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

Berkant Sezer, DDS, MSc, and Tuğba Aydoğdu, DDS

ABSTRACT

Introduction: This study aims to evaluate and compare the performance of three advanced chatbots—ChatGPT-4 Omni (ChatGPT-4o), DeepSeek, and Gemini Advanced—on answering questions related to pulp therapies for immature permanent teeth. The primary outcomes assessed were accuracy, completeness, and readability, while secondary outcomes focused on response time and potential correlations between these parameters. **Methods:** A total of 21 questions were developed based on clinical resources provided by the American Association of Endodontists, including position statements, clinical considerations, and treatment options guides, and assessed by three experienced pediatric dentists and three endodontists. Accuracy and completeness scores, as well as response times, were recorded, and readability was evaluated using Flesch Kincaid Reading Ease Score, Flesch Kincaid Grade Level, Gunning Fog Score, SMOG Index, and Coleman Liau Index. **Results:** Results revealed significant differences in accuracy ($P < .05$) and completeness ($P < .05$) scores among the chatbots, with ChatGPT-4o and DeepSeek outperforming Gemini Advanced in both categories. Significant differences in response times were also observed, with Gemini Advanced providing the quickest responses ($P < .001$). Additionally, correlations were found between accuracy and completeness scores (ρ: .719, $P < .001$), while response time showed a positive correlation with completeness (ρ: .144, $P < .05$). No significant correlation was found between accuracy and readability ($P > .05$). **Conclusions:** ChatGPT-4o and DeepSeek demonstrated superior performance in terms of accuracy and completeness when compared to Gemini Advanced. Regarding readability, DeepSeek scored the highest, while ChatGPT-4o showed the lowest. These findings highlight the importance of considering both the quality and readability of artificial intelligence-driven responses, in addition to response time, in clinical applications. *(J Endod 2025;51:1675–1684.)*

KEY WORDS

Artificial intelligence; ChatGPT; DeepSeek; endodontics; Gemini; large language models

SIGNIFICANCE

This study highlights the strengths and limitations of artificial intelligence chatbots in endodontics, emphasizing their potential to support evidence-based decision-making and enhance patient education through accurate, comprehensive, and readable responses tailored to varying health literacy levels.

Artificial intelligence (AI) refers to systems designed to replicate human cognitive functions such as learning, reasoning, and decision-making[1]. Among its most widely used applications in healthcare are chatbots powered by large language models (LLMs), which can interpret user input and generate human-like responses through natural language processing[2]. In medicine and dentistry, these tools are

From the Department of Pediatric Dentistry, School of Dentistry, Çanakkale Onsekiz Mart University, Çanakkale, Türkiye

Address requests for reprints to Berkant Sezer, Department of Pediatric Dentistry, School of Dentistry, Çanakkale Onsekiz Mart University, Çanakkale 17100, Türkiye.
E-mail address: dt.berkantsezer@gmail.com or berkant.sezer@comu.edu.tr
0099-2399/$ - see front matter

Copyright © 2025 American Association of Endodontists.
https://doi.org/10.1016/j.joen.2025.08.011

Downloaded for Berkant Sezer (dt.berkantsezer@gmail.com) at Çanakkale Onsekiz Mart University from ClinicalKey.com by Elsevier on December 10, 2025. For personal use only. No other uses without permission. Copyright ©2025. Elsevier Inc. All rights reserved.

increasingly used to assist in clinical decision-making, improve workflow efficiency, and enhance patient communication[1,2].

In recent years, multiple LLM-based chatbots have been developed, each with varying capabilities. Notably, OpenAI introduced ChatGPT-4 Omni (ChatGPT-4o)[3], a highly capable generative model for text-based interaction. Google launched the Gemini models (1.0 and 1.5), with advanced multimodal features[4], while DeepSeek emerged as a promising open source alternative[5]. The expanding role of such models in medicine and dentistry necessitates a clearer understanding of their clinical reliability and usability across different specialties[6,7].

AI has been increasingly integrated into various dental specialties, including restorative dentistry, orthodontics, oral and maxillofacial surgery, and periodontology, to support disease detection, treatment planning, and prognosis prediction[8-11]. In endodontics, AI has demonstrated significant potential in assisting with the detection of root fractures and periapical pathologies, determining working length, analyzing root morphology, identifying apical foramina, and predicting disease outcomes[8,12,13]. Recently, AI-based chatbots—powered by LLMs—have also emerged as promising tools in the field of endodontics. These systems can support clinicians by providing instant access to theoretical and clinical knowledge, aiding in decision-making, and offering preliminary responses to common patient queries[14]. Furthermore, they hold potential as innovative educational tools for dental students and professionals in endodontic training programs[15].

Immature permanent teeth requiring pulp therapy present a uniquely complex endodontic scenario due to their open apices, the need for continued root development, and the increased risk of treatment failure. Therefore, evaluating AI performance in this context allows for a nuanced assessment of clinical adequacy. Accuracy and completeness are crucial for ensuring patient safety and evidence-based decision-making, while response time and readability reflect real-world usability, especially in educational and patient-facing settings. Although previous studies have evaluated the performance of different AI models across various endodontic topics, to the best of our knowledge, no prior research has specifically assessed AI effectiveness in pulp therapy for immature permanent teeth. This study addresses these gaps by offering a multidimensional comparison of three advanced AI models—ChatGPT-4o, DeepSeek, and Gemini Advanced—across four key metrics. By integrating clinical complexity with a comprehensive evaluation framework, the present study aims to bridge the gap between technical capabilities and clinical applicability.

This study aims to conduct a comparative evaluation of the accuracy, completeness, response time, and readability of the answers provided by three advanced AI-powered chatbots—ChatGPT-4o, DeepSeek, and Gemini Advanced—in response to questions concerning pulp therapy for immature permanent teeth. The objective is to comprehensively assess the adequacy of clinical information offered by these models. Accordingly, the null hypothesis of this study is: There is no statistically significant difference in the accuracy, completeness, response time, and readability scores among ChatGPT-4o, DeepSeek, and Gemini Advanced.

MATERIALS AND METHODS

Study Design and Ethical Considerations

This study was designed as a comparative cross-sectional analysis involving the evaluation of LLM-based chatbots. The reporting of this study adheres to the TRIPOD-LLM (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis using Large Language Models) guidelines. Data collection was conducted in February 2025 using ChatGPT-4o, DeepSeek, and Gemini Advanced. Since this study did not involve human or animal participants, it was exempt from institutional review board approval.

