Journal of International
Dental Sciences
Uluslararası Diş Hekimliği Bilimleri Dergisi

Research Araştırma

Evaluation of Artificial Intelligence Chatbots In The Management of Primary Tooth Traumas: A Comparative Analysis

Süt Dişi Travmalarının Yönetiminde Yapay Zekâ Sohbet Botlarının Değerlendirilmesi: Karşılaştırmalı Bir Analiz

Mihriban GÖKCEK TARAÇ[1]
ORCID: 0000-0003-3960-8518

[1]Karabük University, Faculty of Dentistry,
Department of Pediatric Dentistry,
Karabük, Turkey

ABSTRACT

Objective: This study aimed to evaluate the reliability and consistency of four artificial intelligence (AI) chatbots—ChatGPT 3.5, Google Gemini, Bing, and Claude AI—as public sources of information on the management of primary tooth trauma.

Materials and Method: A total of 31 dichotomous questions were developed based on common issues and concerns related to dental trauma, particularly those frequently raised by parents. Each question, sequentially presented to the four AI chatbots, was repeated three times daily, with a one-hour interval between repetitions, over a five-day period, to assess the reliability and reproducibility of responses. Accuracy was determined by calculating the proportion of correct responses, with 95% confidence intervals estimated using the Wald binomial method. Reliability was assessed using Fleiss' kappa coefficient.

Results: All AI chatbots demonstrated high accuracy. Bing emerged as the most accurate model, achieving an accuracy rate of 96.34%, while Claude had the lowest accuracy at 88.17%. Consistency was classified as "almost perfect" for ChatGPT, Bing, and Gemini, whereas Claude exhibited a "substantial" level of agreement. These findings underscore the relative performance of AI models in tasks requiring high accuracy and reliability.

Conclusion: Among the four AI chatbots evaluated, Bing delivered the most accurate and consistent responses. ChatGPT 3.5 and Gemini followed closely in terms of performance, whereas Claude lagged behind, exhibiting lower accuracy and consistency. These results emphasize the importance of critically evaluating AI-based systems for their potential use in clinical applications. Continuous improvements and updates are essential to enhance their reliability and ensure their effectiveness as public information tools.

Key Words: Artificial Intelligence, ChatGPT, Dental Trauma.

ÖZ

Amaç: Bu çalışma, dört yapay zeka sohbet botunun (ChatGPT 3.5, Google Gemini, Bing ve Claude AI) süt dişi travmasının yönetimiyle ilgili kamuya açık bilgi kaynakları olarak güvenilirliğini ve tutarlılığını değerlendirmeyi amaçlamıştır.

Gereç ve Yöntemler: Ebeveynlerin dental travmalar hakkında en sık sorduğu sorular temel alınarak, "Evet" veya "Hayır" şeklinde yanıtlanabilen 31 soru hazırlanmıştır. Her soru, dört Yapay Zeka sohbet botuna sırasıyla yöneltilmiş ve yanıtların güvenilirliğini ve tekrarlanabilirliğini değerlendirmek amacıyla beş gün boyunca, günde üç kez, birer saat arayla tekrarlanmıştır. Doğruluk, doğru yanıtların oranı hesaplanarak belirlenmiş ve %95 güven aralıkları Wald binom yöntemi kullanılarak tahmin edilmiştir. Güvenilirlik, Fleiss'in kappa katsayısı ile değerlendirilmiştir.

Bulgular: Tüm Yapay Zeka sohbet botları yüksek doğruluk sergilemiştir. Bing, %96,34 doğruluk oranı ile en doğru model olarak öne çıkarken, Claude %88,17 doğruluk oranı ile en düşük performansı göstermiştir. Tutarlılık açısından ChatGPT, Bing ve Gemini "neredeyse mükemmel" düzeyde uyum gösterirken, Claude "önemli" düzeyde bir uyum sergilemiştir. Bu bulgular, yüksek doğruluk ve güvenilirlik gerektiren görevlerde Yapay Zeka modellerinin göreceli performansını vurgulamaktadır.

