HEALTH SCIENCE MONITOR

Health Science Monitor

Health Sci Mon 2025; 3: e251201
https://doi.org/10.5577/hesmon.e251201

Original Article Open Access

Comparison of ChatGPT-5, Gemini 2.5 Pro, and DeepSeek-V3.1 chatbot responses to dental avulsion injuries

Özge İlter Er[1], Özkan Adıgüzel[2], Makbule Taşyürek[2], Hatice Ortaç[3]

[1] Dicle University, Faculty of Dentistry, Department of Pediatric Dentistry, Diyarbakır, Türkiye
[2] Dicle University, Faculty of Dentistry, Department of Endodontics, Diyarbakır, Türkiye
[3] Dicle University, Faculty of Medicine, Department of Biostatistics, Diyarbakır, Türkiye

Received: 26 November 2025
Accepted: 16 December 2025
Published: 25 December 2025

Correspondence:

Dr. Özge İLTER ER
Dicle University, Faculty of Dentistry, Department of Pediatric Dentistry, Diyarbakır, Türkiye.
E-mail: ozgltr@icloud.com

How to cite this article:

İlter Er Ö, Adıgüzel Ö, Taşyürek M, Ortaç H. Comparison of ChatGPT-5, Gemini 2.5 Pro, and DeepSeek-V3.1 chatbot responses to dental avulsion injuries. Health Sci Mon. 2025;3: e251201. https://doi.org/10.5577/hesmon.e251201

Abstract

Aim: The aim of this study was to compare three large language models (LLMs); ChatGPT-5 (OpenAI), Gemini 2.5 Pro (Google), and DeepSeek-V3.1 (Hangzhou DeepSeek AI), in terms of their ability to provide information on questions asked about avulsion, a type of traumatic dental injury.

Methods: In the study, 25 open-ended questions prepared based on the guidelines of the International Association of Dental Traumatology (IADT) were directed to large language models in their "Think / DeepThink / Reasoning" modes. The responses were evaluated by two experienced experts using a five-point Likert scale for accuracy and a three-point Likert scale for completeness (exhaustiveness). Readability was analyzed using the Flesch Reading Ease Score (FRES), Flesch-Kincaid Grade Level (FKGL), Gunning Fog Index (GFI), Simplified Measure of Gobbledygook (SMOG), and Coleman-Liau Index (CLI). Inter-rater reliability was assessed, and ANOVA/Kruskal-Wallis and post-hoc Dunn-Bonferroni tests were used for group comparisons, based on the data's normality.

Results: The results showed that all three models demonstrated high evaluator agreement in terms of both accuracy and completeness. Significant differences were observed between the models in terms of accuracy scores (p=0.016). In subgroup analyses, the ChatGPT-5 model's accuracy score was significantly higher than that of the DeepSeek-V3.1 model (p=0.020). Significant differences were observed between the models in terms of completeness scores (p=0.006). In the subgroup analyses, the completeness scores of the ChatGPT-5 and Gemini 2.5 Pro models were significantly higher than those of the DeepSeek-V3.1 model (p=0.007 and p=0.046, respectively). In terms of readability, the DeepSeek-V3.1 model achieved higher FRES scores and lower FKGL, GFI, and SMOG scores. Overall, ChatGPT-5 demonstrated the highest performance in terms of accuracy and completeness, while DeepSeek-V3.1 achieved the highest readability values.

Conclusion: The results show that ChatGPT-5 is the most reliable model in terms of clinical accuracy and content integrity, whereas DeepSeek-V3.1 produces more linguistically understandable responses. However, no model can completely replace human expertise. These systems should be considered tools to support clinical decision-making processes and should be used under physician supervision, especially in pediatric emergencies.

Keywords: Artificial intelligence, large language models, chatbot, ChatGPT, Gemini, DeepSeek, avulsion

Introduction

Transformer models are modern artificial intelligence architectures developed in the field of natural language processing (NLP) to understand the context of language and make accurate predictions. These models are built on a transformer architecture that relies on a self-attention mechanism to analyze inputs, establish a meaningful context, and generate appropriate outputs (Vaswani, 2017). The transformer structure has created a turning point in large language models (LLMs) owing to its effective use of the attention mechanism and ability to process inputs in parallel on a large scale. This allows the evaluation of the relationships between words in broad contexts and creates a rich statistical representation of the structural and semantic patterns of language (1).

Large language models (LLMs) are generative mathematical models trained on hundreds of terabytes of textual data (articles, books, and other online content) that can sample from the statistical distributions of tokens in public text, including words, graphics, individual characters, and punctuation marks (2). However, most chatbots are proprietary models, and their functions are largely unknown (3). AI-powered technologies have attracted considerable interest in recent years due to their potential to enhance treatment, education, and research activities in various specialties of medicine and dentistry (4).

In November 2022, a prototype version of ChatGPT (OpenAI, San Francisco, CA, USA), which has over 175 billion parameters, was released (5). ChatGPT (Chat Generative Pre-trained Transformer), one of the most innovative LLM models developed by OpenAI, can generate text by gathering information from various data sources, such as articles, books, journals, and open-access websites, translate between languages, and answer questions using information obtained from these sources (6). ChatGPT is a chatbot derived from artificial intelligence that uses Deep Learning (DL) technology for information processing and has the ability to provide information on a wide variety of topics, from science and technology to current events (7). The release of GPT-5 in 2025 is considered a potential turning point for the use of artificial intelligence in education. OpenAI's GPT-5 model is described as "our best AI system to date" and represents a significant leap in intelligence compared to its predecessors. GPT-5 combines state-of-the-art performance in areas such as mathematics, coding, writing, health, and even visual understanding with architectural innovations aimed at making the model more versatile and reliable (8).

Gemini is an LLM program developed by Google DeepMind (Google, Menlo Park, CA, USA) and released on December 19, 2024. It combines language understanding with advanced reasoning and is designed for tasks such as answering questions, explaining concepts, and solving problems across various subjects (9). Gemini 2.5 was released on March 26, 2025. According to the manufacturer, it is a thinking model designed to solve increasingly complex problems. Gemini 2.5 models can pass thoughts through a logic filter before responding, and are therefore considered to provide enhanced performance and increased accuracy (10).

DeepSeek (Hangzhou DeepSeek Artificial Intelligence Basic Technology Research Co., Ltd., Beijing, China), a China-based AI model that emphasizes transparency and reproducibility, made waves worldwide with its chatbot launched in December 2024, owing to its low cost and high download rates (10). Evaluations conducted on DeepSeek revealed that it can perform at a level competitive with AI applications such as ChatGPT, a pioneering LLM program. All these developments have made the DeepSeek AI application popular and attracted worldwide attention. Similar to ChatGPT, DeepSeek is also thought to have the potential to transform the medical field in particular (11). DeepSeek-V3.1 stands out as an important milestone symbolizing the transition to the "agent era" in artificial intelligence. The hybrid inference architecture of the model combines the "Think" and "Non-Think" modes into a single structure, providing high adaptability to different usage scenarios. The newly developed "Think" mode offers a significant increase in speed and efficiency compared to the previous version, DeepSeek-R1-0528. This enables more effective results, especially in time-sensitive tasks. Following the training process, comprehensive optimizations noticeably enhanced the model's capabilities in vehicle operation, multi-step problem solving, and complex task planning. According to the developer, DeepSeek-V3.1 is not merely a performance upgrade; it also represents a new standard in artificial intelligence research, owing to its advanced autonomous agent capabilities (11).

Avulsion, defined as the complete dislodgement of a tooth from its socket, occurs in 0.5-16% of all dental injuries. Avulsion of permanent teeth is a serious dental injury. Prompt and accurate emergency management is essential for achieving the best outcomes after this injury (12, 13). The ideal approach after avulsion is to immediately reposition the tooth. A quick and correct approach after avulsion leads to a successful treatment outcome. For the reimplantation of an avulsed tooth, it is ideal for the tooth to remain outside the mouth for as short a time as possible. This is because the survival of periodontal ligament cells depends on the time spent outside of the mouth. At the same time, the storage conditions of the tooth outside the mouth are also important for the survival of periodontal cells. After 30 min of exposure to dry conditions, most periodontal ligament cells cannot survive. Even if the tooth is stored in a suitable storage solution for more than 60 min, periodontal ligament cells cannot remain viable (14, 15).