Question Development and Content

The questions used in this study were developed based on clinical resources provided by the American Association of Endodontists (AAE),[16-18] including position statements, clinical considerations, and treatment options guides, along with related review articles[19-23]. These resources outline best practices in the pulp therapies of immature permanent teeth. Initially, a total of 26 questions were formulated to assess the accuracy, completeness, and readability of responses generated by ChatGPT-4o, DeepSeek, and Gemini Advanced.

To ensure content validity, the initial question set was reviewed by a panel consisting of three pediatric dentists and three endodontists, each with a minimum of 10 years of academic and clinical experience. The review process involved two rounds. In the first round, each expert independently assessed the clarity, clinical relevance, and evidence-based appropriateness of the 26 proposed questions. Questions that were redundant, lacked support in the current literature, or were deemed ambiguous were flagged for revision or removal. In the second round, the refined question set was re-evaluated to confirm content relevance and clarity. The Content Validity Index (CVI) was calculated for each item (I-CVI) and for the entire set (S-CVI). Items with I-CVI values below 0.78 were revised or excluded. The final 21-question set achieved a scale-level CVI (S-CVI) of 0.92, indicating excellent content validity. The complete list of questions is presented in Table 1.

Primary and Secondary Outcomes

The primary outcomes of this study were to compare the accuracy and completeness of responses generated by ChatGPT-4o, DeepSeek, and Gemini Advanced for questions related to the pulp therapy for immature permanent teeth. The secondary outcomes included the readability of AI-generated responses, the correlation between accuracy, completeness, readability, and response times, as well as the comparison of response times across the three chatbots.

Querying Procedure

The last premium version of ChatGPT (ChatGPT-4o; OpenAI, San Francisco, USA), DeepSeek (V3; Hangzhou DeepSeek Artificial Intelligence Basic Technology Research Co., Ltd.; Beijing, China), and Gemini Advanced (Google, Menlo Park, USA) were utilized in this study. Each question was entered into the chatbots individually and in English to ensure precise interpretation and reduce contextual interference. To eliminate any potential memory effects, each query was processed in an independent session. The original wording, punctuation, and syntax of the questions were strictly maintained during input. No modifications, pretesting, or prompt optimization were applied, ensuring that the study conditions closely resembled real-world use. Furthermore, no additional system prompts, role assignments, or context-setting instructions were used; each chatbot received only the original question text as input. All models operated under default settings, with no adjustments to temperature, top-p, max tokens, or other parameters. In rare cases where a response was incomplete or clearly hallucinatory (fewer than 5% of total queries), the same question was resubmitted once in a new session, and the more complete and clinically relevant response was retained for analysis.

Downloaded for Berkant Sezer (dt.berkantsezer@gmail.com) at Çanakkale Onsekiz Mart University from ClinicalKey.com by Elsevier on December 10, 2025. For personal use only. No other uses without permission. Copyright ©2025. Elsevier Inc. All rights reserved.

TABLE 1 - Questions List

Questions
1. What are the treatment approaches for immature permanent teeth exhibiting symptoms of reversible pulpitis?
2. What are the treatment approaches for immature permanent teeth exhibiting symptoms of irreversible pulpitis or pulp necrosis?
3. What is a protective liner for immature permanent teeth, and under what circumstances is it applied?
4. What is apexogenesis for immature permanent teeth, and what treatment methods are associated with this procedure?
5. What are the mechanisms of action of pulp capping materials utilized in apexogenesis for immature permanent teeth?
6. What clinical and radiographic criteria should be considered when assessing the success of apexogenesis for immature permanent teeth?
7. What is partial pulpotomy for immature permanent teeth, and how is it performed?
8. What are the fundamental biological and clinical differences between regenerative endodontic treatment and apexification in immature permanent teeth?
9. What is apexification for immature permanent teeth, and under what circumstances is it indicated?
10. What factors should be considered when determining the optimal number of treatment sessions for apexification for immature permanent teeth?
11. What is Frank's method in apexification for immature permanent teeth?
12. Which materials are most commonly used for apexification procedures in immature permanent teeth?
13. What are the advantages and disadvantages of mineral trioxide aggregate and calcium silicate-based materials in apexification for immature permanent teeth?
14. Does apexification contribute to an increase in root wall thickness in immature permanent teeth?
15. What defines regenerative endodontic treatment for immature permanent teeth?
16. Which characteristics are ideal for regenerative endodontic treatment for immature permanent teeth?
17. How is the regenerative endodontic treatment protocol recommended by the American Association of Endodontists implemented for immature permanent teeth?
18. Which biomaterials are commonly utilized in regenerative endodontic treatment for immature permanent teeth?
19. What are the most frequently encountered complications in regenerative endodontic treatment in immature permanent teeth, and how can they be prevented?
20. What is the role of ethylenediaminetetraacetic acid in the regenerative endodontic treatment protocol for immature permanent teeth?
21. According to the American Association of Endodontists, what are regenerative endodontic treatment's primary, secondary, and tertiary objectives?

To ensure consistency, a single observer submitted each question to all chatbots twice, with a 1-week interval between attempts. The initial responses were used for primary analysis, while the second set was analyzed to assess reliability. Agreement between the two query attempts was evaluated using Cohen's Kappa (κ), yielding the following values: ChatGPT-4o ($\kappa = .88$), DeepSeek ($\kappa = .85$), and Gemini Advanced ($\kappa = .87$). Additionally, the time taken for each chatbot to process the questions, from input to completion of the response, was recorded in seconds using an online stopwatch.

Responses from all three chatbots were recorded by the same researcher in a Microsoft Word document. To minimize the risk of evaluator bias and ensure objectivity, no information regarding which response belonged to which chatbot was included in the files.

