applied sciences

MDPI

Article

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

Berkant Sezer * and Tuğba Aydoğdu

Department of Pediatric Dentistry, School of Dentistry, Çanakkale Onsekiz Mart University, 17020 Çanakkale, Türkiye
* Correspondence: dt.berkantsezer@gmail.com or berkant.sezer@comu.edu.tr

Abstract

This study aimed to evaluate and compare the performance of four advanced artificial intelligence-powered chatbots—ChatGPT-4 Omni (ChatGPT-4o), DeepSeek, Gemini Advanced, and Claude 3.7 Sonnet—in responding to questions related to traumatic dental injuries (TDIs) in the primary dentition. The assessment focused on accuracy, completeness, readability, and response time, aligning with the 2020 International Association of Dental Traumatology guidelines. Twenty-five open-ended TDI questions were submitted to each model in two separate sessions. Responses were anonymized and evaluated by four pediatric dentists. Accuracy and completeness were rated using Likert scales; readability was assessed using five standard indices; and response times were recorded in seconds. ChatGPT-4o demonstrated significantly higher accuracy than Gemini Advanced ($p = 0.005$), while DeepSeek outperformed Gemini Advanced in completeness ($p = 0.010$). Response times differed significantly ($p < 0.001$), with DeepSeek being the slowest and ChatGPT-4o and Gemini Advanced being the fastest. DeepSeek produced the most readable outputs relatively, though none met public readability standards. Claude 3.7 generated the most complex texts ($p < 0.001$). A strong correlation existed between accuracy and completeness ($\rho = 0.701$, $p < 0.001$). These findings emphasize the cautious integration of artificial intelligence chatbots into pediatric dental care due to varied performance. Clinical accuracy, completeness, and readability are critical when offering information aligned with guidelines to support decisions in dental trauma management.

Keywords: artificial intelligence; chatbot; dental trauma; traumatic dental injuries; primary dentition; large language models; dental trauma; ChatGPT; DeepSeek; Claude; Gemini; accuracy; completeness

check for updates

Academic Editor: Vittorio Checchi

Received: 18 May 2025
Revised: 27 June 2025
Accepted: 7 July 2025
Published: 11 July 2025

Citation: 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*, 7778. https://doi.org/10.3390/app15147778

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Artificial intelligence (AI) encompasses a wide range of technologies that enable computer systems to perform tasks traditionally associated with human intelligence—such as learning, reasoning, interpretation, and data analysis—without direct human intervention [1]. A subset of AI, machine learning (ML), refers to systems that can identify patterns and learn from data without relying solely on predefined algorithms or manual input [2]. Recent advances in AI and ML have significantly contributed to fields such

as natural language processing (NLP) and deep learning, leading to the development of models capable of analyzing and interpreting complex textual information. These models have demonstrated substantial potential in healthcare applications, including responding to medical inquiries, assisting in diagnoses, and supporting treatment planning [3].

Chat Generative Pre-trained Transformer (ChatGPT), developed by OpenAI and released in November 2022, is based on the GPT-4 architecture and is considered one of the most advanced autoregressive language models to date [4–6]. Similarly, Google introduced its AI chatbot Bard in 2023, followed by the release of Gemini in 2024—an advanced multimodal platform capable of processing complex data types, including visual and graphical inputs [7,8]. Developed in China, the AI model DeepSeek-R1 garnered global attention following the launch of its chatbot in December 2024. The model is distinguished by its focus on transparency, reproducibility, accessibility, and affordability, which have contributed to its rapid and widespread adoption worldwide [9,10]. In the same year, Anthropic released Claude 3.7 Sonnet, a model designed to achieve high performance in language comprehension and multimodal data analysis [11,12].

Traumatic dental injuries (TDIs) are highly prevalent among children and young adults, accounting for approximately 5% of all bodily injuries. One in four school-aged children experiences at least one dental trauma, while nearly one-third of adults report a history of trauma to their permanent teeth—most of which occur before the age of 19 [13]. In early childhood, falls and collisions are common due to the ongoing development of motor control and coordination. These everyday incidents frequently result in injuries to the primary dentition. However, not all oral injuries are accidental; some may be caused by adverse events such as child abuse, motor vehicle accidents, or other external causes [14]. Mismanagement of TDIs often results from inadequate initial clinical assessment. Prompt intervention and appropriate treatment are crucial for improving the prognosis of traumatized teeth [15]. Accurate diagnosis and timely management of acute TDIs are essential for ensuring the long-term survival of affected teeth and for supporting the normal development of the dentoalveolar complex [16]. Successful treatment outcomes depend not only on the clinician's expertise but also on how promptly the injury is recognized and managed at the time of occurrence. Despite the critical importance of timely decision-making, diagnostic and management errors remain common in TDI cases, especially when providers lack immediate access to guidelines or expert consultation. In this context, AI tools—particularly large language models (LLMs)—have emerged as potential clinical support systems in dental traumatology [17,18]. In pediatric dentistry, however, the number of AI validation studies remains scarce. While a few recent papers have explored LLMs' performance in general dental trauma scenarios [15,16,19,20], there is a notable gap in research specifically targeting TDIs in the primary dentition. Given the anatomical and clinical differences between primary and permanent teeth—as well as the specific decision-making nuances involved—this area warrants further investigation.

To build a clinically meaningful understanding of AI utility in pediatric dental trauma, it is essential to move beyond general evaluations and examine model performance within the specific anatomical and diagnostic context of primary teeth. This distinction is critical because the anatomical and treatment considerations for primary teeth differ significantly from those of permanent dentition. Moreover, early childhood presents unique challenges in communication, diagnosis, and parental anxiety, which increase the demand for accessible, accurate, and comprehensible information. Although LLMs are increasingly consulted by the general public, this study specifically aimed to evaluate their utility in clinical scenarios involving pediatric dental trauma, with a primary focus on professional users—such as general dentists, emergency physicians, or non-specialist clinicians—who may encounter such cases in urgent care settings. Therefore, a focused assessment of LLM performance in

this pediatric dental context is warranted to determine whether these tools can serve as a reliable adjunct when immediate access to professional consultation is unavailable.

In pediatric dental trauma care, decision-making often unfolds under time pressure, heightened emotional stress, and limited specialist access. Therefore, evaluating AI-generated responses through a multidimensional lens—combining accuracy, completeness, readability, and response time—aligns with the complex real-world needs of this field. While accuracy ensures factual reliability, completeness safeguards clinical safety by minimizing omissions. Readability is crucial for facilitating effective communication with non-specialist caregivers, and response time reflects the practical feasibility of real-time AI assistance in urgent situations. Together, these parameters reflect a holistic framework for assessing AI's potential to serve as a supportive tool in pediatric dental trauma management, rather than merely as an informational resource.

This study aims to evaluate and compare the performance of four advanced LLMs—ChatGPT-4 Omni (ChatGPT-4o), DeepSeek, Gemini Advanced, and Claude 3.7 Sonnet—in answering questions related to TDIs in the primary dentition. The evaluation is based on five key criteria: accuracy, completeness, readability, response time, and adherence to the most recent (2020) guideline of the International Association of Dental Traumatology (IADT) [21]. Accordingly, the following null hypotheses were formulated:

- H_{01} (Accuracy): There is no statistically significant difference among ChatGPT-4o, DeepSeek, Gemini Advanced, and Claude 3.7 Sonnet in terms of the accuracy of their responses;
- H_{02} (Completeness): There is no statistically significant difference among the models in terms of the completeness of the provided information;
- H_{03} (Readability): There is no statistically significant difference among the models regarding the readability of their responses;
- H_{04} (Response time): There is no statistically significant difference among the models in terms of their average response time.

2. Materials and Methods

2.1. Study Design and Ethical Approval

This study was designed as a comparative cross-sectional investigation aimed at evaluating and contrasting the performance of four advanced AI models—ChatGPT-4o, DeepSeek, Gemini Advanced, and Claude 3.7—in the context of TDIs affecting the primary dentition. These models were selected based on their current prominence, accessibility, and advanced NLP capabilities. ChatGPT-4o and Gemini Advanced are widely recognized for their real-time usability and integration into major platforms, while Claude 3.7 and DeepSeek represent newer, emerging alternatives that offer strong performance and competitive reasoning abilities. The methodology adhered to the updated guidelines of the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement and followed the recommendations of the EQUATOR Network. Data collection was completed in April 2025. As the study did not involve human subjects, clinical interventions, or the use of patient data, ethical approval by an institutional review board was not required.

Among the publicly accessible advanced LLMs available at the time of data collection, ChatGPT-4o, DeepSeek, Gemini Advanced, and Claude 3.7 were selected based on their usability, multimodal support, and emerging use in healthcare contexts. While other models such as Meta's LLaMA 3 or Mistral were technically available, they were not user-facing or optimized for conversational clinical queries. Similarly, Microsoft's Copilot platform is built on GPT-4 but does not provide transparent standalone outputs suitable for reproducible evaluation. A summary comparison of these models is presented in Table 1.

Table 1. Overview of large language models considered for inclusion in the study and rationale for final selection.

Model	Developer	Release Date	Multimodal	Public Access	Healthcare Use Cases	Included
ChatGPT-4o	OpenAI	May 2024	Yes	Yes (paid)	Extensive use in medical Q&A	✓
DeepSeek	DeepSeek AI	Dec 2024	No	Yes (free)	Growing number of studies in China	✓
Gemini Advanced	Google	Feb 2024	Yes	Yes (paid)	Clinical scenario support (limited)	✓
Claude 3.7 Sonnet	Anthropic	Mar 2024	Yes	Yes (paid)	Used in diagnostic case evaluations	✓
Copilot	Microsoft	Mar 2024	Yes	Limited	Common in documentation, no output	✗
Mistral	Mistral AI	Nov 2023	No	Yes (open)	No known medical applications	✗
LLaMA 3	Meta	Apr 2024	No	Yes (open)	Research-focused, not user-facing	✗

ChatGPT-4o: ChatGPT-4 Omni, AI: Artificial Intelligence, Q&A: Question and Answer, Dec: December, Feb: February, Mar: March, Nov: November, Apr: April.

2.2. Question Development and Content

The questions used in this study were developed based on the most recent clinical guideline on TDIs in the primary dentition, published by the IADT [21]. Two experienced pediatric dentists independently reviewed the guideline and collaboratively designed an initial pool of 33 questions intended to evaluate the performance of advanced AI models in terms of accuracy, completeness, response time, and readability.

To ensure clarity and eliminate redundancy, a second independent review was conducted by two additional pediatric dentists. During this phase, questions that overlapped in content or could be merged without compromising clinical relevance were identified and revised. As a result of this refinement process, a final set of 25 distinct questions was established. These standardized questions were subsequently submitted to each of the four AI models for evaluation. To enhance content validity, the final set of 25 questions was subjected to an informal pilot review by the expert panel. Although no formal content validity index or inter-rater reliability metrics were calculated at this stage, the panel confirmed the clarity, representativeness, and clinical relevance of each item through iterative discussion and consensus. This process ensured that the question set appropriately covered the key domains of TDIs in primary dentition. While the inclusion or exclusion of certain questions may influence specific performance scores to some extent, the current set was intentionally designed to reflect a broad and clinically relevant sample of real-world scenarios. Therefore, the core trends observed in model behavior—particularly regarding accuracy and guideline adherence—are expected to remain robust even with minor modifications to the item set. The complete list of questions is presented in Table 2, while the questions along with the corresponding responses generated by the AI models are provided in the Supplementary Materials.