Sonuç: Değerlendirilen dört Yapay Zeka sohbet botu arasında Bing, en doğru ve tutarlı yanıtları sağlamıştır. ChatGPT 3.5 ve Gemini, performans açısından Bing'i yakından takip ederken, Claude daha düşük doğruluk ve tutarlılık göstererek geride kalmıştır. Bu sonuçlar, klinik uygulamalarda potansiyel kullanımları açısından Yapay Zeka tabanlı sistemlerin eleştirel bir şekilde değerlendirilmesinin önemini ortaya koymaktadır. Güvenilirliklerini artırmak ve kamuya açık bilgi araçları olarak etkinliklerini sağlamak için sürekli iyileştirmeler ve güncellemeler gereklidir.

Anahtar Kelimeler: Yapay Zeka, ChatGPT, Dental Travma.

Geliş tarihi / *Received: 14.02.2025*
Kabul tarihi / *Accepted: 11.09.2024*

İletişim Adresi/*Corresponding Adress:*
Mihriban GÖKCEK TARAÇ,
Karabük University, Faculty of Dentistry,
Department of Pediatric Dentistry,
Karabük, Turkey
E-posta/e-mail: gokcekmihriban@karabuk.edu.tr

INTRODUCTION

Technological innovations are at the forefront of recent global developments, with rapid advancements in the field of artificial intelligence (AI) playing a pivotal role, particularly in the healthcare sector. Artificial intelligence can be defined as a collection of technologies capable of mimicking various human attributes, such as performing tasks, learning, reasoning, interpreting, and data planning. These tasks are carried out through computer systems that operate autonomously (1). Chatbots based on AI have significantly transformed digital communication, enhancing the quality of human interaction. Using deep learning algorithms, these chatbots are extensively trained on large datasets, continuously improving the accuracy of responses while simulating human neural networks to enhance relevance (2). Leading platforms in this field include GPT-3.5, developed by OpenAI Inc., Google Gemini by Google LLC, Microsoft Corporation's Bing, and the Claude model by Anthropic PBC (3). In recent years, AI integration has marked a significant advancement in dentistry, providing powerful tools for the prediction, diagnosis, and development of treatment plans for dental diseases (4). The ChatGPT series is regarded as one of the most advanced natural language processing models available to the public today (5). As a prominent example of an AI application, ChatGPT is an advanced chatbot capable of fulfilling a variety of text-based tasks, such as answering questions, engaging in dialogues, preparing content, and providing responses. It also successfully performs tasks, such as taking exams and translating between different languages, offering valuable assistance in complex discussions regarding efficiency and guidance (6,7). As of May 2024, ChatGPT boasts approximately 200 million active users per month, positioning it among the top-ranked and most popular AI chatbots (3). Other notable platforms include Google's Gemini, with 200 million users, Microsoft Corporation's Bing, which has 1.2 billion users, and Anthropic PBC's Claude AI, released in March 2023, which claims higher accuracy than ChatGPT. Since its launch, Claude AI has gained 1.7 million active monthly users (8). These applications are equipped with advanced natural language processing and translation capabilities, having undergone extensive training in vocabulary and language. Additionally, they offer text generation functions (9). These tools can assist patients by answering their questions, helping them understand treatments, potential side effects or complications, prognoses, and expected outcomes (10). Traumatic dental injuries (TDIs) are common among children and young adults, accounting for 5% of all injuries. It has been reported that TDIs affect 25% of school-aged children and typically occur before the age of 19 (11). These injuries are the second most prevalent oral health issue after dental caries and are often a leading cause of emergencies in the oral region (12). Timely and accurate guidance is crucial for providing appropriate treatment in dental trauma cases (3). However, because of difficulties in securing timely appointments or accessing expert consultations, individuals have increasingly turned to AI chatbots for advice (13). Consequently, the tendency to seek advice online after dental trauma has notably increased in recent years (3). The field of dentistry has undergone significant transformation in recent decades, with AI-based technologies playing a critical role in this shift (14). Unlike search engines that provide general information from various sources on a specific topic, AI-powered chatbots present information in a conversational style, making it easier to understand complex subjects (15,16). Although there has been growing interest in the use of chatbots in medical research, concerns persist regarding the accuracy and reliability of the health information provided by these chatbots (17). Although previous studies have investigated the validity and reliability of AI-powered chatbots, a limited number of studies have examined the accuracy of information related to dental trauma. To the best of our knowledge, no study has focused on the treatment of dental trauma in primary teeth. Therefore, the aim of this study is to evaluate the reliability and consistency of AI chatbots as a public source of information on dental trauma. The null hypothesis is that AI applications provide accurate and consistent recommendations regarding the treatment of dental trauma in primary teeth.

