Check for updates

Dental Traumatology

WILEY

Dental Traumatology

ORIGINAL ARTICLE OPEN ACCESS

Does Sequential Use of Multiple Chatbots Influence Emergency Guidance for Dental Avulsion?

Mehmet Buldur[1] | Berkant Sezer[2]

[1]Department of Restorative Dentistry, School of Dentistry, Çanakkale Onsekiz Mart University, Çanakkale, Türkiye | [2]Department of Pediatric Dentistry, School of Dentistry, Çanakkale Onsekiz Mart University, Çanakkale, Türkiye

Correspondence: Berkant Sezer (dt.berkantsezer@gmail.com; berkant.sezer@comu.edu.tr)

Received: 5 February 2026 | **Revised:** 3 March 2026 | **Accepted:** 18 April 2026

Keywords: chatbot | dental avulsion | dental trauma | large language models | traumatic dental injuries

ABSTRACT

Background/Aim: Artificial intelligence–based chatbot systems are increasingly used for dental emergency guidance, yet the clinical value of sequential consultation remains unclear. This study investigated whether response order influences the quality of chatbot-generated answers to permanent tooth avulsion questions.

Material and Methods: Eight frequently asked questions covering key decision points in avulsion management were developed based on International Association of Dental Traumatology guidelines. Three publicly accessible chatbots—ChatGPT (GPT-5), Gemini (Gemini 1.5 Pro), and DeepSeek (DeepSeek-R1)—were queried using a standardized sequential prompt design. For each question, all chatbot order permutations were applied so that each system generated first-, second-, and third-position responses. Second- and third-position prompts required evaluation of prior responses and correction of missing or incorrect information, thereby creating a controlled "maximum correction opportunity" condition. Responses were generated in independent sessions and evaluated by three expert dentists blinded to model identity using predefined criteria: accuracy, safety, completeness, correction ability, clarity, and usefulness. Composite scores were calculated within seven conceptual evaluation frameworks, and response-order effects were tested using the Friedman test with Kendall's W.

Results: Across all seven conceptual evaluation frameworks, composite quality scores did not differ significantly according to response order (all $p > 0.05$), with Kendall's W indicating negligible effect sizes. No consistent improvement was observed in accuracy, safety, completeness, or correction ability in second- or third-position responses. Model-specific analyses confirmed the absence of response-order effects across all chatbot platforms.

Conclusions: Under structured evaluative conditions permitting explicit revision, sequential consultation of multiple chatbot systems did not improve the quality, safety, or guideline alignment of information provided for dental avulsion management. These findings suggest that response-order effects alone may have limited corrective impact in this context and should be interpreted within the boundaries of controlled experimental modeling.

1 | Introduction

Dental trauma is an important clinical problem, particularly prevalent among children and young adults, and can lead to long-term functional and esthetic consequences [1, 2]. Among these injuries, permanent tooth avulsion—characterized by the complete displacement of the tooth from the alveolar socket—represents one of the most severe forms of dental trauma, with prognosis largely dependent on on-site first aid management and time-related factors [1, 3]. Incorrect or delayed interventions can rapidly compromise periodontal ligament viability, thereby adversely affecting the long-term

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2026 The Author(s). *Dental Traumatology* published by John Wiley & Sons Ltd.

survival of the replanted tooth [2, 4]. Consequently, the management of avulsion involves several guideline-based critical steps, including proper handling of the tooth, appropriate cleaning, storage in a suitable medium, replantation as soon as possible, and post-replantation follow-up [3, 4]. The International Association of Dental Traumatology (IADT) has standardized these steps in its current evidence-based guidelines and has particularly emphasized the decisive role of time and storage conditions [3, 4].

Permanent tooth avulsion is most often recognized at the scene by individuals who are not dental professionals, and initial management is commonly performed by parents, teachers, or sports coaches [5–7]. However, the literature indicates that knowledge regarding the emergency management of avulsion is frequently inadequate among these groups, with incorrect practices—such as handling the tooth by the root, aggressive cleaning with tap water, or the inadvertent replantation of primary teeth—being widely reported [5–7]. This knowledge gap can lead to a rapid loss of periodontal ligament viability, particularly in time-critical avulsion cases, thereby directly and adversely affecting prognosis [1–3]. Digital health tools, including artificial intelligence–based chatbot systems, are increasingly accessible in emergency contexts; therefore, the accuracy and safety of the information they provide warrant careful evaluation [8–10].

In recent years, a growing number of studies have examined the performance of artificial intelligence–based chatbots in the context of dental trauma. These investigations have predominantly evaluated chatbot outputs as isolated responses, focusing on dimensions such as guideline alignment, factual accuracy, completeness, and safety [11, 12]. While these studies have provided valuable insight into baseline chatbot performance, they have generally treated responses as independent outputs, without accounting for the potential influence of prior contextual exposure. Consequently, the dynamic behavior of large language models under sequential exposure conditions remains insufficiently explored within dental traumatology research. Nevertheless, these studies are largely based on the assumption that responses generated by artificial intelligence are independent of prior inputs and contextual influences. This assumption conflicts with contemporary theoretical and methodological evidence demonstrating that large language models can modify their outputs by taking previous responses into account [13, 14]. Accordingly, the potential influence of response generation order and contextual exposure represents a distinct experimental variable that warrants systematic investigation. Rather than presuming specific user behaviors, the present study focuses on isolating whether exposure to prior artificial intelligence-generated responses enables measurable corrective modification in subsequent outputs under controlled conditions [15–17].