Table 2. Questions list.

	Questions.
1.	In which types of traumatic dental injuries in the primary dentition is radiographic examination recommended, and which types of radiographic approaches should be used?
2.	Following which types of traumatic dental injuries in the primary dentition is the risk of developmental anomalies in the permanent teeth higher, and what types of pathologies may occur in the permanent teeth as a result of dental trauma?
3.	According to the current International Association of Dental Traumatology guidelines, what is recommended when the root of a traumatized primary tooth is displaced toward the permanent tooth germ in cases of intrusion and lateral luxation injuries?

Table 2. *Cont.*

	Questions.
4.	What recommendations should be given to parents regarding the management of acute symptoms following traumatic dental injuries such as intrusion, lateral luxation, and root fractures in primary teeth?
5.	In cases of avulsion of a primary tooth, should the tooth be replanted?
6.	In which situations may systemic antibiotic use be necessary following traumatic dental injuries to primary teeth?
7.	How long does it typically take for an intruded primary tooth to reposition spontaneously?
8.	What procedures should be performed if there is contamination in the trauma area in cases of traumatic dental injuries to the primary dentition?
9.	How should oral hygiene be maintained at home following traumatic dental injuries to the primary dentition, and what care instructions should be provided to parents?
10.	According to the current International Association of Dental Traumatology guidelines, what clinical approach should be followed in cases of intrusion injuries in the primary dentition?
11.	According to the current International Association of Dental Traumatology guidelines, what clinical approach should be followed in cases of lateral luxation injuries in the primary dentition?
12.	According to the current International Association of Dental Traumatology guidelines, what clinical approach should be followed in cases of extrusion injuries in the primary dentition?
13.	According to the current International Association of Dental Traumatology guidelines, how should the frequency and duration of follow-up be determined based on the type of injury in the primary dentition?
14.	According to the current International Association of Dental Traumatology guidelines, what is the recommended treatment approach when an incisor in the primary dentition presents with an enamel fracture?
15.	According to the current International Association of Dental Traumatology guidelines, what evaluation process should be followed if the dental trauma history does not match the clinical findings in cases involving the primary dentition?
16.	In traumatic dental injuries involving the primary dentition, which type of tissue injury is most commonly observed?
17.	According to the current International Association of Dental Traumatology guidelines, what is the recommended treatment approach for a primary tooth with a complicated crown fracture?
18.	According to the current International Association of Dental Traumatology guidelines, what is the recommended treatment approach for a primary tooth with a crown and root fracture?
19.	According to the current International Association of Dental Traumatology guidelines, what are the recommended treatment approaches for a primary tooth with a root fracture, depending on whether the coronal fragment is displaced or not?
20.	According to the current International Association of Dental Traumatology guidelines, what is the recommended treatment approach when an alveolar fracture occurs in the primary dentition?
21.	According to the current International Association of Dental Traumatology guidelines, what is the recommended treatment approach when there is bleeding from the gingival sulcus without displacement or occlusal interference following trauma?

Table 2. *Cont.*

Questions.	
22.	According to the current International Association of Dental Traumatology guidelines, what is the recommended treatment approach for a traumatized primary maxillary central incisor with mild extrusion (less than 3 mm) and slight mobility?
23.	According to the current International Association of Dental Traumatology guidelines, what is the recommended treatment approach for a traumatized primary maxillary central incisor with severe extrusion (greater than 3 mm) and advanced mobility?
24.	When the root of an intruded primary tooth is displaced buccally, how does it appear radiographically, and what is the recommended treatment approach according to the current International Association of Dental Traumatology guidelines?
25.	When the root of an intruded primary tooth is displaced palatally, how does it appear radiographically, and what is the recommended treatment approach according to the current International Association of Dental Traumatology guidelines?

2.3. Primary and Secondary Outcomes

The primary objective of this study was to compare the performance of four advanced AI models—ChatGPT-4o, DeepSeek, Gemini Advanced, and Claude 3.7—in responding to questions regarding TDIs in the primary dentition. Specifically, the responses were evaluated for accuracy, completeness, and readability using predefined standardized criteria.

Secondary outcomes involved exploring potential correlations among these three performance metrics and the response time of each model. Additionally, a comparative analysis was conducted to examine differences in response times across the four AI models. A comprehensive overview of the study workflow, including all major stages from question development to statistical analysis, is provided in Figure 1.

2.4. Querying Procedure

The latest premium versions of ChatGPT-4o (OpenAI, San Francisco, CA, USA), DeepSeek (Hangzhou DeepSeek Artificial Intelligence Basic Technology Research Co., Ltd.; Beijing, China), Gemini Advanced (Google, Menlo Park, CA, USA), and Claude 3.7 Sonnet (Anthropic, San Francisco, CA, USA) were utilized in this study. Each of the 25 standardized questions was entered individually into the chat interfaces of the four models, using English to ensure accurate interpretation and minimize linguistic ambiguity. To prevent contextual carryover and memory effects, each question was submitted in a new, independent session. Before querying, browser cookies were cleared and each model was accessed through a separate user account to eliminate any potential influence from prior interactions or personalization.

All questions were input without any alterations to wording, punctuation, or syntax. No prompt engineering, pre-testing, or optimization strategies were applied in order to simulate a typical end-user experience.

To ensure procedural consistency, a single researcher (T.A.) submitted each question to all four chatbots on two separate occasions, with a one-week interval between the sessions. The initial responses were used for the primary analysis, while the second set was used to assess consistency and reliability. Agreement between the two rounds of responses was evaluated using Cohen's Kappa (κ), yielding the following values: ChatGPT-4o ($\kappa = 0.97$), DeepSeek ($\kappa = 0.89$), Gemini Advanced ($\kappa = 0.95$), and Claude 3.7 ($\kappa = 0.92$).

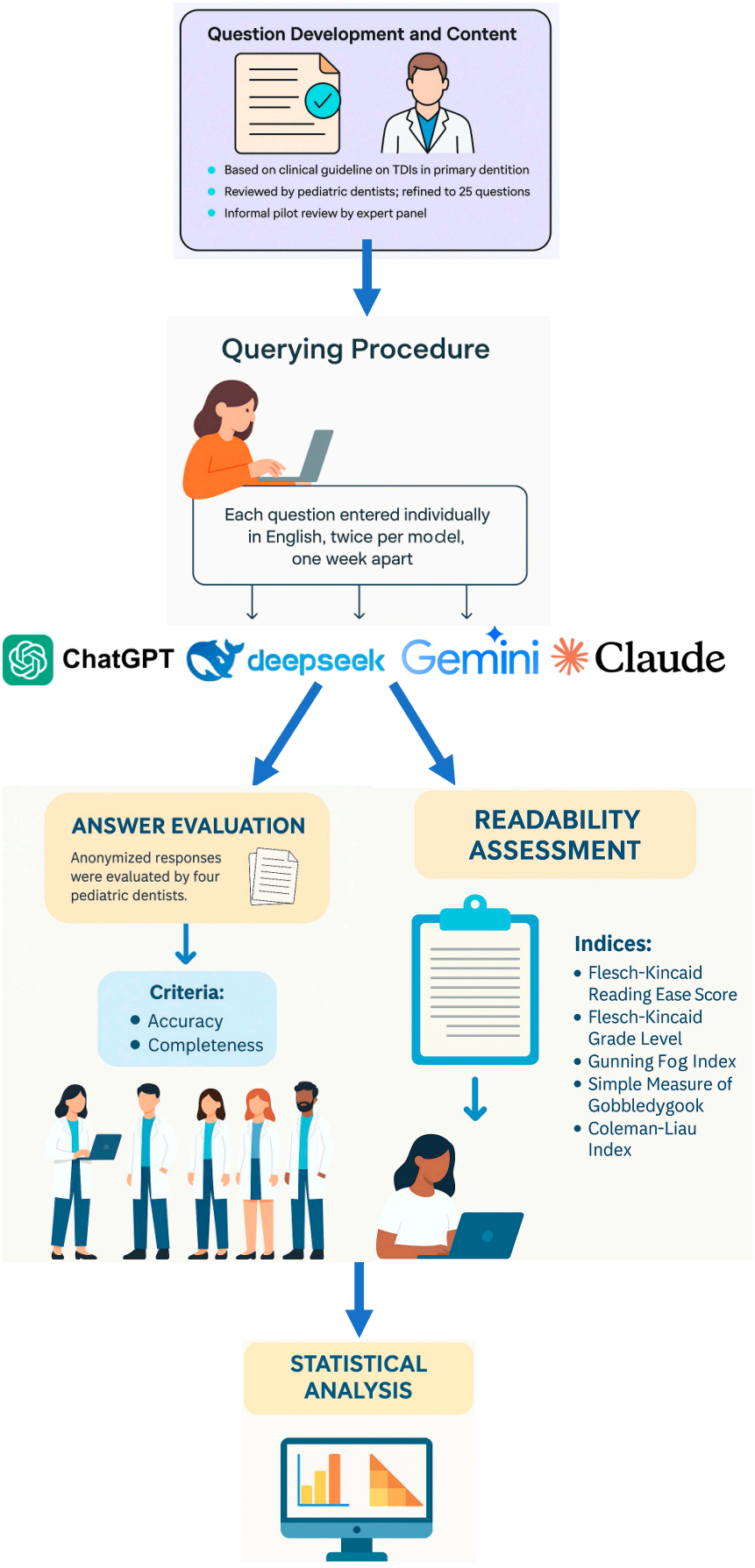


Figure 1. Overview of the study design and evaluation workflow.

Response time for each model—measured from the moment the question was entered to the completion of the response—was recorded in seconds using an online stopwatch. All outputs were compiled by the same researcher in a Microsoft Excel spreadsheet. To eliminate potential rater bias, all responses were anonymized, and no chatbot identifiers were included during the subsequent evaluation phase.

Additionally, all collected responses were independently reviewed twice by the principal investigator (B.S.), with a one-week interval between reviews, to verify the absence of hallucinated or fabricated information.

2.5. Answers Evaluation

After collecting the responses from all four AI models, the anonymized outputs were distributed to four pediatric dentists, each with at least 10 years of academic and clinical experience. To ensure unbiased evaluation, the identity of the generating model was concealed for all responses. Evaluators were provided with the current IADT guidelines on TDIs in primary teeth [21], which served as the reference standard for their assessments. In addition to the anonymized responses and the guidelines, a standardized Microsoft Excel spreadsheet was supplied to each evaluator for structured scoring.

To assess intra-rater reliability, each expert independently evaluated the full set of responses twice, with a one-week interval between assessments. In cases where significant discrepancies occurred between the two rounds, the corresponding question-response pairs were re-sent for clarification and re-evaluation. The final score from this third review was accepted as the definitive rating for each item. Additionally, if disagreements persisted or if scoring discrepancies greater than one point remained after the third review, a structured consensus meeting was conducted involving all four pediatric dentists. During these sessions, evaluators discussed their rationale in a blinded manner and reached a final score through majority agreement. This procedure ensured scoring consistency and helped minimize subjectivity across all AI-generated responses.