Artificial intelligence chatbots can also be easily used in various health areas, such as helping users assess their need to consult a healthcare professional, identify symptoms, and provide clinical decision support (16). Although the use of artificial intelligence, especially GPTs, in health-related topics is becoming widespread, there are concerns regarding its validity and reliability (17).

Decision-making in cases of avulsion in children often occurs under time pressure, increased emotional stress, and limited access to specialists. At the same time, in cases of avulsion, it may not always be possible for parents or healthcare professionals to access a dentist specializing in this field (endodontist, pediatric dentist) (18). Therefore, evaluating AI-generated responses from a multidimensional perspective—accuracy, completeness, and readability—and discussing their advantages and disadvantages can greatly contribute to this topic (17). Accuracy ensures factual reliability, while completeness minimizes omissions and maintains clinical safety. Readability is crucial for facilitating effective communication with non-expert caregivers. Together, these parameters reflect a holistic framework for evaluating AI's potential to serve as a supportive tool in the management of dental trauma in children, rather than merely as a source of information.

The aim of this study was to evaluate and compare the accuracy, completeness, and readability performance of three advanced LLMs (ChatGPT-5, DeepSeek V3.1, Gemini 2.5 Pro) in answering questions related to avulsion from traumatic dental injuries.

Materials and Methods

Ethical committee approval was not required for this research, as the study did not involve any patient data or personal information and aimed only to review publicly available and freely accessible online information sources. However, ethical principles were adhered to in the study, and research integrity and data confidentiality were ensured.

In this study, an experienced pedodontist and endodontist prepared 25 open-ended questions about avulsion, taking into account the guidelines of the International Association of Dental Traumatology (IADT), and the answer sheets were organized according to the IADT guidelines (14). The questions and answers were saved in two separate Microsoft Word (Office 365, V16.0; Microsoft Corp., Redmond, WA, US) documents.

These questions covered first aid, reimplantation, splinting, endodontic treatment, medication use, follow-up and complications, special cases (primary teeth and teeth that have been dry for a long time), and patient information (Table 1).

Table 1. List of questions asked of the chatbots.

Questions	
1.	What are the recommended instructions for an avulsed permanent tooth?
2.	Should a tooth that comes out in milk be rinsed before being placed back in the socket? If so, what irrigation solution should be used?
3.	How long should the splinting period be for a tooth with an open apex that is reimplanted after avulsion?
4.	How long should the splinting period be for a closed-apex tooth reimplanted after avulsion?
5.	What is the ideal splint material for splinting an avulsed tooth?
6.	When should root canal treatment be initiated after reimplantation of an avulsed tooth with a closed apex?
7.	When should endodontic treatment be initiated for an avulsed tooth with an open apex?
8.	Should a tooth with an extraoral dry time of more than 1 hour be soaked in any solution before reimplantation?
9.	What is the best storage medium for preserving an avulsed tooth?
10.	What type of systemic medication is prescribed after tooth replantation?
11.	What will be the duration of the follow-up period for an avulsed tooth?
12.	How should the socket be treated before reimplantation of an avulsed tooth?
13.	What are the possible complications that may occur in an avulsed tooth?
14.	What is the treatment for an avulsed primary tooth?
15.	How long should the splinting period be for an avulsed tooth that has been dry outside the mouth for more than 60 minutes?
16.	What recommendations should be given to the patient after reimplantation of an avulsed tooth?
17.	At what intervals should radiographic follow-up of an avulsed tooth be performed?
18.	Should root canal treatment (RCT) following avulsion be performed in a single session or multiple sessions?
19.	What intracanal medicament is employed in endodontic therapy after tooth avulsion?
20.	What are the clinical signs of ankylosis in an avulsed tooth?
21.	Is pulp canal obliteration an expected outcome after tooth replantation with an open apex?
22.	When is the decision for crown removal made in the case of an avulsed tooth?
23.	What treatment options should be considered following the failure of reimplantation of an avulsed tooth?
24.	Should the patient be offered tetanus vaccination after avulsion?
25.	Should the position of an avulsed tooth that was reimplanted in the wrong position be changed?
Total 25 Questions	

These 25 open-ended questions were asked of three different LLMs: ChatGPT-5 Think mode, DeepSeek V3.1 DeepThink mode, and Gemini 2.5 Pro Reasoning mode.

The questions were submitted to three different large language models with newly created accounts on the same computer, on the same day, and in the same order by an independent volunteer who was not involved in the study on September 15, 2025. The models used were ChatGPT-5's Think mode (OpenAI, [https://chatgpt.com/], accessed September 15, 2025, paid), Gemini 2.5 Pro's Reasoning mode (Google, [https://gemini.google.com/, paid], accessed September 15, 2025), and DeepSeek V3.1's DeepThink mode (DeepSeek Inc., Hangzhou, China) [https://chat.deepseek.com/], accessed September 15, 2025, free].

Access was performed solely via the web interface of the relevant platforms using a computer. It was designed to reflect the real clinical environment conditions as accurately as possible. No preliminary tests, rapid adjustments, or additional interventions were performed. Interactions with the models were limited to predefined English questions prepared for research purposes. No additional instructions, interventions, or alternative methods of use were applied. The researchers did not alter any parameters related to the production process (e.g., temperature or sampling settings).

Questions were asked in English, and responses were received in English. Before each question-and-answer session, all search histories and cookies were cleared from the participants' computers. To minimize the influence of previous responses, a new chat window was created for each session. Each response was saved in a Microsoft Word document (Office 365, V16.0; Microsoft Corp., Redmond, WA, US). The responses of each chatbot were color-coded before being sent to researchers. Thus, the researchers did not know the name of the chatbot they were evaluating. Finally, the color codes corresponding to the chatbots were identified.

In a review conducted by experts with at least 10 years of experience, the accuracy of the LLM responses was evaluated using a five-point Likert scale. To ensure impartiality and minimize the risk of bias, the process was conducted using a double-blind design (19, 20).

Evaluation Scale:

1. Very poor – Responses were not factually accurate, and important details were omitted or inconsistent.
2. Poor – Some elements were partially correct, but the core content was largely missing or unclear.
3. Average – Both accurate and missing/unclear information were present, and the responses were partially accurate.
4. Good – Responses were mostly accurate and consistent, with only minor omissions.
5. Excellent – Extremely accurate, comprehensive, consistent, and well-structured responses were provided.

In another review conducted by experts, the completeness of the LLM responses was rated using a three-point Likert scale (19, 20).

1 = Incomplete: The fundamental aspects of the question were not addressed.

2 = Adequate: All the important components of the question were addressed.

3 = Comprehensive: The response addressed all necessary elements and included additional, relevant insights.

The readability of the responses generated by the LLMs was evaluated using Readable Software (https://readable.com), an online analysis tool. The English responses to the questions were copied and pasted into the website individually, and each question was analyzed separately. Five established metrics were used as a basis for this review: the Flesch Reading Ease Score (FRES), Flesch–Kincaid grade level (FKGL), Gunning Fog Index (GFI), Coleman–Liau Index (CLI), and Simple Measure of Gobbledygook (SMOG) (10, 21).

The Flesch Reading Ease Score (FRES) measures text readability by assigning higher scores to more accessible and understandable content. For example, a score between 90 and 100 indicates a very easy-to-read text, while a score below 30 indicates an academically challenging text. This formula is widely used by healthcare organizations to create easy-to-read patient documents (Table 2) (22).

The Flesch-Kincaid Grade Level (FKGL) determines the relevant school grade level of a text in the US. For example, a score of 8.0 indicates that the text is appropriate for 8th-grade students (Table 2) (23).

The Simple Measure of Gobbledygook (SMOG) estimates the required education level based on the number of complex words. It is used to assess the readability of health and educational materials (24).

The Gunning Fog Index (GFI) calculates the grade level by considering the prevalence of long words and complex sentence structures. It is commonly used to assess the readability of business and technical writing

(25). The index typically ranges from 1 to 18 or higher. The values represent the number of years of education required for the reader to understand the text. A score of approximately 8 is considered appropriate for the general public, while texts with scores above 17 are generally aimed at graduate-level readers (26).

