IDR International Dental Research

Int Dent Res 2025; 15(3): 123-135
https://doi.org/10.5577/intdentres.662

Original Article Open Access

Comparative evaluation of the responses from ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.1 chatbots to patient inquiries about endodontic treatment in terms of accuracy, understandability, and readability

Makbule Taşyürek[1], Özkan Adıgüzel[1,2], Mustafa Gündoğar[3], Myroslav Goncharuk-Khomyn[4], Hatice Ortaç[5]

[1] Dicle University, Faculty of Dentistry, Department of Endodontics, Diyarbakır, Türkiye
[2] Nelcotech R&D Center, Department of Dentistry, Sheridan, WY, USA
[3] İstanbul Medipol University, Faculty of Dentistry, Department of Endodontics, İstanbul, Türkiye
[4] Uzhhorod National University, Faculty of Dentistry, Department of Restorative Dentistry, Uzhhorod, Ukraine
[5] Dicle University, Faculty of Medicine, Department of Biostatistics, Diyarbakır, Türkiye

Received: 14 September 2025
Accepted: 28 September 2025

Correspondence:

Dr. Makbule TAŞYÜREK
Dicle University, Faculty of Dentistry, Department of Endodontics, Diyarbakır, Türkiye.
E-mail: makbule.tasyurek@dicle.edu.tr

How to cite this article:

Taşyürek M, Adıgüzel Ö, Gündoğan M, Goncharuk-Khomyn M, Ortaç H. Comparative evaluation of the responses from ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.1 chatbots to patient inquiries about endodontic treatment in terms of accuracy, understandability, and readability. Int Dent Res. 2025;15(3):123-135. https://doi.org/10.5577/intdentres.662

Abstract

Aim: The aim of this study is to compare three large language models (LLMs) (ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.1) in terms of accuracy, understandability, and readability, based on the answers provided to frequently asked endodontic questions.

Methodology: Thirty open-ended frequently asked questions were generated using the AlsoAsked and AnswerThePublic websites. Two experienced endodontists scored the accuracy of the responses using a 5-point Likert scale. The understandability of the responses was analyzed using the Patient Education Materials Assessment Tool for Printed Materials (PEMAT-P) tool. Readability was assessed using the Flesch-Kincaid Readability Score (FRES), Flesch-Kincaid Grade Level (FKGL), Gunning Fog Index (GFI), Simple Measure of Gobbledygook (SMOG), and Coleman-Liau Index (CLI). Group comparisons were performed using the ANOVA/Kruskal-Wallis test, followed by post-hoc Dunn-Bonferroni tests.

Results: Inter-rater agreement was excellent (accuracy ICC: 0.908–0.917; reliability ICC: 0.992–0.995; all $p < 0.001$). A significant difference was found between the models in terms of accuracy ($p < 0.001$): DeepSeek-V3.1 (4.63 ± 0.81) scored the highest and performed significantly better than ChatGPT-5 (3.93 ± 0.79) and Gemini 2.5 Flash (3.67 ± 0.76). No significant difference between ChatGPT-5 and Gemini 2.5 Flash ($p > 0.05$). The understandability (PEMAT-P) scores were similar ($p = 0.683$), and all models scored above 70% (ChatGPT-5, 77.46%; Gemini, 76.04%; DeepSeek-V3.1, 77.57%). Differences were found in readability metrics: DeepSeek-V3.1 scored higher than ChatGPT-5 in FRES ($p = 0.044$); Gemini scored higher than DeepSeek-V3.1 in FKGL ($p = 0.001$); in GFI, Gemini 2.5 Flash scored higher than both ChatGPT-5 ($p = 0.036$) and DeepSeek-V3.1 ($p < 0.001$); in SMOG, Gemini outperformed DeepSeek-V3.1 ($p = 0.003$); and in CLI, ChatGPT-5 was higher than DeepSeek-V3.1 ($p = 0.004$). No significant correlation was found between readability and understandability ($p > 0.05$).

Conclusion: DeepSeek-V3.1 outperformed ChatGPT-5 and Gemini 2.5 Flash in terms of accuracy. While all models produced similar scores above the PEMAT-P understandability threshold (70%), there were significant differences in readability metrics; furthermore, no model consistently reached the recommended 6th–8th grade level.

Keywords: Artificial intelligence, large language models (LLMs), ChatGPT, Gemini, DeepSeek, patient questions in endodontics, accuracy, understandability, PEMAT-P, readability

Introduction

In the context of modern healthcare, the era in which patients were solely dependent on the medical information and advice they received from their dentists has ended. With the widespread use of the Internet, individuals can now access comprehensive information about symptoms, diseases, and treatment methods through a simple online search; thus, they are becoming more informed, engaged, and proactive in their healthcare decision-making processes (1).

The Internet has fundamentally changed the way patients access their health information, leading to the empowerment of patients and reshaping the dynamics of doctor-patient communication (2). However, the unreliability of all information available online poses a significant problem. Encountering content that lacks a scientific basis or is inaccurate can cause confusion among patients and lead to misdiagnosis (3). Therefore, the spread of incorrect or harmful information about endodontic treatment online can be a significant cause for concern. Moreover, this situation can further increase patients' fear of the dentist, negatively affect the process, make it difficult for them to access accurate information, and hinder their ability to develop a positive attitude toward treatment.

Large language models (LLMs) are advanced artificial intelligence systems designed to generate human-like text. They are trained on massive amounts of data and can understand and generate natural language for various tasks, such as translation, summarization, and conversation. Users provide keywords or lists of questions, and the LLMs generate content on these topics. The user interface typically follows a conversational structure that iterates between user questions or inputs and system responses or outputs. This design considers previous interactions to effectively mimic human conversations (4).

