of essays in the medical field has not yet been evaluated; in addition to essay structure and language, the evaluation of essay scripts in the field of OMS would require a level of understanding of dentistry, medicine and surgery.

Therefore, this study aimed to evaluate the reliability of ChatGPT for AES in OMS examinations for final-year dental undergraduate students compared to human assessors. Our null hypothesis was that there would be no difference in the scores between the ChatGPT and human assessors. The research question for the study was as follows: How reliable is ChatGPT when used for AES in OMS examinations compared to human assessors?

Materials and methods

This study was conducted in the Faculty of Dentistry, National University of Singapore, under the Department of Oral and Maxillofacial Surgery. The study received ethical approval from the university's Institutional Review Board (REF: IRB-2023–1051) and was conducted and drafted with guidance from the education interventions critical appraisal worksheet introduced by BestBETs [18].

Sample size calculation for this study was based on the formula provided by Viechtbauer et al.: $n = \ln(1-\gamma) / \ln(1-\pi)$, where n , γ and π represent the sample size, significance level and level of confidence respectively [19]. Based on a 5% margin of error, a 95% confidence level and a 50% outcome response, it was calculated that a minimum sample size of 59 subjects was required. Ultimately,

the study recruited 69 participants, all of whom were final-year undergraduate dental students. A closed-book OMS examination was conducted on the Exemplify platform (ExamSoft Worldwide Inc., Texas, USA) as a part of the end-of-module assessment. The examination comprised two open-ended essay questions based on the topics taught in the module (Table 1).

Creation of standardized assessment

An assessment rubric was created for each question through discussion and collaboration of a workgroup comprising four assessors involved in the study. All members of the work group were academic staff from the faculty (I.I., B.Q., L.Z., T.J.H.S.) (Supplementary Tables S1 and S2) [20]. An analytic rubric was generated using the strategy outlined by Popham [21]. The process involved a discussion within the workgroup to agree on the learning outcomes of the essay questions. Two authors (I. I. and B. Q.) independently generated the rubric criteria and descriptions for Question 1 (Infection). Similarly, for Question 2 (Trauma), the rubric criteria and descriptions were generated independently by two authors (I.I. and T.J.H.S.). The rubrics were revised until a consensus was reached between each pair. In the event of any disagreement, a third author (L.Z.) provided their opinion to aid in decision making.

Marking categories of Poor (0 marks), Satisfactory (2 marks), Good (3 marks), and Outstanding (4 marks) were allocated to each criterion, with a maximum of 4 marks attainable from each criterion. A criterion for overall essay structure and language was also included, with a maximum attainable 5 marks from this criterion. The highest score for each question was 40.

Model answers to the essays were prepared by another author (C.W.Y.), who did not participate in the creation of the rubrics. Using the rubrics as a reference, the author modified the model answer to create 5 variants of the answers such that each variant fell within different score ranges of 0–10, 11–20, 21–30, 31–40, 41–50.

Subsequently, three authors (B. Q., L. Z., and T.J.H.S.) graded the essays using the prepared rubrics. Revisions to the rubrics were made with consensus by all three authors, a process that also helped calibrate these three authors for manual essay scoring.

AES with ChatGPT

Essay scoring was performed using ChatGPT (GPT-4, released March 14, 2023) by one assessor who did not participate in the manual essay scoring exercise (I.I.). Prompts were generated based on a guideline by Giray, and the components of Instruction, Context, Input Data and Output Indication as discussed in the guideline were included in each prompt (Supplementary Tables 3 and 4) [22]. A prompt template was generated for each question by one assessor (I.I.) with advice from two experts in prompt engineering, based on the marking rubric. The criterion and point allocation were clearly written in prose and point forms. For the fine-tuning process, the prompts were input into ChatGPT using variants of the model answers provided by C.W.Y. Minor adjustments were made to the wording of certain parts of the prompts as necessary to correct any potential misinterpretations of the prompts by the ChatGPT. Each time, the prompt was entered into a new chat in the ChatGPT in a browser where the browser history and cookies were cleared. Subsequently, finalized prompts (Supplementary Tables 3 and 4) were used to score the student essays. AES scores were not used to calculate students' actual essay scores.

Manual essay scoring

Manual essay scoring was completed independently by three assessors (B.Q., L.Z., and T.J.H.S.) using the assessment rubrics (Supplementary Tables S1 and S2). Calibration was performed during the rubric creation stage. The essays were anonymized to prevent bias during the marking process. The assessors recorded the marks allocated to each criterion, as well as the overall score of each essay, on a pre-prepared Excel spreadsheet. Scoring was performed separately and independently by all assessors before the final collation by a research team member (I.I.) for statistical analyses. The student was considered 'able to briefly mention' a criterion if they did not mention any of the keywords of the points within the criterion. The student was considered 'able to elaborate on' a point within the criterion if they were able to mention the keywords of that point as stated in the rubric, and were thus awarded higher marks in accordance with the rubric (e.g. the student was given a higher mark if they were able to mention the need to check for dyspnea and dysphagia, instead of simply mentioning a need to check the patient's airway). Grading was performed with only

Table 1 Essay examination questions

#	Question
1	A male adult presents with a buccal extraoral swelling. He reports pain over a decayed lower molar with foul taste in his mouth. On examination you notice an extraoral swelling and pus discharge intraorally. How will you manage this case?
2	A young male construction worker fell from a height of 10 m. He was conscious when brought to the Accident and Emergency Department. You are called to assess his orofacial trauma. Discuss your management of the patient

whole marks as specified in the rubrics, and assessors were not allowed to give half marks or subscores.