Two key dimensions were employed in the manual evaluation process: accuracy and completeness.

(i) **Accuracy** was assessed using a five-point Likert scale [22,23]:

1 = *Poor*: The response lacked factual accuracy and coherence; most essential details were missing.
2 = *Fair*: Some accurate elements were present, but important content was either missing or unclear.
3 = *Moderate*: Partially accurate, with a mix of correct and missing or vague information.
4 = *Good*: Mostly accurate and coherent, with only minor omissions.
5 = *Excellent*: Highly accurate, comprehensive, and well-organized response.

(ii) **Completeness** was rated using a three-point Likert scale [22,23]:

1 = *Incomplete*: Key aspects of the question were not addressed.
2 = *Adequate*: All major components of the question were addressed.
3 = *Comprehensive*: The response addressed all required elements and included additional relevant insights.

To further clarify how expert ratings were applied in practice, the following example illustrates the evaluation process for the question:

> *"In which types of traumatic dental injuries in the primary dentition is radiographic examination recommended, and which types of radiographic approaches should be used?"*

A response that included specific types of TDIs where radiographs are indicated (e.g., intrusion, lateral luxation, suspected root fractures), explained the rationale, and referenced

appropriate imaging modalities such as periapical or occlusal radiographs—while also addressing the Acceptable, being Indication-oriented and Patient-specific (ALADAIP) principle—was scored as:

- Accuracy: 5 (Excellent)
- Completeness: 3 (Comprehensive)

A response that listed some appropriate injury types (e.g., intrusion, avulsion) and mentioned "periapical X-ray" without elaboration or missed key concepts like ALADAIP was scored as:

- Accuracy: 3–4 (Moderate to Good)
- Completeness: 2 (Adequate)

A response that recommended radiographs for all types of injuries indiscriminately or suggested using cone beam computed tomography (CBCT) without justification or context was scored as:

- Accuracy: 2 (Fair)
- Completeness: 1 (Incomplete)

This example demonstrates how reviewers applied the Likert scales based on guideline adherence, depth of explanation, and the inclusion of critical concepts such as appropriate imaging types and justification for use.

To assess the reliability of expert ratings, Fleiss' kappa statistics were computed across all four evaluators for both accuracy and completeness scores. The resulting kappa values were 0.79 for accuracy and 0.73 for completeness, indicating substantial inter-rater agreement. These results support the robustness of the subjective evaluation process used in this study.

In parallel with expert evaluation, the readability of each chatbot's response was assessed using the online tool https://readable.com. Readability scores were determined based on five established metrics:

- **Flesch-Kincaid Reading Ease Score (FRES)**: Calculated using the average number of syllables per word and words per sentence. The FRES produces a score ranging from 0 to 100, with higher values indicating easier-to-read text [24,25].
- **Flesch-Kincaid Grade Level (FKGL)**: Estimates the United States school grade level required to comprehend the text, based on syllables per word and sentence length [26].
- **Gunning Fog Index (GFI)**: Provides a score typically between 6 and 17, where 6 corresponds to the reading level of an 11–12-year-old, 12 to a high school graduate, and 17 to a college graduate [24].
- **Simple Measure of Gobbledygook (SMOG) Index**: Considered the gold standard for evaluating healthcare materials, as it estimates the reading level required for 100% comprehension [25].
- **Coleman-Liau Index (CLI)**: Calculates readability based on the average number of characters per 100 words and the average sentence length, producing a United States grade-level score; lower values denote easier readability [26].

These readability metrics were used to assess the accessibility of AI-generated content for general audiences, with a particular focus on healthcare-related communication. Clinical readability guidelines, including those from the National Institutes of Health and the American Medical Association, recommend that patient-facing materials be written at a sixth- to eighth-grade reading level. Accordingly, we used the following thresholds: FKGL ≤ 8, SMOG < 9, GFI ≤ 12, CLI ≤ 8, and FRES ≥ 60 to determine the proportion of chatbot responses that met these standards [24–26].

2.6. Statistical Analysis

All data were recorded in Microsoft Excel (version 16; Microsoft Corp., Redmond, WA, USA) and analyzed using IBM SPSS Statistics (version 26; IBM Corp., Chicago, IL, USA). Correlation heatmaps were generated using GraphPad Prism software (version 10; GraphPad Software, San Diego, CA, USA). Descriptive statistics for continuous variables were reported as means ± standard deviations and as medians with minimum and maximum values to provide a comprehensive summary of the data distribution. The normality of data was assessed using the Shapiro–Wilk test, as well as skewness and kurtosis values. Based on these assessments, it was determined that accuracy, completeness, response time, and CLI scores were normally distributed, whereas FRES, FKGL, GFI, and SMOG Index scores did not follow a normal distribution. Accordingly, statistical comparisons were conducted using tests appropriate for each variable's distribution. For normally distributed variables, a one-way analysis of variance (ANOVA) was performed, followed by Tukey's post hoc test when significant differences were found. For non-normally distributed variables, the Kruskal–Wallis test was used, with Dunn's post hoc test applied to identify significant pairwise differences. To evaluate the strength and direction of associations among evaluation metrics (accuracy, completeness, response time, and readability scores), Spearman's rank correlation coefficient (ρ) was employed due to the presence of non-normally distributed variables. Additionally, effect sizes for between-group comparisons were calculated using eta-squared (η^2) for both parametric and non-parametric analyses in order to assess the practical significance of observed differences. A p-value of less than 0.05 was considered statistically significant for all analyses.

3. Results

The accuracy and completeness scores, along with the response times of ChatGPT-4o, Deepseek, Gemini Advanced, and Claude 3.7 in answering the 25 questions on TDIs in primary dentition, are summarized in Table 3. Statistically significant differences were observed in the accuracy scores among the four chatbots (p = 0.005). Post hoc analysis indicated that ChatGPT-4o (4.74 ± 0.66) achieved significantly higher accuracy scores than Gemini Advanced (4.3 ± 1.01). Similarly, statistically significant differences were found in the completeness scores (p = 0.010), with DeepSeek (2.78 ± 0.46) scoring significantly higher than Gemini Advanced (2.56 ± 0.57). Significant variation was also observed in response times across the models (p < 0.001). DeepSeek (69 ± 21 s) had the longest response time, while ChatGPT-4o (9.37 ± 3.78 s) and Gemini Advanced (7.32 ± 2.43 s) generated the fastest responses. Figure 2 presents the accuracy scores of the four AI models, Figure 3 displays their completeness scores, and Figure 4 illustrates their response times.

Table 3. Comparison of accuracy and completeness scores and response time for ChatGPT-4 Omni, DeepSeek, Gemini Advanced, and Claude 3.7 in answering traumatic dental injuries in primary dentition questions.

		Chatbot				
		ChatGPT-4 Omni	**DeepSeek**	**Gemini Advanced**	**Claude 3.7**	***p*-Value ***
Accuracy	Mean ± SD Median (min.–max.)	4.74 ± 0.66 [a] 5 (2–5)	4.42 ± 0.92 [ab] 5 (2–5)	4.3 ± 1.01 [b] 5 (1–5)	4.44 ± 0.97 [ab] 5 (1–5)	**0.005**
Completeness	Mean ± SD Median (min.–max.)	2.74 ± 0.44 [ab] 3 (2–3)	2.78 ± 0.46 [a] 3 (1–3)	2.56 ± 0.57 [b] 3 (1–3)	2.74 ± 0.52 [ab] 3 (1–3)	**0.010**
Response Time	Mean ± SD Median (min.–max.)	9.37 ± 3.78 [a] 8.87 (4.7–23.32)	69 ± 21 [b] 63.61 (43.55–128)	7.32 ± 2.43 [a] 7.15 (3.56–13.53)	14.1 ± 4.22 [c] 12.53 (8.56–25.03)	**<0.001**

SD: standard deviation, min: minimum, max: maximum. The response time was measured in seconds, with a maximum of two decimal places. * One-way ANOVA with post hoc Tukey's test. Different superscript lower letters indicate a significant difference within relevant rows (p < 0.05).

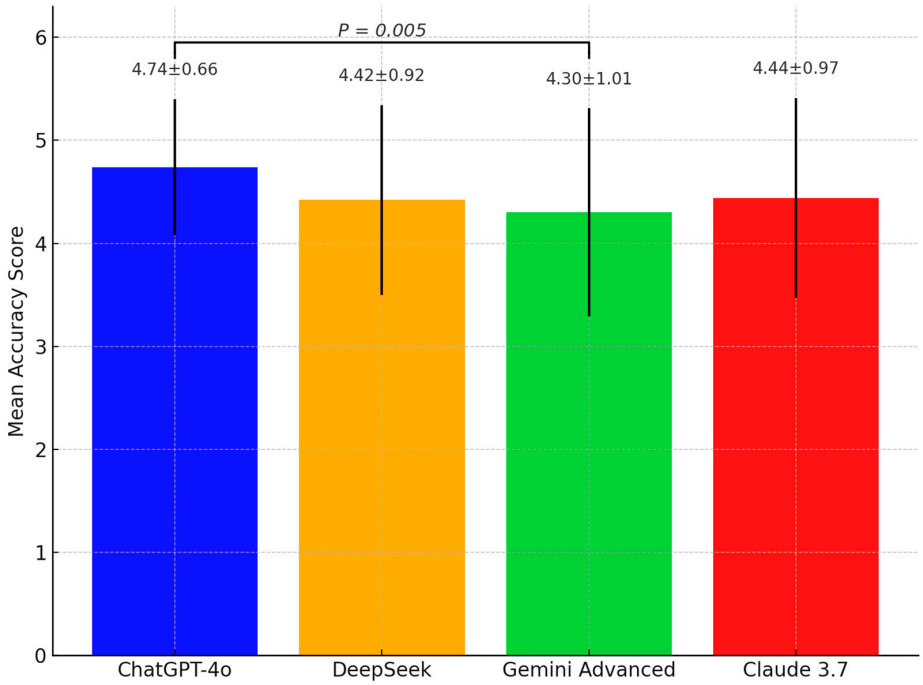


Figure 2. Mean accuracy scores of ChatGPT-4o, DeepSeek, Gemini Advanced, and Claude 3.7 in responding to questions related to traumatic dental injuries in primary dentition.

In addition to statistical significance, effect sizes were calculated using eta-squared (η^2) to quantify the magnitude of between-group differences. The η^2 values for accuracy (0.032) and completeness (0.028) were small, while the response time dimension demonstrated a very large effect ($\eta^2 = 0.846$), underscoring substantial variation in processing latency across chatbot models. These results provide not only statistically significant but also practically meaningful distinctions among the evaluated models.