Chat Generative Pre-Trained Transformer (ChatGPT; OpenAI, San Francisco, CA, USA) is a generative artificial intelligence chatbot developed by OpenAI and released on November 30, 2022 (5). ChatGPT-5, introduced and made available on August 7, 2025, was described by its developers as an intelligent and efficient foundational model. According to the manufacturer's statement, ChatGPT-5 is positioned as the most advanced model developed in the field of healthcare and provides strong support for raising users' health awareness. Demonstrating clear superiority over previous versions in HealthBench performance evaluations, ChatGPT-5 stands out because of its ability to flag potential risks in advance. It also stands out for its ability to ask users additional questions to generate more qualified and useful responses. Although it is not intended to replace a medical professional, ChatGPT-5 was designed to assist in interpreting results, asking the right questions, and making informed decisions. It has been reported that ChatGPT-5 significantly reduces the hallucination rate compared to previous models, particularly in evaluation frameworks (such as LongFact and FactScore) that involve complex, multidimensional, and open-ended questions where factual accuracy is critical (6).

ChatGPT-5, which advances from previous ChatGPT models, presents a unified model architecture that incorporates real-time routing, improved reasoning, and significant enhancements in factual accuracy and multimodal functionality. ChatGPT-5 demonstrates enhanced instruction-following, increased reliability, and reduced hallucination rates across domains, including coding, healthcare, and writing. In the health domain, ChatGPT-5 scored 46.2% on HealthBench Hard, significantly outperforming previous models (6).

ChatGPT has significant potential to address the healthcare needs of disadvantaged communities. First, its multilingual communication capabilities can reduce structural barriers to accessing health information by overcoming language barriers to access. This makes it easier for individuals, especially in areas with limited access to healthcare services, to obtain accurate and reliable information about their health status. Furthermore, considering the complexity of healthcare systems in developing countries, ChatGPT can increase the rate of healthcare utilization by facilitating the understanding of processes, such as insurance registration, appointment scheduling, and medication use. Overall, the AI-based support offered by ChatGPT can contribute to the democratization of health information (7, 8).

Gemini is an LLM developed by Google DeepMind (Mountain View, CA, USA) and was released on December 19, 2024. It combines language understanding with advanced reasoning and is designed for tasks such as answering questions, explaining concepts, and solving problems across various subjects (9). Gemini 2.5, introduced on March 26, 2025, is described by its manufacturer as an enhanced “thinking model” designed to solve increasingly complex problems. This model family has a structure that can evaluate information by first passing it through a cognitive logic filter, rather than directly transferring it during the response generation process. Thus, Gemini 2.5 not only provides performance gains but also significantly increases the level of accuracy and reliability (10).

DeepSeek (Hangzhou DeepSeek AI Basic Tech. Res. Co., Ltd.; Beijing, China) is a new LLM developed by a Chinese research team. It focuses specifically on academic and professional use cases, namely, science, technology, engineering, and mathematics (STEM) fields. DeepSeek can solve mathematical problems, generate codes, and provide support for research-related questions. It also excels in tasks requiring professional knowledge and logical reasoning, such as those in science, mathematics, and programming (11). According to the manufacturer's statement, DeepSeek-V3.1 is a significant innovation symbolizing the transition to the “agent era” in artificial intelligence. The model combines the ‘Think’ and “Non-Think” modes into a single structure owing to its hybrid inference architecture, which adapts to flexible usage scenarios. The new “Think” mode offers higher speed and

efficiency compared with the previous version, DeepSeek-R1-0528, making it particularly suitable for time-critical applications. Post-training improvements significantly enhanced the vehicle-handling capabilities of the model and its ability to perform multi-step tasks. In this context, the manufacturer emphasizes that DeepSeek-V3.1 not only offers increased performance but also sets a new standard in AI research with its advanced agent capabilities (12).

Endodontics is a specialized field of dentistry that focuses on the prevention, diagnosis, and treatment of diseases of the dental pulp and its surrounding tissues (13). This discipline involves complex clinical procedures that often require advanced knowledge and skill. The symptoms of endodontic diseases can significantly affect not only oral and dental health but also the overall quality of life of individuals. Therefore, endodontics plays a critical role in protecting patients' health, both functionally and aesthetically (14).

The adequacy of LLMs in making definitive diagnoses or recommending treatments in the field of endodontics is debatable. This is because making the right decision in this field relies not only on theoretical knowledge but also on clinical examination findings, imaging results, and the dentist's experience. However, LLMs can play a valuable guiding role in making the process more understandable for patients, ensuring that they ask the right questions, and strengthening communication with dentists.

The aim of this study is to compare three different LLMs (ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.1) in terms of accuracy, understandability (PEMAT-P), and readability for frequently asked questions by patients regarding endodontic treatment. The findings aim to identify the current limitations of using LLMs for patient education and guide users in employing emerging AI-based chatbots from a more informed and critical perspective. The evaluation was based on three key criteria: accuracy, understandability (PEMAT-P), and readability. Accordingly, the following null hypotheses were proposed:

- H_{01} (Accuracy): There is no statistically significant difference among ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.1 chatbots in terms of the accuracy of their responses.
- H_{02} (Understandability [PEMAT-P]): There is no statistically significant difference among the models in terms of the understandability (PEMAT-P) of the provided information.
- H_{03} (Readability): There is no statistically significant difference among the models regarding the readability of their responses.

Materials And Methods

Ethical committee approval was not required for this research because the study did not involve any patient information or data; it only aimed to analyze freely accessible online information sources.

To learn what questions patients might ask about endodontic treatment, we used popular question-answer sites, including Also Asked (alsoasked.com) in our online research (15) and Answer the Public (answerthepublic.com) (16). We used these tools. These tools are commonly used to identify the most frequently asked questions and topics that attract users' interest. Keywords used in the search include "root canal treatment and pain" and "root canal treatment and tooth abscess."

Thirty open-ended questions generated by online tools were compiled by one of the authors into a Microsoft Word document (Microsoft Corp., Microsoft Office 365, V 16.74, Redmond, WA, USA) (Table 1) and subsequently reviewed by two authors to identify and remove duplicate or semantically identical questions. Furthermore, all included questions were examined for grammatical accuracy, and those containing linguistic or stylistic errors were revised to ensure clarity.

Table 1. List of questions asked of the chatbots.