Contemporary large language models are capable of incorporating previous responses as contextual input, thereby refining or reinforcing subsequent outputs, and accumulated contextual information has been shown to play a decisive role in shaping response generation processes [13, 14]. Because large language models condition their outputs on preceding text, sequential exposure to prior responses represents a theoretically meaningful variable independent of specific user behavior patterns. Such contextual accumulation may influence whether subsequent responses demonstrate corrective modification or reinforcement of prior content. However, the existing literature indicates that the sequential contextual effects of responses to the same clinical question have not been treated as a distinct variable in study designs, nor has the impact of response generation order on response quality been systematically evaluated [11, 12, 15]. To our knowledge, response-generation order has not been systematically examined as an independent experimental variable within dental trauma chatbot research. By focusing on response-order mechanisms rather than isolated performance comparison, the present study extends prior work toward a mechanistic analysis of contextual influence in chatbot-generated clinical guidance. Sequential artificial intelligence responses may therefore give rise to two opposing clinical effects: correction, through the critical appraisal of prior responses and completion of missing information, or anchoring, through the uncritical reinforcement of incorrect or incomplete recommendations. In time-dependent clinical situations with low tolerance for error, such as dental avulsion, this balance between correction and anchoring constitutes a critical determinant of patient safety and the clinical acceptability of sequential artificial intelligence responses [2–4, 13, 14].

The aim of this study was to examine the effect of response generation order (first, second, or third response) on the quality of answers provided by artificial intelligence–based chatbot systems to questions related to permanent tooth avulsion. The study employed a controlled experimental design to isolate response generation order as an independent mechanistic variable. Responses were evaluated by blinded experts according to predefined criteria: accuracy, safety, completeness, correction ability, clarity, and usefulness. The null hypotheses were as follows: ($H0_1$) There is no statistically significant difference in response quality according to response generation order. ($H0_2$) The effect of response generation order on response quality does not differ across artificial intelligence models.

2 | Materials and Methods

2.1 | Study Design

This study is based on the evaluation of artificial intelligence–generated outputs obtained using standardized prompts. No human participants, patient data, personal information, or clinical interventions were involved. All data were collected from publicly available and accessible artificial intelligence chatbot platforms. Therefore, ethical committee approval was not required for this study.

The research process was conducted in accordance with contemporary transparent reporting principles for chatbot health advice studies. The manuscript was reviewed and refined in light of the CHART (Chatbot Assessment Reporting Tool) [18] statement to enhance clarity, reproducibility, and methodological transparency. Data collection was carried out during a continuous sampling period in November 2025 in order to minimize variability arising from potential updates to the artificial intelligence models.

16009657, 2026, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/edt.70072 by Berkant Sezer - Canakkale Onsekiz Mart Uni , Wiley Online Library on [07/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2.2 | Artificial Intelligence Chatbot Systems

In this study, three different artificial intelligence–based chatbot systems were evaluated: ChatGPT (OpenAI, San Francisco, CA, USA), Gemini (Google LLC, Mountain View, CA, USA), and DeepSeek (DeepSeek AI, Hangzhou, China). These chatbot systems were included because they are built on contemporary large language model architectures, are publicly accessible, and are widely used for health-related information queries.

All chatbot systems were assessed using the most up-to-date publicly available versions accessible to users at the time of data collection, through the official web-based application interfaces provided by their developers. During response generation, default user settings for each platform were maintained; no API access, custom system prompts, or developer-level model configurations were applied. This approach was adopted to ensure comparability across systems and to maintain standardized experimental conditions while preserving default user-level interaction parameters.

2.3 | Questions, Prompt Design, and Sequential Artificial Intelligence Interaction

In this study, eight frequently asked questions (FAQs) addressing clinically critical decision points in the emergency management of permanent tooth avulsion were utilized (Table 1). The question set was structured to encompass key domains, including initial management, appropriate handling and cleaning of the tooth, suitable storage media, the critical time window for replantation, management of primary teeth, and post-replantation care. The questions were developed based on the IADT guidelines and were finalized through consensus among three expert dentists with experience in traumatic dental injuries. Question wording was simplified to ensure clarity and accessibility while maintaining clinical relevance and consistency across chatbot platforms, and was designed to avoid the use of medical jargon.

To evaluate the effect of response order on response quality, a controlled question–answer sequence was established for each question, in which the ChatGPT-5, Gemini 1.5 Pro, and DeepSeek-R1 chatbot systems provided responses in the first, second, and third positions, respectively (Table 1). Through this design, each artificial intelligence chatbot participated in all possible response orders and was exposed only to the predefined sequential contextual information. This approach was intended to ensure that any observed effects on response quality could be attributed solely to response order and context accumulation, independent of uncontrolled prior interactions.

Response generation was conducted using a three-stage structured prompt strategy. For each question and response generation process, a new and independent chat session was initiated to prevent the influence of prior conversation history; each session was started with a standardized introductory prompt.

Across all stages, response templates were predefined to ensure standardization, and the full text of all generated responses is provided in the Supporting Information.

- *"Ignore all prior context and treat this as a new session. Respond only based on the information provided here."*
- Prompt for 1. Chatbot: *"Question: [insert FAQ here]"*

  Please provide a concise and clear response in 2–3 sentences.
- Prompt for 2. Chatbot: *"Question: [insert FAQ here] Previous response: [Chatbot 1's response]"*

  Provide your own response and briefly evaluate the previous answer by indicating whether you agree with it, add missing

TABLE 1 | Frequently asked questions and dialogue variation structure used in the study.

Questions	C1	C2	C3	C4	C5	C6
Q1. My tooth just fell out completely. What should I do right away?	B → C → A	C → A → B	B → A → C	A → C → B	C → B → A	A → B → C
Q2. Can I put the tooth back into the socket by myself?	C → B → A	B → A → C	C → A → B	A → B → C	B → C → A	A → C → B
Q3. If I pick up the tooth, should I hold it by the root or by the top (crown)?	C → A → B	C → B → A	A → C → B	B → A → C	A → B → C	B → C → A
Q4. If the tooth is dirty, how should I clean it?	A → C → B	A → B → C	B → C → A	B → A → C	C → B → A	C → A → B
Q5. If I cannot put the tooth back right away, in what liquid should I keep it?	C → B → A	A → C → B	C → A → B	B → C → A	B → A → C	A → B → C
Q6. How much time do I have to put the tooth back in?	A → B → C	C → B → A	C → A → B	A → C → B	B → C → A	B → A → C
Q7. My child's baby tooth has fallen out. Should I put it back?	C → B → A	A → B → C	B → A → C	C → A → B	A → C → B	B → C → A
Q8. After the tooth has been placed back, what should I do next?	C → A → B	A → C → B	A → B → C	B → C → A	B → A → C	C → B → A

Abbreviations: A, ChatGPT-5; B, Google Gemini 1.5 Pro; C, Deepseek-R1.