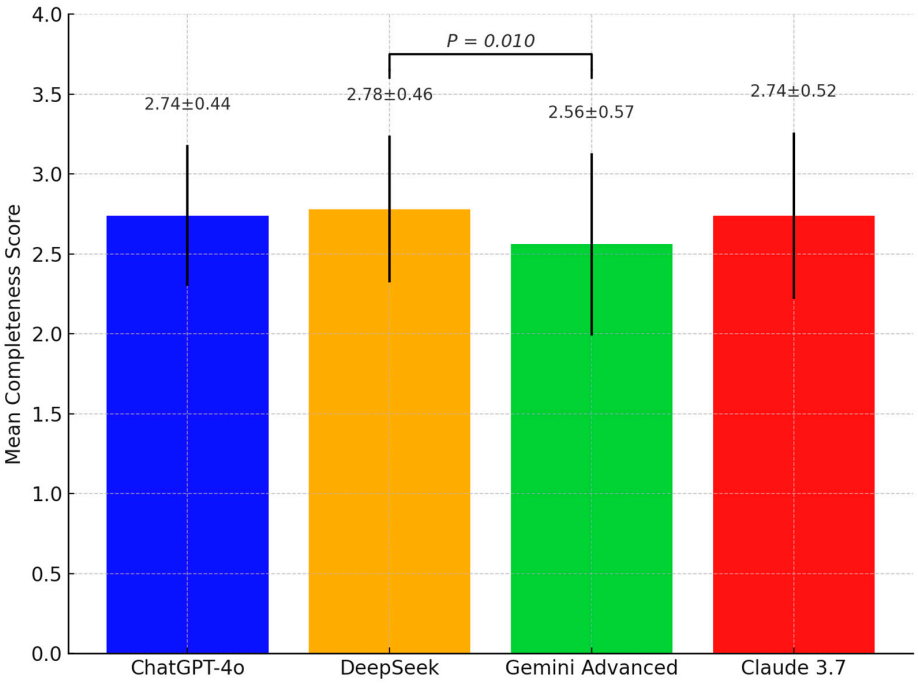


Figure 3. Mean completeness scores of ChatGPT-4o, DeepSeek, Gemini Advanced, and Claude 3.7 in responding to questions related to traumatic dental injuries in primary dentition.

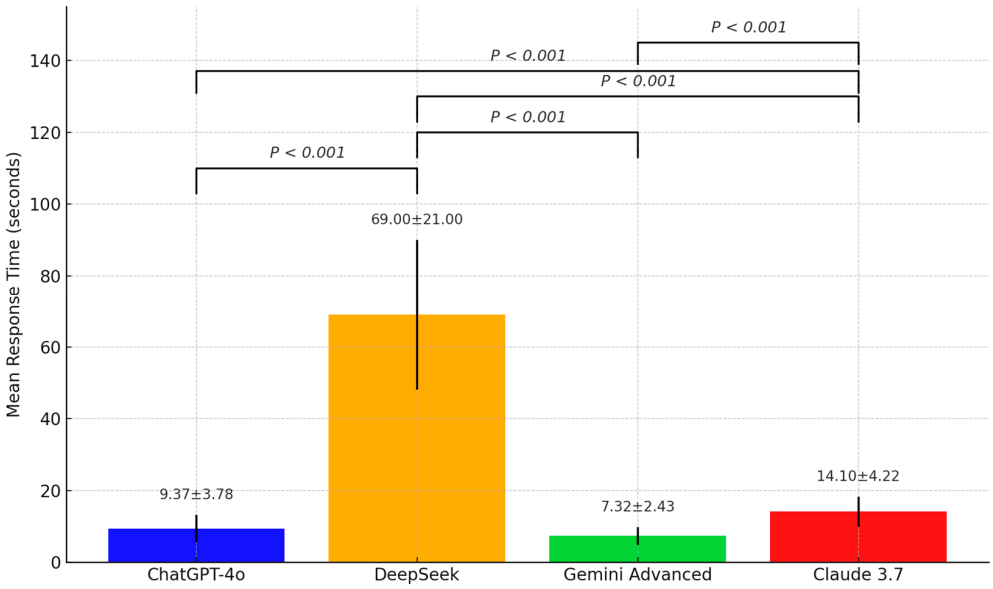


Figure 4. Mean response times of ChatGPT-4o, DeepSeek, Gemini Advanced, and Claude 3.7 in responding to questions related to traumatic dental injuries in primary dentition.

The readability scores of the four chatbots, based on various readability assessment criteria, are presented in Table 4. According to the FRES, Deepseek (41) achieved the highest median score, while Claude 3.7 (4.5) recorded the lowest, with a statistically significant difference ($p < 0.001$). In terms of the FKGL, Claude 3.7 (19.8) demonstrated a significantly higher median grade level, whereas DeepSeek (11.1) had the lowest ($p < 0.001$). For the GFI and SMOG Index, Claude 3.7 also produced significantly higher median scores than the other chatbots (24 and 18.5, respectively) ($p < 0.001$). Regarding the CLI, DeepSeek (15.49 ± 1.35) showed a significantly lower mean score compared to the other chatbots ($p < 0.001$).

Table 4. Comparison of readability scores for ChatGPT-4 Omni, DeepSeek, Gemini Advanced, and Claude 3.7 in answering traumatic dental injuries in primary dentition questions.

		Chatbot				
		ChatGPT-4 Omni	**DeepSeek**	**Gemini Advanced**	**Claude 3.7**	***p*-Value**
FRES	Mean ± SD Median (min.–max.)	18.16 ± 18.92 24 (−46.1–46.8) [a]	39.69 ± 8.71 41 (20.9–58.6) [b]	26.89 ± 11.14 24.7 (9.9–50.3) [c]	−15.23 ± 53.88 4.5 (−171.7–32.9) [d]	**<0.001** [†]
FKGL	Mean ± SD Median (min.–max.)	16.72 ± 5.61 15.8 (10.9–39.8) [a]	11.33 ± 1.53 11.1 (7.9–14.3) [b]	14.28 ± 2.12 14.1 (10.2–17.9) [c]	29.5 ± 21.25 19.8 (12.8–90.1) [d]	**<0.001** [†]
GFI	Mean ± SD Median (min.–max.)	19.71 ± 6.2 18.9 (11.7–45.2) [a]	15 ± 2.03 15 (9.7–17.8) [b]	17.66 ± 2.91 17.1 (13–22.6) [c]	33.23 ± 21.61 24 (15.1–96.4) [d]	**<0.001** [†]
SMOG	Mean ± SD Median (min.–max.)	15.1 ± 4.09 15.1 (9.9–31.1) [a]	11.01 ± 1.38 11 (7.6–12.9) [b]	13.51 ± 2.02 13.3 (9.7–17.6) [c]	21.36 ± 8.99 18.5 (11.6–47.1) [d]	**<0.001** [†]
CLI	Mean ± SD Median (min.–max.)	16.67 ± 2.05 [ac] 16.1 (14.1–20.6)	15.49 ± 1.35 [b] 15.5 (13.1–18.2)	16.2 ± 1.87 [c] 16.6 (10.5–19.3)	16.94 ± 2.11 [a] 17.1 (11.9–21.6)	**<0.001** *

SD: standard deviation, min: minimum, max: maximum, FRES: Flesch–Kincaid Reading Ease Score, FKGL: Flesch–Kincaid Grade Level, GFI: Gunning Fog Index, SMOG: Simple Measure of Gobbledygook Index, CLI: Coleman–Liau Index. [†] Kruskal–Wallis with post hoc Dunn test. * One-way ANOVA with post hoc Tukey's test. Different superscript lower letters indicate a significant difference within relevant rows ($p < 0.05$).

Eta-squared (η^2) values were calculated for each of the five readability indices to determine the magnitude of inter-model differences. All indices showed meaningful variation. The FRES ($\eta^2 = 0.325$), FKGL ($\eta^2 = 0.284$), GFI ($\eta^2 = 0.279$), and SMOG ($\eta^2 = 0.363$) indices all

indicated large effect sizes, suggesting considerable divergence in the linguistic complexity of chatbot outputs. The CLI yielded a moderate effect size ($\eta^2 = 0.081$), supporting the presence of stylistic and syntactic differences in how models construct their responses. These effect size findings reinforce the clinical importance of readability variations, especially when chatbot outputs are evaluated for end-user comprehension in health contexts.

In addition to reporting mean readability scores, we analyzed the proportion of chatbot responses that met commonly accepted clinical readability thresholds. According to health communication guidelines, educational materials intended for patients should ideally fall within the sixth- to eighth-grade reading level (e.g., FKGL ≤ 8, CLI ≤ 8, SMOG < 9, GFI ≤ 12, FRES ≥ 60). Among the evaluated models, DeepSeek exhibited the highest clinical readability compliance, with 7 to 10 out of 25 responses meeting the standard across different indices. In contrast, ChatGPT-4o and Claude 3.7 met these thresholds in only 0 to 1 response, indicating a tendency toward higher complexity. Gemini Advanced displayed intermediate performance, meeting select thresholds in 3 to 6 cases depending on the metric. These results suggest that while some models generate accurate and complete responses, their readability may limit their practical value for non-expert caregivers or emergency settings.

The results of the Spearman correlation analysis assessing relationships among all variables are presented in Table 5 and Figure 5. A statistically significant positive correlation was found between accuracy and completeness scores (ρ: 0.701, $p < 0.001$). No significant correlation was observed between response time and accuracy ($p = 0.570$), whereas a weak but significant positive correlation was identified between response time and completeness (ρ: 0.112, $p = 0.025$). No significant associations were found between accuracy and readability scores ($p > 0.05$) or between completeness and readability scores ($p > 0.05$). However, statistically significant positive or negative correlations were observed among all readability indices and scoring systems used to evaluate readability ($p < 0.001$).

Table 5. Spearman correlation analysis between accuracy and completeness scores, response time, and readability scores for ChatGPT-4 Omni, DeepSeek, Gemini Advanced, and Claude 3.7.

Variables		Accuracy	Completeness	Response Time	FRES	FKGL	GFI	SMOG	CLI
Accuracy	***ρ***		**0.701**	−0.028	−0.013	0.017	−0.019	−0.016	−0.064
	***p*-value ***		**<0.001**	0.570	0.790	0.728	0.710	0.753	0.198
Completeness	***ρ***	**0.701**		**0.112**	−0.054	0.057	0.070	0.049	−0.043
	***p*-value ***	**<0.001**		**0.025**	0.280	0.258	0.165	0.331	0.391
Response Time	***ρ***	−0.028	**0.112**		**0.314**	**−0.367**	**−0.318**	**−0.357**	**−0.248**
	***p*-value ***	0.570	**0.025**		**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
FRES	***ρ***	−0.013	−0.054	**0.314**		**−0.961**	**−0.914**	**−0.928**	**−0.639**
	***p*-value ***	0.790	0.280	**<0.001**		**<0.001**	**<0.001**	**< 0.001**	**<0.001**
FKGL	***ρ***	0.017	0.057	**−0.367**	**−0.961**		**0.956**	**0.986**	**0.500**
	***p*-value ***	0.728	0.258	**<0.001**	**<0.001**		**<0.001**	**<0.001**	**<0.001**
GFI	***ρ***	−0.019	0.070	**−0.318**	**−0.914**	**0.956**		**0.971**	**0.476**
	***p*-value ***	0.710	0.165	**<0.001**	**<0.001**	**<0.001**		**<0.001**	**<0.001**
SMOG	***ρ***	−0.016	0.049	**−0.357**	**−0.928**	**0.986**	**0.971**		**0.463**
	***p*-value ***	0.753	0.331	**<0.001**	**<0.001**	**<0.001**	**<0.001**		**<0.001**
CLI	***ρ***	−0.064	−0.043	**−0.248**	**−0.639**	**0.500**	**0.476**	**0.463**	
	***p*-value ***	0.198	0.391	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	

FRES: Flesch–Kincaid Reading Ease Score, FKGL: Flesch–Kincaid Grade Level, GFI: Gunning Fog Index, SMOG: Simple Measure of Gobbledygook Index, CLI: Coleman–Liau Index. * Spearman correlation analysis. Bold values mean statistically significant ($p < 0.05$).