Questions
1. How long does root canal treatment take?
2. How many sessions does root canal treatment require?
3. Can root canal treatment be performed in a single session?
4. What is done during the second session of root canal treatment?
5. Can I undergo root canal treatment if I have a tooth abscess?
6. Can infection occur again after root canal treatment?
7. Will I feel pain when the nerves in my tooth are removed during the root canal treatment?
8. Should I take antibiotics before undergoing root canal treatment?
9. What should I do if the temporary filling falls out during the root canal treatment?
10. Are antibiotics used after root canal treatment?
11. Can a tooth with root canal treatment decay?
12. Can root canal treatment be performed while taking antibiotics?
13. What happens to my tooth if I do not undergo root canal treatment?
14. What helps with toothache caused by an abscess?
15. Can an abscess form in a tooth that has undergone root canal treatment?
16. How many hours after receiving root canal treatment should I eat?
17. Can I smoke after having root canal treatment?
18. How long does the numbness last after root canal treatment?
19. How many years will a tooth that has undergone root canal treatment remain healthy?
20. Should I undergo root canal treatment or have the tooth extracted?
21. Should I undergo root canal treatment or an implant?
22. Why does a tooth that has had root canal treatment hurt later on?
23. What happens if the root canal treatment fails?
24. Should you brush your teeth after root canal treatment?
25. Does root canal treatment darken the color of teeth?
26. Is root canal treatment harmful during pregnancy?

27.	Can root canal treatment be performed twice on the same tooth?
28.	What are the disadvantages of root canal treatment?
29.	Is there pain during the second session of root canal treatment?
30.	What is placed inside a tooth that has undergone root canal treatment?
Total 30 Questions	

The questions were posed by an independent volunteer who did not participate in the study on September 2, 2025, using newly created accounts on the same computer as the original study. All chatbots (ChatGPT-5 [Thinking mode], Gemini 2.5 Flash [Deep Research mode], and DeepSeek-V3.1 [Deep Think mode]) were asked questions in the specified order via their official web addresses. To prevent the LLMs from being influenced by previous questions and answers, each question was asked in a separate chat session. At the beginning of each session, the system memory was reset, deleting past conversations to ensure that the model focused only on the current input. This method aims to minimize the possibility of contextual bias and prejudice arising from sequential interactions. Furthermore, the models were only allowed to provide a single response, and it was not possible to modify or regenerate the initial responses. The responses were recorded in a Microsoft Word document.

To ensure that the researchers remained impartial when evaluating the responses, each chatbot response was marked with a different color code. This allowed researchers to evaluate responses without knowing which response belonged to which chatbot. After theevaluation process was completed, the color codes were revealed to indicate which chatbots they represented.

The LLM responses were independently evaluated by two experienced endodontists. They were given time to assess the quality of their responses. To create a consistent and systematic scoring process, all evaluators were provided with a standardized Microsoft Excel scoring form (Microsoft Corp., Microsoft 365, Version 16.74, Redmond, WA, USA).

In the evaluation conducted by experts, a five-point Likert scale was used to examine the accuracy of the LLM responses. The evaluation process was conducted using a double-blind design to minimize the risk of bias.

- 1= Very poor: Responses were not factual or consistent; key details were missing.
- 2= Poor: Some correct elements are present, but important content is missing or unclear.
- 3= Average: Partially correct, containing both correct and missing or unclear information.
- 4= Good: Mostly correct and consistent, with only minor omissions.
- 5=Excellent: Extremely accurate, comprehensive, and well-organized response.

The readability of each LLM's responses was examined using an online tool called Readable software (https://readable.com). The readability assessment was based on five different established criteria, and scores were determined according to these metrics. Schematic representation of the study is graphically illustrated in Figure 1.

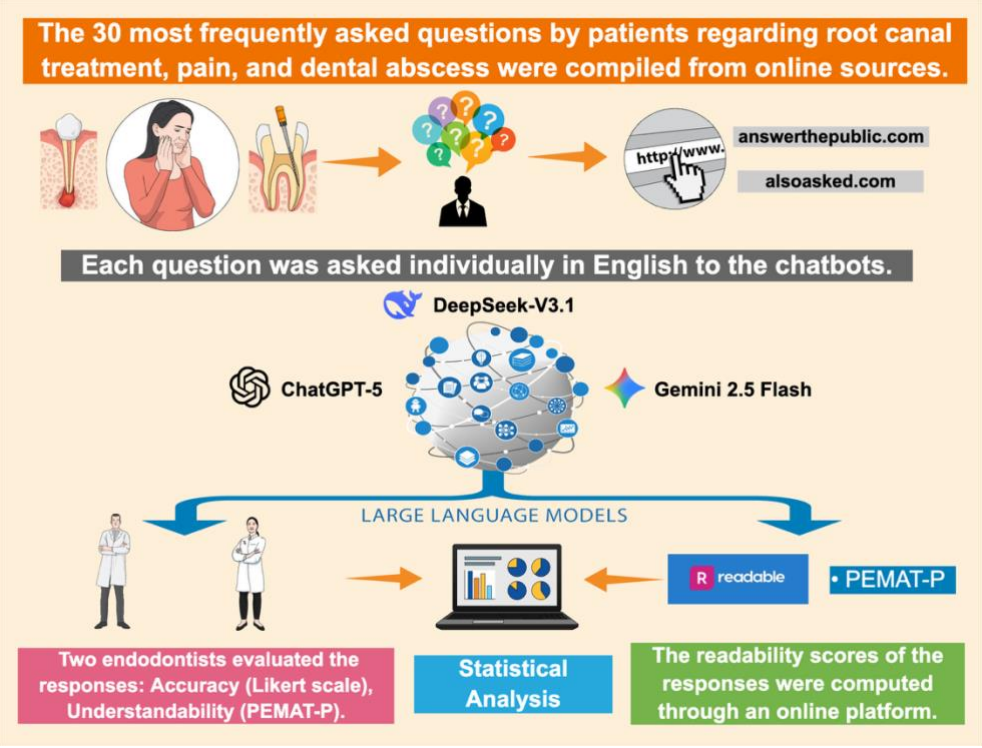

Figure 1. Schematic representation of the study.