Data synthesis

The scores given out of 40 per essay by each assessor were compiled. Data analyses were subsequently performed using SPSS® version 29.0.1.0(171) (IBM Corporation, New York, United States). For each essay question, correlations between the essay scores given by each assessor were analyzed and displayed using the inter-item correlation matrix. A correlation coefficient value (r) of 0.90–1.00 was indicative of a very strong, 0.70–0.89 indicative of strong, 0.40–0.69 moderate, 0.10–0.39 weak and < 0.10 negligible positive correlation between the scorers [23]. The cutoff p -value for the significance level was set at $p < 0.05$. The intraclass correlation coefficient (ICC) and Cronbach's α were then calculated between all assessors to assess the inter-rater agreement and reliability, respectively [24]. The ICC was interpreted on a scale of 0 to 1.00, with a higher value indicating a higher level of agreement in scores given by the scorers to each student. A value less than 0.40 was indicative of poor, 0.40–0.59 fair, 0.60–0.74 good, and 0.75–1.00 excellent agreement [25]. Using Cronbach's α , reliability was expressed on a range from 0 to 1.00, with a higher number indicating a higher level of consistency between the scorers in their scores given across the students. The reliability was considered 'Less Reliable' if the score was less 0.20, 'Rather Reliable' if the score was 0.20–0.40, 'Quite Reliable' if 0.40–0.60, 'Reliable' if 0.60–0.80 and 'Very Reliable' if 0.80–1.00 [26].

Similarly, the mean scores of the three manual scorers were calculated for each question. The mean manual scores were then analyzed for correlations with AES scores by using Pearson's correlation coefficient. Student's t -test was also used to analyze any significant differences in mean scores between manual scoring and AES. A p -value of < 0.05 was required to conclude the presence of a statistically different score between the groups.

Results

All final-year dental undergraduate students (69/69, 100%) had their essays graded by all manual scorers and AES as part of the study. Table 2 shows the mean scores for each individual assessor as well as the mean scores for the three manual scorers (Scorers 1, 2, and 3).

Analysis of correlation and agreement between all scorers

The inter-item correlation matrices and their respective p -values are listed in Table 3. For Question 1, there was a strong positive correlation between the scores provided

Table 3 Inter-item correlation matrix and significance

Question 1					
		Scorer 1	Scorer 2	Scorer 3	AES
Scorer 1	r	1.000			
	p				
Scorer 2	r	0.848	1.000		
	p	< 0.001*			
Scorer 3	r	0.797	0.773	1.000	
	p	< 0.001*	< 0.001*		
AES	r	0.810	0.753	0.752	1.000
	p	< 0.001*	< 0.001*	< 0.001*	
Question 2					
		Scorer 1	Scorer 2	Scorer 3	AES
Scorer 1	r	1.000			
	p				
Scorer 2	r	0.829	1.000		
	p	< 0.001*			
Scorer 3	r	0.756	0.655	1.000	
	p	< 0.001*	< 0.001*		
AES	r	0.571	0.527	0.542	1.000
	p	< 0.001*	< 0.001*	< 0.001*	

A strong correlation was found between all groups for Question 1, and a strong to moderate correlation was found between the groups for Question 2

AES Automated essay scoring

*signifies significant difference

Table 2 Mean scores and standard deviations (S.D.) for each assessor. Manual scoring was performed by Scorers 1, 2, and 3, while AES was performed by a separate team member (I.I.). The mean scores of Scorers 1, 2, and 3 were calculated to obtain the Combined Manual Scores. AES showed a significantly lower mean score than manual scoring for Question 2, but not for Question 1

Scorer	Question 1			Question 2		
	Mean	S.D	p	Mean	S.D	p
Scorer 1	18.72	6.010		23.23	4.446	
Scorer 2	11.71	4.288		21.51	4.928	
Scorer 3	14.12	5.731		24.58	4.587	
Combined Manual Scores	14.85	4.988	0.726	23.11	4.241	< 0.001*
AES	14.54	5.490		18.62	4.044	

by each assessor (Scorers 1, 2, 3, and AES), with r -values ranging from 0.752–0.848. All p -values were <0.001 , indicating a significant positive correlation between all assessors. For Question 2, there was a strong positive correlation between Scorers 1 and 2 ($r=0.829$) and Scorers 1 and 3 ($r=0.756$). There was a moderate positive correlation between Scorers 2 and 3 ($r=0.655$), as well as between AES and all manual scores (r -values ranging from 0.527 to 0.571). Similarly, all p -values were <0.001 , indicative of a significant positive correlation between all scorers.

For the analysis of inter-rater agreement, ICC values of 0.858 (95% CI 0.628 – 0.933) and 0.794 (95% CI 0.563 – 0.892) were obtained for Questions 1 and 2, respectively, both of which were indicative of excellent inter-rater agreement. Cronbach's α was 0.932 for Question 1 and 0.881 for Question 2, both of which were 'Very Reliable'.

Analysis of correlation between manual scoring versus AES

The results of the Student's t -test comparing the test score values from manual scoring and AES are shown in Table 2. For Question 1, the mean manual scores (14.85 ± 4.988) were slightly higher than those of the AES (14.54 ± 5.490). However, these differences were not statistically significant ($p > 0.05$). For Question 2, the mean manual scores (23.11 ± 4.241) were also higher than those of the AES (18.62 ± 4.044); this difference was statistically significant ($p < 0.001$).

The results of the Pearson's correlation coefficient calculations are shown in Table 4. For Question 1, there was a strong and significant positive correlation between manual scoring and AES ($r=0.829$, $p < 0.001$). For Question 2, there was a moderate and statistically significant positive correlation between the two groups ($r=0.599$, $p < 0.001$).

Qualitative feedback from AES

Figures 1, 2 and 3 show three examples of essay feedback and scoring provided by ChatGPT. ChatGPT provided feedback in a concise and systematic manner. Scores were clearly provided for each of the criteria listed in the assessment rubric. This was followed by in-depth feedback on the points within the criterion that the student

had discussed and failed to mention. ChatGPT was able to differentiate between a student who briefly mentioned a key point and a student who provided better elaboration on the same point by allocating them two or three marks, respectively.

