BJOS
Brazilian Journal of Oral Sciences

Original Research

Volume 24
2025
e254954

Performance of ChatGPT in dentistry: a cross-sectional, multi-specialty and multi-centric study

Priyanshu Kumar Shrivastava[1], Arpita Rai[2], Ranjit J Injety[3], Sanjay Singh[4], Ashish Jain[5], Amit Vasant Mahuli[6], Anita Parushetti[7], Anka Sharma[2], Arvind Sivakumar[8,9], Bindiya Narang[10], Farheen Sultan[11], Gaurav Shah[12], Gokul Sridharan[13], Jeyaseelan Augustine[14], Madhu Ranjan[15], Neelam Singh[16], Nishant Mehta[17], Nishat Sultan[18], Panchali Batra[19], Sangita Singh[20], Sapna Gokul[21], Sayani Roy[20], Shabina Sachdeva[22], Sharmila Tapashetti[23], Simpy Amit Mahuli[24], Sridhar Kannan[25], Sugandha Verma[2], Tushar[15], Vijay Yadav[26], Vivek Gupta[27], Deborah Sybil[4*]

[1] Faculty of Dentistry, Jamia Millia Islamia, New Delhi, India.

[2] Department of Oral Medicine and Radiology, Dental College, Rajendra Institute of Medical Sciences, Ranchi, India.

[3] Department of Neurology, Christian Medical College & Hospital, Ludhiana, Punjab, India.

[4] Department of Oral and Maxillofacial Surgery, Faculty of Dentistry, Jamia Millia Islamia, New Delhi, India.

[5] Department of Periodontics, Dr. HSJ Institute of Dental Sciences & Hospital, Chandigarh, India.

[6] Department of Public Health Dentistry, Dental College, Rajendra Institute of Medical Sciences, Ranchi, India.

[7] Department of Oral and Maxillofacial Surgery, Malla Reddy Institute of Dental Sciences, Hyderabad, India.

[8] Reface-Dentofacial Centre of Excellence, Kilpauk, Chennai, India.

[9] Smile Train Leadership Centre, Bhagwan Mahavir Jain Hospital, Bengaluru, India.

[10] Department of Oral Pathology and Microbiology, Sinhgad Dental College and Hospital, Pune, Maharashtra, India.

[11] Department of Pediatric and Preventive Dentistry, Maulana Azad Institute of Dental Sciences, New Delhi, India.

[12] Department of Oral and Maxillofacial Surgery, Geetanjali Dental and Research Institute, Udaipur, Rajasthan, India.

[13] Department of Oral Pathology and Microbiology, Dr. G. D. Pol Foundation YMT Dental College and Hospital, Kharghar, Navi Mumbai.

[14] Department of Oral Pathology and Microbiology, Maulana Azad Institute of Dental Sciences, New Delhi, India.

[15] Department of Prosthodontics, Dental College, Rajendra Institute of Medical Sciences, Ranchi, India.

[16] Department of Conservative Dentistry and Endodontics, Faculty of Dentistry, Jamia Millia Islamia, New Delhi, India.

[17] Department of Public Health Dentistry, Oral Health Science Centre, PGIMER, Chandigarh, India.

[18] Department of Periodontics, Faculty of Dentistry, Jamia Millia Islamia, New Delhi, India.

[19] Department of Orthodontics, Faculty of Dentistry, Jamia Millia Islamia, New Delhi, India.

[20] Department of Pediatric and Preventive Dentistry, Dental College, Rajendra Institute of Medical Sciences, Ranchi, India.

[21] Department of Periodontics, Nair Hospital Dental College, Mumbai, Maharashtra, India.

[22] Department of Prosthodontics, Faculty of Dentistry, Jamia Millia Islamia, New Delhi, India.

[23] Department of Conservative Dentistry and Endodontics, SDM College of Dental Sciences, Dharwad, Karnataka, India.

[24] Dental College, Rajendra Institute of Medical Sciences, Ranchi, India.

[25] Department of Orthodontics, Sudha Rustagi College of Dental Sciences and Research, Faridabad, India.

[26] Department of Dentistry, All India Institute of Medical Sciences, Bibinagar, India.

[27] Department of Periodontics, Dental College, Rajendra Institute of Medical Sciences, Ranchi, India.