16009657, 2026, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/edt.70072 by Berkant Sezer - Canakkale Onsekiz Mart Uni , Wiley Online Library on [07/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

information, or correct any inaccuracies. Maximum length: 3–4 sentences.

- Prompt for 3. Chatbot: *Question: [insert FAQ here] Previous responses:*

  [Chatbot 1's response]

  [Chatbot 2's response]

  Provide your own response and evaluate the two previous answers. Maximum length: 3–4 sentences.

2.4 | Evaluation Criteria and Expert Scoring

All generated responses were evaluated by three independent expert dentists with experience in traumatic dental injuries. The evaluators were blinded to the identity of the artificial intelligence chatbot that generated each response, and all dialogues ($n = 48$) were presented to them in a randomized order. Response evaluation was performed using a predefined and structured scoring framework (Table 2). This framework comprised six primary criteria: accuracy, safety, completeness, clarity, usefulness, and correction ability. Each criterion was scored using scales appropriate to the nature of the construct being assessed. Correction ability was evaluated only for the second and third responses; by definition, the first response received a score of 0, as it did not have any preceding content to evaluate or correct. All evaluation criteria and scoring definitions are summarized in Table 2.

The evaluation framework was developed to reflect a multidimensional assessment of clinical information quality rather than relying solely on factual correctness. In line with principles consistent with interpretive and expert-based assessment approaches [19, 20], the selected domains (accuracy, safety, completeness, correction ability, clarity, and usefulness) were conceptualized to capture both technical guideline alignment and practical decision relevance. The inclusion of correction ability as a distinct evaluative dimension was intended to assess adaptive reformulation in response to prior contextual exposure. Expert evaluators were instructed to assess each response holistically within the predefined avulsion management framework, thereby integrating structured scoring with interpretive clinical reasoning. This approach aligns with contemporary recommendations for comprehensive assessment of artificial intelligence-generated health information beyond single-metric evaluation [18, 21, 22].

To allow response quality to be evaluated separately across different clinical and theoretical priorities, the scoring criteria were grouped into seven distinct conceptual evaluation packages. These packages were defined as "Traumatology Guideline–Aligned," "Patient Safety–Critical Risk," "Guideline Execution Chain," "Learning & Adaptation," "Time-Critical Decision," "User Comprehension & Actionability," and "Worst-Case Harm Minimization" (Table 3).

The clinical priorities encompassed by these conceptual packages were determined based on evaluation approaches recommended in the literature on dental trauma management, patient safety, and artificial intelligence–supported health communication. The weighting schemes applied to each package were defined a priori—independent of the study outcomes—through consensus among three expert dentists with experience in dental trauma.

2.5 | Inter-Rater Reliability

Inter-rater reliability was analyzed using statistical methods appropriate to the scale types of the evaluation criteria. For criteria assessed using ordinal scoring (accuracy, safety, completeness, clarity, usefulness, and correction ability), inter-rater reliability was calculated using the intraclass correlation coefficient (ICC) based on a two-way random-effects model with absolute agreement.

2.6 | Statistical Analysis

Inter-rater reliability was assessed based on composite scores using the intraclass correlation coefficient for average measures with absolute agreement, derived from a two-way random-effects model (ICC[2,k]). To evaluate response quality in a multidimensional manner, composite scores were calculated using the criteria of accuracy, safety, completeness, correction ability, clarity, and usefulness. Question-level composite scores were analyzed separately within

TABLE 2 | Evaluation criteria and scoring framework for artificial intelligence-generated responses.

Criterion	Scale type	Score range	Description
Accuracy	Likert	1–5	Degree of factual correctness and consistency with IADT dental trauma guidelines
Safety	Likert	1–3	Potential risk of harm or unsafe clinical guidance
Completeness	Likert	1–3	Extent to which essential dental trauma management elements were addressed
Clarity	Likert	1–3	Ease of understanding and suitability for a layperson
Usefulness	Likert	1–3	Practical actionability and decision-support value
Correction Ability[a]	Directional	–1/0/+1	Ability to correct, ignore, or reinforce previous errors

[a]Correction Ability was applied only to second and third Chatbot responses.

16009657, 2026, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/edt.70072 by Berkant Sezer - Canakkale Onsekiz Mart Uni , Wiley Online Library on [07/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 3 | Conceptual weighting frameworks (compact transposed format).

Domain/ framework	Traumatology guideline–aligned	Patient safety–critical risk	Guideline execution chain	Learning and adaptation	Time-critical decision	User comprehension and actionability	Worst-case harm minimization
Accuracy	0.15	0.10	0.20	0.15	0.30*	0.15	0.15
Safety	0.25	0.35*	0.25	0.10	0.25	0.20	0.30*
Completeness	0.20	0.20	0.35*	0.25	0.20	0.05	0.20
Correction Ability	0.30*	0.25	0.10	0.40*	0.15	0.05	0.25
Clarity	0.05	0.05	0.05	0.05	0.05	0.30*	0.05
Usefulness	0.05	0.05	0.05	0.05	0.05	0.25	0.05

Note: Weighting schemes were defined a priori to reflect trauma guideline principles and alternative clinically motivated perspectives and were not adjusted to optimize statistical significance. Asterisks indicate the highest-weighted component within each domain.

seven conceptual frameworks: Traumatology Guideline–Aligned, Patient Safety–Critical Risk, Guideline Execution Chain, Learning & Adaptation, Time-Critical Decision, User Comprehension & Actionability, and Worst-Case Harm Minimization. Differences related to response generation order were evaluated using nonparametric methods appropriate for dependent measurements; specifically, the Friedman test was applied. The level of statistical significance was set at $p < 0.05$. Effect sizes were reported using Kendall's W coefficient and interpreted according to thresholds recommended in the literature for small, moderate, and large effects. In addition, to assess the effect of response order at the model level, analyses were repeated separately for each chatbot model. All statistical analyses were conducted using IBM SPSS Statistics version 26.0 (IBM Corp., Armonk, NY, USA).