The Coleman-Liau Index (CLI) estimates the appropriate grade level by analyzing the ratio of letters to sentences in the text. It is used to evaluate the readability of texts written in digital environments (27). A schematic representation of this study is shown in Fig. 1.

Table 2. The Flesch Reading Ease Score (FRES) and Flesch-Kincaid Grade Level (FKGL) were graded (10).

FRES	Reading Level	FKGL	Estimated Reading Graduate Level
90	Very Easy	5	5th Grade
80-89	Easy	6	6th Grade
70-79	Fairly Easy	7	7th Grade
60-69	Standard	8-9	8th-9th Grade
50-59	Fairly Difficult	10-12	High School
30-49	Difficult	13-16	College Level
0-29	Very Difficult	>16	College Graduate

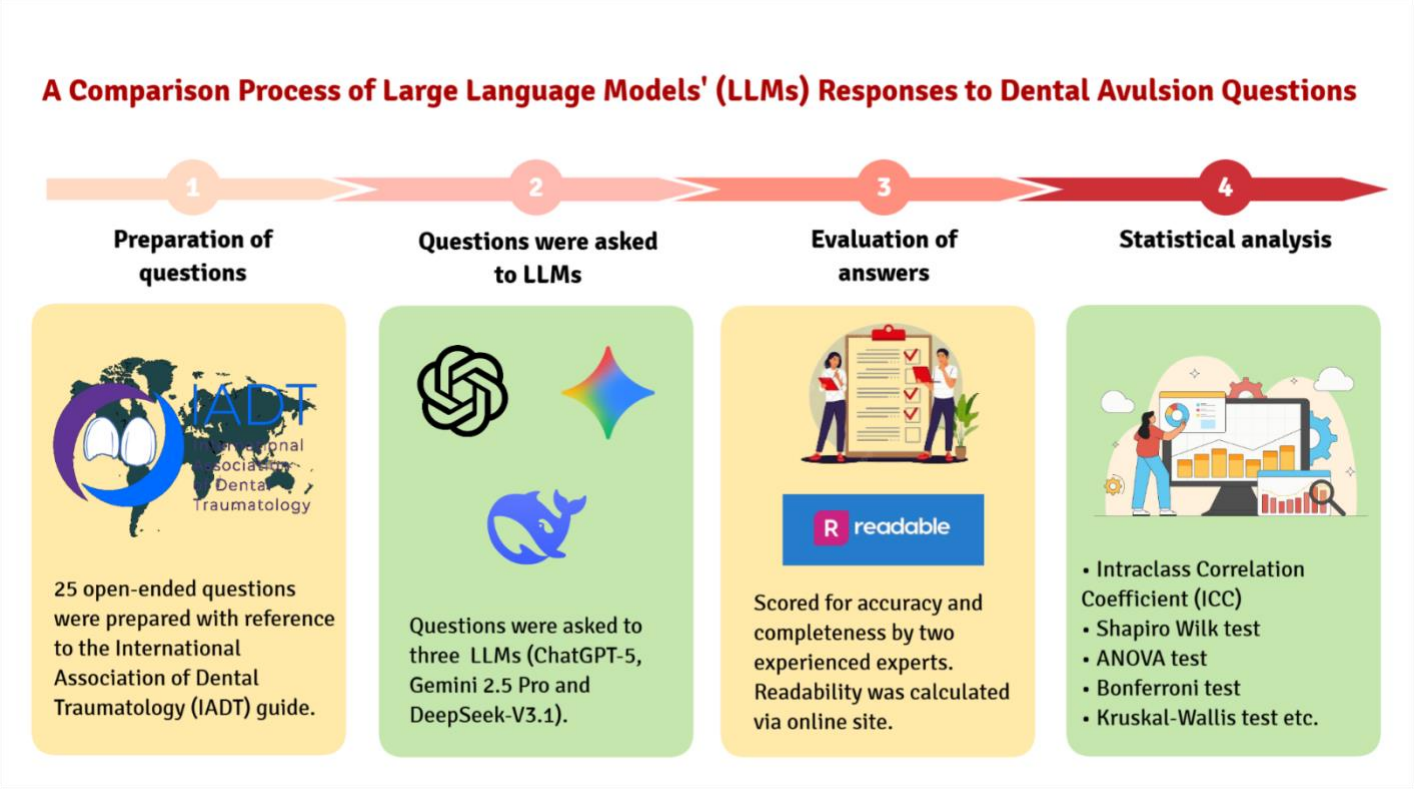

Figure 1. Schematic representation of the study.

Statistical analysis

The Intraclass Correlation Coefficient (ICC) was calculated to assess the agreement between the two evaluators' scores. The normality of the continuous variables was examined using the Shapiro-Wilk test. According to the normality test results, variables that showed normality were expressed as mean ± standard deviation, while variables that did not show normality were expressed as median (minimum: maximum) values. Based on the normality test results, ANOVA was used to

compare groups when there were more than two groups and a normal distribution was observed, while the Kruskal-Wallis test was used when a normal distribution was not observed. Following the ANOVA test, subgroup analyses were performed using the Bonferroni test in cases of overall significance. Following the Kruskal-Wallis test, subgroup analyses were performed using the Dunn-Bonferroni test in cases of overall significance. The relationships between the scores were examined using correlation analysis, and Pearson and Spearman's correlation coefficients were calculated. The SPSS software (IBM Corp., Armonk, NY, USA; IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY, USA) was used for statistical analyses, and the type I error rate was set at 5%.

Results

The consistency between the two evaluators in terms of accuracy scores was assessed using a single-measure two-way consistency type [ICC(2,1), two-way random-effects, absolute agreement, single measures] coefficient for all the artificial intelligence models. The inter-rater reliability was excellent for all models. For ChatGPT, ICC(2,1) = 0.968 (95% CI: 0.929-0.986, p<0.001), ICC(2,1) = 0.971 (95% CI: 0.935-0.987, p<0.001) for Gemini, and ICC(2,1) = 0.986 (95% CI: 0.969-0.994, p<0.001) for DeepSeek.

Table 3. Consistency values between the two evaluators in terms of accuracy scores.

	ICC (2,1)	95% CI	p-value
Model			
ChatGPT-5	0.968	0.929-0.986	<0.001
Gemini 2.5 Pro	0.971	0.935-0.987	<0.001
DeepSeek-V3.1	0.986	0.969-0.994	<0.001

CI: Confidence Interval

Regarding completeness scores, the consistency between the two evaluators was assessed for all artificial intelligence models using a single-measure two-way consistency type coefficient [ICC(2,1), two-way random-effects, absolute agreement, single measures]. According to the analysis results, the inter-evaluator reliability for all the models was excellent. For ChatGPT, ICC(2,1) = 0.854 (95% CI: 0.698-0.933, p < 0.001), ICC(2,1) = 0.855 (95% CI: 0.700-0.934, p < 0.001) for Gemini, and ICC(2,1) = 0.941 (95% CI: 0.871-0.974, p < 0.001) for DeepSeek.

Table 4. Consistency values between the two evaluators in terms of completeness of scores.

	ICC (2,1)	95% CI	p-value
Model			
ChatGPT-5	0.854	0.698-0.933	<0.001
Gemini 2.5 Pro	0.855	0.700-0.934	<0.001
DeepSeek-V3.1	0.941	0.871-0.974	<0.001

CI: Confidence Interval

In the analysis of avulsion, the accuracy scores of responses to open-ended questions across three different artificial intelligence models were compared, and significant differences were found between the models in terms of these scores (p=0.016). In the subgroup analyses, it was determined that the median accuracy value of the ChatGPT-5 model was 5 (Q1:5 - Q3:5), while the DeepSeek-V3.1 model had a median value of 4 (Q1:1.5 – Q3:5). The ChatGPT-5 model had a higher accuracy score than the DeepSeek-V3.1 model (p=0.020) (Table 5) (Fig. 2).