Answers Evaluation

After collecting the responses from all three chatbots, the response files, along with the clinical resources and articles used for question formulation, were sent to three pediatric dentists and three endodontists, each with at least 10 years of academic and clinical experience, for evaluation. Importantly, these evaluators were independent from the expert group that contributed to the formulation of the questions. The materials provided included the chatbot responses, the full list of questions, and the reference documents used during question creation, such as position papers and consensus statements published by AAE. To ensure the reliability of the evaluation, the same group of evaluators was asked to re-assess the questions and responses—presented in a randomized order—two weeks after the initial evaluation. In cases of discrepancies between the initial and follow-up assessments, the specific questions and responses were resent for further analysis. The final scores from this third evaluation round were considered the definitive scores for each respective response. All evaluations were performed individually, without any discussion or consensus meetings among the experts, ensuring independent scoring.

The evaluators were asked to assess the chatbot responses based on two distinct criteria: accuracy and completeness. (i) A five-point Likert scale was utilized to evaluate accuracy. The scale was defined as follows: 1 indicated poor accuracy with a lack of flow, where most information was missing; 2 indicated generally poor accuracy with some information provided but significant gaps; 3 indicated moderate accuracy, with some important details covered but others lacking; 4 indicated good accuracy and flow, with most relevant information present; and 5 indicated excellent accuracy and flow[24,25]. (ii) A three-point Likert scale was applied to assess completeness. The definitions for this scale were: 1 indicated an incomplete response that addressed only some aspects of the question with key elements missing; 2 indicated an adequate response that covered all aspects of the question and included the necessary information for completeness; and 3 indicated a comprehensive response that addressed all aspects of the question and provided additional context or information beyond expectations[24,25].

Simultaneously with the evaluators' assessment process, the readability levels of the chatbot responses were evaluated. For this purpose, a dedicated website was used for data entry in the readability evaluation (https://readable.com/). The readability scores were determined using the Flesch-Kincaid Reading Ease Score, Flesch-Kincaid Grade Level, Gunning Fog Score, SMOG Index, and Coleman-Liau Index. While the evaluated questions were designed for professional use, readability assessment was incorporated to explore how comprehensible the AI-generated responses would be to users with varying levels of health literacy, including patients and dental students. Given that LLMs are increasingly used as sources of health

Downloaded for Berkant Sezer (dt.berkantsezer@gmail.com) at Çanakkale Onsekiz Mart University from ClinicalKey.com by Elsevier on December 10, 2025. For personal use only. No other uses without permission. Copyright ©2025. Elsevier Inc. All rights reserved.

TABLE 2 - Comparison of Accuracy and Completeness Scores, and Response Time for ChatGPT-4o, Deepseek, and Gemini Advanced in Answering Pulp Therapy for Immature Permanent Teeth Questions

Evaluation parameter	Chatbot			*P* value*
	ChatGPT-4o	Deepseek	Gemini advanced	
Accuracy				
Mean ± SD	4.57 ± 0.66[a]	4.43 ± 0.79[a]	4.29 ± 0.94[b]	**<.05**
Median (min.-max.)	5 (3–5)	5 (3–5)	5 (2–5)	
Completeness				
Mean ± SD	2.71 ± 0.45[a]	2.67 ± 0.47[a]	2.52 ± 0.59[b]	**<.05**
Median (min.-max.)	3 (2–3)	3 (2–3)	3 (1–3)	
Response time				
Mean ± SD	25.93 ± 8.53[a]	86.1 ± 16.83[b]	6.13 ± 1.22[c]	**<.001**
Median (min.-max.)	25.71 (8.1–39.29)	86.22 (49.6–118.75)	5.83 (3.52–8.7)	

max, maximum; min, minimum; SD, standard deviation.
The response time was measured in seconds, with a maximum of two decimal places. Bold values mean statistically significant ($P < .05$).
*One-way ANOVA with posthoc Tukey's test. Different superscript lower letters indicate significant difference within relevant rows.

information by the general public, we considered readability to be a relevant secondary parameter.

The Flesch Reading Ease Score is a widely used metric for determining the difficulty level of understanding a text. This readability test assigns a score on a scale from 0 to 100. Lower scores (0–10) indicate that the content is extremely difficult to understand, often requiring a higher education level, such as university training. On the other hand, higher scores (90–100) suggest that the text is easy to read, with an average 11-year-old student able to understand the material without difficulty. The Flesch-Kincaid Grade Level measures the education level needed for an individual to understand a particular text. This score is determined by analyzing the number of words, sentences, and syllables in the document. The outcome reflects the reading difficulty and corresponding U.S. grade level of the text[1]. The Gunning Fog Score is a readability measure that calculates the difficulty of a text based on the length of words and sentences. The index typically ranges from 1 to 18 or higher. The values represent the number of years of education needed for a reader to comprehend the text. A score around 8 is considered suitable for general public understanding, while texts scoring above 17 are typically aimed at graduate-level readers[26]. The SMOG Index estimates the number of years of education required for an average person to understand a text. It is commonly used in the healthcare industry and is regarded as one of the most precise readability metrics. The SMOG formula involves counting words with three or more syllables in three 10-sentence samples, taking the square root of this count, and adding three to determine the grade level[27]. The Coleman-Liau Index evaluates the readability of a text by using the number of sentences and letters as variables. It is often used alongside other readability scores, particularly for analyzing medical and legal documents[28].

Statistical Analysis

Data were entered using Microsoft Excel (version 16; Microsoft Corp., WA, USA). Continuous variables were presented as means with standard deviations, along with medians, minimum, and maximum values. The normality of the data was assessed using the Kolmogorov-Smirnov test, and since a significant portion of the data adhered to a normal distribution, parametric tests were employed. Differences in accuracy, completeness, response time, and readability scores between the chatbots were evaluated using a One-way ANOVA. In cases where significant differences were detected, post-hoc analysis using Tukey's test was performed to identify which groups differed. Pearson's correlation coefficient was applied to examine the relationships between parameters. Statistical analyses were conducted using the Statistical Package for Social Sciences, version 26 (IBM; Chicago, IL, USA), with a significance level set at .05.

RESULTS

The accuracy and completeness scores, along with the response times of ChatGPT-4o, DeepSeek, and Gemini Advanced to the 21 questions regarding pulp therapy for immature permanent teeth, are presented in Table 2. Statistically significant differences were observed in the accuracy scores between the three chatbots ($P < .05$). Post-hoc analysis revealed that ChatGPT-4o (4.57 ± 0.66) and DeepSeek (4.43 ± 0.79) demonstrated significantly higher accuracy scores compared to Gemini Advanced (4.29 ± 0.94). Similarly, statistically significant differences were found in the completeness scores of the chatbots ($P < .05$), with ChatGPT-4o (2.71 ± 0.45) and DeepSeek (2.67 ± 0.47) achieving significantly higher completeness scores than Gemini Advanced (2.52 ± 0.59). Additionally, there were significant differences in the response times among the chatbots ($P < .001$). DeepSeek (86.1 ± 16.83 seconds) was the chatbot with the longest response time, while Gemini Advanced (6.13 ± 1.22 seconds) provided the quickest responses.