3 | Results

Inter-rater reliability was found to be at a "good" level (ICC = 0.76). Accordingly, the mean scores assigned by the three evaluators for each response were used in subsequent statistical analyses. Although no statistically significant differences related to response order were identified, question-level descriptive analyses were conducted within each conceptual evaluation package. The median (IQR) values of composite scores calculated using package-specific a priori weights revealed distinct quality profiles across questions (Table 4).

Across all packages, Q4 was characterized by relatively lower median composite scores, whereas Q6 and Q7 achieved the highest median values in most packages. In particular, within the Learning & Adaptation and User Comprehension & Actionability packages, narrower IQR ranges were observed for these questions, indicating a more consistent distribution of response quality. In contrast, wider IQR ranges were observed for Q1 and Q2 in some packages, indicating relatively greater heterogeneity in response quality. In addition to composite safety scores, a categorical examination of safety ratings (1–3 scale) was performed to provide a clinically oriented assessment. Across all responses and evaluators (8 questions × 6 response-order permutations × 3 response positions × 3 evaluators; $n = 432$ individual safety ratings), 18 (4.2%) were classified as clinically unsafe (score = 1). These unsafe ratings were predominantly concentrated in the cleaning-related scenario (Q4; 11/18) and primarily involved variation in rinsing recommendations. No unsafe responses were identified for primary tooth management (Q7), storage medium selection (Q5), or replantation timing (Q6). Furthermore, no monotonic decrease in unsafe ratings was observed across first-, second-, and third-position responses.

Across all seven conceptual frameworks, no statistically significant effect of response generation order (first, second, or third response) on composite scores was identified (all Friedman tests, $p > 0.05$; Kendall's W values ranging from negligible to small effect sizes) (Table 5). This lack of a response-order effect was consistently observed across all frameworks emphasizing guideline adherence, patient safety, execution chain, learning and adaptation, time-critical decision-making, user comprehension and actionability, and worst-case harm minimization. Notably, even in frameworks in which correction ability was assigned a higher a priori weight, no improvement in performance was observed with subsequent responses.

The effect of response order on composite scores was analyzed separately for each artificial intelligence model. Across all conceptual evaluation packages, no statistically significant differences were found between response order and composite scores for any of the three models (all comparisons, $p > 0.05$). Kendall's W values calculated for all conceptual packages were low, indicating that effect sizes associated with response order were negligible or small (Table 5).

4 | Discussion

The present study examined whether the sequential use of multiple chatbot systems leads to progressive improvement in the quality of clinical information provided for a time-critical dental emergency. Unlike most previous investigations that have focused primarily on the intrinsic accuracy or completeness of individual chatbot outputs, this study experimentally isolated response generation order as an independent variable under controlled conditions, rather than directly replicating spontaneous

16009657, 2026, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/edt.70072 by Berkant Sezer - Canakkale Onsekiz Mart Uni , Wiley Online Library on [07/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 4 | Question-level median (IQR) composite scores by conceptual framework.

	Traumatology guideline–aligned	Patient safety–critical risk	Guideline execution chain	Learning and adaptation	Time-critical decision	User comprehension and actionability	Worst-case harm minimization	Overall (equal-weight composite)
Q1	2.23 (2.02–2.41)	2.28 (2.07–2.45)	2.35 (2.13–2.53)	2.92 (2.73–3.04)	2.46 (2.23–2.63)	2.87 (2.64–3.05)	2.38 (2.18–2.55)	2.67 (2.51–2.82)
Q2	2.32 (1.89–2.46)	2.33 (1.93–2.47)	2.40 (1.98–2.57)	2.98 (2.75–3.06)	2.50 (2.05–2.67)	2.92 (2.70–3.08)	2.43 (2.00–2.58)	2.69 (2.36–2.82)
Q3	2.32 (2.23–2.33)	2.37 (2.25–2.38)	2.43 (2.31–2.45)	3.05 (2.93–3.13)	2.53 (2.41–2.55)	2.98 (2.87–3.06)	2.47 (2.36–2.48)	2.72 (2.63–2.78)
Q4	1.99 (1.89–2.07)	2.03 (1.94–2.12)	2.07 (1.96–2.16)	2.76 (2.70–2.89)	2.17 (2.05–2.27)	2.70 (2.63–2.84)	2.11 (2.00–2.20)	2.42 (2.33–2.50)
Q5	2.36 (2.30–2.40)	2.41 (2.38–2.45)	2.47 (2.45–2.51)	3.05 (3.01–3.15)	2.57 (2.54–2.61)	3.00 (2.96–3.08)	2.51 (2.45–2.55)	2.78 (2.72–2.83)
Q6	2.30 (2.12–2.40)	2.35 (2.20–2.47)	2.41 (2.24–2.51)	3.03 (2.91–3.11)	2.52 (2.35–2.60)	2.96 (2.85–3.05)	2.45 (2.26–2.53)	2.72 (2.51–2.78)
Q7	2.40 (2.22–2.40)	2.45 (2.27–2.45)	2.51 (2.33–2.51)	3.15 (3.10–3.15)	2.61 (2.43–2.61)	3.08 (3.03–3.08)	2.55 (2.37–2.55)	2.83 (2.68–2.83)
Q8	2.37 (2.22–2.50)	2.42 (2.25–2.56)	2.47 (2.31–2.61)	3.05 (3.00–3.10)	2.58 (2.43–2.72)	3.00 (2.95–3.05)	2.52 (2.37–2.63)	2.78 (2.67–2.83)

16009657, 2026, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/edt.70072 by Berkant Sezer - Canakkale Onsekiz Mart Uni , Wiley Online Library on [07/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 5 | Effect of response order (first–second–third) on composite scores and effect of response order according to artificial intelligence models.