One limitation of ChatGPT that was identified during the scoring process was its inability to identify content that was not relevant to the essay or that was factually incorrect. This was despite the assessment rubric specifying that incorrect statements should be given 0 marks for that criterion. For example, a student who included points about incision and drainage also incorrectly stated that bone scraping to induce bleeding and packing of local hemostatic agents should be performed. Although these statements were factually incorrect, ChatGPT was unable to identify this and still awarded student marks for the point. Manual assessors were able to spot this and subsequently penalized the student for the mistake.

Discussion

Since its recent rise in popularity, many people have been eager to tap into the potential of large language models, such as ChatGPT. In their review, Khan et al. discussed the growing role of ChatGPT in medical education, with promising uses for the creation of case studies and content such as quizzes and flashcards for self-directed practice [9]. As an LLM, the ability of ChatGPT to thoroughly evaluate sentence structure and clarity may allow it to confront the task of automated essay marking.

Advantages of ChatGPT in AES

This study found significant correlations and excellent inter-rater agreement between ChatGPT and manual scorers, and the mean scores between both groups showed strong to moderate correlations for both essay questions. This suggests that AES has the potential to provide a level of essay marking similar to that of the educators in our faculty. Similar positive findings were reflected in previous studies that compared manual and automated essay scoring ($r=0.532$ – 0.766) [6]. However, there is still a need to further fine-tune the scoring system such that the score provided by AES deviates as little as possible from human scoring. For instance, the mean AES score was lower than that of manual scoring by 5 marks for Question 2. Although the difference may not seem large, it may potentially increase or decrease the final performance grade of students.

Apart from a decent level of reliability in manual essay scoring, there are many other benefits to using ChatGPT for AES. Compared to humans, the response time to prompts is much faster and can thus increase productivity and reduce the burden of a large workload on educational assessors [27]. In addition, ChatGPT

Table 4 Correlation between mean essay scores by manual scorers and AES. A strong and moderate correlation was found between the two groups for Questions 1 and 2 respectively

	Correlation (r)	
	r	p
Question 1	0.829	<0.001
Question 2	0.599	<0.001

Given the grading rubric and the student's essay, let's evaluate the response according to each criterion and provide a score along with feedback for each category.

1. **Determine Severity / Establishing Airway:** The student did not explicitly mention the key points such as checking for dyspnea, dysphagia, odynophagia, the tripod position, orthopnea, hoarseness of voice, swollen tongue, use of SpO2 monitoring, or trismus. Thus, the student does not touch on the key points under this criterion.

Score: 0 points

Feedback: The essay does not address the initial assessment of the patient's airway and the severity of their condition. Key aspects like breathing difficulty, swallowing issues, or potential airway obstruction signs were not discussed. Future essays should include these critical assessments to ensure a comprehensive approach to patient care.

2. **Evaluate Host Defences:** The student mentions the importance of assessing the patient's medical history for conditions that could make the patient immunocompromised, such as renal disease and diabetes, but does not elaborate on the other conditions listed.

Score: 2 points

Feedback: You have made a start by mentioning diabetes and renal disease, which are important in assessing a patient's immune defense systems. However, there is a need to address other conditions such as use of corticosteroids, transplant patients, malignancy, malnutrition, HIV, and alcoholism to fully cover the spectrum of immunocompromising conditions.

3. **Determine Setting of Care:** The student does not address the considerations for hospitalization directly, such as checking vital signs, the need for general anesthesia, concerns for airway compromise, or uncontrolled medical conditions.

Score: 0 points

Feedback: The essay lacks an assessment of the factors that influence the decision to hospitalize a patient or treat them on an outpatient basis. Including points such as vital sign assessment and the potential need for general anesthesia would be crucial to determining the appropriate setting of care.

4. **Investigations:** The student mentions the use of radiographs to identify the source of infection but does not cover the general investigations like blood tests or the local dental sensibility tests.

Score: 2 points

Feedback: You have correctly identified the need for radiographic investigations, which is commendable. However, to provide a comprehensive diagnostic approach, including blood tests and specific dental tests would enhance your management plan.

5. **Treat Surgically 1:** The student's essay does not cover the necessary surgical considerations such as local vs. general anesthesia, deep fascial space involvement, or trismus.

Score: 0 points

Feedback: The surgical treatment plan needs to include considerations for anesthesia, potential involvement of deep fascial spaces, and issues like trismus that can affect surgical approach. These details are essential for a well-rounded surgical plan.

6. **Treat Surgically 2:** The student briefly mentions the need to drain pus and extract the tooth but does not elaborate on the surgical approach or the details of the incision and drainage process.

Score: 2 points

Feedback: While you touch on drainage and tooth extraction, expanding on the surgical approaches and detailing the incision and drainage process would greatly improve the quality of your essay.

7. **Treat Surgically 3:** The student mentions tooth extraction but does not provide detailed information on the removal of the source of infection such as tooth extraction or root canal treatment.

Score: 2 points

Feedback: It is good that you mention tooth extraction, but elaborating on the methods and justifications for removing the source of infection, including considerations for tooth extraction or root canal treatment, would better meet the criteria for this section.

8. **Support Medically:** The student does not discuss the need for good hydration, nutrition, or the importance of follow-up after surgery and discharge.

Score: 0 points

Feedback: Postoperative care such as ensuring proper hydration and nutrition and the rationale behind these measures are missing from your essay. Including these aspects along with the need for follow-up would constitute a more comprehensive treatment plan.

9. **Specific Antibiotic Therapy:** The student mentions antibiotic prescriptions but does not discuss culture and sensitivity, empirical antibiotics, or the adjustment of antibiotics based on culture results.