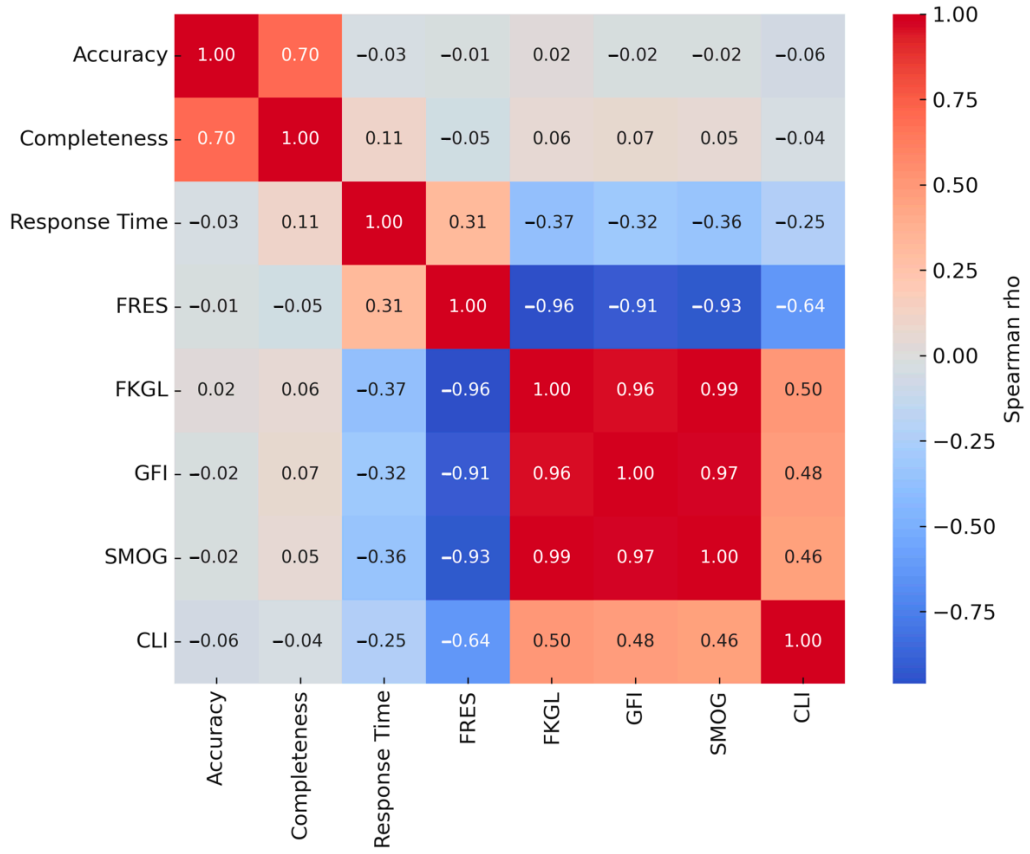


Figure 5. Correlation heatmap illustrating the relationships among accuracy, completeness, response time, and readability scores based on Spearman's correlation coefficients (FRES: Flesch–Kincaid Reading Ease Score, FKGL: Flesch–Kincaid Grade Level, GFI: Gunning Fog Index, SMOG: Simple Measure of Gobbledygook Index, CLI: Coleman–Liau Index).

4. Discussion

This study comparatively evaluated the performance of four state-of-the-art LLMs—ChatGPT-4o, DeepSeek, Gemini Advanced, and Claude 3.7—in generating responses to questions on TDIs in the primary dentition. The models were assessed using four key criteria: accuracy, completeness, response time, and readability. Statistical analyses revealed significant differences among the models across all evaluated parameters. Specifically, performance differences were statistically significant in terms of accuracy, completeness, response time, and readability (all $p < 0.05$), resulting in the rejection of all four null hypotheses. These findings demonstrate that the evaluated LLMs vary considerably in how they generate responses, with no single model consistently outperforming the others across all domains. Therefore, while all four models show potential in supporting clinical comprehension and decision-making regarding TDIs in children, their performance is highly context-dependent—highlighting the importance of selecting an appropriate model based on specific clinical or educational needs.

4.1. Comparison with Previous Studies

4.1.1. Accuracy

Several studies have evaluated the accuracy of responses generated by various LLM-based chatbots in relation to dental trauma, often using the IADT guidelines as a reference. In a study conducted by Ozden et al. [15], ChatGPT and Google Bard were assessed using 25 dichotomous questions on dental trauma, and both chatbots demonstrated an overall accuracy rate of 57.5% across 4500 responses. In another study by Johnson et al. [19],

the validity and reliability of chatbot-generated answers to frequently asked questions about dental trauma were evaluated across four chatbots (ChatGPT-3.5, Google Gemini, Bing, and Claude), with Claude significantly outperforming the others in both reliability and validity. Another investigation compared five models (ChatGPT-4, ChatGPT-3.5, two versions of Microsoft Copilot, and Google Gemini) using multiple-choice, fill-in-the-blank, and dichotomous formats [20]. Although no statistically significant differences were found among the models, their accuracy rates ranged from 46.7% to 76.7%, with ChatGPT versions and Gemini models performing better on multiple-choice and fill-in-the-blank formats than on dichotomous questions [20]. In a separate study, ChatGPT-3.5 was tested using 45 self-generated questions related to general dental trauma, intrusion, and avulsion and was reported to produce highly reliable responses [27]. In a study by Guven et al. [28], ChatGPT-3.5, ChatGPT-4, and Gemini were compared in their responses to dental trauma questions, with ChatGPT-4 and Gemini demonstrating significantly higher accuracy than ChatGPT-3.5. In the present study, four different chatbots were evaluated based on their responses to 25 open-ended questions on TDIs in the primary dentition, with assessments conducted by independent pediatric dentists in accordance with the IADT guidelines. ChatGPT-4o, DeepSeek, and Claude 3.7 produced significantly more accurate responses than Gemini Advanced. Although there were no statistically significant differences in accuracy among the top three models, ChatGPT-4o achieved the highest overall accuracy. Although the differences in mean accuracy scores between the top-performing models may seem numerically small, they can still carry clinical significance—particularly in pediatric dental trauma cases. Even slight improvements in accuracy may help reduce the risk of misinformation or misinterpretation, especially when such tools are used by non-specialists, such as parents or general practitioners. Furthermore, when these tools are used repeatedly over time, small gains in accuracy may accumulate to support better clinical decisions and more effective patient education. Therefore, while statistical significance does not necessarily imply clinical importance, small but consistent differences in accuracy should not be underestimated. These findings are consistent with prior research indicating that the accuracy of chatbot-generated content can be influenced by factors such as question format, content type, and evaluation methodology. As such, the relative performance of chatbots in dental contexts appears to be context-dependent and may vary across studies.

4.1.2. Completeness

In the present study, completeness was defined as the extent to which the chatbot-generated responses covered the information provided in the IADT guidelines. To ensure standardized evaluation, assessors were supplied with the current IADT guidelines and instructed to rate each response accordingly. Although no previous studies have specifically assessed completeness in chatbot responses related to dental trauma, several studies have examined this parameter in other areas of dentistry. For example, in a study involving dentistry-related questions posed to ChatGPT-3.5, expert evaluators reported a mean completeness score of 2.07 out of 3 (median: 2) [29]. Similarly, when ChatGPT was asked open-ended theoretical and case-based questions on interceptive orthodontics, the model achieved an average score of 2.4 out of 3 [30]. In the study by Alsayed et al. [31], ChatGPT was evaluated using 50 questions across oral surgery, preventive dentistry, and oral cancer. The corresponding completeness scores were 3.2/5 for oral surgery, 4.1/5 for preventive dentistry, and 3.3/5 for oral cancer. Another investigation, which included 144 clinical questions from various head and neck surgery subspecialties and 15 comprehensive clinical scenarios, reported a median completeness score of 3 out of 3 for open-ended questions [32]. As seen in most of these studies, ChatGPT and its various versions have been the primary focus, and completeness has often been measured using a 3-point Likert scale—similar to

the approach adopted in the present study. Consistent with previous findings, ChatGPT-4o achieved the highest completeness score in our analysis, with a mean of 2.74 out of 3 and a median of 3. Although the other chatbots also demonstrated mean scores above 2.5, Gemini Advanced obtained the lowest average score, which was significantly lower than those of the other models. Furthermore, our results revealed a statistically significant positive correlation between accuracy and completeness. In line with this, Gemini Advanced exhibited the lowest performance in both domains, with statistically significant differences observed in comparison to the other chatbots.

4.1.3. Response Time

In this study, the response times of ChatGPT-4o, DeepSeek, Gemini Advanced, and Claude 3.7 were compared in the context of questions concerning TDIs in the primary dentition, and statistically significant differences were observed among the models. Gemini Advanced and ChatGPT-4o delivered the fastest responses, whereas DeepSeek exhibited the slowest—potentially limiting its applicability in time-sensitive clinical situations. Claude 3.7 showed intermediate performance in terms of response time. Notably, Spearman correlation analysis revealed no significant relationship between response time and accuracy. Although shorter response times may be advantageous in urgent clinical scenarios, they might compromise the depth and completeness of information. For example, despite its quicker response, Gemini Advanced scored significantly lower in completeness than DeepSeek, suggesting a potential trade-off between speed and informational richness. This finding is consistent with previous studies. A study conducted in Türkiye comparing ChatGPT-3.5, ChatGPT-4o, Google Bard, and Microsoft Copilot on multiple-choice questions from the Dental Specialty Examination reported ChatGPT-3.5 as the fastest, while ChatGPT-4o had the longest response times [33]. Similarly, another study evaluating ChatGPT-3.5 and ChatGPT-4 on open-ended periodontal surgery questions found that ChatGPT-4 generally required more time to formulate responses [34]. These results underscore the necessity of balancing response speed with content quality when incorporating AI chatbots into clinical dentistry. While rapid feedback is desirable—especially in emergency scenarios such as pediatric dental trauma—accuracy and completeness are ultimately more critical for ensuring optimal patient outcomes. Moreover, in pediatric dentistry, fast response times may help alleviate parental anxiety or support early triage decisions in acute situations. However, whether quicker responses truly enhance patient education or clinical workflow remains underexplored and should be investigated in future research. Therefore, model selection for clinical use should be based on a comprehensive evaluation that considers not only speed but also the clarity, depth, and reliability of the information provided. To the best of our knowledge, this is the first study to report on the response times of the DeepSeek and Claude 3.7 models in the context of pediatric dental trauma, offering novel insights into their clinical utility.