Corresponding author:
Dr. Deborah Sybil, Professor,
Department of Oral and Maxillofacial Surgery, Faculty of Dentistry, Jamia Millia Islamia, Maulana Mohammad Ali Jauhar Marg, Jamia Nagar, New Delhi, India.
Email: dsybilg@gmail.com

Editor: Dr. Altair A. Del Bel Cury

Received: November 07, 2023

Accepted: January 1, 2024

Artificial Intelligence-based language model, ChatGPT, has gained significant traction due to its communicative interface and the relevance of the responses generated. This tool exhibits tremendous potential to be utilized in dentistry for dental education, and possibly as a clinical decision support system. Hence, it is imperative to evaluate the accuracy of the model in relation to the responses generated for dental-related queries and outline the limitations for its use in clinical practice. This study aims to evaluate the performance of ChatGPT-generated responses to questions from multiple dental specialties for their accuracy, completeness, and relevance. **Methods:** This multi-centric study involved 27 subject experts from nine dental specialties of various institutions and 2 heads of institutions. A total of 243 questions were formulated and the answers generated by ChatGPT (version: 3.5) were rated in terms of accuracy (6-point Likert), completeness (4-point Likert), and relevance (5-point Likert). **Results:** The mean accuracy of the ChatGPT-generated answers was 4.61 (SD 1.575), with a median of 5.33. For completeness, the mean score was 2.01 (SD 0.793), and the median was 2.33. Regarding relevance, a mean of 3.13 (SD 1.590) and a median of 3.67 were obtained. The highest ratings were observed for answers related to Oral Medicine and Radiology, as well as for open-ended questions, and questions labelled as easy in terms of difficulty. **Conclusion:** The promising results observed in the study promote the application of ChatGPT for retrieving dental information. However, it is crucial to exercise caution and seek advice from a qualified healthcare for dental health-related queries. Further large-scale testing of the model is necessary before incorporating it into dental clinical practice.

Keywords: Artificial intelligence. Decision support systems, clinical. Education, dental. Dentistry. Large language models. Natural language processing.

Introduction

Utilization of Artificial Intelligence (AI)-powered tools in the delivery of healthcare has seen a recent surge in popularity. From diagnosis of dental conditions to disease prediction, prognostic evaluation, and clinical decision-making, AI has a vast array of applications in dentistry[1]. One of the AI-powered tools that has had a tremendous impact on the general population and the healthcare industry is the Chat Generative pre-training transformer or ChatGPT. Recently launched in November 2022, ChatGPT is a Large Language Model (LLM) developed by OpenAI (OpenAI, L.L.C., San Francisco, CA, USA) which has the potential to revolutionize the field of dentistry[2].

ChatGPT, a major breakthrough in the field of Natural Language Processing (NLP) mimics human language processing abilities using deep learning and neural networks to generate human-like text. Owing to its training on massive datasets involving textbooks, research papers, and webpages, it could process an array of information and provide concise and appropriate responses to human prompts. ChatGPT version 3.5 is easily accessible and free for the public to use, further contributing to its success. Recently ChatGPT has been utilized in healthcare for generating literature reviews and summaries, analysis of clinical datasets, scientific and academic writing, documentation and streamlining clinical workflow[3].

The dental applications of ChatGPT remain to be studied comprehensively given its recent launch. However, few studies have briefly discussed the role of the language model in dental telemedicine, treatment planning, text mining from electronic health records, dental education, and patient education[4,5]. In dental education, ChatGPT proposes to facilitate the teaching of complex dental procedures by generating step-by-step instructions, creating interactive educational content through quizzes and flashcards, generating case scenarios to improve basic knowledge of students and assisting dental students in proficient communication with patients through the generation of simulated patient interactions[3].

Although ChatGPT has shown tremendous potential in other areas, the accuracy and relevance of the responses to dentistry-related queries remain in question. Given the escalated utilization and integration of AI in dental education[6], it is pertinent that the information presented by the AI-driven chatbot maintains a high standard of quality with a comprehensive assessment of all associated limitations. Therefore, this study aims to evaluate the performance of ChatGPT-generated responses to questions from the nine dental specialties in relation to their accuracy, completeness, and relevance, recognizing its potential as an educational resource for students and professionals. By employing a structured assessment framework and involving a diverse panel of subject experts, this research contributes substantively to the ongoing discourse on the integration of AI in dental pedagogy, ensuring that the quality of information disseminated aligns with the rigorous standards of dental education.

Material and Methods

A multi-centric observational study was carried out involving major dental specialties. The study was exempted from review by the Institutional Ethics Committee since no

data was collected from patients or external sources. A total of 27 subject experts from different universities across the country were approached by the senior investigators for collecting questions from 9 dental specialties and for rating the answers generated by ChatGPT (version: 3.5). 2 heads of institutions were also invited to re-evaluate a few randomly selected AI-generated responses.