Score: 2 points

Feedback: Prescribing antibiotics is important, but it is equally important to mention the need for culture and sensitivity tests to tailor antibiotic therapy. Addressing this will improve your management of bacterial infections.

10. **Overall Structure and Language:** The student's essay lacks a clear introduction, and while there is an attempt to address the topic, there are gaps in the flow and structure.

Score: 2 points

Feedback: Your essay demonstrates an attempt at discussing the case but lacks a clear introduction and conclusion. To improve, focus on a more structured approach

Fig. 1 Example #1 of a marked essay with feedback from ChatGPT for Question 1

Given the grading rubric and the student's essay, let's evaluate the response according to each criterion and provide a score along with feedback for each category.

1. **Determine Severity / Establishing Airway:** The student did not mention assessing difficulty breathing (dyspnea), difficulty swallowing (dysphagia), pain on swallowing (odynophagia), the tripod position, orthopnea, hoarseness of voice, or use of pulse oximetry.

Score: 0 points

Feedback: The essay lacks any discussion on the initial assessment of the airway, which is critical in oral surgery cases. It's important to include an evaluation of the patient's ability to breathe, swallow, and the presence of any voice changes. Future essays should address these points to ensure patient safety.

2. **Evaluate Host Defences:** The student mentioned checking the medical history but did not elaborate on specific conditions that affect host defenses.

Score: 2 points

Feedback: While you did mention the importance of medical history, a more detailed discussion on conditions that can compromise the patient's immune system, such as diabetes or the use of corticosteroids, would enhance the quality of your essay.

3. **Determine Setting of Care:** The student did not discuss vital signs, general anesthesia, airway compromise, uncontrolled medical conditions, or malnutrition and dehydration directly.

Score: 2 points

Feedback: In managing this case, consideration of whether to hospitalize the patient based on vital signs and other factors is essential. Include these details to ensure comprehensive care planning.

4. **Investigations:** The student mentioned ordering a dental panoramic tomogram and periapical radiograph, which falls under local investigations. They did not mention general investigations such as blood tests.

Score: 3 points

Feedback: You addressed local investigations well by indicating the need for dental imaging. However, discussing general investigations would provide a complete picture of the patient's health status and should not be overlooked.

5. **Treat Surgically 1:** The essay mentions extraction and debridement but does not cover considerations for treatment under local vs. general anesthesia or elaborate on the need to establish an airway.

Score: 2 points

Feedback: It's crucial to discuss the factors influencing the choice of anesthesia, especially when there's a risk of airway compromise due to swelling or infection.

6. **Treat Surgically 2:** The essay mentions debridement and clearing the socket but does not provide details on the incision and drainage process, use of anesthesia during the procedure, or insertion of an open drain.

Score: 2 points

Feedback: A more thorough explanation of the surgical approach, including the steps of incision and drainage, would enhance your essay. These details are important for surgical planning and patient care.

7. **Treat Surgically 3:** The student elaborates on the need to remove the source of the infection, including whether to extract the tooth or perform root canal treatment, and mentions debridement.

Score: 3 points

Feedback: You've done well to address the need for removing the source of the infection. Your discussion on treatment options and postoperative care is good.

8. **Support Medically:** The student prescribes medications for infection and pain but does not elaborate on the importance of hydration and nutrition or the need for close follow-up.

Score: 2 points

Feedback: While prescribing medication is important, don't forget to elaborate on the supportive medical care like ensuring hydration and nutrition, which are critical for patient recovery.

9. **Specific Antibiotic Therapy:** The student discusses sending pus for culture and sensitivity and prescribing empirical antibiotics but does not mention adjusting therapy based on culture results.

Score: 3 points

Feedback: You've outlined a good initial approach with empirical antibiotics, but make sure to discuss the importance of tailoring antibiotic therapy based on culture and sensitivity results for complete management.

10. **Overall Structure and Language:** The essay lacks clear structure and proper use of English, with multiple grammatical and spelling issues. There is no clear introduction or conclusion.

Score: 2 points

Feedback: Attention to essay structure, grammar, and spelling is as important as content. A clear introduction, logical flow, and a conclusion would greatly improve the presentation of your essay.

Final Feedback: Your essay provides a reasonable approach to managing the clinical scenario presented, with appropriate emphasis on the treatment of infection and pain. However, it's important to follow the structure of a well-organized essay, with a clear introduction, discussion of key points, and a conclusion. Be sure to address all critical aspects of patient assessment, investigations, and treatment planning in future essays.

Fig. 2 Example #2 of a marked essay with feedback from ChatGPT for Question 1

To evaluate the student's essay according to the provided grading rubric, I will address each category in turn, assign a score based on the rubric's criteria, and provide feedback.

1. **Determine Severity / Establishing Airway:** The essay does not discuss initial patient assessment and ensuring the patient's airway is patent. None of the key points such as difficulty breathing, tripod position, hoarseness of voice, pulse oximetry monitoring, or trismus are mentioned.

Score: 0 points

Feedback: The essay failed to address the critical initial steps in patient assessment, specifically the importance of securing the airway and checking for signs of respiratory distress or obstruction. It's essential to include an evaluation for dyspnea, dysphagia, odynophagia, orthopnea, voice changes, and use of pulse oximetry.

2. **Evaluate Host Defences:** The student does not discuss assessing the patient's medical history for conditions that could render them immunocompromised.

Score: 0 points

Feedback: The essay lacks any mention of the patient's medical history which is vital for identifying potential immunocompromised states. Consider including a discussion on conditions like diabetes, corticosteroid use, transplantation status, malignancy, malnutrition, HIV status, chronic renal disease, and alcohol use.