4.1.4. Readability

Effective health communication necessitates presenting information in language that is accessible to individuals with varying levels of literacy. In the present study, the readability of responses generated by four AI chatbots was evaluated using five established indices. The findings indicated that, for all models, the readability scores exceeded the sixth-grade threshold recommended by the United States National Institutes of Health, suggesting limited accessibility for the general population. Among the evaluated chatbots, Claude 3.7 produced the most complex responses, characterized by the lowest FRES and highest FKGL scores—indicating a need for advanced academic literacy. In contrast, DeepSeek achieved the highest FRES and lowest SMOG Index scores, suggesting relatively higher accessibility.

However, even DeepSeek's responses were classified as "difficult," corresponding to a high school reading level. ChatGPT-4o and Gemini Advanced generated content requiring college-level literacy, while Claude 3.7's outputs demanded even more advanced reading skills. These findings are consistent with previous studies in various dental domains, where chatbot-generated content often required high literacy levels. For instance, research in oral cancer, endodontics, and orthodontics has shown that models such as ChatGPT-3.5, ChatGPT-4, and Claude frequently produce complex text that limits accessibility [35–37]. Similarly, a study evaluating the responses of ChatGPT-3.5, ChatGPT-4.0, and Google Gemini in responses to questions on TDIs reported that all models generated highly complex texts requiring at least a college-level reading ability [28]. Another investigation focusing specifically on TDIs found that responses from ChatGPT-3.5 and ChatGPT-4.0 produced responses understandable only to individuals with high school or higher education levels [16]. While Gemini has shown comparatively higher readability in some studies, its outputs have often been criticized for lacking informational depth and completeness. Notably, the present study is among the first to assess the readability of chatbot-generated content specific to TDIs in the primary dentition. Our analysis revealed no statistically significant correlation between readability and accuracy across the evaluated models. This finding suggests that a more readable response does not necessarily equate to a more clinically accurate one. One possible explanation is that simplifying language to improve readability may inadvertently lead to a loss of technical precision—especially in complex clinical contexts such as TDIs in the primary dentition. Conversely, highly accurate responses may incorporate specialized terminology and longer sentence structures, thereby reducing readability. This divergence underscores a key limitation in current LLMs: their inability to simultaneously optimize for clinical accuracy and linguistic accessibility. It also highlights the challenge of balancing these two dimensions in AI-generated patient education materials or clinical support tools, particularly when intended for non-specialist users such as caregivers or general practitioners.

While readability scores provide valuable insight into the accessibility of chatbot-generated content, their interpretation must be contextualized within the intended use case. In our study, most chatbot responses did not meet the standard sixth- to eighth-grade readability thresholds commonly recommended for patient education materials. For instance, ChatGPT-4o and Claude 3.7 rarely produced responses below a FKGL of 8 or a SMOG Index below 9. However, this should not be viewed strictly as a limitation. The questions posed in this study were primarily technical in nature and designed to simulate real-world clinical decision-making scenarios in pediatric dental trauma. As such, they demanded comprehensive and nuanced responses that inherently elevate linguistic complexity. These results therefore suggest that while certain LLMs may not currently optimize readability for laypersons, they may be appropriately tailored for professional or academic audiences seeking detailed, evidence-informed responses.

4.1.5. Potential Determinants of Performance Differences

The observed differences in chatbot performance likely reflect variations in model architecture, training datasets, and update schedules. For instance, ChatGPT-4o and Claude 3.7 are known to leverage advanced multi-modal and retrieval-augmented architectures, which may enhance their contextual coherence and factual accuracy [38]. Gemini Advanced, while optimized for general performance, may exhibit weaker alignment with domain-specific medical guidelines due to training corpus composition [39]. Similarly, DeepSeek's lower readability and response latency might stem from model scale, decoding strategies, or Application Programming Interface throttling behaviors [40]. These architectural and training-related nuances—although proprietary—are essential considerations

when interpreting benchmark results, as they suggest that output quality is not solely a function of question complexity, but also of design decisions embedded in each LLM's development pipeline.

Although our study utilized two independent sessions to assess response consistency, it is important to acknowledge that chatbot outputs may still be subject to temporal variability. Factors such as stochastic decoding (e.g., temperature and top-k sampling), silent Application Programming Interface updates, and evolving model versions can lead to variability in repeated queries—even when prompts are identical. This phenomenon, often referred to as "model drift," complicates the reproducibility of chatbot-based studies and suggests that future research should include time-stamped interactions, model versioning, and possibly prompt seeding for standardization. While we attempted to minimize variability by spacing sessions within a short timeframe and avoiding known model update periods, these uncontrollable dynamics remain a key limitation in LLM benchmarking.

The observed variability in LLM outputs—particularly when responding to identical prompts across sessions—may reflect broader challenges in dynamic decision-making and internal calibration mechanisms intrinsic to large neural models. These systems, like multilayer neural networks used in physical robotics, learn complex non-linear relationships and evolve based on training data, update frequency, and operational context. Similar to how model-free approaches can capture dynamic tool behaviors in robotics without pre-defined physical equations [41], LLMs attempt to align their responses with linguistic, semantic, and clinical patterns in a non-deterministic fashion. Incorporating insights from such calibration frameworks may enrich our understanding of how LLMs adapt—or fail to adapt—to structured guideline-based scenarios in healthcare.

Although this study evaluated LLM outputs solely based on textual prompts and responses, the future integration of multimodal clinical data—including radiographs, intraoral images, and patient histories—represents a promising direction for enhancing AI-driven decision support in pediatric dentistry. In robotics and sensor-rich environments, models that integrate heterogeneous inputs, such as recurrent neural networks for gesture recognition [42], have demonstrated improved contextual sensitivity and responsiveness. Analogously, LLMs capable of harmonizing multiple data streams could facilitate more accurate and clinically aligned responses in complex diagnostic or triage scenarios.

4.2. Guideline Adherence and Clinical Consistency in Chatbot Responses

In the present study, the 2020 IADT guideline [21] served as the reference standard for evaluating both chatbot-generated responses and expert assessments. Notably, some chatbots failed to provide answers consistent with this guideline or appeared to rely on outdated versions of earlier recommendations. Such discrepancies are not merely technical oversights but raise serious concerns regarding the trustworthiness of AI-generated clinical content. For example, in response to questions about radiographic indications, only ChatGPT-4o correctly referenced the most up-to-date principle—ALADAIP. In contrast, Claude 3.7 and Gemini Advanced cited the outdated "As Low As Reasonably Achievable" (ALARA) principle, which has since been replaced in pediatric dental radiology contexts. DeepSeek emphasized minimizing radiation exposure but did not reference any specific guideline. Furthermore, although the current IADT guidelines advise that CBCT should be reserved for well-justified clinical indications [21], only Gemini Advanced and Claude 3.7 accurately acknowledged this restriction. The remaining models omitted CBCT-specific recommendations in their responses. These inconsistencies highlight a critical limitation in LLM-based tools: the risk of disseminating outdated or inaccurate clinical recommendations. Such limitations may undermine clinician and patient trust—especially in pediatric

care settings where strict adherence to current guidelines is essential for ensuring both safety and efficacy.

In response to the question regarding the clinical management of intruded or laterally luxated primary teeth with root displacement toward the developing permanent tooth germ, ChatGPT-4o and Gemini Advanced provided recommendations consistent with the current IADT guidelines. In contrast, DeepSeek and Claude 3.7 offered outdated responses that did not align with contemporary clinical standards. When asked about the expected timeframe for spontaneous repositioning of an intruded primary tooth, Claude 3.7 provided the most guideline-consistent answer. However, regarding clinical management, both ChatGPT-4o and DeepSeek incorrectly recommended extraction when the root was directed toward the permanent successor—contradicting the IADT's recommendation of an observation-based approach. Conversely, Claude 3.7 and Gemini Advanced adhered more closely to current guidelines by recommending clinical monitoring and reserving extraction for cases with clear indications. For lateral luxation injuries, ChatGPT-4o accurately indicated that splinting is recommended in cases of severe displacement and correctly specified the appropriate duration, in accordance with the IADT guidelines. DeepSeek stated that splinting is generally not advised in primary teeth but noted that it may be considered in cases of excessive mobility. Gemini Advanced and Claude 3.7, however, did not mention splinting or provide any guidance regarding its duration.

Regarding questions assessing the recommended follow-up intervals for various types of TDIs, all chatbots exhibited varying degrees of inconsistency with the current IADT guidelines. However, when asked about the appropriate clinical response in cases where there is a discrepancy between the patient's history and clinical findings, all chatbots correctly identified the potential for child abuse and emphasized the importance of mandatory reporting in suspected cases. This demonstrates that the models showed sensitivity to child protection issues and adhered to the ethical and legal principles outlined in the IADT guidelines—highlighting the potential of AI technologies to integrate ethical awareness. In contrast, the clinical reliability of these models remains limited when it comes to more nuanced recommendations. For example, only ChatGPT-4o correctly indicated that splinting is required in root fractures involving coronal fragment displacement and specified the appropriate duration. Other models failed to address this essential aspect of trauma management, reflecting a lack of alignment with current evidence-based practices. Several factors may explain these inconsistencies. One key limitation is the models' reliance on training data that may not include the most recent clinical guidelines, combined with the absence of real-time updating mechanisms. Additionally, guideline documents such as the IADT protocol may not be openly indexed or adequately represented in training datasets—particularly in the case of proprietary or specialty-specific content. These issues highlight the urgent need for model developers to integrate verified, up-to-date clinical resources and to establish systems for regular updates and validation. To mitigate the risk of outdated or erroneous advice in real-world use, several safeguards are warranted. These include the incorporation of structured clinical guideline databases into model training pipelines, the implementation of transparent update logs, and periodic auditing of chatbot outputs by subject-matter experts. Such strategies would improve both the clinical relevance and trustworthiness of AI-powered chatbots in pediatric dental trauma management.

While the 2020 IADT guideline served as the primary reference standard in this study, it is important to acknowledge both its strengths and limitations. The guideline is internationally recognized, developed by a multidisciplinary panel, and provides structured recommendations based on clinical scenarios. However, some recommendations allow for practitioner discretion and may lack detailed visual aids or age-specific nuances—factors that could affect chatbot interpretation. Additionally, the complexity of certain

clinical conditions is simplified in the guideline, which may lead to varied outputs among AI models when faced with ambiguities. Despite these limitations, the IADT guideline remains the most authoritative and comprehensive framework currently available for managing TDIs in primary dentition.