Questions Dataset:

Three subject experts who were distinguished academicians working as faculty members in various dental institutes, each from the nine dental specialties of Conservative dentistry and Endodontics, Oral and Maxillofacial Surgery (OMFS), Oral Medicine and Radiology (OMDR), Oral Pathology, Orthodontics, Paediatric and Preventive dentistry, Periodontics, Prosthodontics, and Public Health Dentistry (PHD), were invited to participate in the study. Each subject expert formulated a set of 9 random specialty-specific clinical as well as non-clinical questions relevant to the dental curriculum in the English language. A specific set of guidelines were given to the experts that included formulating questions with unambiguous answers and involving core concepts of various dental specialties. The set entailed 3 binary questions with yes/no; right/wrong; and true/false as answers, 3 multiple-choice questions or one-word answer questions, and 3 open-ended questions with descriptive answers. The rationale behind choosing these types of questions was the incorporation of a variety of questions usually used within the dental curriculum to test the students. Hence, 27 questions were invited from each specialty, totalling 243 random questions from the nine dental specialties. The subject experts assigned the difficulty level for each question as Easy, Medium, and Hard. The rationale behind this label was to compare the responses of ChatGPT across each category and whether the model was accurate in answering difficult dental-related queries. Each question and the difficulty level for that question was internally validated through discussion between the three experts and any duplicate question was substituted.

Generation of answers:

The set of questions was entered on the web version of ChatGPT 3.5 (Mar14 version) at http://chat.openai.com by a single investigator to ensure consistency. Separate chats were created for each specialty and the following prompt was entered to ensure specificity and appropriateness of the answers generated: "Please answer the following questions of (name of Dental Specialty). Be specific and incorporate any standard guidelines applicable". All the answers generated by the chatbot were tabulated specialty-wise and sent for evaluation. The list of questions and the responses generated by ChatGPT have been uploaded as a supplementary file. [Supplementary file 1]

Performance evaluation:

All 27 responses generated for a specialty were rated by the three subject experts and a mean score was obtained. Additionally, 3 open-ended questions with descriptive answers were randomly selected from each specialty and sent to 2 heads of institutions for rating to determine inter-observer agreement. The answers generated

by ChatGPT were rated in terms of Accuracy (6-point Likert Scale with the following description 1: Completely incorrect, 2: More incorrect than correct, 3:Approximately equal correct and incorrect, 4: More correct than incorrect, 5: Nearly all correct, and 6: correct; Completeness (3-point Likert Scale with the description as 1: Incomplete (Addresses some aspects of the question, but significant parts are missing), 2: Adequate (Addresses all aspects of the question and provides the minimum amount of information required to be considered complete), 3: Comprehensive (Addresses all aspects of the question and provides additional information or context beyond what was expected); and Relevance in relation to the length of answer generated was tested on a 5-point Likert Scale ranging from 1: Very irrelevant (Lengthy answer with no relevance to the question asked), 2: Irrelevant, 3: Quite relevant, 4: Relevant, to 5: Very relevant (Concise and to-the-point answer). Completely incorrect (accuracy=1) answers were not graded for completeness, and incomplete answers were not graded for relevance (Graded as 0 on the scale). The scales used in the present study were adapted from a previously conducted study of a similar design[7].

Statistical analysis:

The average score for each subject was descriptively (mean, standard deviation) listed and then compared based on difficulty levels (easy, medium and hard), category (accuracy, completeness, relevance) and type of questions (multiple-choice questions, binary and descriptive) using t-test. Mann-Whitney U or Kruskal Wallis testing was done when comparing the groups. Responses for selected questions graded by faculty and the deans were compared using the paired sample t-test. The comparisons were considered significant if $p<0.05$ at 95% CI. All statistical analysis was performed using SPSS software version 21.0 (SPSS Inc).

Results

Overall performance

The mean and median for accuracy of AI-generated answers (n=243) were found to be 4.61 (SD 1.575) and 5.33 respectively. The mean value for completeness was 2.01 (SD 0.793) and a median of 2.33 was noted. A mean score of3.13 (SD 1.590) and a median of 3.67 was obtained for the relevance of the AI-generated answers. 11 answers were rated as completely inaccurate by all three experts (mean accuracy score=1). 49 answers were rated completely inaccurate by at least one subject expert. 68 AI-generated answers were rated as completely accurate (mean accuracy score=6) by all three subject experts.