3. **Determine Setting of Care:** The essay does not discuss whether to hospitalize the patient or treat them as an outpatient, nor does it mention vital signs, the need for general anesthesia, concerns for airway compromise, uncontrolled medical conditions, or malnutrition/dehydration.

Score: 0 points

Feedback: The management of the case needs to address the criteria for determining the appropriate setting of care. Mentioning the assessment of vital signs and the potential need for hospitalization would strengthen this section.

4. **Investigations:** The student briefly mentions the use of radiographs like CT scans to assess the spread of infection, which addresses one of the key points under local investigations.

Score: 2 points

Feedback: While you touched on the use of a CT scan to evaluate the infection's extent, it would be beneficial to discuss other general and local investigations, such as blood tests and dental sensibility tests.

5. **Treat Surgically 1:** The essay briefly mentions incision and drainage and the use of local or general anesthesia but does not elaborate on any of the specific considerations listed.

Score: 0 points

Feedback: Surgical treatment options were mentioned but not elaborated upon. Future essays should delve into the considerations for using local versus general anesthesia and address concerns such as airway management and the involvement of deep fascial spaces.

6. **Treat Surgically 2:** The student discusses the incision and drainage process, including giving local anesthesia, making an incision, blunt dissection, wound irrigation, and insertion of a drain, which covers several key points and shows a good understanding of the surgical approach.

Score: 3 points

Feedback: Excellent coverage of the incision and drainage process. To improve, ensure that the approach (intraoral versus extraoral) and rationale behind the choice are also discussed.

7. **Treat Surgically 3:** The student briefly discusses tooth extraction and the removal of the source of infection, addressing the criterion but without detailed elaboration.

Score: 2 points

Feedback: You have recognized the need to remove the source of infection. To enhance your response, include a more detailed explanation of the decision-making process between tooth extraction and root canal treatment.

8. **Support Medically:** The essay does not mention the need for good hydration and nutrition or the importance of close follow-up after surgery.

Score: 0 points

Feedback: It is crucial to discuss the medical support required post-surgery, emphasizing hydration, nutrition, and follow-up care. These elements are vital for patient recovery and should not be omitted.

9. **Specific Antibiotic Therapy:** The student mentions the use of empirical antibiotics and specifies antibiotics like clindamycin and amoxicillin but does not elaborate on the importance of culture and sensitivity testing or the adjustment of antibiotic therapy based on these results.

Score: 2 points

Feedback: While empirical antibiotic use is addressed, the essay should emphasize the importance of culture and sensitivity tests in guiding specific antibiotic therapy. Mentioning the need to adjust the antibiotic regimen based on these tests would provide a more comprehensive discussion.

10. **Overall Structure and Language:** The essay is structured in prose form, with a clear introduction to the topic and various paragraphs discussing different criteria. There are minimal grammar or spelling issues, but the essay could benefit from clearer transitions and a more defined conclusion.

Score: 3 points

Feedback: The essay is well-structured with a clear introduction and minimal language errors. However, improving the flow of ideas and including a conclusive paragraph would enhance readability and the strength of the argument.

Overall Feedback: The essay demonstrates a basic understanding of surgical treatment and antibiotic therapy but lacks a comprehensive approach to patient assessment, setting of care, medical support, and fails to mention crucial elements such as ensuring a patent airway and evaluating host defenses. To improve, it is important to cover all aspects of patient care, including preoperative assessment, considerations for inpatient versus outpatient treatment, and postoperative support. Additionally, ensure to elaborate on each point for a more in-depth analysis and provide a clear conclusion to summarize the management plan.

Fig. 3 Example #3 of a marked essay with feedback from ChatGPT for Question 1

can provide individualized feedback for each essay (Figs. 1, 2 and 3). This helps provide students with comments specific to their essays, a feat that is difficult to achieve for a single educator teaching a large class size.

Similar to previous systems designed for AES, machine scoring is beneficial for removing human inconsistencies that can result from fatigue, mood swings, or bias [10]. ChatGPT is no exception. Furthermore, ChatGPT is more widely accessible than the conventional AES systems. Its software runs online instead of requiring downloads on a computer, and its user interface is simple to use. With GPT-3.5 being free to use and GPT-4 being 20 USD per month, it is also relatively inexpensive.

Marking the essay is only part of the equation, and the next step is to allow the students to know what went wrong and why. Nicol and Macfarlane described seven principles for good feedback. ChatGPT can fulfil most of these principles, namely, facilitating self-assessment, encouraging teacher and peer dialogue, clarifying what good performance is, providing opportunities to close the gap between current and desired performance, and delivering high-quality information to students [28]. In this study, the feedback given by ChatGPT was categorized based on the rubric, and elaboration was provided for each criterion on the points the student mentioned and did not mention. By highlighting the ideal answer and where the student can improve, ChatGPT can clarify performance goals and provide opportunities to close the gap between the student's current and desired performance. This creates opportunities for self-directed learning and the utilization of blended learning environments. Students can use ChatGPT to review their preparation on topics, self-grade their essays, and receive instant feedback. Furthermore, the simple and interactive nature of the software encourages dialogue, as it can readily respond to any clarification the student wants to make. The importance of effective feedback has been demonstrated to be an essential component in medical education, in terms of enhancing the knowledge of the student without developing negative emotions [29, 30].

These potential advantages of engaging ChatGPT for student assessments play well into the humanistic learning theory of medical education [31, 32]. Self-directed learning allows students the freedom to learn at their own pace, with educators simply providing a conducive environment and the goals that the student should achieve. ChatGPT has the potential to supplement the role of the educator in self-directed learning, as it can be readily available to provide constructive and tailored feedback for assignments whenever the student is ready for it. This removes the burden that assignment deadlines place on students, which can allow them a greater sense

of independence and control over their learning, and lead to greater self-motivation and self-fulfillment.