4.3. Ethical and Legal Considerations in the Clinical Use of Chatbots

Despite the promise of LLMs in enhancing clinical decision-making and health communication, their deployment also carries significant ethical and medico-legal risks—particularly in pediatric dental trauma settings. Misinformation resulting from hallucinated or outdated content may mislead caregivers or non-specialist users, potentially delaying appropriate clinical intervention or causing harm if interpreted without proper oversight. Over-reliance on AI tools may also reduce critical thinking or clinician vigilance, especially when users assume the outputs are inherently accurate. These concerns are magnified in pediatric care, where vulnerable populations are at stake. Moreover, the lack of transparency in how LLMs generate responses complicates issues of accountability and informed consent. While AI chatbots can serve as valuable support tools, they should never be viewed as replacements for clinician expertise. Their use must be strictly supervised, supported by safeguards such as warning prompts, usage disclaimers, and routine expert validation. To ensure patient safety and legal compliance, the future development and deployment of AI systems in dentistry must be guided by ethical frameworks, regulatory standards, and clinician-led evaluation processes. Only through such responsible integration can AI tools contribute meaningfully and safely to pediatric dental trauma management.

4.4. Study Contributions and Novelty

This study stands out as one of the few comprehensive investigations exploring the potential of AI-powered chatbots in dental practice from a multidimensional perspective. While most of the existing literature has focused on widely used models, this research offers a comparative evaluation of four advanced LLMs, including Claude 3.7—which has not been previously analyzed in this context—and DeepSeek, a relatively underexplored model. To the best of our knowledge, no prior study has evaluated the performance of DeepSeek in dental trauma or pediatric dentistry, and only one study [19] has included Claude in a dental trauma-related analysis. Therefore, the inclusion of these two models provides novel insights by expanding the scope beyond commonly assessed LLMs such as ChatGPT and Gemini. This study offers the first comparative evaluation of DeepSeek and Claude 3.7 in a pediatric dental trauma setting, establishing a new benchmark for accuracy, completeness, and readability.

4.5. Strengths, Limitations, and Methodological Challenges

Unlike earlier studies that focused either on general dental trauma or solely on ChatGPT models, the present work offers a broader and clinically specific comparison by targeting primary dentition and incorporating less-explored LLMs. Methodologically, the study adopted a human-centered and balanced approach. Five distinct readability indices (FRES, FKGL, GFI, SMOG, and CLI) were used to enhance objectivity. Each question was submitted to all four models in two separate sessions, enabling assessment of response consistency over time. The questions and scoring criteria were grounded in the most recent, evidence-based IADT guidelines. Although the question set was standardized and guideline-based, we acknowledge that modifying the number or scope of items—such as adding or replacing questions—could affect performance scores. However, the overall trends in model behavior, particularly regarding accuracy and guideline adherence, are expected to remain robust due to the systematic design and expert validation of the original set. The inclusion of response time and readability further allowed for a truly multidimen-

sional evaluation. Nevertheless, this study has several limitations. First, it assessed only four general-purpose AI models, and their performance may vary in other clinical contexts. Second, the exclusion of domain-specific models like Med-PaLM or BioGPT—due to their restricted availability—may limit the generalizability of the findings to highly specialized or clinical-only scenarios. However, this study focused on evaluating publicly accessible, general-purpose models that are more commonly used by non-specialist clinicians or caregivers in real-world pediatric dental trauma situations. Third, the exclusive use of English-language inputs and outputs may limit the generalizability of our findings. While English is widely used in international research and forms a core part of LLM training corpora, pediatric dental practice often involves multilingual interactions with patients and caregivers. LLMs may exhibit variable performance across different languages due to disparities in training data volume, grammar structure, and semantic complexity. Future investigations should explore the multilingual robustness of these models to ensure broader applicability in global health contexts. Lastly, given the rapid pace of AI development, these findings represent a snapshot in time and may not reflect future model updates or improvements.

In addition to the limitations acknowledged, this study faced several practical and methodological challenges. First, aligning the evaluation criteria across diverse AI models proved difficult—especially given differences in language generation patterns and interface design [43]. Second, although the questions were based on standardized guidelines, LLMs sometimes interpreted open-ended prompts inconsistently, introducing variability not solely attributable to model quality [44]. This issue has also been reported in similar chatbot assessment studies [15,19]. Third, readability assessment—though comprehensive in our design—relied on linguistic indices that may not fully capture real-world comprehension, a limitation noted by prior reviews [45,46]. Finally, balancing objectivity and clinical relevance in expert scoring posed a challenge, as subjective interpretation of nuanced responses could vary slightly among evaluators. These challenges underscore the complexity of evaluating AI-generated medical content and the need for refined, standardized frameworks in future studies.

To mitigate the risks associated with overreliance on LLMs in pediatric dental settings, clinicians should adopt a cautious and structured approach. LLM-generated responses should never replace clinical expertise but may serve as adjunctive tools for triage or patient education when verified against trusted sources. For safe integration into practice, we recommend verifying AI outputs using updated clinical guidelines, avoiding sole reliance in urgent or ambiguous cases, and maintaining transparency when LLMs are used as part of patient communication. Furthermore, institutional policies and training modules could help practitioners better understand LLM capabilities and boundaries, fostering responsible use in pediatric dental care. A summary of recommended safeguards and responsible use strategies is provided in Table 6.

Table 6. Suggested safeguards and responsible use strategies for integrating large language models into pediatric dental triage and communication.

Recommended Use of LLMs in Pediatric Dental Triage	
Appropriate Scenarios	General education, FAQs, non-urgent symptom triage
Avoid Use When	Complex trauma, consent decisions, emergency care
Clinician Role	Always verify content, never rely solely on AI
Preferred Conditions	Use validated prompts, in native language if possible
Documentation	Clearly document source and AI involvement

LLMs: Large language models, AI: Artificial intelligence, FAQs: Frequently asked questions.

5. Conclusions

This study provides a comprehensive evaluation of four advanced AI chatbots—ChatGPT-4o, DeepSeek, Gemini Advanced, and Claude 3.7—in responding to questions related to TDIs in the primary dentition. The findings revealed considerable variability among the models in terms of accuracy, completeness, readability, and response time. Importantly, none of the evaluated chatbots consistently adhered to the 2020 IADT guideline, highlighting the potential clinical risks of relying on AI-generated advice without expert supervision. While this study was designed for a professional audience, the observed complexity of language remains a concern, even in clinical settings, as it may hinder comprehension and usability—especially in high-stress situations. Moreover, inappropriate or outdated recommendations—particularly in pediatric dental trauma cases—could lead to suboptimal decision-making or clinical mismanagement if not carefully interpreted by qualified personnel. Therefore, AI chatbots should be viewed as supportive tools rather than standalone sources of clinical guidance, particularly in pediatric dentistry. To enhance their utility in professional education, early triage, and communication, continuous validation, careful integration of updated guidelines, and risk-aware implementation strategies are essential.

Future research should explore hybrid evaluation frameworks that combine textual and visual inputs (e.g., radiographic data) to more accurately simulate real-world clinical scenarios. Additionally, the development of specialty-specific LLMs, trained on updated and guideline-indexed datasets, should be prioritized. Comparative studies across different languages and healthcare systems could shed light on the global applicability and accessibility of AI tools in pediatric dental care. Finally, co-design methodologies involving both clinicians and AI developers may facilitate the creation of more user-centered and clinically trustworthy AI interfaces.

The findings of this study offer preliminary insights into how different LLMs perform across clinically relevant domains. These results can inform the selection, refinement, and integration of AI tools in both dental education and practice. By identifying model-specific strengths and limitations, our findings may guide future adoption strategies, curricular development, and clinical decision-support implementations.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app15147778/s1, Table S1: The questions asked to the chatbots and the responses provided by the chatbots in the study.

Author Contributions: Conceptualization: B.S. and T.A. Methodology: B.S. Formal analysis: B.S. Writing—original draft preparation: B.S. and T.A. Writing—review and editing: B.S. and T.A. Supervision and integrity of the data and analysis: B.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author (B.S.) upon reasonable request.

Acknowledgments: The figures in the article were created with the assistance of ChatGPT-4 Omni (OpenAI, San Francisco, CA, USA) large language model-based artificial intelligence software.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Brandini, D.A.; Sonoda, C.K.; Takamiya, A.S.; Monteiro, D.R. Use of ChatGPT in dental traumatology. *Aust. Endod. J.* **2024**, *50*, 464–465. [CrossRef] [PubMed]
2. Bagde, H.; Dhopte, A.; Alam, M.K.; Basri, R. A systematic review and meta-analysis on ChatGPT and its utilization in medical and dental research. *Heliyon* **2023**, *9*, e23050. [CrossRef]
3. Kinikoglu, I. Evaluating ChatGPT and Google Gemini Performance and Implications in Turkish Dental Education. *Cureus* **2025**, *17*, e77292. [CrossRef] [PubMed]
4. Grzybowski, A.; Pawlikowska-Łagód, K.; Lambert, W.C. A History of Artificial Intelligence. *Clin. Dermatol.* **2024**, *42*, 221–229. [CrossRef] [PubMed]
5. Horiuchi, D.; Tatekawa, H.; Oura, T.; Oue, S.; Walston, S.L.; Takita, H.; Matsushita, S.; Mitsuyama, Y.; Shimono, T.; Miki, Y.; et al. Comparing the Diagnostic Performance of GPT-4-based ChatGPT, GPT-4V-based ChatGPT, and Radiologists in Challenging Neuroradiology Cases. *Clin. Neuroradiol.* **2024**, *34*, 779–787. [CrossRef]
6. Baker, H.P.; Dwyer, E.; Kalidoss, S.; Hynes, K.; Wolf, J.; Strelzow, J.A. ChatGPT's Ability to Assist with Clinical Documentation: A Randomized Controlled Trial. *J. Am. Acad. Orthop. Surg.* **2024**, *32*, 123–129. [CrossRef]
7. Carlà, M.M.; Gambini, G.; Baldascino, A.; Boselli, F.; Giannuzzi, F.; Margollicci, F.; Rizzo, S. Large language models as assistance for glaucoma surgical cases: A ChatGPT vs. Google Gemini comparison. *Graefes Arch. Clin. Exp. Ophthalmol.* **2024**, *262*, 2945–2959. [CrossRef]
8. Sismanoglu, S.; Capan, B.S. Performance of artificial intelligence on Turkish dental specialization exam: Can ChatGPT-4.0 and gemini advanced achieve comparable results to humans? *BMC Med. Educ.* **2025**, *25*, 214. [CrossRef]
9. Kaygisiz, Ö.F.; Teke, M.T. Can deepseek and ChatGPT be used in the diagnosis of oral pathologies? *BMC Oral Health* **2025**, *25*, 638. [CrossRef]
10. Gibney, E. China's cheap, open AI model DeepSeek thrills scientists. *Nature* **2025**, *638*, 13–14. [CrossRef]
11. Oura, T.; Tatekawa, H.; Horiuchi, D.; Matsushita, S.; Takita, H.; Atsukawa, N.; Mitsuyama, Y.; Yoshida, A.; Murai, K.; Tanaka, R.; et al. Diagnostic accuracy of vision-language models on Japanese diagnostic radiology, nuclear medicine, and interventional radiology specialty board examinations. *Jpn. J. Radiol.* **2024**, *42*, 1392–1398. [CrossRef] [PubMed]
12. Kurokawa, R.; Ohizumi, Y.; Kanzawa, J.; Kurokawa, M.; Sonoda, Y.; Nakamura, Y.; Kiguchi, T.; Gonoi, W.; Abe, O. Diagnostic performances of Claude 3 Opus and Claude 3.5 Sonnet from patient history and key images in Radiology's "Diagnosis Please" cases. *Jpn. J. Radiol.* **2024**, *42*, 1399–1402. [CrossRef] [PubMed]
13. Levin, L.; Day, P.F.; Hicks, L.; O'Connell, A.; Fouad, A.F.; Bourguignon, C.; Abbott, P.V. International Association of Dental Traumatology guidelines for the management of traumatic dental injuries: General introduction. *Dent. Traumatol.* **2020**, *36*, 309–313. [CrossRef] [PubMed]
14. Bulut, E.; Güçlü, Z.A. Evaluation of primary teeth affected by dental trauma in patients visiting a university clinic, Part 1: Epidemiology. *Clin. Oral Investig.* **2022**, *26*, 6783–6794. [CrossRef]
15. Ozden, I.; Gokyar, M.; Ozden, M.E.; Sazak Ovecoglu, H. Assessment of artificial intelligence applications in responding to dental trauma. *Dent. Traumatol.* **2024**, *40*, 722–729. [CrossRef]
16. Öztürk, Z.; Bal, C.; Çelikkaya, B.N. Evaluation of Information Provided by ChatGPT Versions on Traumatic Dental Injuries for Dental Students and Professionals. *Dent. Traumatol.* 2025, *in press*. [CrossRef]
17. Bubna, D.P.; Felipe de Jesus Freitas, P.; Ferraz, A.X.; Abuabara, A.; Baratto-Filho, F.; Marques de Mattos de Araujo, B.; Kuchler, E.C.; Roskamp, L.; Deliga Schroder, A.G.; Miranda de Araujo, C. Dental Trauma Evo—Development of an Artificial Intelligence-Powered Chatbot to Support Professional Management of Dental Trauma. *J. Endod.* 2025, *in press*. [CrossRef]
18. Gökcek Taraç, M.; Nale, T. Artificial intelligence in pediatric dental trauma: Do artificial intelligence chatbots address parental concerns effectively? *BMC Oral Health* **2025**, *25*, 736. [CrossRef]
19. Johnson, A.J.; Singh, T.K.; 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.* **2025**, *41*, 187–193. [CrossRef]
20. Kuru, H.E.; Aşık, A.; Demir, D.M. Can Artificial Intelligence Language Models Effectively Address Dental Trauma Questions? *Dent. Traumatol.* 2025, *in press*. [CrossRef]
21. Day, P.F.; Flores, M.T.; O'Connell, A.C.; Abbott, P.V.; Tsilingaridis, G.; Fouad, A.F.; Cohenca, N.; Lauridsen, E.; Bourguignon, C.; Hicks, L.; 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*, 343–359. [CrossRef] [PubMed]
22. De Vito, A.; Colpani, A.; Moi, G.; Babudieri, S.; Calcagno, A.; Calvino, V.; Ceccarelli, M.; Colpani, G.; d'Ettorre, G.; Di Biagio, A.; et al. Assessing ChatGPT's Potential in HIV Prevention Communication: A Comprehensive Evaluation of Accuracy, Completeness, and Inclusivity. *AIDS Behav.* **2024**, *28*, 2746–2754. [CrossRef] [PubMed]
23. Coskun, B.N.; 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*, 509–515. [CrossRef] [PubMed]