Conceptual package	χ^2 (df = 2)	*p*	Kendall's W
Traumatology guideline–aligned	0.023	0.989	0.0005
Patient safety–critical risk	0.023	0.989	0.0005
Guideline execution chain	0.442	0.802	0.0092
Learning and adaptation	0.023	0.989	0.0005
Time-critical decision	1.326	0.515	0.0276
User comprehension and actionability	0.738	0.691	0.0154
Worst-case harm minimization	0.207	0.902	0.0043

Model	χ^2 range	*p*	Kendall's W
Model A	0.00–1.87	0.39–1.00	0.000–0.117
Model B	0.22–1.07	0.59–0.90	0.014–0.067
Model C	0.07–1.56	0.46–0.96	0.005–0.097

Note: Response order (first, second, and third responses) was treated as a within-subject factor and analyzed using the Friedman test (χ^2, df = 2). Blocks were defined as the combination of question and Chatbot model. Kendall's W represents the corresponding effect size estimates for the Friedman tests. For each model, the effect of response order was evaluated separately across seven conceptual packages using Friedman tests. The χ^2 and Kendall's W ranges represent the minimum and maximum values of the test statistics obtained across the seven conceptual packages for each model.

real-world user behavior. Using tooth avulsion as a high-risk and guideline-sensitive clinical scenario, the results consistently demonstrated that response order—whether a chatbot generated the first, second, or third answer—did not significantly influence composite performance scores. This pattern was observed across all seven conceptual evaluation frameworks, including those that explicitly prioritized correction ability and adaptive learning.

One plausible explanation for the absence of a response-order effect is the phenomenon of informational anchoring. Even when explicitly instructed to evaluate and revise previous responses, second- and third-stage chatbot outputs tended to align with the initial informational frame rather than fundamentally restructure it. Large language models are optimized to produce coherent and contextually consistent responses, which may inadvertently promote convergence rather than contradiction when prior information is available [23, 24]. This tendency mirrors the anchoring bias described in cognitive psychology, whereby early information disproportionately shapes subsequent judgments and decisions [25]. In conversational artificial intelligence settings, such anchoring effects may limit the system's capacity for genuine error correction, particularly when revisions are constrained to brief evaluative statements.

The structured sequential prompting strategy employed in this study provides additional insight into this mechanism. By explicitly instructing subsequent models to evaluate and revise prior responses, the design created conditions theoretically favorable to corrective reformulation. However, the persistence of convergence toward the initial informational frame suggests that anchoring dynamics may operate even under deliberate evaluative engagement. Rather than merely reflecting a methodological constraint, this pattern may indicate an inherent tendency of large language models to privilege contextual coherence over contradiction when exposed to preceding content [26]. From this perspective, the absence of progressive improvement can be interpreted not only as a lack of corrective amplification, but as empirical support for the stabilizing effect of informational anchoring in controlled sequential response exposure [27].

Another contributing factor may relate to the intrinsic characteristics of guideline-driven dental emergencies. Management of tooth avulsion is governed by well-established and highly standardized clinical protocols, which define critical steps such as handling of the avulsed tooth, appropriate storage media, replantation timing, and clear contraindications [3]. In such contexts, first-stage chatbot responses may already approximate guideline-consistent recommendations, resulting in a ceiling effect that restricts the potential for measurable improvement in subsequent answers [28].

Position-specific descriptive analyses, however, did not demonstrate saturation of first-position responses near the theoretical upper bound of the composite scoring framework. The mean composite score for first-position responses was 2.55, compared with 2.70 and 2.71 for second- and third-position responses, respectively. These values indicate that baseline responses were not clustered near the maximum achievable range and that measurable variability remained present across response positions. Furthermore, the effect sizes associated with response order were consistently negligible (Kendall's $W < 0.03$ across conceptual frameworks). Such extremely small effect sizes suggest that the absence of statistically significant differences is unlikely to be attributable solely to insufficient statistical power. While the possibility of very small undetected effects cannot be entirely excluded, the magnitude of the observed W values indicates that any response-order influence, if present, is minimal and unlikely to carry meaningful practical relevance.

Importantly, categorical examination of safety ratings demonstrated that clinically unsafe responses were uncommon and scenario-specific rather than systematic. Unsafe ratings accounted for 4.2% of all safety evaluations and were predominantly confined to the cleaning-related scenario, primarily involving variation in rinsing recommendations. No unsafe responses were identified in scenarios involving primary tooth management, storage medium selection, or replantation timing, indicating the absence of catastrophic guideline violations. Moreover, unsafe ratings did not demonstrate a consistent reduction across sequential response positions, suggesting that structured sequential exposure to prior responses did not materially alter the occurrence of clinically unsafe advice. Even when numerically limited, unsafe outputs in time-dependent

16009657, 2026, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/edt.70072 by Berkant Sezer - Canakkale Onsekiz Mart Uni , Wiley Online Library on [07/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

dental emergencies warrant careful consideration. In avulsion management, inappropriate cleaning or handling recommendations may compromise periodontal ligament viability and thereby influence long-term tooth survival. The absence of a consistent reduction in unsafe responses across sequential positions indicates that repeated or comparative chatbot consultation cannot be assumed to reliably mitigate clinically relevant risk under structured evaluative conditions.