Potential pitfalls of ChatGPT

Potential pitfalls associated with the use of ChatGPT were identified. First, the ability to achieve reliable scores relies heavily on a well-created marking rubric with clearly defined terms. In this study, the correlations between scorers were stronger for Question 1 compared to Question 2, and the mean scores between the AES and manual scorers were also significantly different for Question 2, but not for Question 1. The lower reliability of the AES for Question 2 may be attributed to its broader nature, use of more complex medical terms, and lengthier scoring rubrics. The broad nature of the question left more room for individual interpretation and variation between humans and AES. The ability of ChatGPT to provide accurate answers may be reduced with lengthier prompts and conversations [27]. Furthermore, with less specific instructions or complex medical jargon, both automated systems and human scorers may interpret rubrics differently, resulting in varied scores across the board [10, 33, 34]. The authors thus recommend that to circumvent this, the use of ChatGPT for essay scoring should be restricted to questions that are less broad (e.g. shorter essays), or by breaking the task into multiple prompts for each individual criterion to reduce variations in interpretation [27, 35]. Furthermore, the rubrics should contain concise and explicit instructions with appropriate grammar and vocabulary to avoid misinterpretation by both ChatGPT and human scorers, and provide a brief explanation to specify what certain medical terms mean (e.g. writing 'pulse oximetry (SpO₂) monitoring' instead of only 'SpO₂') for better contextualization [35, 36].

Second, prompt engineering is a critical step in producing the desired outcome from ChatGPT [27]. A prompt that is too ambiguous or lacks context can lead to a response that is incomplete, generic, or irrelevant, and a prompt that exhibits bias runs the risk of bias reinforcement in the given reply [22, 34]. Phrasing the prompt must also be carefully checked for spelling, grammatical mistakes, or inconsistencies, since ChatGPT uses the prompt's phrasing literally. For example, a prompt that reads 'give 3 marks if the student covers one or more coverage points' will result in ChatGPT only giving the marks if multiple points are covered, because of the plural nature of the word 'points'.

Third, irrelevant content may not be penalized during the essay-marking process. Students may 'trick' the AES by producing a lengthier essay to hit more relevant points and increase their score. This may result in essays of lower quality with multiple incorrect or nonsensical

statements still rewarded with higher scores [10]. Our assessment rubric attempts to penalize the student with 0 marks if incorrect statements on the criterion are made; however, none of the students were penalized. This issue may be resolved as ChatGPT rapidly and continuously gains more medical and dental knowledge. Although data to support the competence of AI in medical education are sparse, the quality of the medical knowledge that ChatGPT already has is sufficient to achieve a passing mark at the USMLE [5, 37]. In dentistry, when used to disseminate information on endodontics to patients, ChatGPT was found to provide detailed answers with an overall validity of 95% [38]. Over time, LLMs such as ChatGPT may be able to identify when students are not factually correct.

Other comments

The lack of human emotion in machine scoring can be both an advantage and a disadvantage. AES can provide feedback that is entirely factual and less biased than humans, and grades are objective and final [39]. However, human empathy is an essential quality that ChatGPT does not possess. One principle of good feedback is to encourage and motivate students to provide positive learning experiences and build self-esteem [28]. While ChatGPT can provide constructive feedback, it will not be able to replace the compassion, empathy, or emotional intelligence possessed by a quality educator possesses [40]. In our study, ChatGPT awarded lower mean scores of 14.54/40 (36.4%) and 18.62/40 (46.5%) compared to manual scoring for both questions. Although objective, some may view automated scoring as harsh because it provided failing grades to an average student.

This study demonstrates the ability of GPT-4 to evaluate essays without any specialized training or prompting. One long prompt was used to score each essay. Although more technical prompting methods, such as chain of thought, could be deployed, our single prompt method makes the method scalable and easier to adopt. As discussed earlier, ChatGPT is the most reliable when prompts are short and specific [34]. Hence, each prompt should ideally task ChatGPT to score only one or two criteria, rather than the entire rubric of the 10 criteria. However, in a class of 70, the assessors are required to run 700 prompts per question, which is impractical and unnecessary. With only one prompt, a good correlation was still found between the AES and manual scoring. It is likely that further exploration and experimentation with prompting techniques can improve the output.

While LLMs have the potential to revolutionize education in healthcare, some precautions must be taken. Artificial Hallucination is a widely described phenomenon; ChatGPT may generate seemingly genuine but

inaccurate information [41–43]. Hallucinations have been attributed to biases and limitations of training data as well as algorithmic limitations [2]. Similarly, randomness of the generated responses has been observed; while it is useful for generating creative content, this may be an issue when ChatGPT is employed for topics requiring scientific or factual content [44]. Thus, LLMs are not infallible and still require human subject matter experts to validate the generated content. Finally, it is essential that educators play an active role in driving the development of dedicated training models to ensure consistency, continuity, and accountability, as overreliance on a corporate-controlled model may place educators at the mercy of algorithm changes.

The ethical implications of using ChatGPT in medical and dental education also need to be explored. As much as LLMs can provide convenience to both students and educators, privacy and data security remain a concern [45]. Robust university privacy policies and informed consent procedures should be in place for the protection of student data prior to the use of ChatGPT as part of student assessment. Furthermore, if LLMs like ChatGPT were to be used for grading examinations in the future, issues revolving around fairness and transparency of the grading process need to be resolved [46]. GPT-4 may have provided harsh scores in this study, possibly due to some shortfall in understanding certain phrases the students have written; models used in assessments will thus require sufficient training in the field of healthcare to properly acquire the relevant medical knowledge and hence understand and grade essays fairly.