Specialty-wise performance

AI-generated answers were found to be the most accurate, complete, and relevant for OMDR [mean accuracy 5.63 (SD 0.854), completeness 2.36 (SD 0.497), relevance of 4.04 (SD 1.149)] whereas the least accurate results were observed for orthodontics-related questions [mean accuracy 3.36 (SD 1.821), completeness 1.28 (SD 0.918), relevance 1.68 (SD 1.743)]. [Figure 1] The performance rating for all dental specialties is listed in Table 1.

Table 1. Dental specialty-wise performance of ChatGPT

Specialty	Rating	n	Minimum	Maximum	Mean	Std. Deviation
Oral Pathology	Accuracy	27	1	6	4.41	1.701
	Completeness		0	3	1.94	.801
	Relevance		0	5	3.10	1.622
Oral and Maxillofacial Surgery	Accuracy	27	2	6	5.19	1.099
	Completeness		1	3	2.20	.483
	Relevance		1	5	3.32	1.256
Periodontics	Accuracy	27	1	6	4.67	1.720
	Completeness		0	3	2.10	.895
	Relevance		0	5	3.15	1.567
Pedodontics	Accuracy	27	2	6	4.58	.932
	Completeness		1	3	2.11	.462
	Relevance		0	5	3.43	1.065
Oral Medicine and Radiology	Accuracy	27	2	6	5.63	.854
	Completeness		1	3	2.36	.497
	Relevance		0	5	4.04	1.149
Conservative Dentistry and Endodontics	Accuracy	27	1	6	4.21	1.558
	Completeness		0	3	1.80	.813
	Relevance		0	5	2.46	1.491
Prosthodontics	Accuracy	27	1	6	4.47	1.693
	Completeness		0	3	2.02	.947
	Relevance		0	5	3.16	1.794
Orthodontics	Accuracy	27	1	6	3.36	1.821
	Completeness		0	3	1.28	.918
	Relevance		0	5	1.68	1.743
Public Health Dentistry	Accuracy	27	2	6	5.02	1.536
	Completeness		1	3	2.31	.666
	Relevance		1	5	3.80	1.344

Difficulty-wise performance

35.4% (n=86) of the questions were easy, 40.7% (n=99) were of medium level, and 23.9% (n=58) were hard. The distribution of difficulty levels with respect to different dental specialties has been depicted in the figure. [Figure 2] An example of an easy question from Orthodontics was "What are clear aligners?", of a medium was "What is wagon wheel effect in straight wire orthodontic appliance?" and hard was "What is Eastman correction in orthodontic cephalometry?". The mean score for easy questions was found to be the highest in terms of accuracy [mean 5.03 (SD 1.475)], completeness [mean 2.15 (SD 0.760)], and relevance [mean 3.48 (SD 1.437)]. Medium and hard questions were almost of comparable accuracy, completeness, and relevance.

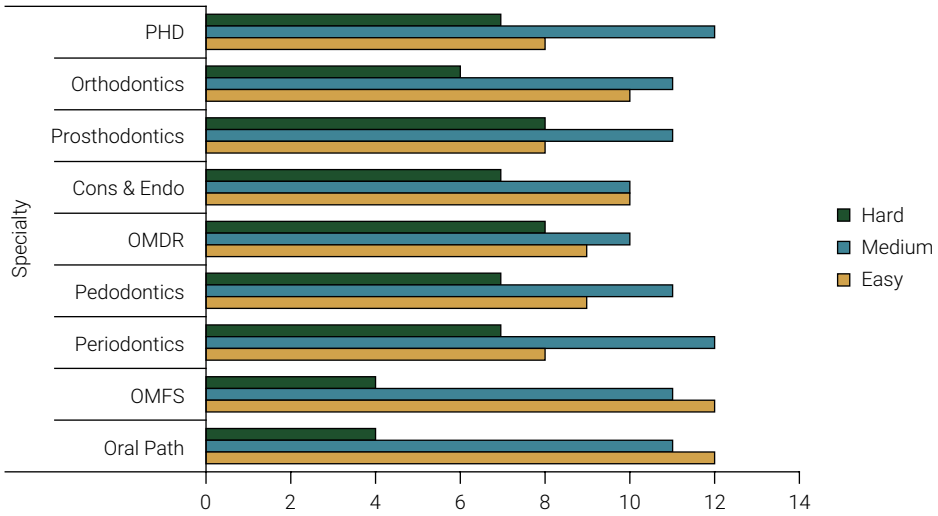

Figure 1. Distribution of the difficulty level of questions across different dental specialties.