The readability scores of the three chatbots, based on various readability assessment criteria and scoring systems, are presented in Table 3. According to the Flesch-Kincaid Reading Ease Score, DeepSeek (32.9 ± 6.6) achieved a significantly higher score, while ChatGPT-4o (23.24 ± 10.65) obtained the lowest score ($P < .001$). When evaluating the Flesch-Kincaid Grade Level, ChatGPT-4o (13.91 ± 3.18) was found to have a significantly higher grade level, whereas DeepSeek (11.93 ± 0.99) exhibited the lowest grade level ($P < .001$). Regarding the Gunning Fog Index, Gemini Advanced (15.94 ± 1.86) had a significantly higher score compared to DeepSeek (14.63 ± 1.2) ($P < .001$). Analysis of the SMOG Index means revealed that both ChatGPT-4o (12.13 ± 2.81) and Gemini Advanced (12.63 ± 1.28) had significantly higher scores than DeepSeek (10.93 ± 0.8) ($P < .001$). No significant differences were observed in the readability scores of the chatbots based on the Coleman-Liau Index ($P > .05$).

The results of the Pearson correlation analysis conducted to determine the correlations among all variables are presented in Table 4. A statistically significant positive correlation was found between accuracy and completeness scores (ρ: .719, $P < .001$). No correlation was identified between response time and accuracy ($P > .05$), but a positive and significant correlation was observed between response time and completeness (ρ: .144, $P < .05$). No relationship was found between accuracy and readability ($P > .05$); however, a statistically significant negative correlation was determined between the Gunning Fog score and completeness (ρ: −.106, $P < .05$). Furthermore, statistically significant positive or negative correlations were identified between all the readability

Downloaded for Berkant Sezer (dt.berkantsezer@gmail.com) at Çanakkale Onsekiz Mart University from ClinicalKey.com by Elsevier on December 10, 2025. For personal use only. No other uses without permission. Copyright ©2025. Elsevier Inc. All rights reserved.

TABLE 3 - Comparison of Readability Scores for ChatGPT-4o, Deepseek, and Gemini Advanced in Answering Pulp Therapy for Immature Permanent Teeth Questions

Readability score/index	Chatbot			*P* value*
	ChatGPT-4o	Deepseek	Gemini advanced	
Flesch Kincaid Reading Ease Score				
Mean ± SD	23.24 ± 10.65[a]	32.9 ± 6.6[b]	26.04 ± 9.41[c]	**<.001**
Median (min.-max.)	24.2 (−3.5 to 41.5)	32.3 (21–44.3)	25.1 (8.3–44.3)	
Flesch Kincaid Grade Level				
Mean ± SD	13.91 ± 3.18[a]	11.93 ± 0.99[b]	13.82 ± 1.82[c]	**<.001**
Median (min.-max.)	13.4 (10–22.7)	12.1 (10.1–13.6)	13.7 (10.7–16.5)	
Gunning Fog Score				
Mean ± SD	15.22 ± 3.64[ab]	14.63 ± 1.2[a]	15.94 ± 1.86[b]	**<.001**
Median (min.-max.)	14.6 (−10.4 to 24.9)	14.4 (12.4–17.2)	16 (13.3–19.1)	
SMOG Index				
Mean ± SD	12.13 ± 2.81[a]	10.93 ± 0.8[b]	12.63 ± 1.28[a]	**<.001**
Median (min.-max.)	10.9 (8.9–19.7)	10.9 (9.3–12.8)	12.7 (10.4–14.6)	
Coleman Liau Index				
Mean ± SD	16.43 ± 1.61	16.29 ± 1.34	16.3 ± 1.43	>.05
Median (min.-max.)	16.9 (13.2–19.1)	16.4 (14–18.2)	16.1 (13.7–19.4)	

max, maximum; min, minimum; SD, standard deviation.
Different superscript lower letters indicate significant difference within relevant rows. Bold values mean statistically significant ($P < .05$).
*One-way ANOVA with posthoc Tukey's test.

indices and scoring systems used for the assessment of readability ($P < .001$). A heatmap visualizing the correlations among all variables is provided in Figure 1.

DISCUSSION

This study evaluated the performance of three advanced AI-powered LLMs—ChatGPT-4o, DeepSeek, and Gemini Advanced—in generating responses to questions about pulp therapy for immature permanent teeth, using four key parameters: accuracy, completeness, response time, and readability. The results revealed statistically significant differences among the models, leading to the rejection of the null hypothesis for multiple criteria. These findings highlight that while all three models offer valuable support in clinical contexts, they differ notably in performance characteristics, with each exhibiting distinct strengths and limitations.

In this study, accuracy was defined as the extent to which an AI model provided scientifically sound, clinically relevant, and up-to-date information in response to a given question. The responses were evaluated against current literature and clinical resources, including consensus-based documents published by the AAE. Accuracy was determined based on alignment with these references, and responses were deemed accurate only if they were free from incorrect, incomplete, or misleading content. Previous studies have similarly assessed the accuracy of AI-based chatbots in dentistry. Ekmekci and Durmazpinar[29] reported that ChatGPT-4 with a PDF plugin achieved the highest accuracy (98.1%) in regenerative endodontics, while Gemini performed the worst (48%). Künzle and Paris[30] found that ChatGPT-4o outperformed other models—including Gemini 1.0 and ChatGPT-3.5—in answering undergraduate endodontic and restorative dentistry exam questions. In a diagnostic task involving text-based clinical cases, DeepSeek surpassed ChatGPT-4o (91.6% vs. 88.9%), and both significantly outperformed the average accuracy of medical journal readers (47.8%)[31]. These findings align with the present study, reinforcing that ChatGPT-4o and DeepSeek demonstrate superior accuracy compared to Gemini in dental contexts.