MATERIAL AND METHODS

Study Design

In this study, the accuracy and consistency of the responses provided by AI chatbots—ChatGPT 3.5, Google Gemini, Bing, and Claude—were evaluated in the context of managing TDIs in primary teeth.

Question Preparation

The questions used in this study were based on the guidelines established by the International Association of Dental Traumatology (IADT), specifically the 2020 guidelines for the management of TDIs in primary dentition (18). A total of 39 questions and their answers were formulated, addressing common issues and concerns related to dental trauma, particularly those frequently raised by parents. All questions were designed to have a yes/no format to minimize the risk of ambiguous or partially correct answers. A pilot study was conducted with 15 pediatric dentists to ensure the clarity of the questions. On the basis of their feedback, any unclear or potentially confusing sections were revised, and 31 questions were finalized (Appendix 1).

Extent of Knowledge Required by Questions

The questions were categorized based on the level of knowledge required to answer them: simple, moderate, and difficult. The classification was based on a survey conducted with 73 fifth-year dental students. Students were asked to respond to the questions in a yes/no format. On the basis of their responses, questions with more than 60% correct answers were classified as simple, those with 30%–60% correct answers were classified as moderate, and those with less than 30% correct answers were categorized as difficult. The impact of the level of required knowledge on the accuracy and consistency of AI responses was subsequently evaluated.

Chatbot Processing

New accounts were created for each AI platform to prevent any influence from prior searches. The questions were asked sequentially to the four AI chatbots (ChatGPT 3.5, Google Gemini, Bing, and Claude AI) and were framed with the following instruction: “As an experienced pediatric dentist, please answer each of the following questions regarding the management of traumatic dental injuries in primary teeth with either a yes or no.” All questions were asked by a single person to ensure consistency in the responses. Each question was asked three times at the same time of the day, with a one-hour interval between each repetition, to evaluate the reliability and reproducibility of the answers. This procedure was repeated for five days. The “new conversation” option was selected each time to ensure the independence of the questions, resulting in a total of 1,860 responses.

Accuracy of Responses

Accuracy refers to the alignment of the AI chatbot responses with an external standard (the reference answers), which, in this study, was based on the IADT guidelines for the management of TDIs in primary teeth (18).

Reliability and Consistency of Responses

Reliability refers to the consistency of the responses produced by the chatbot under the same or similar conditions. By contrast, consistency refers to whether the AI tool produces the same or similar answers over different time periods. In this study, consistency was assessed by comparing responses from the same AI platform over different times.

Ethical Approval

As this study did not involve human or animal subjects, ethical approval was not required. The data collected were obtained in full compliance with the terms of service of the relevant AI platforms.

Statistical Analysis

All responses were recorded in an Excel spreadsheet (Microsoft, Redmond, WA, USA) and analyzed using statistical software (IBM SPSS Statistics 30.0.0.0 and R Programming 4.4.2). An accuracy analysis was conducted for the AI applications. Accuracy was measured as the proportion of correct responses, and 95% confidence intervals (CIs) were calculated using the Wald binomial method. Reliability was evaluated using Fleiss’ kappa coefficient, which measures the level of agreement across responses over time and among multiple “raters”.