Flesch-Kincaid Readability Score (FRES): Calculated using the average number of words per sentence and the average number of syllables per word, FRES provides a score between 0 and 100, with higher values indicating easier-to-read text (Table 2) (17, 18).

Flesch-Kincaid Grade Level (FKGL): Estimates the grade level in the United States required to understand the text based on the number of syllables per word and sentence length (Table 2) (19).

Table 2. Interpretation of the Flesch Reading Ease Score (FRES) and Flesch-Kincaid Grade Level (FKGL).

FRES	Reading Level	FKGL	Estimated Reading Graduate Level
90-100	Very Easy	5	5th Grade
80-89	Easy	6	6th Grade
70-79	Fairly Easy	7	7th Grade
60-69	Standard	8-9	8th-9th Grade
50-59	Fairly Difficult	10-12	High School
30-49	Difficult	13-16	College Level
0-29	Very Difficult	>16	College Graduate

Gunning Fog Index (GFI): Typically scores between 6 and 17; 6 corresponds to the reading level of an 11-12-year-old child, 12 to a high school graduate, and 17 to a college graduate (17).

Simple Measure of Gobbledygook (SMOG): It is considered the gold standard for evaluating health materials because it estimates the reading level required for 100% comprehension (18).

Coleman-Liau Index (CLI): Calculates readability based on the average number of characters per 100 words and the average sentence length, and assigns a score according to the grade level in the United States; lower values indicate easier readability (19).

The main readability indexes frequently used in health communication research and their calculation formulas are presented in Table 3 (20). These formulas estimate the grade level or complexity of a text based on sentence length, word structure, and counts of syllables or letters.

The clinical readability guidelines published by the National Institutes of Health (NIH) and the American Medical Association (AMA) recommend that patient-facing materials be written at a 6th to 8th grade level. The following threshold values were used to determine the readability rate of chatbot responses that met these standards (17-19) (Table 4).

Table 3. Readability indexes and calculation formulas.

Indexes	Formulas
FRES	$206.835 - 1.015 \times (\text{total words} / \text{total sentences}) - 84.6 \times (\text{total syllables} / \text{total words})$
FKGL	$0.39 \times (\text{total words} / \text{total sentences}) + 11.8 \times (\text{total syllables} / \text{total words}) - 15.59$
SMOG	$1.0430 \times \sqrt{[\text{Number of polysyllabic words} \times (30 / \text{Number of sentences})]} + 3.1291$
GFI	$0.4 \times [(\text{total words} / \text{total sentences}) + 100 \times (\text{complex words} / \text{total words})]$
CLI	$(0.0588 \times L) - (0.296 \times S) - 15.8$

L=average number of letters per 100 words; S = average number of sentences per 100 words

Table 4. Criteria and threshold values used to evaluate the readability of chatbot responses.

Criterion	Threshold Value
Flesch Reading Ease (FRES)	≥ 60
Flesch-Kincaid Grade Level (FKGL)	≤ 8
Simple Measure of Gobbledygook (SMOG)	< 9
Gunning Fog Index (GFI)	≤ 8
Coleman-Liau Index (CLI)	≤ 8

To assess the suitability of the responses generated by ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.1 as educational materials for all patients in terms of understandability and applicability, the Patient Education Materials Assessment Tool for Printed Materials (PEMAT-P), a validated measurement tool, was used (21).

The PEMAT-P consists of 24 questions divided into two subsections that measure understandability and applicability. Since the evaluation was conducted without visual aids and chatbots cannot generate visual content, questions related to visual elements were marked as "inapplicable" and excluded from the analysis. Two independent authors scored all responses

provided by each chatbot according to the PEMAT-P user's guide. Each question's answer was scored using a rating system based on the items specified in the guide: 0 for "disagree," 1 for "agree," and N/A for "inapplicable.'' (22).

PEMAT Understandability Score Calculation Method
Understandability Score (%) = (Total Score / Total Valid Items) × 100 (23)
· **Total Score:** The number of items marked "Agree." Each 'Agree' response is worth 1 point (23).
· **Total Valid Items:** The number of items remaining after excluding all items marked "Not Applicable (N/A)." (23).

Table 5 presents the score distributions of the understandability items. Accordingly, 12 items were marked as "Agree," each valued at 1 point, resulting in a total score of 12 points. Three items were marked as "Disagree" and calculated as 0 points. One item was considered "Not Applicable (N/A)" and was therefore excluded from the assessment. Based on these results, the total score was determined as 12. Table 6 shows the steps for calculating the PEMAT understandability score. In the first step, the total score was found to be 12. In the second step, the number of valid items was calculated as 15 by subtracting 1 N/A from the 16 items. In the final step, the total score (12) was divided by the number of valid items (15) and multiplied by 100, resulting in an understandability score of 80% (23).

Table 5. Score distribution of the understandability items.

Article Type	Number of Articles	Coefficient	Points Calculation
I agree	12	1	12 × 1 = 12
I disagree	3	0	3 × 0 = 0
Not applicable (N/A)	1	-	Excluded
TOPLAM			12

Table 6. Steps for calculating the PEMAT understandability score.

Step	Explanation	Calculation	Result
Step 1	Find the total score	12 items × 1 point	12 points
Step 2	Calculate the number of valid items	16 items - 1 N/A = 15 items	15 items
Step 3	Calculate the score	(12 ÷ 15) × 100	%80

The scores provided by the two evaluators were transferred to the table, resulting in two separate scores for each question. The consistency between these scores was then examined, and once consistency was established, the final result was determined based on the score of a single evaluator. Importantly, the authors who performed this evaluation were different individuals from the researcher who directed the questions to the chatbots. This ensured an unbiased and independent evaluation of the results. Ultimately, the understandability scores were calculated as percentages, with higher percentages indicating better levels of understandability.