The structure and length constraints imposed on chatbot responses may have further influenced correction dynamics [29]. In the present study, responses were intentionally limited to short and concise formats to ensure standardized comparison across systems while maintaining consistent user-level interaction parameters. However, effective correction in avulsion management often requires reconstruction of an entire decision chain rather than isolated modifications. Partial revisions within a narrow word limit may therefore be insufficient to meaningfully alter clinical guidance. This observation aligns with emerging evidence suggesting that complex clinical reasoning tasks benefit from structured, stepwise, or checklist-based response formats rather than brief narrative outputs [30, 31].

When contextualized within the existing literature, the findings of this study offer an important extension to prior work on artificial intelligence performance in dentistry and health communication. Most published studies evaluating chatbot performance have assessed responses in isolation, primarily focusing on dimensions such as factual accuracy, completeness, readability, and alignment with established clinical guidelines [32–34]. In the context of dental trauma, including tooth avulsion, existing investigations have largely examined whether chatbot-generated answers correctly reflect guideline-based emergency management recommendations, such as appropriate handling of the avulsed tooth, storage conditions, and timing of replantation [11, 35–37]. These studies typically compare single-response outputs either across different chatbot models or against expert-derived reference answers, without considering how responses evolve under controlled sequential exposure conditions. As a result, while valuable insights have been gained regarding the baseline reliability of chatbot-generated information for dental avulsion scenarios, the potential impact of response-order effects within structured experimental sequencing has remained largely unexplored.

The clinical and societal implications of these findings are noteworthy. Digital health information sources, including conversational artificial intelligence systems, are increasingly accessible in emergency contexts [38]. However, the present results suggest that within a controlled sequential response framework, exposure to multiple chatbot outputs did not yield measurable improvement in guidance quality and may instead promote apparent confirmation without substantive corrective gain. In time-critical scenarios such as permanent tooth avulsion, excessive reliance on sequential artificial intelligence–generated outputs under evaluative exposure conditions may not confer additional clinical benefit. These findings support positioning artificial intelligence–generated responses as supplementary educational tools rather than authoritative decision-making resources, particularly in emergency care contexts. Hybrid approaches integrating chatbot-generated information with professional oversight may represent a more realistic and safer pathway for clinical decision support in emergency contexts.

Several methodological strengths support the robustness of this study. The full permutation of chatbot response sequences enabled isolation of response-order effects independent of model identity. Blinded expert evaluation and randomized presentation minimized assessment bias, while the a priori definition of seven conceptual weighting frameworks allowed results to be interpreted from multiple clinically relevant perspectives. Importantly, correction ability was operationalized as a distinct evaluative domain rather than being implicitly subsumed under accuracy or completeness, representing a conceptual advancement over prior chatbot evaluation approaches. Unlike most previously published chatbot studies in dental traumatology, which have primarily assessed isolated responses and baseline performance metrics, the present study experimentally isolated response-generation order as an independent variable. By incorporating full permutation sequencing and structured contextual exposure, this design enabled mechanistic evaluation of response-order dynamics rather than simple cross-model comparison. Nevertheless, certain limitations warrant consideration. Notably, the structured sequential prompting strategy may have influenced corrective capacity in both amplifying and attenuating directions. The explicit instruction given to second- and third-position responses to evaluate prior answers was intentionally implemented to isolate response-order effects under standardized conditions; however, such explicit evaluative framing does not fully replicate naturalistic emergency user behavior. Accordingly, the findings should be interpreted with consideration of the structured experimental design and its implications for ecological validity. The study examined a limited number of questions within a single clinical scenario, which may restrict generalizability to other dental emergencies or broader medical contexts. The short response format, while methodologically justified, may have constrained the depth of possible correction. Additionally, all interactions were conducted in a single language and during a single data collection period, and contextual variables that may influence real-time information processing—such as emotional stress, incomplete information, or dynamic follow-up interaction—were not modeled within the controlled experimental framework. Although effect sizes associated with response order were consistently negligible, the finite number of question–response permutations examined in this study should be considered when interpreting the precision of response-order estimates. While the observed effects were minimal in magnitude, future studies employing broader scenario sets or alternative experimental designs may further clarify the stability of response-order patterns across different clinical contexts. These limitations, however, also delineate clear and relevant directions for future research.

Future investigations should explore whether allowing unrestricted or checklist-based responses enhances corrective potential, and whether sequential prompting within the same model differs from cross-model interaction. Qualitative analysis of specific high-risk errors—such as inappropriate replantation of primary teeth or omission of critical time-dependent steps—may

16009657, 2026, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/edt.70072 by Berkant Sezer - Canakkale Onsekiz Mart Uni , Wiley Online Library on [07/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

provide deeper insight into the nature of artificial intelligence correction failures. Expanding analyses to multilingual settings, varying levels of health literacy, and simulated high-stress user scenarios may further clarify how conversational artificial intelligence can be safely and effectively integrated into emergency health information pathways.

5 | Conclusion

This study demonstrates that, within a controlled experimental framework, sequential exposure to multiple chatbot-generated responses does not inherently enhance the quality, safety, or guideline alignment of information provided for the management of tooth avulsion. Despite explicit opportunities for evaluation and correction, response order and cross-model exposure did not lead to progressive improvement in chatbot-generated guidance. These findings indicate that response-generation order alone may have limited corrective impact under structured evaluative conditions.

From a clinical perspective, the results suggest that structured sequential exposure to prior artificial intelligence-generated outputs may not confer measurable benefit in time-critical dental scenarios. While conversational artificial intelligence tools may offer accessible informational support, their outputs should be interpreted as supplementary resources rather than substitutes for professional judgment or evidence-based clinical protocols. The effective management of dental avulsion remains highly dependent on timely, guideline-driven decision-making and expert care.

Overall, these findings underscore the importance of careful contextualization of artificial intelligence–generated information within clinical decision-making frameworks. Future investigations should prioritize patient safety, transparency regarding artificial intelligence limitations, and reinforcement of professional standards, particularly in emergency scenarios where delays or misinformation may have irreversible consequences.