In the analysis of avulsion, the completeness scores of the responses to open-ended questions directed at three different artificial intelligence models were compared, and significant differences were found between the models in terms of these scores (p=0.006). Within the subgroup analyses, it was determined that the ChatGPT-5 model had a median accuracy value of 3 (Q1:3-Q3:3), the Gemini 2.5 Pro model had a median accuracy value of 3 (Q1:3-Q3:3), and the DeepSeek-V3.1 model had a median accuracy value of 2 (Q1:1 – Q3:3). The completeness scores of the ChatGPT-5 and Gemini 2.5 Pro models were higher than those of the DeepSeek-V3.1 model (p=0.007 and p=0.046, respectively) (Table 5) (Fig. 3).

In the analysis of avulsion, the FRES scores for the readability of the responses to open-ended questions directed at three different artificial intelligence models were compared, and FKGL readability references were found between the models in terms of these scores

(p<0.001). Within the subintelligence analyses, the average FRES score of the responses given by DeepSeek-V3.1 (45.41±12.05) to open-ended questions was significantly higher than the average score of the responses given by the ChatGPT-5 (30.42±13.55) model (p<0.001). No significant differences were found between the other pairs of groups (p>0.05) (Table 5) (Fig. 4).

In the analysis of avulsion, the readability FKGL scores of the responses to open-ended questions directed at three different artificial intelligence models were compared, and significant differences were found between the models in terms of these scores (p<0.001). Within the subgroup analyses, the mean FKGL score of the responses provided by ChatGPT-5 (12.44±2.3) and Gemini 2.5 Pro (12.82±2.37) to open-ended questions was found to be significantly higher than the mean score of the responses provided by the DeepSeek-V3.1 (10.21±1.79) model's responses (p=0.002 and p<0.001, respectively). No significant differences were found between the other pairs (p>0.05) (Table 5) (Fig. 5).

In the analysis conducted on avulsion, the readability SMOG scores of the responses to open-ended questions directed at three different artificial intelligence models were compared, and significant differences were found between the models in terms of these scores (p<0.001). Within the subgroup analyses, the mean SMOG scores of the responses to open-ended questions provided by ChatGPT-5 (14.12±1.69) and Gemini 2.5 Pro (15±1.89) were significantly higher than those of the responses provided by the DeepSeek-V3.1 model (12.7 ± 1.36) (p=0.010 and p<0.001, respectively). No significant differences were found between the other pairs (p>0.05) (Table 5) (Fig. 6).

In the analysis conducted on avulsion, the readability GFI scores of the responses to open-ended questions directed at three different artificial intelligence models were compared, and significant differences were found between the models in terms of these scores (p<0.001). Within the scope of subgroup analyses, the mean GFI scores of the responses to open-ended questions provided by ChatGPT-5 (14.92±2.51) and Gemini 2.5 Pro (15.35±2.49) were significantly higher than the mean score of the responses provided by the DeepSeek-V3.1 (11.94±1.72) model (p<0.001 and p<0.001, respectively). No significant differences were found between the other pairs (p>0.05) (Table 5) (Fig. 7).

In the analysis conducted on avulsion, the CLI scores for readability of the responses to open-ended questions directed at three different artificial intelligence models were compared, and significant differences were found between the models in terms of these scores (p<0.001). Within the subgroup analyses, the mean CLI score of ChatGPT-5 (14.88±2.19)'s responses to open-ended questions was found to be significantly higher than the mean CLI scores of the responses from Gemini 2.5 Pro (13.5±1.86) and DeepSeek-V3.1 (12.29±1.92) (p<0.001 and p=0.050, respectively). No significant differences were found between the other pairs (p>0.05) (Table 5) (Fig. 8).

Table 5. Comparison of accuracy, completeness, and readability scores of ChatGPT-5, Gemini 2.5 Pro, and DeepSeek-V3.1.

		Chatbot			
		ChatGPT-5	Gemini 2.5 Pro	DeepSeek-V3.1	p-value
Accuracy	Mean ± SD	4.56±1.16	4.44±1.23	3.60±1.68	0.016[a]
	Median (min.-max.)	5(5-5)	5(4.5-5)	4(1.5-5)	
Completeness	Mean ± SD	2.8±0.58	2.68±0.69	2.2±0.87	0.006[a]
	Median (min.-max.)	3(3-3)	3(3-3)	2(1-3)	
Readability					
FRES	Mean ± SD	30.42±13.55	36.72±12.08	45.41±12.05	<0.001[b]
	Median (min.-max.)	30.5(22.7-36.05)	34.4(27.2-43.25)	45.9(36.3-55.15)	
FKGL	Mean ± SD	12.44±2.3	12.82±2.37	10.21±1.79	<0.001[b]
	Median (min.-max.)	13.1(10.8-13.7)	13.3(11.15-14.8)	10.3(9.2-11.6)	
SMOG	Mean ± SD	14.12±1.69	15±1.89	12.7±1.36	<0.001[b]
	Median (min.-max.)	14.1(12.7-15.35)	15.2(13.45-16.5)	12.8(11.75-13.65)	
GFI	Mean ± SD	14.92±2.51	15.35±2.49	11.94±1.72	<0.001[b]
	Median (min.-max.)	15(13.3-16.8)	15.9(13.1-17.55)	12(10.8-13.15)	
CLI	Mean ± SD	14.88±2.19	13.5±1.86	12.29±1.92	<0.001[b]
	Median (min.-max.)	15(13.6-16.35)	13.3(12.7-14.95)	12.8(10.8-13.55)	

The data are expressed as mean ± standard deviation and median (Q1: Q3).
a: Kruskal-Wallis test, b: ANOVA test

A positive and significant correlation was found between the accuracy and completeness scores (rs= 0.96; p<0.001). The findings indicate that an increase in the accuracy score is associated with an increase in the completeness score (Table 6).

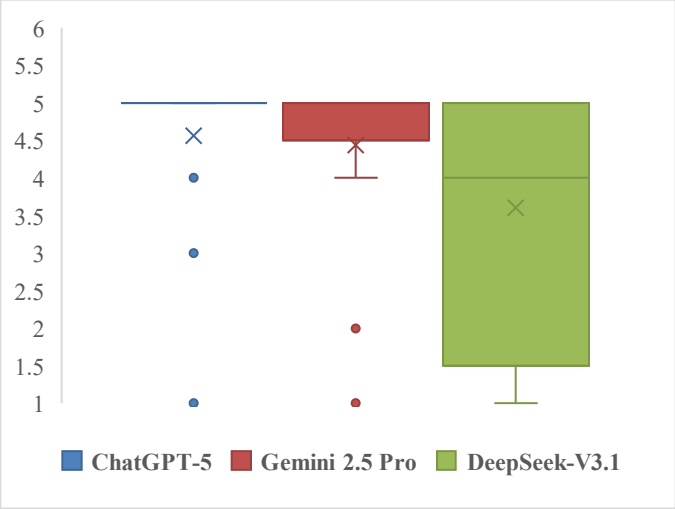

Figure 2. Mean accuracy scores of ChatGPT-5, Gemini 2.5 Pro, and DeepSeek-V3.1 in response to dental avulsion injuries.

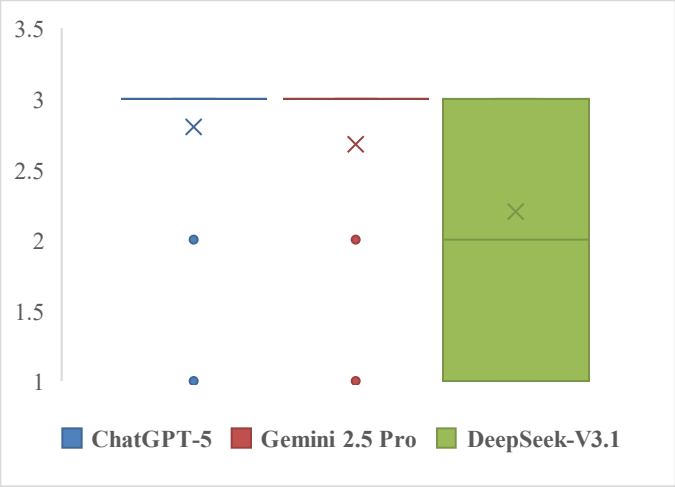

Figure 3. Mean completeness scores of ChatGPT-5, Gemini 2.5 Pro, and DeepSeek-V3.1 in response to dental avulsion injuries.