In this study, completeness refers to the extent to which AI-generated responses address all essential aspects of a clinical question, including evidence-based data, diagnostic criteria, treatment options, and potential outcomes in a well-structured manner. In endodontics—particularly in complex cases such as immature permanent teeth—comprehensive information is critical for accurate decision-making and clinical safety. Incomplete answers may lead to misinterpretation or inadequate treatment planning, highlighting the importance of both accuracy and completeness in AI-assisted dental practice. While several studies have evaluated the completeness of AI responses in various dental contexts, limited data exist specifically within endodontics. For instance, Şişman and Acar[32] reported that both ChatGPT-3.5 and Claude-instant provided responses deemed adequate or comprehensive in over 85% of oral surgery scenarios involving medically complex patients. In another study comparing ChatGPT-4 and Microsoft Copilot, ChatGPT-4 delivered fully comprehensive answers in 53.33% of cases, significantly outperforming Copilot[33]. Similarly, Vaira et al.[34] found that ChatGPT-4 produced more complete informed consent documents than both Gemini Advanced and oral surgery residents. These findings are consistent with the present study, in which ChatGPT-4o demonstrated superior completeness scores compared to Gemini Advanced. However, the scarcity of studies specifically targeting endodontics underscores the need for further research on AI performance in dental specialties.

Readability plays a vital role in health communication, as it reflects how easily the intended audience can comprehend the provided information. It is typically assessed using formulas that measure sentence length, word complexity, and syllable count, with results expressed in either ease scores or grade levels. The Flesch-Kincaid Reading Ease Score and Flesch-Kincaid Grade Level are two widely used tools for this purpose[1]. While no universally accepted standard exists for optimal readability in online health materials, both the American Medical Association and Health Education England recommend presenting such content at or below a sixth-grade reading level to ensure accessibility for the general population[35]. In this study, DeepSeek achieved significantly higher readability ease scores, indicating that

Downloaded for Berkant Sezer (dt.berkantsezer@gmail.com) at Çanakkale Onsekiz Mart University from ClinicalKey.com by Elsevier on December 10, 2025. For personal use only. No other uses without permission. Copyright ©2025. Elsevier Inc. All rights reserved.

TABLE 4 - Pearson Correlation Analysis Between Accuracy and Completeness Scores, Response Time, and Readability Scores for ChatGPT-4o, Deepseek, and Gemini Advanced

Variables	Accuracy	Completeness	Response Time	Flesch Kincaid Reading Ease Score	Flesch Kincaid Grade Level	Gunning Fog Score	SMOG Index	Coleman Liau Index
Accuracy								
ρ		**.719**	.067	−.030	.043	.006	.038	−.049
P value*		**<.001**	>.05	>.05	>.05	>.05	>.05	>.05
Completeness								
ρ	**.719**		**.144**	.017	−.051	**−.106**	−.074	−.006
P value*	**<.001**		**<.05**	>.05	>.05	**<.05**	>.05	>.05
Response time								
ρ	.067	**.144**		**.344**	**−.375**	**−.258**	**−.371**	−.034
P value*	>.05	**<.05**		**<.001**	**<.001**	**<.001**	**<.001**	>.05
Flesch Kincaid Reading Ease Score								
ρ	−.030	.017	**.344**		**−.905**	**−.760**	**−.804**	**−.764**
P value*	>.05	>.05	**<.001**		**<.001**	**<.001**	**<.001**	**<.001**
Flesch Kincaid Grade Level								
ρ	.043	−.051	**−.375**	**−.905**		**.914**	**.957**	**.549**
P value*	>.05	>.05	**<.001**	**<.001**		**<.001**	**< .001**	**<.001**
Gunning Fog Score								
ρ	.006	**−.106**	**−.258**	**−.760**	**.914**		**.931**	**.489**
P value*	>.05	**<.05**	**<.001**	**<.001**	**<.001**		**<.001**	**<.001**
SMOG Index								
ρ	.038	−.074	**−.371**	**−.804**	**.957**	**.931**		**.420**
P value*	>.05	>.05	**<.001**	**<.001**	**<.001**	**<.001**		**<.001**
Coleman Liau Index								
ρ	−.049	−.006	−.034	**−.764**	**.549**	**.489**	**.420**	
P value*	>.05	>.05	>.05	**<.001**	**<.001**	**<.001**	**<.001**	

Bold values mean statistically significant ($P < .05$).
ρ means Pearson's correlation coefficient (rho).
*Pearson correlation analysis.

Downloaded for Berkant Sezer (dt.berkantsezer@gmail.com) at Çanakkale Onsekiz Mart University from ClinicalKey.com by Elsevier on December 10, 2025. For personal use only. No other uses without permission. Copyright ©2025. Elsevier Inc. All rights reserved.