As AI continues to develop, ChatGPT may eventually replace human assessors in essay scoring for dental undergraduate examinations. However, given its current limitations and dependence on a well-formed assessment rubric, relying solely on ChatGPT for exam grading may be inappropriate when the scores can affect the student's overall module scores, career success, and mental health [47]. While this study primarily demonstrates the use of ChatGPT to grade essays, it also points to great potential in using it as an interactive learning tool. A good start for its use is essay assignments on pre-set topics, where students can direct their learning on their own and receive objective feedback on essay structure and content that does not count towards their final scores. Students can use rubrics to practice and gain effective feedback from LLMs in an engaging and stress-free environment. This reduces the burden on educators by easing the time-consuming task of grading essay assignments and allows students the flexibility to complete and grade their assignments whenever they are ready. Furthermore, assignments repeated with new class cohorts can enable

more robust feedback from ChatGPT through machine learning.

Study limitations

The limitations of this study lie in part of its methodology. The study recruited 69 dental undergraduate students; while this is above the minimum calculated sample size of 59, a larger sample size would help to increase the generalizability of the study findings to larger populations of students and a wide scope of topics. The unique field of OMS also requires knowledge of both medical and dental subjects, and hence the results obtained from the use of ChatGPT for essay marking in other medical or dental specialties may differ slightly.

The use of rubrics for manual scoring could also be a potential source of bias. While the rubrics provide a framework for objective assessment, they cannot eliminate the subjectiveness of manual scoring. Variations in the interpretation of the students' answers, leniency errors (whereby one scorer marks more leniently than another) or rater drift (fatigue from assessing many essays may affect leniency of marking and judgment) may still occur. To minimize bias resulting from these errors, multiple assessors were recruited for the manual scoring process and the average scores were used for comparison with AES.

Conclusion

This study investigated the reliability of ChatGPT in essay scoring for OMS examinations, and found positive correlations between ChatGPT and manual essay scoring. However, ChatGPT tended towards stricter scoring and was not capable of penalizing irrelevant or incorrect content. In its present state, GPT-4 should not be used as a standalone tool for teaching or assessment in the field of medical and dental education but can serve as an adjunct to aid students in self-assessment. The importance of proper rubric design to achieve optimal reliability when employing ChatGPT in student assessment cannot be overemphasized.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12909-024-05881-6>.

Supplementary Material 1

Acknowledgements

We would like to extend our gratitude to Mr Paul Timothy Tan Bee Xian and Mr Jonathan Sim for their invaluable advice on the process of prompt engineering for the effective execution of this study.

Authors' contributions

B.Q. contributed in the stages of conceptualization, methodology, study execution, validation, formal analysis and manuscript writing (original draft

and review and editing). L.Z., T.J.H.S. and C.W.Y. contributed in the stages of methodology, study execution, and manuscript writing (review and editing). I.I. contributed in the stages of conceptualization, methodology, study execution, validation, formal analysis, manuscript writing (review and editing) and supervision. All authors provided substantial contributions to this manuscript in the following form:

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

This study was approved by the Institutional Review Board of the university (REF: IRB-2023–1051). The waiver of consent from students was approved by the University's Institutional Review Board, as the scores by ChatGPT were not used as the students' actual grades, and all essay manuscripts were anonymized.

Consent for publication

All the authors reviewed the content of this manuscript and provided consent for publication.

Competing interests

The authors declare no competing interests.

Received: 4 February 2024 Accepted: 9 August 2024

Published online: 03 September 2024

References

1. Floridi L, Chiriatti M. GPT-3: Its nature, scope, limits, and consequences. *Mind Mach.* 2020;30(4):681–94.
2. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, Aziz S, Damseh R, Alabed Alrazak S, Sheikh J. Large language models in medical education: opportunities, challenges, and future directions. *JMIR Med Educ.* 2023;9:e48291.
3. Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F, Gasser U, Groh G, Günemann S, Hüllermeier E, et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn Individ Differ.* 2023;103:102274.
4. Javaid M, Haleem A, Singh RP, Khan S, Khan IH. Unlocking the opportunities through ChatGPT Tool towards ameliorating the education system. *BenchCouncil Transact Benchmarks Standards Eval.* 2023;3(2): 100115.
5. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health.* 2023;2(2):e0000198.
6. Ramesh D, Sanampudi SK. An automated essay scoring systems: a systematic literature review. *Artif Intell Rev.* 2022;55(3):2495–527.
7. Mizumoto A, Eguchi M. Exploring the potential of using an AI language model for automated essay scoring. *Res Methods Appl Linguist.* 2023;2(2): 100050.
8. Erturk S, Tilburg W, Igou E. Off the mark: Repetitive marking undermines essay evaluations due to boredom. *Motiv Emotion* 2022;46.
9. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT - Reshaping medical education and clinical management. *Pak J Med Sci.* 2023;39(2):605–7.
10. Hussein MA, Hassan H, Nassef M. Automated language essay scoring systems: a literature review. *PeerJ Comput Sci.* 2019;5:e208.
11. Blood I. Automated essay scoring: a literature review. *Studies in Applied Linguistics and TESOL* 2011, 11(2).
12. Menezes LDS, Silva TP, Lima Dos Santos MA, Hughes MM, Mariano Souza SDR, Leite Ribeiro PM, Freitas PHL, Takeshita WM: Assessment of landmark detection in cephalometric radiographs with different conditions of