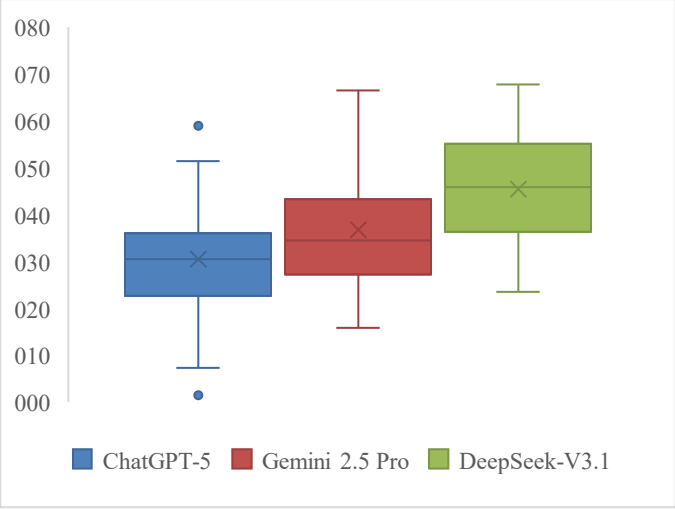

Figure 4. Mean scores of FRES ChatGPT-5, Gemini 2.5 Pro, and DeepSeek-V3.1 in response to dental avulsion injuries.

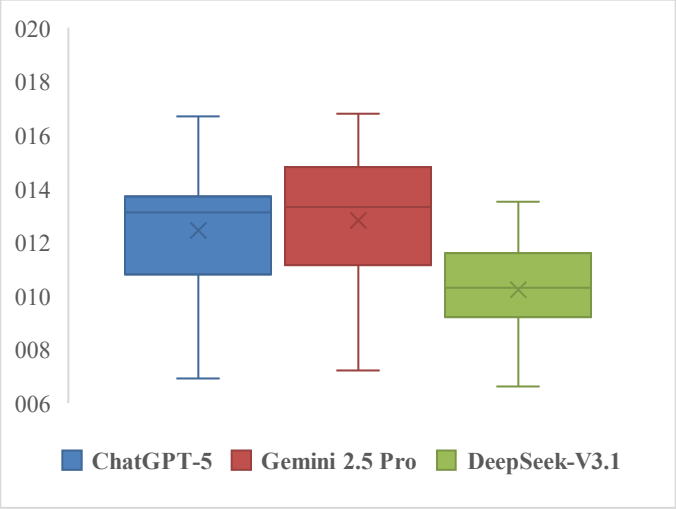

Figure 5. Mean scores of FKGL ChatGPT-5, Gemini 2.5 Pro, and DeepSeek-V3.1 in response to dental avulsion injuries.

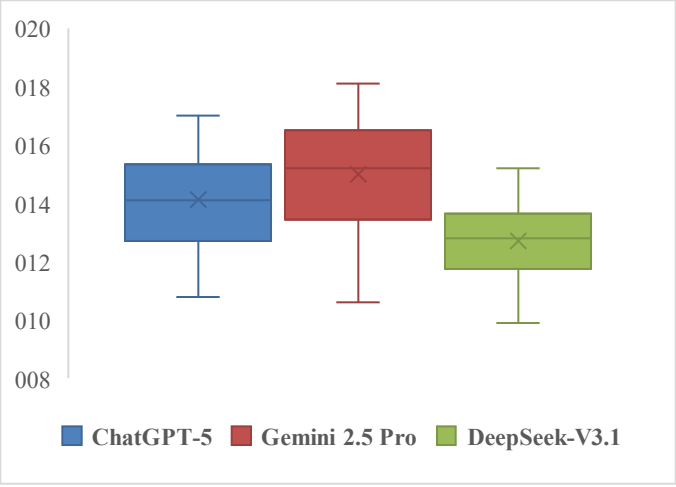

Figure 6. Mean scores of SMOG ChatGPT-5, Gemini 2.5 Pro, and DeepSeek-V3.1 in response to dental avulsion injuries.

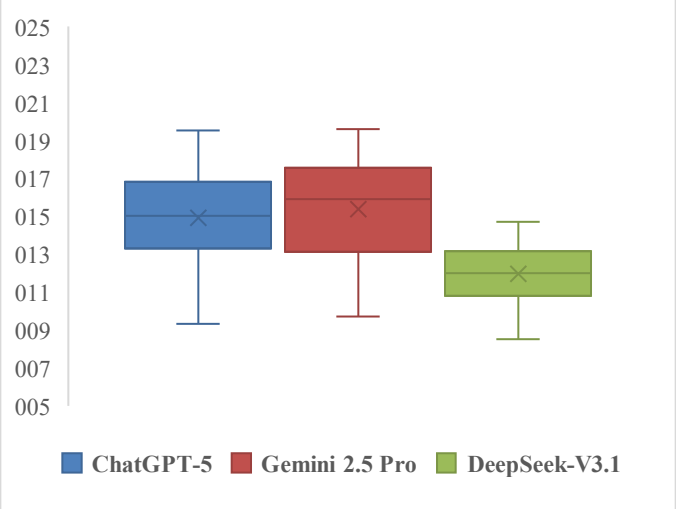

Figure 7. Mean scores of GFI ChatGPT-5, Gemini 2.5 Pro, and DeepSeek-V3.1 in response to dental avulsion injuries.

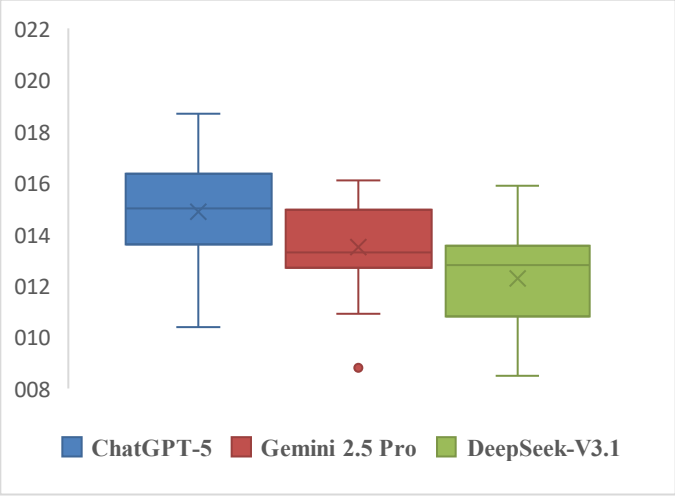

Figure 8. Mean scores of CLI ChatGPT-5, Gemini 2.5 Pro, and DeepSeek-V3.1 in response to dental avulsion injuries.

A significant negative relationship was found between the FRES and FKGL readability scores (rs =-0.93; p<0.001). As the FRES value increases (the reading level becomes more difficult), the FKGL value decreases (reading ease decreases). A significant negative correlation was found between the FRES and SMOG readability scores (rs= -0.74; p<0.001). As the FRES value increases (the reading level becomes more difficult), the SMOG value decreases (the reading ease decreases). A significant negative correlation was found between the FRES and GFI readability scores (rs= -0.78; p<0.001). As the FRES value increases (the reading level becomes more difficult), the GFI value decreases (readability decreases). A significant negative relationship was found between the FRES and CLI readability scores (r(s) =-0.89; p<0.001). As the FRES value increases (the reading level becomes more difficult), the GFI value decreases (reading ease decreases). A significant negative relationship was found between the FKGL and SMOG readability scores (rs= 0.86; p<0.001). This indicates that an increase in the FKGL score is associated with an increase in the SMOG score. A significant negative correlation was found between the FKGL and GFI readability scores (rs=0.87; p<0.001). This indicates that an increase in the FKGL score is associated with an increase in the GFI score. A significant negative correlation was found between the FKGL and CLI readability scores (rs=0.77; p<0.001). This shows that an increase in the FKGL score is associated with an increase in the CLI score. A significant negative correlation was found between the SMOG and GFI readability scores (rs=0.97; p<0.001). This shows that an increase in the SMOG score is also associated with an increase in the GFI score. A significant negative correlation was found between the SMOG and CLI readability scores (rs=0.75; p<0.001). This indicates that an increase in the SMOG score is associated with an increase in the CLI score. A significant negative correlation was found between the GFI and CLI readability scores (rs=0.78; p<0.001). This indicates that an increase in the GFI score is associated with an increase in the CLI score (Table 7).