Appendix 1. Questions and Answers

1. Is the incidence of periodontal tissue injuries higher than that of dental hard tissue injuries in primary teeth?	Yes
2. Are the results of pulp sensitivity tests reliable in primary teeth?	No
3. Should root canal treatment be considered for a primary tooth with yellowish discoloration but no signs of infection following trauma?	No
4. Does a child's level of cooperation influence the choice of treatment for primary tooth trauma?	Yes
5. Does the time remaining until the exfoliation of the primary tooth influence the choice of posttraumatic treatment?	Yes
6. Do treatments for dental trauma in primary and permanent teeth vary?	Yes
7. Should avulsed primary teeth be replanted?	No
8. Can a splint be applied to reposition mobile fragments in primary tooth trauma?	Yes
9. Can primary tooth trauma affect permanent teeth?	Yes
10. Does oral hygiene affect the prognosis of treatment after dental trauma?	Yes
11. Is filling the first treatment option for enamel fractures observed in primary teeth?	No
12. Is it possible for broken fragments of teeth to penetrate into the soft tissues?	Yes
13. Is it possible to reattach broken fragments to the tooth in primary tooth fractures?	Yes
14. Is root canal treatment the only option for managing complicated crown fractures in primary teeth?	No
15. Does the restorability of the crown affect the treatment option for primary teeth with crown-root fractures?	Yes
16. Does the localization of the fracture line in the root fractures of primary teeth affect the treatment option?	Yes
17. Does the mobility of the fractured segment in the root fractures of primary teeth influence the treatment option?	Yes
18. Is it necessary to reposition a displaced but stable coronal fragment in primary teeth with root fractures?	No
19. Is repositioning the fractured fragment necessary in alveolar fractures seen in the primary dentition?	Yes
20. Do all luxation injuries of primary teeth require root canal treatment?	No
21. Is splinting necessary for primary teeth with concussion?	No
22. Is splinting necessary for primary teeth with subluxation?	No
23. Is radiographic evidence observed in primary teeth with extrusion?	Yes
24. Does the degree of extrusion in primary teeth influence the decision for extraction?	Yes
25. Can a primary tooth with extrusion be left to heal spontaneously if there is no interference with occlusion?	Yes
26. Is radiographic evidence observed in primary teeth with lateral luxation?	Yes
27. Can a primary tooth with lateral luxation be left to heal spontaneously if no occlusal interference occurs?	Yes
28. Should a primary tooth that is intruded toward the permanent tooth bud be extracted?	No
29. Should a severely intruded primary tooth be repositioned to its normal position and splinted?	No
30. Is spontaneous eruption expected in cases of severe intrusion of a primary tooth?	Yes
31. Can intrusion and avulsion be clinically confused?	Yes

RESULTS

Accuracy of Responses

Each of the four AI platforms provided answers to a total of 31 questions three times a day over a period of five days. In this method, each AI platform generated 15 responses per question, resulting in a total of 465 responses per platform. The total number of responses from all four platforms was 1,860. An evaluation of all responses showed that the percentage of correct answers provided by the AI applications across all questions was 92.85%, with 1,727 correct responses. The percentage of correct responses for each AI platform is presented in Table 1 (The 95% CIs for these percentages have been calculated).

Table 1. Distribution of Correct Answer Rates by AI Application.

		N	%	CI 95%
Chat-GPT 3.5	Correct responses	444	95.48	(93.6, 97.37)
	Incorrect responses	21	4.52	(2.63, 6.4)
Gemini	Correct responses	425	91.4	(88.85, 93.95)
	Incorrect responses	40	8.6	(6.05, 11.15)
Bing	Correct responses	448	96.34	(94.64, 98.05)
	Incorrect responses	17	3.66	(1.95, 5.36)
Claude	Correct responses	410	88.17	(85.24, 91.11)
	Incorrect responses	55	11.83	(8.89, 14.76)

CI: Confidence Interval, N: Number, %: Percentage

Consistency of Responses

The consistency of the AI applications' responses (measured by Fleiss' kappa) across five days and three different time intervals is compared in Table 2. According to the results, the Bing AI application demonstrated the highest consistency, while the Claude AI application exhibited the lowest consistency. The consistency of ChatGPT, Bing, and Gemini AI applications showed near-perfect agreement, while the consistency of Claude AI was found to be in significant agreement (Table 3) (19,20). The responses provided by the AI applications at different times were found to be meaningfully consistent within each application.