Statistical analysis

SPSS 25.0 software (IBM Corp., IBM SPSS Statistics for Windows, Armonk, NY, USA) was used for statistical analyses, and the type I error rate was set at 5%.

The Intraclass Correlation Coefficient (ICC) was calculated to assess the agreement between the two evaluators' scores.

The normality of the continuous variables was examined using the Shapiro-Wilk test. According to the normality test results, variables that showed normality were expressed as mean ± standard deviation, while variables that did not show normality were expressed as median (minimum: maximum) values. According to the normality test results, the Kruskal-Wallis test was used when there were more than two groups and normality was not observed. Spearman's correlation analysis was performed to assess the relationships among accuracy, PEMAT-P, and readability scores.

Following the Kruskal-Wallis test, subgroup analyses were performed using the Dunn-Bonferroni test in cases of overall significance.

Results

The consistency between the two evaluators in terms of accuracy scores was assessed using a single-measure two-way consistency type coefficient [ICC(2,1), two-way random-effects, absolute agreement, single measures] for all artificial intelligence models. The inter-evaluator reliability was excellent for all models. For ChatGPT-5, ICC(2,1) = 0.917 (95% CI: 0.833-0.959, $p < 0.001$), for Gemini 2.5 Flash, ICC(2,1) = 0.908 (95% CI: 0.817-0.955, $p < 0.001$) and ICC(2,1) = 0.910 (95% CI: 0.820-0.956, $p < 0.001$) for DeepSeek-V3.1 (Table 7).

Table 7. Consistency between the two evaluators in terms of accuracy scores.

	ICC(2,1)	95% CI	*p*-value
Model			
ChatGPT-5	0.917	0.833-0.959	<0.001
Gemini 2.5 Flash	0.908	0.817-0.955	<0.001
DeepSeek-V3.1	0.910	0.820-0.956	<0.001

CI: Confidence Interval

The consistency between the two evaluators in terms of the understandability scores was assessed using a single-measure two-way consistency coefficient [ICC(2,1), two-way random effects, absolute agreement, single measures] for all AI models. The inter-rater reliability was found to be excellent for all models. For ChatGPT-5, ICC (2,1) = 0.995 (95% CI: 0.990-0.998, p < 0.001), for Gemini 2.5 Flash, ICC(2,1) = 0.992 (95% CI: 0.983-0.996, p < 0.001) and ICC(2,1) = 0.993 (95% CI: 0.986-0.997, p < 0.001) for DeepSeek-V3.1 (Table 8).

Table 8. Consistency between the two evaluators in terms of the understandability scores.

	ICC (2,1)	95% CI	p-value
Model			
ChatGPT-5	0.995	0.990-0.998	<0.001
Gemini 2.5 Flash	0.992	0.983-0.996	<0.001
DeepSeek-V3.1	0.993	0.986-0.997	<0.001

Accuracy

In the analysis conducted to guide the use of LLMs with a more informed and critical perspective regarding patients' frequently asked questions about endodontic treatment, the accuracy scores of responses to open-ended questions directed at three different LLMs were compared, and significant differences were found between these scores (p < 0.001) (Table 9) (Fig. 2).

Within the subgroup analyses, it was determined that the average accuracy score of DeepSeek-V3.1 (4.63±0.81) responses to open-ended questions was significantly higher than the average scores of ChatGPT-5 (3.93±0.79) and Gemini 2.5 Flash (3.67±0.76) responses (p < 0.001 and p < 0.001, respectively). No significant differences were found between the other two groups (p > 0.05) (Table 9) (Fig. 2).

Understandability

In the analysis conducted to guide the use of LLMs with a more conscious and critical perspective regarding patients' frequently asked questions about endodontic treatment, the understandability (PEMAT-P) scores of the responses to open-ended questions directed at three different LLMs were compared, and no significant difference was found between these scores (p = 0.683) (Table 7) (Fig. 3).

Readability

It was found that the FRES, FKGL, GFI, SMOG, and CLI scores for the readability of the responses to open-ended questions directed at the three different LLMs showed significant differences (p = 0.024, p = 0.001, p < 0.001, p = 0.004, respectively) (Table 7).

In the subgroup analysis of FRES scores, it was found that the average FRES score of DeepSeek-V3.1 (58.21±6.74) responses to open-ended questions was significantly higher than the average score of ChatGPT-5 (52.64±10.99) responses (p = 0.044). No significant differences were found between the other two groups (p > 0.05) (Table 7) (Fig. 4).

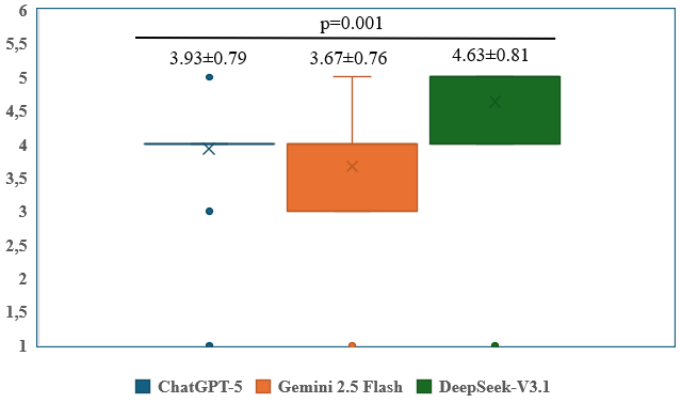

Figure 2. Mean accuracy scores of ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.1 in responding to patient questions regarding endodontic treatment.

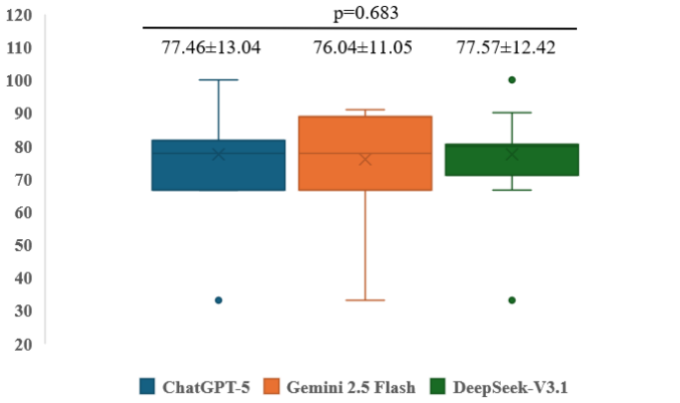

Figure 3. Mean understandability scores (PEMAT-P) of ChatGPT-5, Gemini 2.5 Flash and DeepSeek-V3.1 in responding to questions related patient questions on endodontic treatment.