Table 6. Spearman correlation analysis between the accuracy, completeness, and readability scores of ChatGPT-5, Gemini 2.5 Pro, and DeepSeek-V3.1.

	Accuracy	Completeness
Accuracy		
• rs	-	-
• p-value	-	-
Completeness		
• rs	0.96	-
• p-value	**<0.001**	-
FRES		
• rs	-0.09	-0.12
• p-value	0.436	0.301
FKGL		
• rs	0.09	0.12
• p-value	0.439	0.300
SMOG		
• rs	0.03	0.08
• p-value	0.789	0.490
GFI		
• rs	0.09	0.15
• p-value	0.407	0.204
CLI		
• rs	0.01	0.05
• p-value	0.965	0.691

FKGL: Flesch-Kincaid Grade Level, FRES: Flesch-Kincaid Reading Ease Score, SMOG: Simple Measure of Gobbledygook, GFI: Gunning Fog Index, CLI: Coleman-Liau Index, rs: Spearman's correlation coefficient

Table 7. Spearman correlation analysis of the readability scores of ChatGPT-5, Gemini 2.5 Pro, and DeepSeek-V3.1.

		FRES	FKGL	SMOG	GFI	CLI
FKGL						
•	r	-0.93	-	-	-	-
•	p-value	**<0.001**	-	-	-	-
SMOG						
•	r	-0.74	0.86	-	-	-
•	p-value	**<0.001**	**<0.001**	-	-	-
GFI						
•	r	-0.78	0.87	0.97	-	-
•	p value	**<0.001**	**<0.001**	**<0.001**	-	-
CLI						
•	r	-0.89	0.77	0.75	0.78	-
•	p-value	**<0.001**	**<0.001**	**<0.001**	**<0.001**	-

FKGL: Flesch-Kincaid Grade Level, FRES: Flesch-Kincaid Reading Ease Score, SMOG: Simple Measure of Gobbledygook, GFI: Gunning Fog Index, CLI: Coleman-Liau Index

Discussion

Recent advances in artificial intelligence and human interactions have led to the increasing prevalence of AI-based chatbots. As a result, physicians can rely on AI chatbots as a primary source of information regarding avulsion, a critical dental trauma requiring urgent intervention. Delays in treatment negatively affect prognosis; therefore, ensuring the reliability and consistency of the chatbot responses to relevant questions is crucial. AI chatbots can facilitate treatment by answering physicians' questions and providing information about treatments, potential side effects, complications, prognoses, and outcomes (2, 18, 28).

In this study, the accuracy, completeness, and readability of the responses provided by the ChatGPT-5, DeepSeek-V3.1, and Gemini 2.5 Pro models to questions related to avulsed teeth were statistically evaluated. According to the analysis results, the null hypothesis was rejected for all three criteria, and significant differences were observed.

Each of the prepared questions, as in studies in this field, was directed to the LLM only once in our study (29-31). When the model's response was unsuccessful, follow-up questions were not asked, nor were additional explanations provided. This approach aimed to conduct the evaluation under controlled conditions as much as possible and to ensure the comparability of responses from different models. By simulating scenarios in which dentists seek emergency help for avulsion trauma, this study gained a structure that reflects real-world clinical decision-making environments. Thus, the goal was to evaluate the model performance, not only theoretically, but also in a practical clinical context. Furthermore, the aim of our study was not to evaluate the consistency of LLMs but to examine the accuracy, completeness, and readability of responses to dental trauma, such as avulsion, which develops suddenly and requires urgent intervention.

As noted in the study by Chatzopoulos et al., it has been observed that LLM responses to open-ended clinical questions still need improvement in terms of comprehensiveness, scientific validity, logical consistency, and presentation clarity (32). This finding demonstrates that open-ended questions present an important opportunity to analyze LLMs and evaluate their performance in greater depth. Therefore, open-ended questions were preferred in our study. There are two main reasons for this finding. First, open-ended questions allow us to measure the model's ability to recall information, reason, make logical inferences, and interpret the clinical context holistically. Second, in clinical settings, questions are often posed in an unexpected manner, such as in emergency or complex scenarios involving an avulsion, rather than in a test format. Therefore, such questions allow the model to be evaluated closer to real-world conditions. For these two reasons, an open-ended question format was adopted in our study, and the model's performance was examined in a more comprehensive manner, both scientifically and practically.

The fact that LLMs can generate incorrect information and references has been observed in various studies across different fields (30, 33), and this phenomenon has been termed "hallucination" (34). In one study, to eliminate the so-called hallucinations and

obtain accurate and up-to-date information, questions were prepared solely according to the IADT avulsion injury guidelines, and scoring and evaluation were performed according to these guidelines (31). Similarly, in our study, questions were prepared solely according to the IADT avulsion injury guidelines, and the evaluation and scoring were performed based on these guidelines. Other reasons for using a single source include comparing the reliability of the new LLM versions used to prevent the production of incorrect or inconsistent information, ensuring objectivity in the evaluation, and following a scientific standard, repeatable method.

A previous study evaluating anatomy questions from the Turkish Dental Specialty Exam (DUS) found statistically significant differences in the accuracy of the LLM responses. In this study, ChatGPT-4o showed the highest performance with an accuracy rate of 98.6%, while DeepSeek-v3, Gemini 2.0, and Claude 3.7 Sonnet models achieved the same accuracy level (89.2%) (35). Although the subject area and LLM versions used in our current research are different, significant differences were observed between the accuracy rates of the models. ChatGPT-5's accuracy score was found to be statistically significantly higher than that of DeepSeek-V3.1. In contrast, no significant difference was observed between the accuracy levels of DeepSeek-V3.1 and Gemini 2.5 Pro.

Kaygısız et al. evaluated responses to clinical case scenarios related to oral pathology using two different AI models: DeepSeek-V3 and ChatGPT-4o. The accuracy of the responses was measured using a 5-point Likert scale. The average score for DeepSeek-V3 was 4.02 ± 0.36, while that for ChatGPT-4o was 3.15 ± 0.41. These results indicate that both models performed at a moderate-to-high level. In the overall evaluation, DeepSeek-V3 was found to be statistically more successful than ChatGPT-4o (36). In our study, however, the accuracy score of the ChatGPT-5 model was higher than that of DeepSeek-V3.1. This finding partially contradicts the results of Kaygısız et al. However, this difference may stem from both the change in the subject matter of the study (addressing a situation requiring urgent intervention, such as avulsion, instead of oral pathology) and updates in the versions of the artificial intelligence models used. Therefore, it can be said that the performance of artificial intelligence models is significantly affected not only by the brand or model but also by version differences and the characteristics of the clinical scenario addressed.

In one study, four different artificial intelligence chat models were compared based on their responses to 25 open-ended questions regarding traumatic dental injuries during the deciduous dentition period, as evaluated by independent pediatric dentists, according to the IADT guidelines. This study reported that the ChatGPT-4o, DeepSeek, and Claude 3.7 models produced significantly more accurate responses than Gemini Advanced. Although there was no statistically significant difference in accuracy among the first three models, ChatGPT-4o achieved the highest overall accuracy rate (37). In our current study, although similar next-generation models were used, the accuracy score of the ChatGPT-5 model was significantly higher than that of the DeepSeek-V3.1 model. Although the differences in the average accuracy values of the models were not large in absolute terms, we believe that this difference could have clinically significant consequences in avulsion cases.

In their study comparing the performance of advanced artificial intelligence models in pulp therapy for immature permanent teeth, Sezer and Aydoğdu reported that ChatGPT-4 scored significantly higher than Gemini Advanced in terms of completeness. Our study yielded similar findings: although different versions were used, ChatGPT-5's completeness (accuracy) scores were higher than Gemini 2.5 Pro's. This result is consistent with previous literature data, demonstrating that ChatGPT models can outperform Gemini models, particularly in terms of the comprehensiveness and completeness of the responses (37).