Author Contributions

Conceptualization: M.B.; Methodology: M.B. and B.S.; Validation: M.B. and B.S.; Formal analysis: M.B.; Resources: M.B. and B.S.; Data curation: M.B. and B.S.; Writing – original draft preparation: M.B. and B.S.; Writing – review and editing: M.B. and B.S.; Visualisation: M.B.; Supervision: B.S.; Project administration: M.B.

Acknowledgements

The authors have nothing to report.

Funding

The authors have nothing to report.

Ethics Statement

The authors have nothing to report.

Consent

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

1. L. Andersson, "Epidemiology of Traumatic Dental Injuries," *Journal of Endodontics* 39, no. 3 Suppl (2013): S2–S5, https://doi.org/10.1016/j.joen.2012.11.021.

2. S. Lin, N. Pilosof, M. Karawani, R. Wigler, A. Y. Kaufman, and S. T. Teich, "Occurrence and Timing of Complications Following Traumatic Dental Injuries: A Retrospective Study in a Dental Trauma Department," *Journal of Clinical and Experimental Dentistry* 8, no. 4 (2016): e429–e436, https://doi.org/10.4317/jced.53022.

3. A. F. Fouad, P. V. Abbott, G. Tsilingaridis, et al., "International Association of Dental Traumatology Guidelines for the Management of Traumatic Dental Injuries: 2. Avulsion of Permanent Teeth," *Dental Traumatology: Official Publication of International Association for Dental Traumatology* 36, no. 4 (2020): 331–342, https://doi.org/10.1111/edt.12573.

4. L. Levin, P. F. Day, L. Hicks, et al., "International Association of Dental Traumatology Guidelines for the Management of Traumatic Dental Injuries: General Introduction," *Dental Traumatology: Official Publication of International Association for Dental Traumatology* 36, no. 4 (2020): 309–313, https://doi.org/10.1111/edt.12574.

5. J. D. McIntyre, J. Y. Lee, M. Trope, and W. F. Vann, Jr., "Effectiveness of Dental Trauma Education for Elementary School Staff," *Dental Traumatology: Official Publication of International Association for Dental Traumatology* 24, no. 2 (2008): 146–150, https://doi.org/10.1111/j.1600-9657.2008.00573.x.

6. N. Al Zaher and M. Dashash, "An Educational Intervention for Improving Knowledge of Syrian School Children About Avulsion Using the "Save Your Tooth" Poster," *BMC Oral Health* 21, no. 1 (2021): 24, https://doi.org/10.1186/s12903-020-01380-4.

7. G. H. Ramezani, M. Sohrabi, S. Nasiri, M. J. Kharazifard, S. Ghadimi, and B. Seraj, "Effect of an Educational Pamphlet on Knowledge and Performance of Fitness Trainers About Traumatic Dental Injuries," *Frontiers in Dentistry* 18 (2021): 5, https://doi.org/10.18502/fid.v18i5.5648.

8. T. Nadarzynski, O. Miles, A. Cowie, and D. Ridge, "Acceptability of Artificial Intelligence (AI)-Led Chatbot Services in Healthcare: A Mixed-Methods Study," *DIGITAL HEALTH* 5 (2019): 2055207619871808, https://doi.org/10.1177/2055207619871808.

9. World Health Organization, *Ethics and Governance of Artificial Intelligence for Health* (World Health Organization, 2021), https://www.who.int/publications/i/item/9789240029200.

10. Z. Wang, J. Gao, B. Danek, et al., "Compliance and Factuality of Large Language Models for Clinical Research Document Generation," *Journal of the American Medical Informatics Association: JAMIA* 33 (2025): 563–572, https://doi.org/10.1093/jamia/ocaf174.

11. I. Ozden, M. Gokyar, M. E. Ozden, and H. Sazak Ovecoglu, "Assessment of Artificial Intelligence Applications in Responding to Dental Trauma," *Dental Traumatology: Official Publication of International Association for Dental Traumatology* 40, no. 6 (2024): 722–729, https://doi.org/10.1111/edt.12965.

12. P. Kamath, P. Kamath, S. J. R. Saldanha, et al., "A Brief Exploration of Artificial Intelligence in Dental Healthcare: A Narrative Review," *F1000Research* 13 (2024): 37, https://doi.org/10.12688/f1000research.140481.2.

16009657, 2026, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/edt.70072 by Berkant Sezer - Canakkale Onsekiz Mart Uni , Wiley Online Library on [07/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

13. J. Wei, X. Wang, D. Schuurmans, et al., "Chain-Of-Thought Prompting Elicits Reasoning in Large Language Models," *Advances in Neural Information Processing Systems* 35 (2022): 24824–24837.

14. E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?," in Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT), (2021): 610–623.

15. F. Reis and C. Lenz, "Performance of Artificial Intelligence (AI)-Powered Chatbots in the Assessment of Medical Case Reports: Qualitative Insights From Simulated Scenarios," *Cureus* 16, no. 2 (2024): e53899, https://doi.org/10.7759/cureus.53899.

16. A. Abbasian, Y. Zhang, E. Lehman, M. Wainberg, and K. Vyas, "Foundation Metrics for Evaluating Effectiveness of Healthcare Conversations Powered by Generative Artificial Intelligence," *npj Digital Medicine* 7 (2024): 1074, https://doi.org/10.1038/s41746-024-01074-0.

17. K. Huang, M. Li, J. Chen, Y. Wang, H. Zhang, and Q. Liu, "A Comprehensive Survey on Evaluating Large Language Model Applications in the Medical Industry," *arXiv Preprint* 2404 (2024): 15777, https://arxiv.org/abs/2404.15777.

18. CHART Collaborative, "Reporting Guideline for Chatbot Health Advice Studies: The Chatbot Assessment Reporting Tool (CHART) Statement," *BMJ Medicine* 4, no. 1 (2025): e001632, https://doi.org/10.1136/bmjmed-2025-001632.

