International Journal of Drug Research and Dental Science
Volume 7 Issue 1 (Page: 86-94), 2025 ISSN: 2582-0826

The Role of ChatGPT in Dental Examination A Study on Reliability and Efficiency in Automated Essay Scoring

Dr. Himaja. K[1*], Dr. K. V. R. Pratap[2], Dr. T. Madhavi Padma[3], Dr. Siva Kalyan[4], Dr. Surbhit Singh[5], Dr. V. Srujan Kumar[6]

[1]*Student, Department of Public Health Dentistry, Mamata Dental College, Khammam, India.*
[2]*Professor and HOD, Department of Public Health Dentistry, Mamata Dental College, Khammam, India.*
[3]*Professor, Department of Public Health Dentistry, Mamata Dental College, Khammam, India.*
[4]*Reader, Department of Public Health Dentistry, Mamata Dental College, Khammam, India.*
[5,6]*Senior Lecturer, Department of Public Health Dentistry, Mamata Dental College, Khammam, India.*

Original Article

[*]**Corresponding Author**: Dr. Himaja, Department of Public Health Dentistry, Mamata Dental College, Khammam, India.
E-mail: himajakakuluri@gmail.com
Crossref doi: https://doi.org/10.36437/ijdrd.2025.7.1.F

ABSTRACT

The integration of artificial intelligence (AI) in dental education and assessment has gained significant attention in recent years. This study evaluates the role of ChatGPT in dental examinations, specifically focusing on its reliability and efficiency in automated essay scoring. The research aims to assess how effectively ChatGPT can evaluate dental students' written responses, considering factors such as accuracy, consistency, and grading bias.

A dataset of subjective dental examination answers was analyzed using ChatGPT's natural language processing (NLP) capabilities. The AI-generated scores were compared with manual grading by dental educators, using metrics such as correlation with expert scores, intra-rater reliability, and time efficiency. Results indicate that ChatGPT demonstrates high consistency and efficiency in grading, significantly reducing the time required for evaluation. However, challenges such as contextual misinterpretation, grading fairness, and domain-specific limitations were observed.

This study concludes that ChatGPT has promising potential in automated essay scoring for dental examinations, offering a scalable and time-saving solution. However, human oversight remains essential to ensure clinical relevance and fairness in assessment. Future research should focus on refining AI models to better understand dental-specific terminologies and reasoning for improved accuracy.

Aim

1. Assess the accuracy of ChatGPT's grading compared to manual scoring by dental educators.
2. Analyze the consistency of AI-generated scores across different responses.
3. Evaluate time efficiency, determining whether ChatGPT can reduce the time required for essay evaluation.
4. Identify limitations and challenges, such as contextual misinterpretation or bias in grading.

International Journal of Drug Research and Dental Science
Volume 7 Issue 1 (Page: 86-94), 2025 ISSN: 2582-0826

Objective
1. To analyze the accuracy of ChatGPT's automated essay scoring in dental examinations by comparing AI-generated scores with those given by expert dental educators.
2. To evaluate the reliability of ChatGPT in maintaining consistency across multiple essay responses.
3. To measure the efficiency of ChatGPT in terms of time taken for evaluation compared to manual grading.
Method: A cross sectional survey was conducted among 204 dental students comprising 57 males and 147 females. The survey included 14 questions. The responses were analyzed based on gender and year of study using chi square gets to identify statistically significant differences.

Keywords: Artificial Intelligence, ChatGPT, Dental-specific Terminologies, Intra-rater.

Introduction

The rapid advancement of artificial intelligence (AI) has transformed various sectors, including education and healthcare. In dental education, assessments play a crucial role in evaluating students' understanding of theoretical concepts and clinical decision-making. Traditional grading of essay-based answers is often time-consuming, subjective, and prone to variability among evaluators. The integration of AI-based tools, such as ChatGPT, in automated essay scoring, presents an opportunity to enhance the efficiency and consistency of assessment methods.

ChatGPT, a natural language processing (NLP) model, has shown promise in understanding complex textual inputs and generating meaningful responses. Its potential application in dental examinations raises questions about its reliability, accuracy, and fairness in scoring subjective responses. While AI can assist in reducing the workload of educators, concerns regarding grading bias, contextual understanding, and domain-specific knowledge remain significant.