In a survey conducted among dental professionals regarding knowledge, perception, and management of traumatic tooth avulsion, the majority of participants (87.62%) agreed that a dislodged tooth should be reimplanted. 40.10% knew the ideal transport medium for an avulsed tooth, and 63.86% accepted the critical time window for successful replantation. According to the study results, dentists' knowledge of the clinical management of avulsion is moderate but insufficient, and some aspects of the appropriate treatment protocol for avulsed teeth remain for improvement (38). At this point, the importance of LLM-based artificial intelligence systems increases. This is because clinicians may not always have immediate access to the current literature or accurate information in an emergency. LLMs, which are easily accessible via smartphones and computers, have the potential to bridge this information gap. In time-sensitive cases, such as traumatic tooth avulsion, the ability of LLMs to provide accurate, evidence-based, and practical answers can directly affect treatment success. Therefore, our current study aims to demonstrate that LLMs can be evaluated as potential educational and support tools by examining their information accuracy and clinical guidance capacity in a subject where dentists currently lack sufficient knowledge. The results emphasize the importance of integrating artificial intelligence into dental education and the value of decision support systems in clinical practice.

Readability refers to the extent to which a text can be easily understood by reader. In addition to factors such as the reader's intelligence level, educational background, environment, interests, and reading purpose, the ideas, vocabulary, style, and formal characteristics of the text are also key factors affecting readability (39). In this study, DeepSeek V3.1's median FRES score was found to be 45.9, which was higher than that of other chatbots. This result shows that DeepSeek V3.1 has the highest readability.

The median FKGL scores of all the chatbots ranged from 10.3 to 13.3, corresponding to a fairly advanced reading level. Similar to our study, Civelek et al.'s study

found that the FKGL scores of all chatbots ranged between 9.42 and 10.70, indicating a very difficult reading level (40). These results show that, in general, the responses generated by chatbots require above-average, advanced reading skills.

Chatbots used for patient information and guidance ideally have high FRES and low FKGL scores to be beneficial to the user (41). Although the content of the questions is directed at physicians, avulsion cases require urgent intervention. Therefore, it is crucial that the prepared texts are highly readable; physicians must be able to understand the text easily and make immediate and accurate interventions without wasting time in such situations.

When examining the SMOG scores, the average readability level of the responses provided by ChatGPT-5 and Gemini 2.5 Pro was significantly higher than that of DeepSeek-V3.1. Similarly, GFI scores also supported this result; it was determined that the texts produced by both models (ChatGPT-5 and Gemini 2.5 Pro) contained higher reading difficulty than the responses provided by DeepSeek-V3.1. This suggests that DeepSeek-V3.1 can produce more accessible and comprehensible texts. In the evaluation of the CLI scores, ChatGPT-5's responses scored higher than those of Gemini 2.5 Pro and DeepSeek-V3.1. This result indicates that ChatGPT-5's responses have a relatively more complex structure. Overall, it can be said that there are significant differences in readability levels among the three models, and DeepSeek-V3.1 in particular offers a user-friendly advantage with lower readability index values. Similar to our study, a study evaluating the performance of advanced artificial intelligence models in dental injuries resulting from trauma during the primary dentition period found that DeepSeek produced the most readable outputs (37). This finding supports the potential of DeepSeek to provide more understandable and practical information for physicians, especially in clinical scenarios requiring urgent intervention.

Artificial intelligence chatbots are inherently probabilistic and generate responses by predicting the likelihood of the next word based on statistical patterns in the training data. They operate differently from conventional deterministic software. As a result, asking the same question twice may not always yield the same answers (42, 43). The reason questions related to avulsion were asked only once to the artificial intelligence models in this study is that the focus was on accuracy and completeness rather than on the consistency of responses. Since avulsion is an emergency situation in dentistry requiring immediate clinical intervention, it is critically important for dentists to access the required information quickly and accurately. Therefore, rather than focusing on whether artificial intelligence provides consistent answers to the same question at different times, the study evaluated whether the information provided at the initial moment was accurate, complete, and clinically applicable. Thus, a research method consistent with the need for dentists to make quick decisions in emergencies was used.

Limitations

This study has several limitations. First, the inclusion of only a limited number of large language models (ChatGPT-5, DeepSeek-V3.1, and Gemini 2.5 Pro) limits the generalizability of the results. Furthermore, as the results are limited to the versions available during the data collection process, future updates to the models may yield different findings. The study's focus solely on avulsion may limit the inferences that can be made regarding other dental subspecialties. Furthermore, the questions were only asked once, and the consistency of the models across different sessions was not examined. Therefore, further research supported by repeated measurements is required to evaluate the stable performance of the models. Future studies could overcome these limitations by using broader model groups, different clinical scenarios, and multicenter designs.

Conclusion

This study aimed to comparatively evaluate three LLMs in terms of the accuracy, completeness, and readability of responses to clinical questions related to avulsion in traumatic dental injuries. The findings showed that ChatGPT-5 performed statistically significantly better in terms of accuracy. In the completeness assessment, the ChatGPT-5 and Gemini 2.5 Pro models scored similarly and highly, while DeepSeek-V3.1 lagged behind with lower scores. In the readability analysis, DeepSeek-V3.1 produced texts that were easier to understand, scoring higher on FRES; however, all models exceeded the 6th–8th grade level recommended for patient information materials in terms of FKGL and SMOG scores. These results demonstrate that LLMs have the potential to provide clinical support in situations requiring urgent intervention, such as traumatic tooth avulsion. However, the accuracy and clinical validity of their responses must be evaluated under expert supervision.

While LLM-based systems hold value as a complementary source of information in dental practice based on the current IADT guidelines, they should not be the sole determinant in clinical decision-making processes.

Disclosures

Peer-review: Externally peer-reviewed.

Author Contributions: Concept – Ö.İ.E.; Design – Ö.İ.E., M.T., Ö.A.; Supervision – H.O.; Materials – Ö.A., H.O.; Data collection &/or processing – M.T., Ö.A.; Analysis and/or interpretation – Ö.İ.E., M.T.; Literature search – M.T., Ö.A., H.O.; Writing – Ö.İ.E.; Critical review – H.O.

Conflict of Interest: There is no any conflict of interest.

Funding: There is no any funding.