Performance as per the type of questions:

ChatGPT-generated answers for open-ended questions were found most accurate with a mean accuracy of 4.75 (SD 1.200), followed by binary questions [mean 4.71 (SD 1.680)] and multiple-choice questions [mean 4.39 (SD 1.778)]. Completeness and Relevance of the answers also followed a similar trend with the highest rating received for descriptive answers [mean completeness 2.05 (SD 0.628); mean relevance 3.18 (SD 1.355)]. The most accurate answers to binary questions were seen for OMFS whereas OMDR received the most accurate answers to multiple-choice and open-ended questions. The evaluation of AI-generated responses across all dental specialties in relation to the type of questions entered has been depicted in the table. [Table 2]

Table 2. Performance as per the type of questions across various specialties.

	Type of question	Accuracy	Completeness	Relevance
Oral Pathology	Binary	4.19 (2.174)	1.96 (1.02)	3.26 (1.935)
	MCQ	**4.7 (1.947)**	**2.15 (0.852)**	**3.48 (1.651)**
	Descriptive	4.33 (0.85)	1.7 (0.455)	2.56 (1.247)
Oral & Maxillofacial Surgery	Binary	**5.85 (0.338)**	2.19 (0.377)	**3.59 (0.909)**
	MCQ	5.26 (1.234)	**2.33 (0.527)**	3.48 (1.345)
	Descriptive	4.44 (1.080)	2.07 (0.547)	2.89 (1.472)
Periodontics	Binary	3.78 (2.236)	1.70 (1.160)	2.11 (1.590)
	MCQ	4.63 (1.628)	2.19 (0.973)	3.26 (1.801)
	Descriptive	**5.59 (0.324)**	**2.41 (0.147)**	**4.07 (0.147)**

Continue

Continuation

Pedodontics	Binary	4.70 (1.369)	**2.15 (0.580)**	3.48 (1.365)
	MCQ	4.41 (0.830)	2.04 (0.455)	**3.52 (1.015)**
	Descriptive	4.63 (0.455)	2.15 (0.377)	3.30 (0.873)
Oral Medicine & Radiology	Binary	5.52 (0.852)	2.22 (0.236)	4.19 (0.747)
	MCQ	5.59 (1.222)	2.30 (0.772)	3.67 (1.818)
	Descriptive	**5.78 (0.333)**	**2.56 (0.289)**	**4.26 (0.434)**
Conservative Dentistry & Endodontics	Binary	4.37 (1.611)	1.96 (0.754)	2.22 (1.364)
	MCQ	3.30 (1.837)	1.41 (1.103)	2.00 (1.922)
	Descriptive	**4.96 (0.564)**	**2.04 (0.309)**	**3.15 (0.915)**
Prosthodontics	Binary	**4.89 (1.599)**	**2.26 (0.954)**	**3.81 (1.564)**
	MCQ	4.85 (1.425)	2.26 (0.778)	3.70 (1.359)
	Descriptive	3.67 (1.908)	1.56 (1.014)	1.96 (1.933)
Orthodontics	Binary	3.26 (1.786)	1.15 (0.603)	1.00 (1.323)
	MCQ	3.00 (2.115)	1.11 (1.179)	1.70 (2.065)
	Descriptive	**3.81 (1.651)**	**1.59 (0.909)**	**2.33 (1.691)**
Public Health Dentistry	Binary	**5.81 (0.242)**	**2.63 (0.423)**	**4.44 (0.745)**
	MCQ	3.74 (2.165)	1.93 (0.969)	2.89 (1.929)
	Descriptive	5.52 (0.294)	2.37 (0.200)	4.07 (0.324)

Evaluation of responses by Heads of Institutions:

Two heads of institutions evaluated a set of randomly chosen 27 open-ended questions, 3 from each dental specialty. The mean accuracy was found to be 5.13 (SD 0.804) and a median of 5.50 was noted. Similarly, a mean of 2.33 (SD 0.519) and a median of 2.50 were obtained for completeness, and a mean of 3.89 (SD 1.204) and a median of 4.00 were calculated for the relevance of the AI-generated answers. The dental specialty-wise rating revealed the most accurate answers for Periodontics [mean 5.83 (SD 0.289)] and the least accurate answers were observed for Prosthodontics [mean 4.17 (SD 1.155)]. [Table 3]

Table 3. Dental specialty-wise performance rating by heads of institutions.