Table 2. Consistency of AI Chatbots

	Reliability[a]	Standard Error	*p*	CI 95%[b]
Chat-GPT	0.944	0.0413	$p < 0.0001$	(0.8631, 1)
Bing	0.980	0.0251	$p < 0.0001$	(0.9307, 1)
Gemini	0.940	0.0427	$p < 0.0001$	(0.8564, 1)
Claude	0.793	0.0728	$p < 0.0001$	(0.6504, 0.9356)

[a]Fleiss Kappa Score, [b]Wald Binom Method, CI: Confidence Interval

Table 3. Interpretation of Fleiss Kappa Scores.

Kappa Value	Interpretation
< 0.0	Poor Agreement
0.00–0.2	Weak Agreement
0.21–0.4	Fair Agreement
0.41–0.6	Moderate Agreement
0.61–0.8	Substantial Agreement
0.81–1	Almost Perfect Agreement

Extent of Knowledge Required by the Questions and Accuracy–Consistency

The accuracy and consistency of the responses to questions with varying levels of knowledge requirements were analyzed across four different AI applications. According to the results of the Pearson Chi-Square test, there was a significant association between the knowledge level required by the questions and the accuracy percentage for all four AI applications ($p < 0.05$). As the level of knowledge required by the questions increased, the accuracy of the responses decreased (Table 4). According to the Fleiss Kappa analysis, the consistency of responses provided at different times was higher for questions requiring basic knowledge. However, as the level of knowledge required by the questions increased, the consistency of the responses decreased (Table 5).

Table 4. Accuracy of Chatbots Based on the Level of Knowledge Required.

AI Chatbots	Knowledge Level	Correct Answers	Percentage	Total Number of Answers	CI 95%
ChatGPT 3.5	Basic	150	100	150	-
	Moderate	177	98.3	180	(96.4–100)
	Advanced	117	86.67	135	(84.79–88.53)
Gemini	Basic	150	100	150	-
	Moderate	167	92.77	180	(96.4–100)
	Advanced	108	80	135	(76.22, 83.78)
Bing	Basic	150	100	150	-
	Moderate	179	99.44	180	(98.35, 100)
	Advanced	119	88.15	135	(82.69, 93.60)
Claude	Basic	136	90.66	150	(86.011, 95.32)
	Moderate	167	92.77	180	(88.12, 97.43)
	Advanced	107	79.26	135	(72.42, 86.09)

Table 5. Consistency of Chatbots Based on the Level of Knowledge Required.

AI Chatbots	Knowledge Level	Fleiss Kappa	Standard Error	*p*	CI 95%
ChatGPT 3.5	Basic	1	0	$p < 0.0001$	-
	Moderate	0.94	0.0177		(0.91, 0.97)
	Advanced	0.91	0.0247		(0.86, 0.96)
Gemini	Basic	1	0	$p < 0.0001$	-
	Moderate	0.954	0.0156		(0.923, 0.985)
	Advanced	0.877	0.0283		(0.821, 0.933)
Bing	Basic	1	0	$p < 0.0001$	-
	Moderate	0.977	0.0433		(0.8922, 1)
	Advanced	0.97	0.0569		(0.8585, 1)
Claude	Basic	0.878	0.1035	$p < 0.0001$	(0.6751, 1)
	Moderate	0.76	0.1233		(0.5184, 1)
	Advanced	0.728	0.1483		(0.4373, 1)

CI: Confidence interval

DISCUSSION

Accurate and timely emergency treatment is critical in cases of dental trauma. However, barriers to accessing healthcare services can negatively affect this process (21,22). While search engines, such as Google, are widely used to access information, the use of AI chatbots for such inquiries is becoming increasingly common (5). Many individuals now turn to AI chatbots for detailed and explanatory information.

Consequently, ensuring the validity and reliability of the information provided by these tools is of utmost importance. This study aimed to evaluate the performance of four AI chatbots as public information sources in the management of primary tooth traumas. While a limited number of studies in the literature have assessed the validity and reliability of AI chatbots as information sources in the field of dental trauma (3,5,23,24), no prior research has specifically focused on the treatment of primary tooth traumas or comprehensively evaluated four different AI chatbots.

The questions in this study were designed to align with IADT guidelines, focusing on common scenarios in dental trauma and addressing the topics most frequently asked by parents.