This study aims to evaluate the reliability and efficiency of ChatGPT in automated essay scoring for dental examinations. By comparing AI-generated scores with those given by human examiners, this research seeks to determine whether ChatGPT can be effectively integrated into dental education assessment systems. The study will also explore the challenges and limitations associated with AI-based grading and provide insights into its future potential in improving educational assessments in dentistry.

Methodology

A) Study design and area: A cross-sectional study was carried out at the tertiary care teaching hospital khammam.
B) Study Population: The health care students including those of IV years who responded to the offline paper print questionnaire survey.
C) Study Instrument: A self-administered questionnaire was designed based on knowledge attitude and awareness on the advanced technology had a total 14 questions. Each participant has to fill in their demographic data like Name, age, and year of study. Participants had to select one option from the answers provided against questions the questions were based on knowledge attitude and awareness among dental students.
D) Pilot study: A pilot study was conducted on a group of students to assess the validity and reliability of the study.
E) Sampling method: The sampling method used is a convenience method.
F) Inclusion criteria: The students who were interested in the study and who are willing to participate.
G) Exclusion criteria: students who are not willing to participate are excluded.
H) Organizing the study: The study was designed in a paper-based version of the self-administered

questionnaire of 14 questions focusing on knowledge, and awareness.

Includes the sections of demographic data: Name, Age, Sex, and Year of study demographic information and asked to answer all questions by selecting one option from the provided answers.

I) Statistical analysis: Data from the filled questionnaire was conducted in a tabular form in an Excel worksheet and evaluated for analysis. The analysis was performed by SSPS version 29.

Results

A total of 204 students took part in this with female and male. The age of participants ranged from 18 to 38 years. In this study females for more knowledge regarding the role of chatGPT in dental examination than males. Final years have more knowledge followed by III year students.

AGE					
	N	Minimum	Maximum	Mean	Std. Deviation
AGE	204	18	38	22.33	2.212
Valid N (listwise)	204				

Gender		Frequency	Percent
Valid	MALE	57	27.9
	FEMALE	147	72.1
	Total	204	100.0

Year of Study		Frequency	Percent
Valid			
	I BDS	41	20.0
	II BDS	28	13.7
	III BDS	38	18.6
	IV BDS	97	47.5
	Total	204	100.0

Distribution and comparison of responses based on gender

Item	Response	Males		Females		Chi-Square value	P value
		n	%	n	%		

Q1	1	8	34.8	15	65.2	5.859	0.119
	2	45	27.1	121	72.8		
	3	4	26.7	11	73.3		
Q2	1	9	32.1	19	67.9	1.647	0.800
	2	35	26.3	98	73.6		
	3	13	30.2	30	69.8		
Q3	1	29	27.9	75	72.1	0.81	0.994
	2	7	25.9	20	74.1		
	3	21	28.7	52	71.2		
Q4	1	13	37.1	22	62.9	2.048	0.562
	2	34	26.2	101	73.8		
	3	10	29.4	24	70.6		
Q5	1	15	19.7	61	80.3	10.500	**0.015***
	2	9	60	6	40		
	3	33	30.1	80	69.9		
Q6	1	44	18.1	127	81.9	7.219	0.065
	2	10	50	10	50		
	3	3	23.1	10	76.9		
Q7	1	39	28.9	127	72.1	11.799	0.08
	2	12	52.2	11	47.8		
	3	6	40	9	60		
Q8	1	31	22.1	124	77.9	7.598	**0.054***
	2	9	45	11	55		
	3	7	36.8	12	63.2		
Q9	1	45	25.2	133	74.7	7.974	**0.047***
	2	8	47.1	9	52.9		