Specialty	Rating	n	Minimum	Maximum	Mean	Std. Deviation
Oral Pathology	Accuracy	3	5	6	5.17	.577
	Completeness		3	3	2.67	.289
	Relevance		4	4	3.83	.289
Oral and Maxillofacial Surgery	Accuracy	3	5	6	5.00	.500
	Completeness		2	3	2.17	.289
	Relevance		4	5	4.17	.289

Continue

Continuation

Periodontics	Accuracy	3	6	6	5.83	.289
	Completeness		3	3	2.50	.000
	Relevance		5	5	4.67	.289
Pedodontics	Accuracy	3	5	6	5.33	.289
	Completeness		3	3	2.83	.289
	Relevance		4	5	4.50	.866
Oral Medicine and Radiology	Accuracy	3	5	6	5.50	.500
	Completeness		2	3	2.67	.577
	Relevance		4	5	4.50	.866
Conservative Dentistry and Endodontics	Accuracy	3	5	6	5.50	.500
	Completeness		2	3	2.33	.577
	Relevance		4	5	4.00	.866
Prosthodontics	Accuracy	3	4	6	4.17	1.155
	Completeness		2	2	1.67	.289
	Relevance		1	5	2.50	1.803
Orthodontics	Accuracy	3	3	6	4.33	1.258
	Completeness		1	3	2.00	.866
	Relevance		0	4	2.67	2.309
Public Health Dentistry	Accuracy	3	5	6	5.33	.764
	Completeness		2	3	2.17	.289
	Relevance		4	5	4.17	.577

Agreement between Subject experts and Head of Institutions:

The rating of the 27 randomly selected questions by the heads of the institution was compared with the rating given by the faculty of the respective specialty. The average accuracy rating given by the heads of the institution was 5.13 (SD 0.84) and by the faculty was 4.7 (SD 1.463) which was not statistically significant (p= 0.141). The difference in the rating meant that the faculty and the heads of the institution did not equally agree with the accuracy of the answers provided by ChatGPT. However, both the heads of the institutions and the subject experts rated similarly for the completeness (p= 0.036) and relevance of the answers (p= 0.022).

Discussion

ChatGPT has garnered significant attention owing to its user-friendly platform and interactive interface. Within a relatively short timeframe, it has gained substantial traction and is being widely employed in the healthcare sector for diverse applications. However, the ease of access to the model exposes it to potential misuse and any misinformation provided especially in the context of healthcare can have detrimental

consequences. Moreover, as the platform is open to the general public, patients seeking dental advice might utilize the chatbot to address their dental-related concerns. Consequently, it becomes imperative to subject the model to comprehensive testing, thus identifying the potential limitations and establishing necessary precautions for its application in addressing dental queries.

In terms of accuracy, a relatively high mean score of 4.61 (SD 1.575), and a median of 5.33were noted indicating adequate correctness of the information provided in relation to the dental queries. Completeness, as assessed on a 3-point Likert yielded a mean score of 2.01 (SD 0.793) and a median of 2.33. Hence, while ChatGPT responses were informative, there is a scope for improvement in incorporating comprehensive details. Relevance received a mean score of 3.13 (SD 1.590) and a median of 3.67. Despite the utilization of prompts in each chat session to guide ChatGPT into producing specific results with standard guidelines, a general trend that was observed across the various specialties was the wordiness of the answers generated with no relevance to the questions asked.

Specialty-wise, the model scored highest in OMDR [mean accuracy 5.63 (SD 0.854), completeness 2.36 (SD 0.497), relevance of 4.04 (SD 1.149)]. OMDR constitutes clinical and radiographic diagnosis as well as the medicinal management of all common oral and maxillofacial conditions. Hence, this finding could be extrapolated to the application of ChatGPT as a clinical decision-support system. However, further testing with multiple clinical case scenarios needs to be performed to support its translatability into clinics. The language model scored least in Orthodontics [mean accuracy 3.36 (SD 1.821), completeness 1.28 (SD 0.918), relevance 1.68 (SD 1.743)]. The poor performance could be attributable to the complexity and specificity of orthodontics-related inquiries and the technical nature of the orthodontic concepts requiring precise judgment. Moreover, there is a probability that the dataset used for training ChatGPT might have a limited representation of orthodontics-related content as opposed to other dental specialties. Hence, robust training of the model is required in this sector.