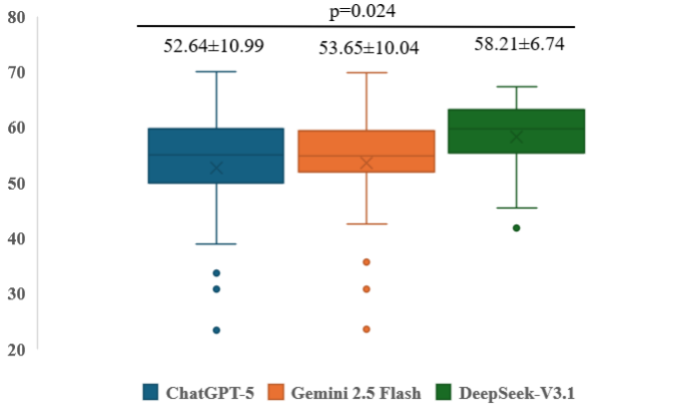

Figure 4. Mean Readability-FRES scores of ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.1 in responding to patient questions on endodontic treatment.

In the subgroup analysis of FKGL scores, it was determined that the average FKGL score of Gemini 2.5 Flash (10.15±2.14) responses to open-ended questions was significantly higher than the average score of DeepSeek-V3.1 (8.48±1.17) responses (p = 0.001). No significant differences were found between the other pairs (p > 0.05) (Table 7) (Fig. 5).

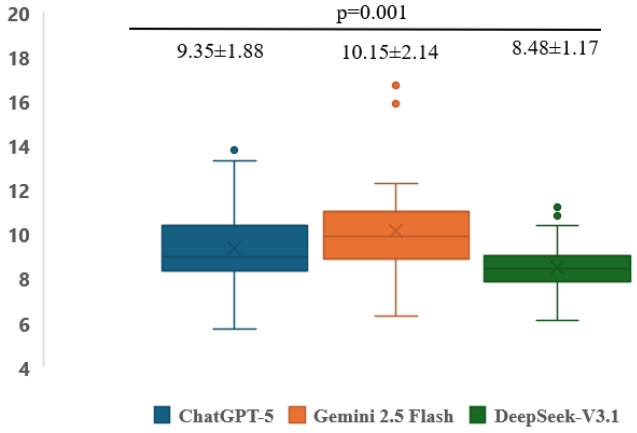

Figure 5. Mean Readability-FKGL scores of ChatGPT-5, Gemini 2.5 Flash and DeepSeek-V3.1 in responding to questions related patient questions on endodontic treatment.

In the subgroup analysis of GFI scores, it was found that the average GFI score of Gemini 2.5 Flash (12.14±2.02) responses to open-ended questions was significantly higher than the average scores of ChatGPT-5 (11.04±2.03) and DeepSeek-V3.1 (10.21±1.15) model's responses (p = 0.036 and p < 0.001, respectively). No significant differences were found between the other pairs of groups (p > 0.05) (Table 9) (Fig. 6).

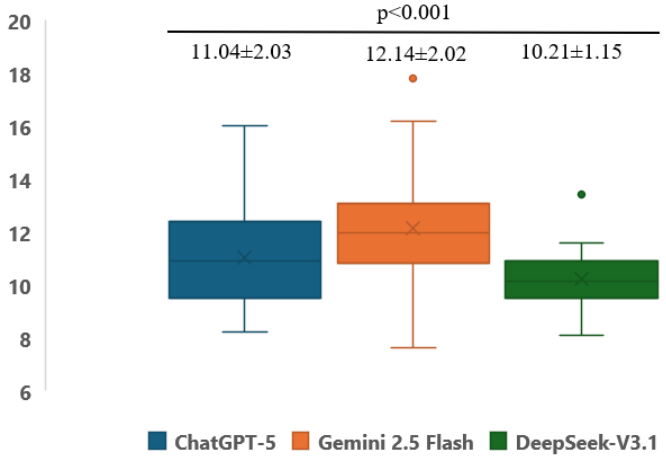

Figure 6. Mean Readability-GFI scores of ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.1 in responding to patient questions on endodontic treatment.

In the subgroup analysis of SMOG scores, the average SMOG score of Gemini 2.5 Flash (12.28±1.50) responses to open-ended questions was found to be significantly higher than the average score of DeepSeek-V3.1 (11.19±0.92) model responses (p = 0.003). No significant differences were found between the other pairs (p > 0.05) (Table 9) (Fig. 7).

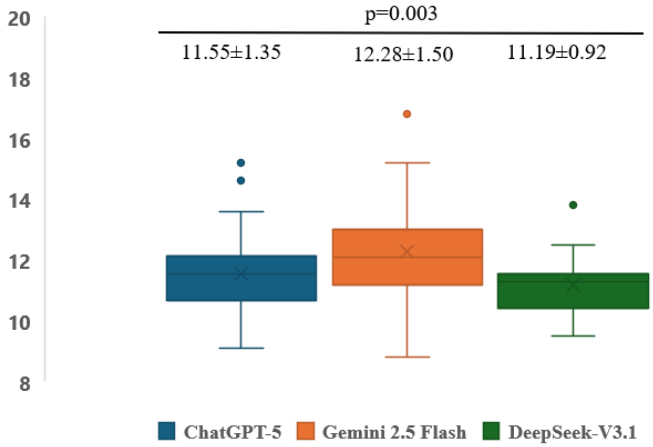

Figure 7. Mean Readability-SMOG scores of ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.1 in responding to patient questions related to endodontic treatment.