- brightness and contrast using the an artificial intelligence software. *Dentomaxillofac Radiol* 2023;20230065.
13. Bennani S, Regnard NE, Ventre J, Lassalle L, Nguyen T, Ducarouge A, Dargent L, Guillo E, Gouhier E, Zaimi SH, et al. Using AI to improve radiologist performance in detection of abnormalities on chest radiographs. *Radiology*. 2023;309(3): e230860.
14. Moussa R, Alghazaly A, Althagafi N, Eshky R, Borzangy S. Effectiveness of virtual reality and interactive simulators on dental education outcomes: systematic review. *Eur J Dent*. 2022;16(1):14–31.
15. Fanizzi C, Carone G, Rocca A, Ayadi R, Petrenko V, Casali C, Rani M, Giachino M, Falsitta LV, Gambatesa E, et al. Simulation to become a better neurosurgeon An international prospective controlled trial: The Passion study. *Brain Spine*. 2024;4:102829.
16. Lovett M, Ahanonu E, Molzahn A, Biffar D, Hamilton A. Optimizing individual wound closure practice using augmented reality: a randomized controlled study. *Cureus*. 2024;16(4):e59296.
17. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. How does ChatGPT perform on the United States medical licensing examination? the implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*. 2023;9:e45312.
18. Educational Intervention Worksheet, BestBets, Accessed 31/03/2024. https://bestbets.org/ca/pdf/educational_intervention.pdf.
19. Viechtbauer W, Smits L, Kotz D, Budé L, Spigt M, Serroyen J, Crutzen R. A simple formula for the calculation of sample size in pilot studies. *J Clin Epidemiol*. 2015;68(11):1375–9.
20. Cox G, Morrison J, Brathwaite B: The Rubric: An Assessment Tool to Guide Students and Markers; 2015.
21. Popham J. W: "What's Wrong—And What's Right—With Rubrics." *Educ Leadersh*. 1997;55(2):72–5.
22. Giray L. Prompt Engineering with ChatGPT: A Guide for Academic Writers. *Ann Biomed Eng*. 2023;51:3.
23. Schober P, Boer C, Schwarte LA. Correlation Coefficients: Appropriate Use and Interpretation. *Anesth Analg*. 2018;126(5):1763–8.
24. Liao SC, Hunt EA, Chen W. Comparison between inter-rater reliability and inter-rater agreement in performance assessment. *Ann Acad Med Singap*. 2010;39(8):613–8.
25. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess*. 1994;6(4):284–90.
26. Hair J, Black W, Babin B, Anderson R: *Multivariate Data Analysis: A Global Perspective*; 2010.
27. Nazir A, Wang Z: A Comprehensive Survey of ChatGPT: Advancements, Applications, Prospects, and Challenges. *Meta Radiol* 2023;1(2).
28. Nicol D, Macfarlane D: Rethinking Formative Assessment in HE: a theoretical model and seven principles of good feedback practice. *IEEE Personal Communications - IEEE Pers Commun* 2004;31.
29. Spooner M, Larkin J, Liew SC, Jaafar MH, McConkey S, Pawlikowska T. "Tell me what is 'better'!" How medical students experience feedback, through the lens of self-regulatory learning. *BMC Med Educ*. 2023;23(1):895.
30. Kornegay JG, Kraut A, Manthey D, Omron R, Caretta-Weyer H, Kuhn G, Martin S, Yarris LM. Feedback in medical education: a critical appraisal. *AEM Educ Train*. 2017;1(2):98–109.
31. Mukhalalati BA, Taylor A. Adult learning theories in context: a quick guide for healthcare professional educators. *J Med Educ Curric Dev*. 2019;6:2382120519840332.
32. Taylor DC, Hamdy H. Adult learning theories: implications for learning and teaching in medical education: AMEE Guide No. 83. *Med Teach*. 2013;35(11):e1561-1572.
33. Chakraborty S, Dann C, Mandal A, Dann B, Paul M, Hafeez-Baig A: Effects of Rubric Quality on Marker Variation in Higher Education. *Studies In Educational Evaluation* 2021;70.
34. Heston T, Khun C. Prompt engineering in medical education. *Int Med Educ*. 2023;2:198–205.
35. Sun GH: Prompt Engineering for Nurse Educators. *Nurse Educ* 2024.
36. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. *J Med Internet Res*. 2023;25:e50638.
37. Sun L, Yin C, Xu Q, Zhao W. Artificial intelligence for healthcare and medical education: a systematic review. *Am J Transl Res*. 2023;15(7):4820–8.
38. Mohammad-Rahimi H, Ourang SA, Pourhoseingholi MA, Dianat O, Dummer PMH, Nosrat A: Validity and reliability of artificial intelligence chatbots as public sources of information on endodontics. *Int Endodontic J* 2023, n/a(n/a).
39. Peng X, Ke D, Xu B: Automated essay scoring based on finite state transducer: towards ASR transcription of oral English speech. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*. Jeju Island, Korea: Association for Computational Linguistics; 2012:50–59.
40. Grassini S. Shaping the future of education: Exploring the Potential and Consequences of AI and ChatGPT in Educational Settings. *Educ Sci*. 2023;13(7):692.
41. Limitations. <https://openai.com/blog/chatgpt>.
42. Sallam M, Salim NA, Barakat M, Al-Tammemi AB. ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations. *Narra J*. 2023;3(1):e103.
43. Deng J, Lin Y. The Benefits and Challenges of ChatGPT: An Overview. *Front Comput Intell Syst*. 2023;2:81–3.
44. Choi W. Assessment of the capacity of ChatGPT as a self-learning tool in medical pharmacology: a study using MCQs. *BMC Med Educ*. 2023;23(1):864.
45. Medina-Romero MÁ, Jinchuña Huallpa J, Flores-Arocupita J, Panduro W, Chauca Huete L, Flores Limo F, Herrera E, Callacna R, Ariza Flores V, Quispe I, et al. Exploring the ethical considerations of using Chat GPT in university education. *Period Eng Nat Sci (PEN)*. 2023;11:105–15.
46. Lee H. The rise of ChatGPT: Exploring its potential in medical education. *Anat Sci Educ*. 2024;17(5):926–31.
47. Steare T, Gutiérrez Muñoz C, Sullivan A, Lewis G. The association between academic pressure and adolescent mental health problems: A systematic review. *J Affect Disord*. 2023;339:302–17.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.