	3	4	44.4	5	55.6		
Q10	1	10	43.5	13	56.5	11.349	**0.010***
	2	44	28.2	123	71.8		
	3	3	21.4	11	78.6		
Q11	1	10	38.5	23	61.5	1.832	0.608
	2	5	21	32	79		
	3	42	35.8	92	64.2		
Q12	1	11	40.7	16	59.3	3.930	0.269
	2	39	25.5	116	73.5		
	3	7	31.8	15	68.2		
Q13	1	7	30.4	16	69.6	1.289	0.732
	2	8	29.6	19	70.4		
	3	17	32.7	35	67.3		
	4	15	24.5	77	75.5		
Q14	1	37	27.2	99	72.7	3.044	0.385
	2	11	40.7	17	59.3		
	3	9	22.5	31	77.5		

P≤0.05 is statistically significant

Distribution and comparison of responses based on year of the study

Item	Response	I BDS		II BDS		III BDS		IV BDS		Chi-Value	P-Value
		n	%	n	%	n	%	n	%		
Q1	1	5	21.7	0	0	2	8.69	16	69.5	15.352	0.223
	2	34	20.4	28	16.8	32	19.2	72	43.3		
	3	2	13.3	0	0	4	26.7	9	60		
Q2	1	23	15.8	24	16.5	25	17.2	73	50.3	28.430	**0.028***
	2	2	15.4	0	0	6	46.2	5	38.5		

	3	16	37.2	4	9.3	7	16.3	16	37.2		
Q3	1	21	15.4	23	15.4	22	13.5	76	52.9	22.944	**0.028***
	2	8	29.6	4	14.8	9	33.3	6	22.2		
	3	12	34.3	1	2.9	7	20	15	42.9		
Q4	1	12	34.3	3	8.6	4	11.4	16	45.7	17.474	0.133
	2	12	28.6	3	7.1	9	21.4	18	42.9		
	3	17	14.7	22	5.9	25	17.6	63	58.8		
Q5	1	25	31.6	11	7.9	10	13.2	35	46.1	17.352	0137
	2	4	26.7	1	6.7	3	20	7	46.7		
	3	12	2.9	16	17.2	25	20.4	55	47.3		
Q6	1	35	19	26	0	29	14.3	81	66.7	14.589	0.265
	2	5	25	0	0	7	35	8	40		
	3	1	7..7	2	15.4	2	15.4	8	61.5		
Q7	1	5	27.8	1	5.6	2	11.1	10	55.6	9.234	0.683
	2	5	21.7	1	4.3	7	30.4	10	43.5		
	3	31	13.3	26	6.7	29	13.3	77	66.7		
Q8	1	7	36.8	0	0	3	15.8	9	47.4	11.801	0.462
	2	31	20	27	5	31	30	77	45		
	3	3	15.8	1	5.3	4	21.1	11	57.9		
Q9	1	4	18.2	0	0	5	22.7	13	59.1	9.958	0.620
	2	4	23.5	1	5.9	5	29.4	7	41.2		
	3	33	33.3	27	0	28	22.2	77	44.4		
Q10	1	2	8.7	0	0	5	21.7	16	69.6	14.840	0250
	2	6	28.6	27	4.8	28	14.3	76	52.4		
	3	33	21.4	1	7.1	5	35.7	5	35.7		
Q11	1	5	19.2	0	0	5	19.2	16	61.5	19.134	0.085

	2	9	45	2	10	4	20	5	25		
	3	27	16.1	26	16.1	29	16.1	76	49.2		
Q12	1	2	7.4	0	0	7	25.9	18	66.7	21.968	**0.038***
	2	8	47.1	0	0	3	17.6	6	35.3		
	3	31	13.6	28	13.6	28	13.6	73	59.1		
Q13	1	3	13	0	0	2	8.7	18	78.3	37.961	**0.001***
	2	9	33.3	1	3.7	9	33.3	7	25.9		
	3	8	15.4	2	3.8	8	15.4	34	65.4		
	4	21	20.6	25	21.6	19	18.6	38	37.3		
Q14	1	30	37.0	24	29.6	4	4.9	23	28.3	15.604	0.210
	2	5	18.5	0	0	6	22.2	16	59.3		
	3	6	6.2	4	4.1	28	29.1	58	60.4		

Discussion

The findings of this study highlight both the potential benefits and limitations of using ChatGPT for automated essay scoring in dental examinations. The comparison between AI-generated scores and human grading suggests that ChatGPT can provide fast and consistent evaluations, reducing the time burden on educators. However, the study also identifies challenges related to accuracy, contextual understanding, and bias in scoring.