In the subgroup analysis of CLI scores, it was found that the average CLI score of ChatGPT-5 (12.01±1.76) responses to open-ended questions was significantly higher than the average score of DeepSeek-V3.1 (10.7±1.15) model responses (p = 0.004). No significant differences were found between the other pairs of groups (p > 0.05) (Table 7) (Fig. 8)

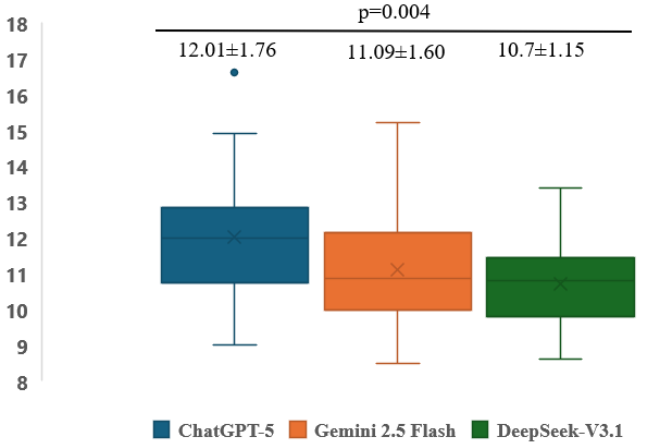

Figure 8. Mean Readability-CLI scores of ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.1 in responding to patient questions on endodontic treatment.

Table 9. Comparison of accuracy, understandability (PEMAT-P), and readability scores for ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.1.

		Chatbot			
		ChatGPT-5	Gemini 2.5 Flash	DeepSeek-V3.1	p-value
Accuracy	Mean ± SD	3.93±0.79	3.67±0.76	4.63±0.81	<0.001[a]
	Median (min.-max.)	4(1-5)	4 (1-5)	5 (1-5)	
Understandability PEMAT-P	Mean ± SD	77.46±13.04	76.04±11.65	77.57±12.42	0.683[a]
	Median (min.-max.)	77.7 (33.3-100)	77.7 (33.3-90.9)	80 (33.3-100)	
Readability					
FRES	Mean ± SD	52.64±10.99	53.65±10.04	58.21±6.74	0.024[a]
	Median (min.-max.)	55 (23.4-70.1)	54.9 (23.6-69.9)	59.8 (41.8-67.3)	
FKGL	Mean ± SD	9.35±1.88	10.15±2.14	8.48±1.17	0.001[a]
	Median (min.-max.)	8.95 (5.7-13.8)	9.9 (6.3-16.7)	8.45 (6.1-11.2)	
GFI	Mean ± SD	11.04±2.03	12.14±2.02	10.21±1.15	<0.001[a]
	Median (min.-max.)	10.9 (8.2-16)	11.95 (7.6-17.8)	10.15 (8.1-13.4)	
SMOG	Mean ± SD	11.55±1.35	12.28±1.50	11.19±0.92	0.003[a]
	Median (min.-max.)	11.55 (9.1-15.2)	12.1 (8.8-16.8)	11.3 (9.5-13.8)	
CLI	Mean ± SD	12.01±1.76	11.09±1.60	10.7±1.15	0.004[b]
	Median (min.-max.)	12 (9-16.6)	10.85 (8.5-15.2)	10.8 (8.6-13.4)	

Data are expressed as mean ± standard deviation and median (minimum: maximum). a: Kruskal-Wallis test, b: ANOVA test

A positive and significant correlation was found between accuracy and PEMAT-P scores (rs = 0.26; p = 0.013). The findings indicate that an increase in accuracy score is also associated with an increase in the PEMAT-P score. A positive and significant correlation was found between the accuracy score and FRES (rs=0.26; p = 0.014). Accordingly, it can be said that as text readability increases (as the FRES value increases), accuracy scores tend to increase. A negative and significant correlation was found between the accuracy and FKGL readability scores (rs = -0.27; p = 0.011). Accordingly, it can be said that as the readability level of the text decreases (as the FKGL value increases), the accuracy scores tend to decrease.

A negative and significant correlation was found between the accuracy score and GFI readability score (rs = -0.25; p = 0.020). Based on this finding, it can be said that as the readability level of the texts decreases (as the GFI value increases), the accuracy scores tend to decrease. A significant negative correlation was found between the accuracy and SMOG readability scores (rs = -0.22; p = 0.036). Accordingly, it can be said that as the readability level of the texts decreases (as the SMOG value increases), the accuracy scores tend to decrease (Table 10).

No significant difference was found between the readability and PEMAT-P scores (p > 0.05) (Table 11).

Table 10. Spearman correlation analysis between accuracy, PEMAT-P, and readability scores for ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.1.

	Accuracy
PEMAT-P	
• r_s	0.26
• p-value	0.013
FRES	
• r_s	0.26
• p-value	0.014
FKGL	
• r_s	-0.27
• p-value	0.011
GFI	
• r_s	-0.25
• p-value	0.020
SMOG	
• r_s	-0.22
• p-value	0.036
CLI	
• r_s	-0.12
• p-value	0.244

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

Table 11. Spearman correlation analysis between PEMAT-P and the readability scores for ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.1.

	PEMAT-P
FRES	
• r_s	0.07
• p-value	0.491
FKGL	
• r_s	-0.06
• p-value	0.557
GFI	
• r_s	-0.03
• p-value	0.768
SMOG	
• r_s	-0.01
• p-value	0.961
CLI	
• r_s	0.05
• p-value	0.654

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

Discussion

The first null hypothesis (H_{01}) was rejected because, when patient questions about endodontic treatment were presented to different LLM models, significant differences were observed in the accuracy of the responses provided by each model.

The second null hypothesis (H_{02}) was also rejected; the analysis revealed statistically significant differences among the responses of different LLM models regarding the understandability of their answers to patient questions about endodontic treatment.

The third null hypothesis (H_{03}) was also rejected, as the results showed significant differences among the responses of the different LLM models in terms of readability when answering patient questions related to endodontic treatment.