The questions were structured to be answered exclusively in a “yes” or “no” format. The responses provided by the AI chatbots were evaluated for accuracy against the answers derived from IADT guidelines. The “yes/no” format was chosen because of the ability of AI chatbots to generate information and references (24). Studies have shown that large language models (LLMs) often have high error rates and may produce incorrect or fabricated responses (25,26). By limiting responses to two options— “yes” or “no”—this study aimed to minimize the risk of additional or fabricated information and ensure the reliability of the collected responses. Ensuring that the questions are clear and comprehensible is crucial to testing the accuracy and validity of the responses obtained from the AI applications. In this study, questions and their answers were designed based on IADT guidelines, a widely accepted resource developed by a comprehensive group of expert dental professionals specializing in dental trauma. The questions addressed common scenarios in dental trauma and topics frequently asked by parents. A pilot study was conducted with 15 pediatric dentists to ensure clarity and comprehensibility. On the basis of the feedback received, ambiguous or potentially confusing elements were revised or removed, resulting in a consensus on 31 finalized questions. This process aimed to minimize errors attributable to question misinterpretation, thereby enhancing the reliability of the responses collected. Variations in network architectures and differences in the size and diversity of data resources lead to AI applications producing different responses to the same questions. These differences highlight the unique strengths and weaknesses of each application (27). In diagnostic studies, an acceptable accuracy threshold is required to exceed 90% (28). In this study, the highest accuracy percentage was observed with Bing, while only Claude failed to achieve a percentage above 90%. Comparing the findings of this study with those of other research on the accuracy of AI applications in dental trauma, Özden et al. reported an accuracy rate of 57.5% for ChatGPT 3.5 and Google Gemini (24). Portilla et al. found that Gemini achieved a general accuracy rate of 80.8% (23). Conversely, Johnson et al. reported Claude AI as the most valid and reliable application, while identifying Bing as the least valid (3). By contrast, this study observed a notably higher overall accuracy rate of 92.85%. The improvement in accuracy rate represents a significant advancement in AI applications. The evolving capabilities of chatbot response generation may explain differences in the quality and reliability of responses across studies conducted at different times (29). Furthermore, Bing’s superior accuracy can be attributed to Microsoft’s robust infrastructure. Microsoft’s extensive and continuously updated knowledge base provides Bing with a comprehensive source of information, enabling it to deliver highly accurate, thorough, and up-to-date responses on a wide range of topics. In this study, the AI application with the lowest accuracy rate was Claude. This contrasts with the findings of Johnson et al., who, in their study evaluating the validity of responses from four AI applications on dental trauma, reported that Claude AI provided significantly more valid answers than other chatbot applications did (3). This discrepancy between the two studies may stem from the differences in the types of questions posed to the AI models. In Johnson’s study, open-ended questions requiring text-based responses were used, whereas this study employed questions with yes or no answers. Claude excels in understanding and generating human language, and it can provide successful text-based responses. Its comprehensive answers made it more valid than ChatGPT and Gemini in handling traditional questions. Özden et al. reported that they were unable to achieve a sufficient level of consistency when evaluating ChatGPT and Google Bard (Gemini) (24). Mohammad-Rahimi et al. measured the consistency of Google Bard’s (Gemini) responses to frequently asked questions in the field of endodontics using Cronbach’s alpha scale and achieved a result of 0.703, which they considered to demonstrate acceptable reliability (30). However, they also emphasized that the LLM tends to provide different phrasing with each repetition of the questions. In this study, satisfactory consistency was observed with all AI robots, except for Claude AI (greater than 90%). Bing showed the highest overall consistency, while Claude AI exhibited the lowest. Nevertheless, the responses obtained at different times remained internally consistent. Similar to this study, Portilla et al. reported high consistency with the Gemini application (Fleiss kappa = 0.941) (23). The improvement in consistency among AI applications could be attributed to updates made to these models. Specifically, the new version of Gemini introduced a new architecture and significant developments in training and service infrastructure, enhancing efficiency, reasoning, and long-context performance (31). In this study, it was observed that as the level of knowledge required by the questions increased, the accuracy and consistency of the AI applications’ responses decreased. This trend may be attributed to the limited information density in the datasets on which the AI models were pre-trained or to the specific context or expertise required for more complex questions. Therefore, developing chatbots that are carefully trained on verified and high-quality databases is essential to fully benefit from AI chatbots. This approach will ensure that users can quickly find the necessary information while also guaranteeing accuracy by limiting the results to the reliable contents of the databases. In this way, the spread of unreliable information or content from unknown sources can be prevented, ensuring that users access accurate answers (23). This study has several important limitations. First, the questions used in the study were designed to be answered with a simple "yes" or "no." This approach limited the responses, preventing unnecessary or irrelevant information from emerging. However, this type of closed-ended question does not fully reflect the complexity of interactions between dentists and patients or real-life scenarios in dental practice. Second, the continuous updating of data sources by AI applications has created a dynamic process that may lead to variations in the study results across different time periods. This factor made it difficult to compare the data obtained with other studies and introduced notable variability in the results. In this regard, the nature of AI systems is considered a factor that could affect the validity of the findings.