References

1. Arılı Öztürk E, Turan Gökduman C, Çanakçi BC. Evaluation of the performance of ChatGPT-4 and ChatGPT-4o as a learning tool in endodontics. Int Endod J. 2025. https://doi.org/10.1111/iej.14217
2. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. J Esthet Restor Dent. 2023;35(7):1098-1102. https://doi.org/10.1111/jerd.13046
3. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930-1940. https://doi.org/10.1038/s41591-023-02448-8
4. Munir I. Artificial intelligence ChatGPT in medicine. Can it be the friend you are looking for? BMANA J. 2023:1-4.
5. Fatani B. ChatGPT for future medical and dental research. Cureus. 2023;15(4):e37983. https://doi.org/10.7759/cureus.37983
6. Haupt CE, Marks M. AI-generated medical advice—GPT and beyond. JAMA. 2023;329(16):1349-1350. https://doi.org/10.1001/jama.2023.5721
7. Chakraborty C, Pal S, Bhattacharya M, Dash S, Lee SS. Overview of chatbots with special emphasis on artificial intelligence-enabled ChatGPT in medical science. Front Artif Intell. 2023;6:1237704. https://doi.org/10.3389/frai.2023.1237704
8. Choi WC, Chang CI. ChatGPT-5 in education: new capabilities and opportunities for teaching and learning. Preprints. 2025. https://doi.org/10.20944/preprints202508.0684.v1
9. Alhur A. Redefining healthcare with artificial intelligence (AI): the contributions of ChatGPT, Gemini, and Co-pilot. Cureus. 2024;16(4):e58892. https://doi.org/10.7759/cureus.58892
10. Taşyürek M, Adıgüzel Ö, Gündoğar M, Goncharuk-Khomyn M, Ortaç H. Comparative evaluation of the responses from ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.1 chatbots to patient inquiries about endodontic treatment in terms of accuracy, understandability, and readability. Int Dent Res. 2025;15(Advanced Online).
11. Paiva L, Luijten G, Puladi B, & Egger J. How does deepseek-r1 perform on usmle?. 2025. medRxiv. https://doi.org/10.1101/2025.02.06.25321749
12. Bal C, Öztürk Z, Atak D, Yaşar M, Aksoy M. Evaluation of parents' awareness levels of traumatic dental injuries: the ToothSOS mobile application and "Save Your Tooth" posters. Int Dent Res. 2023;13(2):60-66. https://doi.org/10.5577/idr.2023.vol13.no2.3
13. Andreasen JO, Andreasen FM, Andersson L. Textbook and Color Atlas of Traumatic Injuries to the Teeth. Hoboken: Wiley-Blackwell; 2018.
14. Fouad AF, Abbott PV, Tsilingaridis G, Cohenca N, Lauridsen E, Bourguignon C, et al. International Association of Dental Traumatology guidelines for the management of traumatic dental injuries: 2. Avulsion of permanent teeth. Dent Traumatol. 2020;36(4):331-342. https://doi.org/10.1111/edt.12573
15. Soğukpınar Önsüren A, İpek S. Are student knowledge levels and attitudes about avulsion dental trauma adequate in the faculty of medicine? Int Dent Res. 2021;11(Suppl 1):143-151.
16. Mustuloğlu Ş, Deniz BP. Evaluation of chatbots in the emergency management of avulsion injuries. Dent Traumatol. 2025;41(4):437-444. https://doi.org/10.1111/edt.13041
17. Rokhshad R, Zhang P, Mohammad-Rahimi H, Pitchika V, Entezari N, Schwendicke F. Accuracy and consistency of chatbots versus clinicians for answering pediatric dentistry questions: a pilot study. J Dent. 2024;144:104938. https://doi.org/10.1016/j.jdent.2024.104938
18. Ozden I, Gokyar M, Ozden ME, Sazak Ovecoglu H. Assessment of artificial intelligence applications in responding to dental trauma. Dent Traumatol. 2024;40(6):722-729. https://doi.org/10.1111/edt.12958
19. De Vito A, Colpani A, Moi G, Babudieri S, Calcagno A, Calvino V, et al. Assessing ChatGPT's potential in HIV prevention communication: a comprehensive evaluation of accuracy, completeness, and inclusivity. AIDS Behav. 2024;28(8):2746-2754. https://doi.org/10.1007/s10461-024-04358-9
20. Coskun BN, Yagiz B, Ocakoglu G, Dalkilic E, Pehlivan Y. Assessing the accuracy and completeness of artificial intelligence language models in providing information on methotrexate use. Rheumatol Int. 2024;44(3):509-515. https://doi.org/10.1007/s00296-023-05283-9
21. Ponnudurai G, Caszo B, Gnanou J. Readability analysis as a tool for evaluating English proficiency in first-year medical students. BMC Med Educ. 2025;25(1):945. https://doi.org/10.1186/s12916-025-05892-3
22. Flesch R. A new readability yardstick. J Appl Psychol. 1948;32(3):221-233. https://doi.org/10.1037/h0057532
23. Kincaid JP, Fishburne RP Jr, Rogers RL, Chissom BS. Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for navy enlisted personnel. 1975.
24. McLaughlin GH. SMOG grading—a new readability formula. J Read. 1969;12(8):639-646.
25. Gunning R. The Technique of Clear Writing. New York: McGraw-Hill; 1952.
26. Olszewski R, Watros K, Mańczak M, Owoc J, Jeziorski K, Brzeziński J. Assessing the response quality and readability of chatbots in cardiovascular health, oncology, and psoriasis: a comparative study. Int J Med Inform. 2024;190:105562. https://doi.org/10.1016/j.ijmedinf.2024.105562
27. Coleman M, Liau TL. A computer readability formula designed for machine scoring. J Appl Psychol. 1975;60(2): 283-284.https://doi.org/10.1037/h0076540
28. Ghanem YK, Rouhi AD, Al-Houssan A, Saleh Z, Moccia MC, Joshi H, et al. Dr. Google to Dr. ChatGPT: assessing the content and quality of artificial intelligence-generated medical information on appendicitis. Surg Endosc. 2024;38(5):2887-2893. https://doi.org/10.1007/s00464-024-10885-3
29. Shrivastava P, Rai A, Injety R, Singh S, Jain A, Mahuli A, et al. Performance of ChatGPT in dentistry. Braz. J. Oral Sci. 2025;24:e254954. https://doi.org/10.20396/bjos.v24i00.8674954
30. Suárez A, Díaz-Flores García V, Algar J, Gómez Sánchez M, Llorente de Pedro M, Freire Y. Unveiling the ChatGPT phenomenon: evaluating the consistency and accuracy of endodontic question answers. Int Endod J. 2024;57(1):108-113. https://doi.org/10.1111/iej.13952
31. Tokgöz Kaplan T, Cankar M. Evidence-based potential of generative artificial intelligence large language models on dental avulsion: ChatGPT versus Gemini. Dent Traumatol. 2025;41(2):178-186. https://doi.org/10.1111/edt.12971
32. Chatzopoulos GS, Koidou VP, Tsalikis L, Kaklamanos EG. Large language models in periodontology: assessing their performance in clinically relevant questions. J Prosthet Dent. 2024;132(6):1329.e1-1329.e6. https://doi.org/10.1016/j.prosdent.2024.07.017

33. Acar AH. Can natural language processing serve as a consultant in oral surgery? J Stomatol Oral Maxillofac Surg. 2024;125(3):101724. https://doi.org/10.1016/j.jormas.2023.101724
34. Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of generative pre-trained transformer 3 (GPT-3) in healthcare delivery. NPJ Digit Med. 2021;4(1):93. https://doi.org/10.1038/s41746-021-00464-x
35. Tassoker M. Who knows anatomy best? A comparative study of ChatGPT-4o, DeepSeek, Gemini, and Claude. Clin Anat. 2025. https://doi.org/10.1002/ca.70012
36. Kaygisiz ÖF, Teke MT. Can DeepSeek and ChatGPT be used in the diagnosis of oral pathologies? BMC Oral Health. 2025;25(1):638. https://doi.org/10.1186/s12903-025-05892-3
37. Sezer B, Aydoğdu T. Performance of advanced artificial intelligence models in traumatic dental injuries in primary dentition: a comparative evaluation of ChatGPT-4 Omni, DeepSeek, Gemini Advanced, and Claude 3.7 in terms of accuracy, completeness, response time, and readability. Appl Sci. 2025;15(14):7778. https://doi.org/10.3390/app15147778
38. Al-Huthaifi BH, Ghwainem AA, Alqarni AS, Alshehri BY, Almnea RA, Alelyani AA, et al. Knowledge, perception, and management toward traumatic tooth avulsion among dental professionals: a cross-sectional study. BMC Med Educ. 2025; 25(1):1206. https://doi.org/10.1186/s12909-025-07791-7
39. Momenaei B, Wakabayashi T, Shahlaee A, Durrani AF, Pandit SA, Wang K, et al. Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases. Ophthalmol Retina. 2023;7(10):862-868. https://doi.org/10.1016/j.oret.2023.05.014
40. Özcivelek T, Özcan B. Comparative evaluation of responses from DeepSeek-R1, ChatGPT-o1, ChatGPT-4, and dental GPT chatbots to patient inquiries about dental and maxillofacial prostheses. BMC Oral Health. 2025;25(1):871. https://doi.org/10.1186/s12903-025-06267-w
41. Helvacioglu-Yigit D, Demirturk H. Evaluating artificial intelligence chatbots for patient education in oral and maxillofacial radiology. Oral Surg Oral Med Oral Pathol Oral Radiol. 2025;139(6):750-759. https://doi.org/10.1016/j.oooo.2025.01.012
42. Taşyürek M, Adıgüzel Ö, Cangül S, & Ortaç H. Comparative evaluation of the accuracy and completeness of responses provided by ChatGPT-5, Gemini 2.5 Flash, Grok 4, and DeepSeek-R1 to open-ended questions on cracked and fractured teeth. J Med Dent Invest. 2025;6:e250161.
43. Ghahramani Z. Probabilistic machine learning and artificial intelligence. Nature. 2015;521(7553):452-459. https://doi.org/10.1038/nature14541