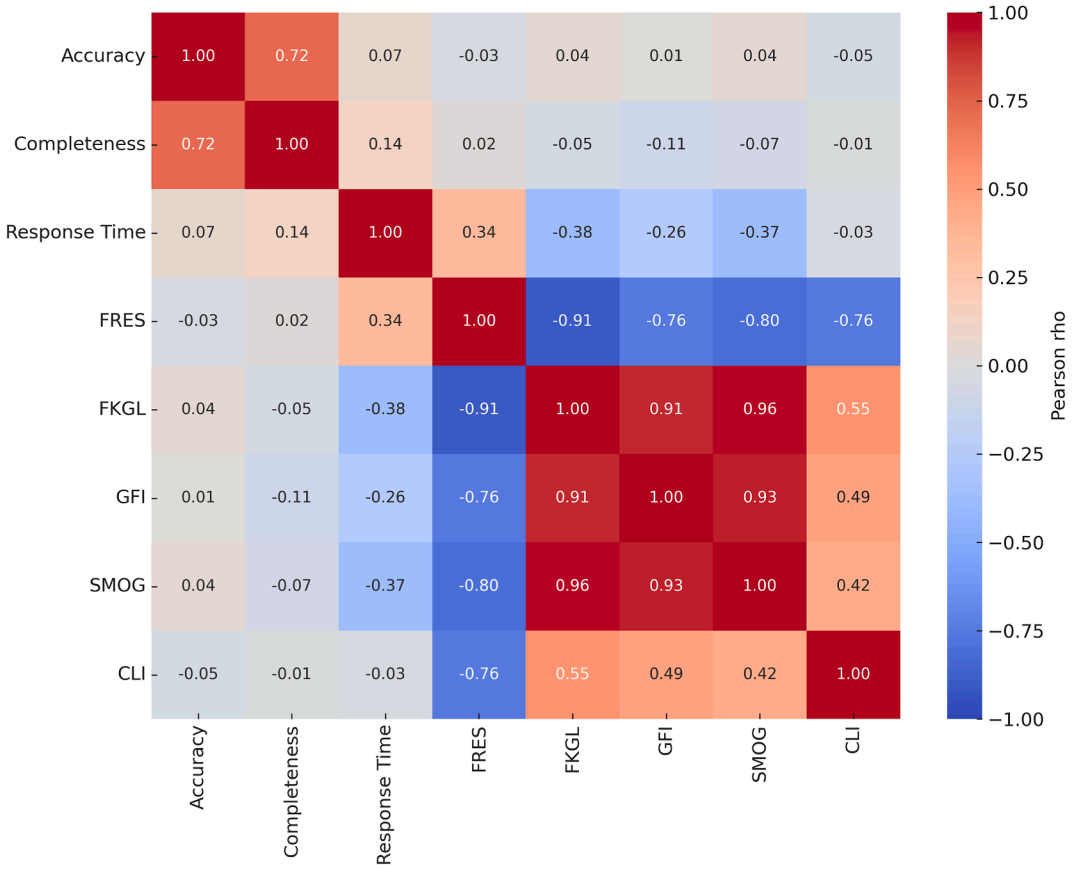


FIGURE 1 – Correlation heatmap illustrating the relationships among accuracy, completeness, response time, and readability scores based on Pearson's correlation coefficients (CLI, Coleman–Liau Index; FKGL, Flesch–Kincaid Grade Level; FRES, Flesch–Kincaid Reading Ease Score, GFI, Gunning Fog Index; SMOG, Simple Measure of Gobbledygook Index).

its content was simpler and more accessible. In contrast, ChatGPT-4o generated responses with the highest Flesch-Kincaid Grade Level, suggesting that its outputs required more advanced literacy skills. Although this may reflect the model's capacity to deliver in-depth clinical information, it could limit accessibility for users with lower health literacy. These findings highlight the trade-off between the depth of information and ease of comprehension in AI-generated health content.

In addition to the Flesch-Kincaid metrics, this study employed the Gunning Fog Index and SMOG Index, both of which estimate the years of formal education required to comprehend a given text based on sentence complexity and word structure[36]. DeepSeek consistently achieved lower scores on both indices, indicating a simpler and more accessible language compared to ChatGPT-4o and Gemini Advanced. The SMOG Index considered a reliable standard in health-related readability assessments[36], further supported the notion that DeepSeek provides more publicly accessible health communication. Interestingly, no significant correlation was observed between accuracy and readability, suggesting that higher quality clinical information does not necessarily equate to easier comprehension. Findings from previous studies corroborate these results. For example, Gemini demonstrated greater readability in a study comparing chatbot responses to frequently asked questions about orthodontic clear aligners, despite other models scoring higher in accuracy[37]. Similarly, research assessing ChatGPT models' responses on traumatic dental injuries noted that although content was generally accurate, it was written at a complexity level more suitable for individuals with advanced education[38]. In the domain of breast imaging, Gemini and Copilot yielded more readable responses, while ChatGPT-4 excelled in accuracy[39]. These patterns mirror the present findings, highlighting a trade-off between clinical precision and public readability, particularly in models like ChatGPT-4o. While this model excels in delivering high-quality, evidence-based content, its complexity may hinder its effectiveness in patient education, where clarity and accessibility are essential.

While previous studies have examined the readability of AI-generated content in various medical fields such as cardiology, orthodontics, and orthopedics, limited attention has been given to endodontics. Existing research in this area has primarily focused on information retrieved through search engines rather than evaluating responses generated by AI-based chatbots. This represents a notable gap in the literature, especially considering the growing use of digital tools in patient education. Clear and accessible AI-generated content on endodontic procedures could enhance patients' understanding, improve health literacy, and support shared decision-making—ultimately contributing to a more satisfactory treatment experience.

Although response time is a noteworthy parameter when evaluating AI models, it alone does not reflect overall performance. While rapid output may be beneficial in time-sensitive situations, it can compromise accuracy and depth, especially in clinical contexts. In the current study, Gemini Advanced demonstrated the shortest response time, whereas DeepSeek, despite being the slowest, produced more comprehensive and readable content. This highlights the importance of evaluating AI

Downloaded for Berkant Sezer (dt.berkantsezer@gmail.com) at Çanakkale Onsekiz Mart University from ClinicalKey.com by Elsevier on December 10, 2025. For personal use only. No other uses without permission. Copyright ©2025. Elsevier Inc. All rights reserved.

performance through a multidimensional lens, balancing speed with informational quality. Similarly, previous research comparing ChatGPT-3.5, ChatGPT-4o, Google Bard, and Microsoft Copilot in answering dental radiology questions from Türkiye's Dental Specialty Admission Exam found that ChatGPT-3.5 responded the fastest, while ChatGPT-4o was the slowest[40]. Another study assessing knowledge and comprehension on periodontal surgery reported that ChatGPT-4o required more time to generate responses than ChatGPT-3.5[41]. These findings align with our results, reinforcing that more advanced models like ChatGPT-4o tend to sacrifice speed for greater accuracy and depth. Notably, no prior studies have evaluated the response time of DeepSeek, which showed the longest latency in our analysis.