CONCLUSION

The results of this study showed that among the four AI chatbots, Bing emerged as the most accurate and consistent in providing responses. Following Bing, ChatGPT 3.5 and Gemini ranked next in terms of performance. However, Claude lagged the other applications, showing lower performance in both accuracy and consistency. These findings highlight the need for careful evaluation when assessing the usability of AI-based systems in clinical applications. While AI systems provide quick responses, their potential to offer incorrect or incomplete information poses a significant risk to the quality of information provided to patients. Continuous improvements and data updates are essential to enhance the reliability and effectiveness of AI chatbots. Furthermore, more research is needed to understand and improve in depth the role of AI technologies in the healthcare sector. Such studies could provide valuable insights into the effectiveness and reliability of AI in clinical settings, contributing to the future development of dental health services.

Acknowledgements

I would like to express my sincere gratitude to the 15 pediatric dentists who participated in the pilot study for their valuable insights, which contributed to the clarity of the questions. Additionally, I extend my heartfelt thanks to the fifth-year dental students who took part in the survey, helping to determine the required knowledge level for the questions. Their time, effort, and support were crucial in the successful execution of this study.

REFERENCES

1. Bagde H, Dhopte A, Alam MK, Basri R. A systematic review and meta-analysis on ChatGPT and its utilization in medical and dental research. Heliyon. 2023;9:e23050.

2. LeCun Y, Bengio Y, Hinton G. Deep Learning. Nature. 2015;28;521(7553):436-44.

3. Johnson AJ, Singh TK, Gupta A, Sankar H, Gill I, Shalini M, Mohan N. Evaluation of validity and reliability of AI Chatbots as public sources of information on dental trauma. Dent Traumatol. 2024 41(2):187-93.

4. Agrawal P, Nikhade P. Artificial intelligence in dentistry: past, present, and future. Cureus. 2022;14:e27405.

5. Guven Y, Ozdemir OT, Kavan MY. Performance of Artificial Intelligence Chatbots in Responding to Patient Queries Related to Traumatic Dental Injuries: A Comparative Study. Dent Traumatol. 2024 22: 1-10.

6. Liu X, Zheng Y, Du Z, Ding M, Qian Y, Yang Z, et al. GPT understands, too. AI Open. 2024;5:208–15.

7. Atkinson CF. Cheap, quick, and rigorous: artificial intelligence and the systematic literature review. Soc Sci Comput Rev. 2023;42:376–93.

8. Safi Z, Abd-Alrazaq A, Khalifa M, Househ M. Technical Aspects of Developing Chatbots for Medical Applications: Scoping Review, J Med Internet Res. 2020;22(12):e19127.

9. Koçyiğit A, Darı AB. ChatGPT in artificial intelligence communication: the future of humanized digitalization. J Strateg Soc Res. 2023;7(2):427–38.

10. Ayers JW, Zhu Z, Poliak A, Leas E, Dredze M, Hogarth M, et al. Evaluating Artificial Intelligence Responses to Public Health Questions. JAMA Network Open. 2023;6(6):e2317517.