24. Worrall, A.P.; Connolly, M.J.; O'Neill, A.; O'Doherty, M.; Thornton, K.P.; McNally, C.; McConkey, S.J.; de Barra, E. Readability of online COVID-19 health information: A comparison between four English speaking countries. *BMC Public Health* **2020**, *20*, 1635. [CrossRef]
25. Cheng, C.; Dunn, M. Health literacy and the Internet: A study on the readability of Australian online health information. *Aust. N. Z. J. Public Health* **2015**, *39*, 309–314. [CrossRef]
26. Uzunçıbuk, H.; Marrapodi, M.M.; Ronsivalle, V.; Cicciù, M.; Minervini, G. Lessons to be learned when designing comprehensible patient-oriented online information about temporomandibular disorders. *J. Oral Rehabil.* **2025**, *52*, 222–229. [CrossRef]
27. Bordin, R.W.; Bartnack, C.C.; Westphalen, V.P.D.; Gasparello, G.G.; Bark, M.J.; Gava, T.N.; Tanaka, O.M. Evaluating generative pretraining transformer reliability in addressing dental trauma: A cross-sectional observational study on avulsion and intrusion. *Saudi Endod. J.* **2025**, *15*, 45–52. [CrossRef]
28. Guven, Y.; Ozdemir, O.T.; Kavan, M.Y. Performance of Artificial Intelligence Chatbots in Responding to Patient Queries Related to Traumatic Dental Injuries: A Comparative Study. *Dent. Traumatol.* **2025**, *41*, 338–347. [CrossRef]
29. Molena, K.F.; Macedo, A.P.; Ijaz, A.; Carvalho, F.K.; Gallo, M.J.D.; Wanderley Garcia de Paula ESilva, F.; de Rossi, A.; Mezzomo, L.A.; Mugayar, L.R.F.; Queiroz, A.M. Assessing the Accuracy, Completeness, and Reliability of Artificial Intelligence-Generated Responses in Dentistry: A Pilot Study Evaluating the ChatGPT Model. *Cureus* **2024**, *16*, e65658. [CrossRef]
30. Hatia, A.; Doldo, T.; Parrini, S.; Chisci, E.; Cipriani, L.; Montagna, L.; Lagana, G.; Guenza, G.; Agosta, E.; Vinjolli, F.; et al. Accuracy and Completeness of ChatGPT-Generated Information on Interceptive Orthodontics: A Multicenter Collaborative Study. *J. Clin. Med.* **2024**, *13*, 735. [CrossRef]
31. Alsayed, A.A.; Aldajani, M.B.; Aljohani, M.H.; Alamri, H.; Alwadi, M.A.; Alshammari, B.Z.; Alshammari, F.R. Assessing the quality of AI information from ChatGPT regarding oral surgery, preventive dentistry, and oral cancer: An exploration study. *Saudi Dent. J.* **2024**, *36*, 1483–1489. [CrossRef] [PubMed]
32. Vaira, L.A.; Lechien, J.R.; Abbate, V.; Allevi, F.; Audino, G.; Beltramini, G.A.; Bergonzani, M.; Bolzoni, A.; Committeri, U.; Crimi, S.; et al. Accuracy of ChatGPT-Generated Information on Head and Neck and Oromaxillofacial Surgery: A Multicenter Collaborative Analysis. *Otolaryngol. Head. Neck. Surg.* **2024**, *170*, 1492–1503. [CrossRef] [PubMed]
33. Tassoker, M. ChatGPT-4 Omni's superiority in answering multiple-choice oral radiology questions. *BMC Oral Health* **2025**, *25*, 173. [CrossRef] [PubMed]
34. Li, C.; Zhang, J.; Abdul-Masih, J.; Zhang, S.; Yang, J. Performance of ChatGPT and Dental Students on Concepts of Periodontal Surgery. *Eur. J. Dent. Educ.* **2025**, *29*, 36–43. [CrossRef]
35. Rokhshad, R.; Khoury, Z.H.; Mohammad-Rahimi, H.; Motie, P.; Price, J.B.; Tavares, T.; Jessri, M.; Bavarian, R.; Sciubba, J.J.; Sultan, A.S. Efficacy and empathy of AI chatbots in answering frequently asked questions on oral oncology. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.* **2025**, *139*, 719–728. [CrossRef]
36. Aljamani, S.; Hassona, Y.; Fansa, H.A.; Saadeh, H.; Jamani, K.D. Evaluating Large Language Models in Addressing Patient Questions on Endodontic Pain: A Comparative Analysis of accessible chatbots. *J. Endod.* 2025, *in press*. [CrossRef]
37. Dursun, D.; Bilici Geçer, R. Can artificial intelligence models serve as patient information consultants in orthodontics? *BMC Med. Inform. Decis. Mak.* **2024**, *24*, 211. [CrossRef]
38. Low, Y.S.; Jackson, M.L.; Hyde, R.J.; Brown, R.E.; Sanghavi, N.M.; Baldwin, J.D.; Pike, C.W.; Muralidharan, J.; Hui, G.; Alexander, N.; et al. Answering real-world clinical questions using large language model, retrieval-augmented generation, and agentic systems. *Digit. Health* **2025**, *11*, 20552076251348850. [CrossRef]
39. Belaroussi, R. Subjective Assessment of a Built Environment by ChatGPT, Gemini and Grok: Comparison with Architecture, Engineering and Construction Expert Perception. *Big Data Cogn. Comput.* **2025**, *9*, 100. [CrossRef]
40. Li, H.; Moon, J.T.; Iyer, D.; Balthazar, P.; Krupinski, E.A.; Bercu, Z.L.; Newsome, J.M.; Banerjee, I.; Gichoya, J.W.; Trivedi, H.M. Decoding radiology reports: Potential application of OpenAI ChatGPT to enhance patient understanding of diagnostic reports. *Clin. Imaging* **2023**, *101*, 137–141. [CrossRef]
41. Su, H.; Qi, W.; Hu, Y.; Sandoval, J.; Zhang, L.; Schmirander, Y.; Chen, G.; Aliverti, A.; Knoll, A.; Ferrigno, G.; et al. Towards Model-Free Tool Dynamic Identification and Calibration Using Multi-Layer Neural Network. *Sensors* **2019**, *19*, 3636. [CrossRef] [PubMed]
42. Qi, W.; Ovur, S.E.; Li, Z.; Marzullo, A.; Song, R. Multi-sensor guided hand gesture recognition for a teleoperated robot using a recurrent neural network. *IEEE Robot. Autom. Lett.* **2021**, *6*, 6039–6045. [CrossRef]
43. Chow, J.C.L.; Wong, V.; Li, K. Generative Pre-Trained Transformer-Empowered Healthcare Conversations: Current Trends, Challenges, and Future Directions in Large Language Model-Enabled Medical Chatbots. *BioMedInformatics* **2024**, *4*, 837–852. [CrossRef]
44. Wang, L.; Li, J.; Zhuang, B.; Huang, S.; Fang, M.; Wang, C.; Li, W.; Zhang, M.; Gong, S. Accuracy of Large Language Models When Answering Clinical Research Questions: Systematic Review and Network Meta-Analysis. *J. Med. Internet Res.* **2025**, *27*, e64486. [CrossRef] [PubMed]

45. Wangsa, K.; Karim, S.; Gide, E.; Elkhodr, M. A Systematic Review and Comprehensive Analysis of Pioneering AI Chatbot Models from Education to Healthcare: ChatGPT, Bard, Llama, Ernie and Grok. *Future Internet* **2024**, *16*, 219. [CrossRef]
46. Aghajani, M.; Maye, E.; Burrell, K.; Kok, C.; Frew, J.W. Evaluating the Quality and Readability of Online Information on Hidradenitis Suppurativa: A Systematic Review. *Clin. Exp. Dermatol.* 2025, *in press*. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.