Our study identified a statistically significant positive correlation between accuracy and completeness scores, indicating that chatbot responses deemed more accurate were also rated as more complete. This finding is consistent with previous research conducted in the field of pediatric dentistry, where the performance of AI models was evaluated based on their responses to questions about traumatic dental injuries in primary dentition[42]. That study similarly reported a strong positive correlation between accuracy and completeness, supporting the notion that these two constructs often align when expert evaluations are applied. In contrast, a study investigating chatbot-generated information about 3D-printed orthodontic appliances found a negative correlation between accuracy and completeness, highlighting that the relationship between these two parameters may vary depending on the clinical context and the specificity of the topic[33]. Specifically, a moderate negative correlation was observed for ChatGPT-4, while a strong negative correlation was identified for Microsoft Copilot, indicating that increased accuracy in these models was associated with reduced completeness in that particular domain[33]. With respect to readability, our analysis revealed no significant correlation between accuracy and readability scores. This result aligns with findings from studies that assessed AI chatbot responses to clinical and patient-facing questions in fields such as myopia[43] and traumatic dental injuries[42], where no meaningful associations were observed between expert-rated content quality and objective readability indices. These findings collectively suggest that higher readability does not necessarily equate to higher clinical accuracy. Simplifying language to improve readability may compromise the precision of information, especially in complex endodontic contexts. Conversely, responses with high clinical validity may contain specialized terminology and complex sentence structures, potentially reducing readability scores. This highlights an important trade-off in the development and evaluation of AI models—balancing technical correctness with linguistic accessibility—especially when the end-users may include both specialists and nonspecialists.

This study presents several notable strengths. Unlike studies focusing on a single model, it offers a comparative evaluation of three advanced AI chatbots, including DeepSeek—a relatively underexplored tool in current literature. By utilizing five different readability indices instead of relying on a single metric, the study ensures a more objective assessment of linguistic complexity. Additionally, posing each question twice enabled evaluation of response consistency, enhancing methodological rigor. The questions were derived from up-to-date, evidence-based clinical resources, and the analysis was multidimensional, covering accuracy, completeness, response time, and readability. As such, this research stands out as one of the few comprehensive studies evaluating the clinical utility of AI-based chatbots in dentistry. However, the study has some limitations. It included a limited number of models, and source verification for chatbot-generated references was not performed—potentially excluding hallucinated citations. The exclusive use of open-ended questions, without image-based or multiple-choice formats, may restrict the scope of evaluation. Furthermore, the study's findings are specific to the model versions available at the time of data collection, and future updates may lead to different outcomes. The focus on pulp treatments for immature permanent teeth limits generalizability across other dental domains. Finally, the absence of visual prompts, such as radiographs, highlights a need for further research into multimodal capabilities of AI in clinical decision support.

CONCLUSION

This study compared three advanced AI-powered chatbots—ChatGPT-4o, DeepSeek, and Gemini Advanced—by evaluating their accuracy, completeness, response time, and readability within the context of pulp therapy for immature permanent teeth. ChatGPT-4o and DeepSeek outperformed Gemini in terms of content accuracy and completeness. Notably, ChatGPT-4o generated detailed, clinically robust responses, but its higher linguistic complexity may limit accessibility for individuals with lower health literacy. Conversely, DeepSeek provided more readable and accessible content, indicating its potential as a valuable tool for patient education and public health communication. Among the three models, DeepSeek consistently required a lower reading level to comprehend, making it more inclusive for a broader audience. Although Gemini Advanced demonstrated the fastest response time, it lagged in clinical accuracy and completeness, and its complex sentence structures may be more suitable for professional use rather than patient-facing applications.

In summary, this study underscores the importance of evaluating AI chatbots not only by the accuracy of their content but also by how effectively they meet the needs of users with varying literacy levels. Future research should expand on these findings by exploring the evolving capabilities of AI tools across different domains of dentistry and their integration into multimodal clinical decision-support systems.

ETHICS

Since this study did not involve human or animal participants, it was exempt from institutional review board approval.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

In writing this paper, ChatGPT-4 software was used only to check grammar and spelling rules. After using this tool, the authors reviewed and edited the content and takes full responsibility for the content of the publication.

ACKNOWLEDGMENTS

The authors deny any conflicts of interest related to this study.

Downloaded for Berkant Sezer (dt.berkantsezer@gmail.com) at Çanakkale Onsekiz Mart University from ClinicalKey.com by Elsevier on December 10, 2025. For personal use only. No other uses without permission. Copyright ©2025. Elsevier Inc. All rights reserved.

REFERENCES

1. Ozdemir ZM, Yapici E. Evaluating the accuracy, reliability, consistency, and readability of different large language models in restorative dentistry. J Esthet Restor Dent 2025;37:1740–52.
2. Reader A, Drum M. A review of ChatGPT as a reliable source of scientific information regarding endodontic local anesthesia. J Endod 2025;51:571–6.
3. Dündar Sarı MB, Sezer B. ChatGPT-4 Omni's accuracy in multiple-choice dentistry questions: a multidisciplinary and bilingual assessment. Essent Dent 2025;4:0029.
4. Taymour N, Fouda SM, Abdelrahaman HH, Hassan MG. Performance of the ChatGPT-3.5, ChatGPT-4, and google gemini large language models in responding to dental implantology inquiries. J Prosthet Dent 2025. epub ahead of print.
5. Temsah A, Alhasan K, Altamimi I, et al. DeepSeek in healthcare: revealing opportunities and steering challenges of a new open-source artificial intelligence frontier. Cureus 2025;17:e79221.
6. Chen ZH, Lin L, Wu CF, et al. Artificial intelligence for assisting cancer diagnosis and treatment in the era of precision medicine. Cancer Commun (Lond) 2021;41:1100–15.
7. Pradhan P. Accuracy of ChatGPT 3.5, 4.0, 4o and gemini in diagnosing oral potentially malignant lesions based on clinical case reports and image recognition. Med Oral Patol Oral Cir Bucal 2025;30:e224–31.
8. Karobari MI, Adil AH, Basheer SN, et al. Evaluation of the diagnostic and prognostic accuracy of artificial intelligence in endodontic dentistry: a comprehensive review of literature. Comput Math Methods Med 2023;2023:7049360.
9. Auconi P, Scazzocchio M, Cozza P, et al. Prediction of class III treatment outcomes through orthodontic data mining. Eur J Orthod 2015;37:257–67.
10. Kim BS, Yeom HG, Lee JH, et al. Deep learning-based prediction of paresthesia after third molar extraction: a preliminary study. Diagnostics (Basel) 2021;11:1572.
11. Ossowska A, Kusiak A, Świetlik D. Artificial intelligence in dentistry-narrative review. Int J Environ Res Public Health 2022;19:3449.
12. Orhan K, Bayrakdar IS, Ezhov M, et al. Evaluation of artificial intelligence for detecting periapical pathosis on cone-beam computed tomography scans. Int Endod J 2020;53:680–9.
13. Setzer FC, Shi KJ, Zhang Z, et al. Artificial intelligence for the computer-aided detection of periapical lesions in cone-beam computed tomographic images. J Endod 2020;46:987–93.
14. Mohammad-Rahimi H, Ourang SA, Pourhoseingholi MA, et al. Validity and reliability of artificial intelligence chatbots as public sources of information on endodontics. Int Endod J 2024;57:305–14.
15. Mohammad-Rahimi H, Setzer FC, Aminoshariae A, et al. Artificial intelligence chatbots in endodontic education-concepts and potential applications. Int Endod J 2025. epub ahead of print.
16. American Association of Endodontists (AAE). Guide clinical endodontics. Available at: https://www.aae.org/specialty/clinical-resources/guide-clinical-endodontics. Accessed February 27, 2025.
17. American Association of Endodontists (AAE). Guidelines position statements. Available at: https://www.aae.org/specialty/clinical-resources/guidelines-position-statements. Accessed February 27, 2025.
18. American Association of Endodontists (AAE). Regenerative endodontics. Available at: https://www.aae.org/specialty/clinical-resources/regenerative-endodontics. Accessed February 27, 2025.
19. Shabahang S. Treatment options: apexogenesis and apexification. J Endod 2013;39:S26–9.
20. Kahler B, Rossi-Fedele G, Chugal N, Lin LM. An evidence-based review of the efficacy of treatment approaches for immature permanent teeth with pulp necrosis. J Endod 2017;43:1052–7.
21. Hargreaves KM, Diogenes A, Teixeira FB. Treatment options: biological basis of regenerative endodontic procedures. J Endod 2013;39:S30–43.
22. Vidal K, Martin G, Lozano O, et al. Apical closure in apexification: a review and case report of apexification treatment of an immature permanent tooth with biodentine. J Endod 2016;42:730–4.
23. Torabinejad M, Nosrat A, Verma P, Udochukwu O. Regenerative endodontic treatment or mineral trioxide aggregate apical plug in teeth with necrotic pulps and open apices: a systematic review and meta-analysis. J Endod 2017;43:1806–20.