In relation to the difficulty level, ChatGPT performed better on easier questions [mean accuracy 5.03 (SD 1.475)] which is suggestive of the limited effectiveness of the model in handling medium or difficult dental-related queries. Although the difference was not very substantial, the model dealt with hard or medium-level questions efficiently. Regarding the types of questions entered, ChatGPT was most accurate in answering open-ended questions [mean 4.75 (SD 1.200)]. Furthermore, superior results were noted for descriptive answers rated by heads of institutions with a mean accuracy of above 5 for each specialty except for Prosthodontics and Orthodontics. Hence, it can be inferred that ChatGPT demonstrates higher proficiency in generating subjective responses as opposed to objective ones. However, it is crucial to acknowledge the potential overestimation of the accuracy scores in open-ended questions that is attributable to several possible explanations when compared with multiple-choice or binary questions offering limited alternatives.

During prompting, few observations were made, including occasional loading errors that necessitated regenerating the response or refreshing the page. Notably, a warning message was displayed at the bottom of the webpage indicating the possibility

of ChatGPT producing inaccurate information about people, places, or facts. This reflects the model's limitations in dealing with eponyms and proper nouns. For instance, in the field of OMFS, when asked about the name of the procedure where the zygomatic arch is fractured to prevent Temporomandibular Joint dislocation, ChatGPT generated the response "zygomatic arch osteotomy", which is indeed the actual technique used. However, the correct answer should have been Dautrey's procedure. Furthermore, the model was sensitive to the prompts entered and minor rephrasing produced entirely different responses.

Additionally, there were instances where ChatGPT exhibited a "hallucination effect" generating completely inaccurate and irrelevant responses, as noted previously in other studies[8]. This was observed in questions related to Conservative Dentistry and Endodontics such as "a condition exhibiting radiographic picture of Bull's eye" and "J-shaped radiolucency", where Central Giant Cell Granuloma and Periapical Cemento-osseous Dysplasia were produced as answers. Hence, a notable limitation was the lack of precautionary warning accompanying the model's responses, and the authoritativeness with which the model presents the answer which could be potentially misleading. Therefore, it is imperative that the model exercises caution when uncertain and avoids attempting to answer rather than providing misinformation.

Incorporating questions of varying difficulty levels in the methodology ensures a robust evaluation, assessing ChatGPT's performance across a spectrum of complexity, thereby providing insights into its proficiency at different skill levels. The inclusion of questions from diverse dental specialties is pivotal as it tests the model's ability to cater to a wide range of topics, demonstrating its versatility and applicability in addressing the diverse needs of dental education. Three random open-ended questions from each of the nine dental specialties were sent for rating to two heads of institutions. This aimed to assess the inter-observer variability in rating descriptive answers. Since there is a lower probability of such variation in the rating of one-word or multiple-choice questions, only descriptive answers were sent for rating. The adoption of a multi-centric study design involving subject experts from various dental specialties and different centers enhances the study's external validity and generalizability. This approach ensures a broader representation of perspectives and expertise, reducing potential biases associated with a single-center study. The inclusion of subject experts from multiple centers brings forth a diversity of teaching methodologies, curricular nuances, and regional variations in dental education. Consequently, the findings become more robust and applicable across a wider spectrum of dental education contexts, contributing to the overall reliability and relevance of the study.

ChatGPT 3.5 is a relatively recent language model and only a few studies have been conducted to assess its feasibility. In a similar study as the present one, the accuracy of the answers generated by ChatGPT in response to 284 medical questions was tested. A mean accuracy of 4.4 and a median of 5 were determined. The study also noted an improvement in the scores when reassessed after 17 days, suggesting the continuously updating nature of the language model[7]. Further, a study evaluated the performance of ChatGPT on the United States Medical Licensing Exams (USMLE). It was observed that the model performed at or

near the passing threshold with a high level of concordance in its explanations[9]. These findings, along with the findings observed in the present study favours the use of ChatGPTin medical and dental education. The advent of newer language models that deciphers image-based input to construct relevant responses further presents an innovative application in clinical-decision making. However, this would require rigorous audit and supervision.

The launch of ChatGPT has elicited a varied response within the scientific community regarding its potential applications in healthcare, accompanied by concerns regarding the misuse of technology. Certain ethical and legal concerns may arise in this context, including plagiarism, authorship concerns, infringement of copyright laws, and accountability of the generated content[10]. There are also certain limitations to ChatGPT that need to be acknowledged. As a text-based AI language model, ChatGPT 3.5 is unable to process visual information, rendering it unsuitable for image-related tasks. This is however addressed in the updated version of ChatGPT 4.0 which incorporates image-based prompts. Secondly, the model's generated output highly depends on the input phrase and minor tweaks, or rephrasing could change the response substantially. As stated earlier, the hallucination effect is a concern whereby plausible-sounding text might be generated with no relevance to the original question. Moreover, the reference source used for generating information is conspicuous, which raises concerns regarding the validity of responses. ChatGPT has a knowledge cutoff dated September 2021. Therefore, the model might not produce accurate information regarding the latest technological updates. Additionally, as with any AI-based technology, significant concerns surrounding user privacy and cybersecurity mandate the need for adequate safeguarding mechanisms in place to protect users' data and prevent potential misuse[2].