1. Reliability and Accuracy

ChatGPT demonstrated high consistency in grading similar responses, indicating strong intra-rater reliability. However, its ability to accurately assess complex dental concepts varied. While it performed well in grading basic theoretical questions, it struggled with responses that required clinical reasoning, case-based analysis, and patient-specific decision-making. This limitation suggests that while AI can assist in grading factual knowledge, it may not fully replace human evaluators for subjective assessments.

2. Efficiency and Time-Saving Potential

One of the major advantages of AI-based grading is its speed. ChatGPT significantly reduced the time required for scoring compared to manual assessment, making it a valuable tool for large-scale evaluations. This can be particularly beneficial in institutions where a high volume of answer scripts needs to be evaluated within short timeframes.

3. Challenges in Contextual Understanding

Despite its linguistic capabilities, ChatGPT sometimes misinterpreted responses, particularly when students used technical dental terminology, abbreviations, or unconventional sentence structures. This indicates the need for further fine-tuning of AI models to better understand domain-specific knowledge.

4. Bias and Fairness in Scoring

AI models, including ChatGPT, can exhibit bias in grading based on how they interpret language

patterns rather than actual knowledge depth. Some responses with complex sentence structures received higher scores than simpler yet equally correct answers. Additionally, ChatGPT may lack the ability to recognize creativity or unique clinical approaches, which are often valued in subjective assessments.

5. Future Implications and Recommendations

To improve AI-driven grading in dental education, the following recommendations should be considered:

- Training AI models on dental-specific datasets to enhance understanding of medical terminology and clinical reasoning.
- Using AI as an adjunct tool rather than a replacement for human grading, ensuring that final scores are reviewed by educators.
- Implementing hybrid grading systems, where AI provides an initial score, followed by human validation for complex responses.
- Regularly updating the AI model based on feedback from dental educators to refine its scoring algorithm.

Conclusion

ChatGPT shows promising potential as a supplementary tool for grading dental examination essays, offering speed and consistency in evaluation. However, its limitations in clinical reasoning and contextual accuracy suggest that human oversight remains essential. Future advancements in AI models, specifically trained for dental education could further improve the reliability of automated scoring systems.

References

1. Floridi, L., Chiriatti, M. GPT-3: Its Nature, Scope, Limits, and Consequences. Minds & Machines 30, 681–694 (2020). https://doi.org/10.1007/s11023-020-09548-1
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. doi: https://doi.org/10.2196/48291
3. Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274
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. doi: http://dx.doi.org/10.1016/j.tbench.2023.100115
5. 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. doi: http://dx.doi.org/10.1016/j.tbench.2023.100115
6. Ramesh, D., Sanampudi, S.K. An automated essay scoring systems: a systematic literature review. Artif Intell Rev 55, 2495–2527 (2022). https://doi.org/10.1007/s10462-021-10068-2
7. Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. Research Methods in Applied

Linguistics, 2(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

8. Erturk, S., van Tilburg, W.A.P. & Igou, E.R. Off the mark: Repetitive marking undermines essay evaluations due to boredom. Motiv Emot 46, 264–275 (2022). https://doi.org/10.1007/s11031-022-09929-2
9. Khan, R. A., Jawaid, M., Khan, A. R., & Sajjad, M. (2023). ChatGPT - Reshaping medical education and clinical management. Pakistan Journal of Medical Sciences, 39(2), 605. https://doi.org/10.12669/pjms.39.2.7653

How to cite this Article: Himaja. K, K. V. R. Pratap, T. Madhavi Padma, Siva Kalyan, Dr. Surbhit Singh, V. Srujan Kumar; *The Role of ChatGPT in Dental Examination A Study on Reliability and Efficiency in Automated Essay Scoring;* Int. J. Drug Res. Dental Sci., 2025; 7(1): 86-94, **doi:** https://doi.org/10.36437/ijdrd.2025.7.1.F
Source of Support: Nil, **Conflict of Interest:** Nil.
Received: 22-1-2025 **Revised:** 04-3-2025 **Accepted:** 07-3-2025