Downloaded for Berkant Sezer (dt.berkantsezer@gmail.com) at Çanakkale Onsekiz Mart University from ClinicalKey.com by Elsevier on December 10, 2025. For personal use only. No other uses without permission. Copyright ©2025. Elsevier Inc. All rights reserved.

24. De Vito A, Colpani A, Moi G, et al. Assessing ChatGPT's potential in HIV prevention communication: a comprehensive evaluation of accuracy, completeness, and inclusivity. AIDS Behav 2024;28:2746–54.

25. B Coskun BN, Yagiz B, Ocakoglu G, et al. Assessing the accuracy and completeness of artificial intelligence language models in providing information on methotrexate use. Rheumatol Int 2024;44:509–15.

26. Olszewski R, Watros K, Mańczak M, et al. Assessing the response quality and readability of chatbots in cardiovascular health, oncology, and psoriasis: a comparative study. Int J Med Inform 2024;190:105562.

27. Elkarmi R, Abu-Ghazaleh S, Sonbol H, et al. ChatGPT for parents' education about early childhood caries: a friend or foe? Int J Paediatr Dent 2025;35:717–24.

28. Ichhpujani P, Parmar UPS, Kumar S. Appropriateness and readability of google bard and ChatGPT-3.5 generated responses for surgical treatment of glaucoma. Rom J Ophthalmol 2024;68:243–8.

29. Ekmekci E, Durmazpinar PM. Evaluation of different artificial intelligence applications in responding to regenerative endodontic procedures. BMC Oral Health 2025;25:53.

30. Künzle P, Paris S. Performance of large language artificial intelligence models on solving restorative dentistry and endodontics student assessments. Clin Oral Investig 2024;28:575.

31. Diniz-Freitas M, Diz-Dios P. DeepSeek: another step forward in the diagnosis of oral lesions. J Dent Sci 2025;20:1904–7.

32. Şişman AÇ, Acar AH. Artificial intelligence-based chatbot assistance in clinical decision-making for medically complex patients in oral surgery: a comparative study. BMC Oral Health 2025;25:351.

33. Yousuf AM, Ikram F, Gulzar M, et al. Performance assessment of artificial intelligence chatbots (ChatGPT-4 and Copilot) for sharing insights on 3D-printed orthodontic appliances: a cross-sectional study. Int Orthod 2025;23:100992.

34. Vaira LA, Lechien JR, Maniaci A, et al. Evaluating AI-generated informed consent documents in oral surgery: a comparative study of ChatGPT-4, bard gemini advanced, and human-written consents. J Craniomaxillofac Surg 2025;53:18–23.

35. Frost JC, Baldwin AJ. Readability of online monkeypox patient education materials: Improved recognition of health literacy is needed for dissemination of infectious disease information. Infect Dis Health 2023;28:88–94.

36. Sharma S, Latif Z, Makuvire TT, et al. Readability and accessibility of patient-education materials for heart failure in the United States. J Card Fail 2025;31:154–7.

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.

38. Öztürk Z, Bal C, Çelikkaya BN. Evaluation of information provided by ChatGPT versions on traumatic dental injuries for dental students and professionals. Dent Traumatol 2025;41:427–36.

39. Tepe M, Emekli E. Assessing the responses of large language models (ChatGPT-4, gemini, and Microsoft Copilot) to frequently asked questions in breast imaging: a study on readability and accuracy. Cureus 2024;16:e59960.

40. Tassoker M. ChatGPT-4 Omni's superiority in answering multiple-choice oral radiology questions. BMC Oral Health 2025;25:173.

41. Li C, Zhang J, Abdul-Masih J, et al. Performance of ChatGPT and dental students on concepts of periodontal surgery. Eur J Dent Educ 2025;29:36–43.

42. 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.

43. Chang LC, Sun CC, Chen TH, et al. Evaluation of the quality and readability of ChatGPT responses to frequently asked questions about myopia in traditional Chinese language. Digit Health 2024;10:20552076241277021.

Downloaded for Berkant Sezer (dt.berkantsezer@gmail.com) at Çanakkale Onsekiz Mart University from ClinicalKey.com by Elsevier on December 10, 2025. For personal use only. No other uses without permission. Copyright ©2025. Elsevier Inc. All rights reserved.