11. Levin L, Day PF, Hicks L, O'Connell A, Fouas AF, Bourguignon C, et al. International Association of Dental Traumatology Guidelines for the Management of Traumatic Dental Injuries: General Introduction. Dental Traumatol. 2020;36(4):309–13.

12. Khan L. Dental Care and Trauma Management in Children and Adolescents. Pediatr Ann. 2019**;**48:e3–e8.

13. Erwin J, Horrell J, Wheat H, Axford N, Burns L, Booth J, et al. Access to Dental Care for Children and Young People in Care and Care Leavers: A Global Scoping Review. Dental Journal. 2024;12(2):37.

14. Shahnavazi M, Mohamadrahimi H. The application of artificial neural networks in the detection of mandibular fractures using panoramic radiography. Dent Res J. 2023;20:27.

15. Pandey S, Sharma S. A comparative study of retrieval-based and generative-based chatbots using deep learning and machine learning. Healthc Anal. 2023;3:100198.

16. Chiesa-Estomba CM, Lechien JR, Vaira LA, Brunet A, Cammaroto G, Mayo-Yanez M, et al. Exploring the potential of Chat-GPT as a supportive tool for sialendoscopy clinical decision making and patient information support. Eur Arch Otorhinolaryngol. 2024;281:2081–6.

17. Beltrami EJ, Grant-Kels JM. Consulting ChatGPT: Ethical dilemmas in language model artificial intelligence. J Am Acad Dermatol. 2024;90(4):879-80.

18. Day PF, Flores MT, O'Connell AC, Abbott PV, Tsilingaridis F, Fouad AF, et al. International Association of Dental Traumatology guidelines for the management of traumatic dental injuries: 3. Injuries in the primary dentition. Dent Traumatol. 2020;36(4):343-359.

19. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.

20. Fleiss JL. Measuring Nominal Scale Agreement Among Many Raters. Psychological Bulletin, 1971;76(5):378–382.

21. Wagle E, Allred EN, Needleman HL. Time delays in treating dental trauma at a children's hospital and private pediatric dental practice. Pediatr Dent. 2014;36(3):216-21.

22. Kayıllıoğlu Zencircioğlu Ö, Eden E, Öcek ZA. Access to health care after dental trauma in children: A quantitative and qualitative evaluation. Dent Traumatol. 2019;35(3):163-70.

23. Portilla ND, Garcia-Font M, Nagendrababu V, Abbott PV, Sanchez JAG, Abella F. Accuracy and Consistency of Gemini Responses Regarding the Management of Traumatized Permanent Teeth. Dent Traumatol. 2024 Oct 26 Epub ahead of print.

24. Ozden I, Gokyar M, Ozden ME, Sazak Ovecoglu H. H. Assessment of artificial intelligence applications in responding to dental trauma. Dent Traumatol. 2024;40(6):722-9.

25. Samaan JS, Yeo YH, Rajeev N, Hawley L, Abel S, Ng WH, et al. Assessing the Accuracy of Responses by the Language Model ChatGPT to Questions Regarding Bariatric Surgery. Obes Surg. 2023;33(6):1790-6.

26. Lahat A, Shachar E, Avidan B, Glicksberg B, Klang E. Evaluating the Utility of a Large Language Model in Answering Common Patients' Gastrointestinal Health-Related Questions: Are We There Yet? Diagnostics. 2023;13(11):1950.

27. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the Performance of Generative AI Large Language Models ChatGPT, Google Bard, and Microsoft Bing Chat in Supporting Evidence-Based Dentistry: Comparative Mixed Methods Study. J Med Internet Res. 2023; 28;25:e51580.

28. Umer F, Habib S. Critical Analysis of Artificial Intelligence in Endodontics: A Scoping Review. J Endod. 2022;48(2):152-60.

29. Sharma D, Vidhate DA, Osei-Asiamah J, Kumari M, Mahajan V, Rajagopal K. Exploring the Evolution of Chatgpt: From Origin to Revolutionary Influence. Educational Administration: Theory and Practice 30(5),2685-92.

30. 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 Endod J. 2024;57(3):305-14.

31. Gemini Team, M. Reid, N. Savinov, et al. "Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context. arXiv. 2024;4: 1–154.