19. S. Reddy, W. Rogers, V. P. Makinen, et al., "Evaluation Framework to Guide Implementation of AI Systems Into Healthcare Settings," *BMJ Health & Care Informatics* 28, no. 1 (2021): e100444, https://doi.org/10.1136/bmjhci-2021-100444.

20. C. Jacob, N. Brasier, E. Laurenzi, et al., "AI for IMPACTS Framework for Evaluating the Long-Term Real-World Impacts of AI-Powered Clinician Tools: Systematic Review and Narrative Synthesis," *Journal of Medical Internet Research* 27 (2025): e67485, https://doi.org/10.2196/67485.

21. A. Landerholm, "AI in Qualitative Health Research Appraisal: Comparative Study," *JMIR Formative Research* 9 (2025): e72815, https://doi.org/10.2196/72815.

22. F. Dellafiore, A. Saba, C. Collaro, and G. Artioli, "Artificial Intelligence in Qualitative Research: Insights From Experts via Reflexive Thematic Analysis," *Qualitative Health Research* 36, no. 2–3 (2026): 145–165, https://doi.org/10.1177/10497323251389800.

23. S. Schmidgall, C. Harris, I. Essien, et al., "Addressing Cognitive Bias in Medical Language Models," *arXiv* 2402 (2024): 08113v3.

24. J. Lou and Y. Sun, "Anchoring Bias in Large Language Models: An Experimental Study," *Journal of Computational Social Science* 9, no. 1 (2026): 11, https://doi.org/10.1007/s42001-025-00435-2.

25. F. Lieder, T. L. Griffiths, Q. J. M. Huys, and N. D. Goodman, "The Anchoring Bias Reflects Rational Use of Cognitive Resources," *Psychonomic Bulletin & Review* 25, no. 1 (2018): 322–349, https://doi.org/10.3758/s13423-017-1286-8.

26. T. Cohen, W. Xu, Y. Guo, S. Pakhomov, and G. Leroy, "Coherence and Comprehensibility: Large Language Models Predict Lay Understanding of Health-Related Content," *Journal of Biomedical Informatics* 161 (2025): 104758, https://doi.org/10.1016/j.jbi.2024.104758.

27. P. J. Chen, I. H. Hsu, Y. Y. Huang, and H. Y. Lee, "Mitigating the Impact of Speech Recognition Errors on Chatbot Using Sequence-To-Sequence Model," *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (2017): 497–503, https://doi.org/10.48550/arXiv.1709.07862.

28. M. Gish, A. Nowominski, and R. Dror, "Breaking the Ceiling: Mitigating Extreme Response Bias in Surveys Using an Open-Ended Adaptive-Testing System and LLM-Based Response Analysis," *AI* 7, no. 2 (2026): 73, https://doi.org/10.20944/preprints202601.0952.v1.

29. U. Gnewuch, S. Morana, M. Adam, and A. Maedche, "Faster Is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction," *Research Paper* 113 (2018): 1–17, https://aisel.aisnet.org/ecis2018_rp/113.

30. R. S. Goodman, J. R. Patrinely, C. A. Stone, Jr., et al., "Accuracy and Reliability of Chatbot Responses to Physician Questions," *JAMA Network Open* 6, no. 10 (2023): e2336483, https://doi.org/10.1001/jamanetworkopen.2023.36483.

31. CHART Collaborative, "Reporting Guidelines for Chatbot Health Advice Studies: Explanation and Elaboration for the Chatbot Assessment Reporting Tool (CHART)," *BMJ (Clinical Research Ed.)* 390 (2025): e083305, https://doi.org/10.1136/bmj-2024-083305.

32. B. Sezer and A. E. Okutan, "Evaluation of ChatGPT-4's Performance on Pediatric Dentistry Questions: Accuracy and Completeness Analysis," *BMC Oral Health* 25, no. 1 (2025): 1427, https://doi.org/10.1186/s12903-025-06791-9.

33. Ş. Mustuloğlu and B. P. Deniz, "Evaluation of Chatbots in the Emergency Management of Avulsion Injuries," *Dental Traumatology: Official Publication of International Association for Dental Traumatology* 41, no. 4 (2025): 437–444, https://doi.org/10.1111/edt.13041.

34. B. Sezer and T. Aydoğdu, "Performance of Advanced Artificial Intelligence Models in Pulp Therapy for Immature Permanent Teeth: A Comparison of ChatGPT-4 Omni, DeepSeek, and Gemini Advanced in Accuracy, Completeness, Response Time, and Readability," *Journal of Endodontics* 51, no. 11 (2025): 1675–1684, https://doi.org/10.1016/j.joen.2025.08.011.

35. B. Sezer and T. Aydoğdu, "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," *Applied Sciences* 15, no. 14 (2025): 7778, https://doi.org/10.3390/app15147778.

36. V. Kumar, A. Sachdeva, S. Sharma, et al., "Performance Comparison of a Domain-Specific Chatbot and General-Purpose Chatbots in Dental Traumatology," *Dental Traumatology: Official Publication of International Association for Dental Traumatology* (2025): 1–13, https://doi.org/10.1111/edt.70039.

37. N. Grinberg, S. Arbel, Y. Y. Boyadjiev, C. Ianculovici, S. Kleinman, and O. Peleg, "The Performance of Artificial Intelligence in Providing Real-Time Aid in Emergency Dental Trauma: A Clinical Validation Study," *Dental Traumatology: Official Publication of International Association for Dental Traumatology* (2025): 1–7, https://doi.org/10.1111/edt.70022.

38. M. Gökcek Taraç and T. Nale, "Artificial Intelligence in Pediatric Dental Trauma: Do Artificial Intelligence Chatbots Address Parental Concerns Effectively?," *BMC Oral Health* 25, no. 1 (2025): 736, https://doi.org/10.1186/s12903-025-06105-z.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1:** edt70072-sup-0001-Supinfo01.docx.

16009657, 2026, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/edt.70072 by Berkant Sezer - Canakkale Onsekiz Mart Uni , Wiley Online Library on [07/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License