Limitations of the study

Although the study was carried out on a large multi-centric scale, a moderate dataset of questions specific to the specialties was utilized for testing. We propose conducting comprehensive specialty-wise testing with a substantial dataset specifically focused on dental inquiries of that specialty. Secondly, the questions provided were self-validated by subject experts and not subjected to an external audit. Thirdly, the rating of the responses was subjective and could have inter-observer variations. To mitigate potential biases, all AI-generated responses of a specialty were rated by three experts, and a mean score was obtained. Further, two heads of institutions re-assessed the descriptive questions from each specialty. The medium used for questions and answers was English, which limits the obtained accuracy data for English speakers only.

Conclusion

ChatGPT demonstrated promising performance in terms of accuracy, completeness, and relevance of the responses generated to dental-related queries, acknowledging its utility in dental education. Furthermore, satisfactory category-wise results obtained promote its application towards answering specific questions from various dental specialties. A scope of improvement was identified in Orthodontics,

where the model exhibited the least accuracy. While the study promotes the use of the language model as a valuable resource for retrieving information regarding dental queries, caution should be exercised due to its potential limitations. Furthermore, future improvements should include rigorous training of the model to encompass the latest technological updates in specialized areas of dentistry, refinement in handling eponyms, and the inclusion of cautionary warnings to mitigate potential errors and prevent the dissemination of misinformation.

Sources of Funding

None.

Conflicts of Interest

None.

Author Contributions

Priyanshu Kumar Shrivastava, Arpita Rai, Ranjit J Injety, Sanjay Singh, Deborah Sybil: Contributed to conception and design, data acquisition, analysis, and interpretation, and drafted the initial versions of the manuscript. **Ashish Jain, Amit Vasant Mahuli, Anita Parushetti, Anka Sharma, Arvind Sivakumar, Bindiya Narang, Farheen Sultan, Gaurav Shah, Gokul Sridharan, Jeyaseelan Augustine, Madhu Ranjan, Neelam Singh, Nishant Mehta, Nishat Sultan, Panchali Batra, Sangita Singh, Sapna Gokul, Sayani Roy, Shabina Sachdeva, Sharmila Tapashetti, Simpy Amit Mahuli, Sridhar Kannan, Sugandha Verma, Tushar, Vijay Yadav, Vivek Gupta:** Contributed to data acquisition, analysis, and interpretation, and critically revised the manuscript for important intellectual content. All authors reviewed the final version of the manuscript to be published. We declare that all authors actively participated in two distinct criteria related to authorship.

References

1. Shan T, Tay FR, Gu L. Application of Artificial Intelligence in Dentistry. J Dent Res. 2021 Mar;100(3):232-244. doi: 10.1177/0022034520969115. Epub 2020 Oct 29.
2. Introducing ChatGPT. c2015 – 2023. OpenAI [cited 2023 May 15]. Available from: https://openai.com/blog/chatgpt.
3. 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 Apr;3(1):e103. doi: 10.52225/narra.v3i1.103.
4. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. J Esthet Restor Dent. 2023 Oct;35(7):1098-102. doi: 10.1111/jerd.13046.
5. Huang H, Zheng O, Wang D, Yin J, Wang Z, Ding S, et al. ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model. Int J Oral Sci. 2023 Jul 28;15(1):29. doi: 10.1038/s41368-023-00239-y.

6. Thurzo A, Strunga M, Urban R, Surovková J, Afrashtehfar KI. Impact of artificial intelligence on dental education: a review and guide for curriculum update. Educ Sci. 2023 Jan;13(2):150. doi: 10.3390/educsci13020150.

7. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, et al. Assessing the accuracy and reliability of ai-generated medical responses: an evaluation of the Chat-GPT Model. Res Sq [Preprint]. 2023 Feb 28:rs.3.rs-2566942. doi: 10.21203/rs.3.rs-2566942/v1.

8. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid Concerns. Healthcare (Basel). 2023 Mar;11(6):887. doi: 10.3390/healthcare11060887.

9. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023 Feb;2(2):e0000198. doi: 10.1371/journal.pdig.0000198.

10. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023 May 4;6:1169595. doi: 10.3389/frai.2023.1169595.