Endodontics is considered one of the most challenging branches of dentistry because of the micro-level precision it requires and the wide variety of methods and protocols involved. Endodontic treatment is becoming increasingly popular because of the longer life expectancy, people's desire to preserve their natural teeth, and growing patient awareness of the benefits of preserving teeth rather than extracting them (24).

A review has shown that root canal treatment is psychologically more stressful for patients than routine restorative or cleaning procedures (25). In recent years, there has been a notable trend of patients turning to LLMs to answer their health-related questions. The increased prevalence of endodontic treatments, coupled with the higher level of anxiety they often induce, has led patients to rely more frequently on these AI-based information sources. Patients can thus obtain detailed information about the treatment, attempt to resolve their pre-procedure uncertainties, and seek reassuring explanations to alleviate their anxiety. However, the accuracy and currency of the information provided by LLM-based systems cannot be guaranteed. Misleading or context-disconnected responses can lead to false expectations in patients and may cause undesirable outcomes in clinical decision-making processes.

Artificial intelligence chatbots have the capacity to process various query formats, including multiple-choice, open-ended, and yes/no formats (26). In our study, we deliberately avoided using multiple-choice and yes/no question formats and designed our questions to be completely open-ended. There are two main reasons for choosing this dataset. First, we believe that these closed-ended formats do not adequately represent the way patients approach dentists with questions in real-life settings. The second and more important reason is that we anticipate that yes/no or multiple-choice questions will be insufficient to capture the complex structure inherent in the clinical decision-making process, which involves uncertainties and multidimensional assessments.

In a study by Kaygısız et al., DeepSeek-V3 was reported to score significantly higher on a 5-point Likert scale compared to ChatGPT-4o in oral pathology patient scenarios (27). In this study, which followed a similar methodology, a 5-point Likert scale was used to evaluate the accuracy of the responses. Our evaluations revealed that the average accuracy score of the new version, DeepSeek-V3.1, was statistically significantly higher than the average of ChatGPT-5, another current model, in the responses to open-ended questions frequently asked by patients regarding endodontic treatment. DeepSeek's accuracy superiority, which surpassed ChatGPT-4o in oral pathology in the previous version (V3) (27), has demonstrated the same success against ChatGPT-5 in a different field, endodontics, with its new version (V3.1). We believe that this demonstrates that DeepSeek's ability to understand and process medical content is robust, regardless of the field.

In Huang et al.'s study, DeepSeek-V3 outperformed ChatGPT-4 and Gemini in the initial accuracy scoring of answering multiple-choice questions on Taiwan's national dental technician licensing examination. The initial overall accuracy for DeepSeek-V3 was 69.6%, whereas the accuracy rates after one, two, and three weeks were 72.7%, 72.2%, and 70.1%, respectively. The initial overall accuracy rate for Gemini was 57.2%. The overall accuracy rates after one, two, and three weeks were 67.0%, 66.0%, and 67.0%, respectively. For ChatGPT-4, the initial overall accuracy rate was 52%. The accuracy rate slightly increased to 56.2% after one week and remained at 56.2% after two weeks. However, after three weeks, the accuracy rate dropped back to the initial rate of 52.1% (11). Our study shows that DeepSeek's superior performance continues even under varying conditions, such as three different LLM versions

(DeepSeek-V3.1, ChatGPT-5, and Gemini 2.5), different areas of expertise, and different types of questions. DeepSeek achieved a significantly higher accuracy score, even within this diversity.

Bashah et al., in their study evaluating DeepSeek, Gemini, ChatGPT-4o, and Perplexity in response to salivary gland cancer, focused on diagnostic criteria, treatment protocols, and follow-up strategies, with a total of 30 questions, 26 Yes/No questions (binary responses), and 4 open-ended questions. DeepSeek achieved the highest accuracy rate at 86.9%, followed by Gemini at 78.9%, ChatGPT-4o at 72.8%, and Perplexity at 71.6% (26). In our current study, the average accuracy score of DeepSeek-V3.1's responses to frequently asked open-ended questions about endodontic treatment was found to be significantly higher than both Gemini 2.5 Flash and ChatGPT-5.

In a study comparing the artificial intelligence platforms "Google Gemini 2.5 Flash, Google Gemini 2.0 Flash, DeepSeek V3, and ChatGPT-4o" in solving multiple-choice questions on different subtopics of anatomy, the overall accuracy rate across all platforms was reported to be 95.9% (28). Google Gemini 2.5 Flash achieved the highest performance ranking; it was followed by ChatGPT-4o, DeepSeek V3, and Google Gemini 2.0 Flash, respectively. In the current study, Gemini 2.5 Flash was compared with the more recent versions, DeepSeek-V3.1 and ChatGPT-5. According to our findings, DeepSeek-V3.1 showed the best performance, followed by ChatGPT-5 in second place and Gemini 2.5 Flash in third place. Singal et al. (28) noted that Gemini 2.5 Flash was not compared with DeepSeek-V3.1 in their study. We believe that this difference stems from the fact that DeepSeek-V3.1 is a more recent version and potentially outperforms Gemini 2.5 Flash in terms of its performance.

Readability refers to the degree to which a text can be easily understood by the reader. Readability is influenced by the reader's intelligence, education level, environment, interests, and purpose, as well as the ideas, vocabulary, style, and formal structure of the text. Chatbots used for patient information and guidance should ideally have high FRE and low FKGL scores to benefit the users (29). In our study, DeepSeek-V3.1 achieved the highest FRES score and the lowest median FKGL score. This result demonstrates that the model offers higher readability than other chatbots.

SMOG is one of the most preferred readability criteria for evaluating the understandability of health materials intended for consumers. The SMOG score corresponds to the amount of education a person typically requires to easily understand a text. Scores between 6 and 7 were considered very easy, scores between 8 and 9 were moderately easy, scores between 10 and 12 were moderately difficult, and scores of 13 or higher were considered difficult to read (30). The results show that Gemini 2.5 Flash has a higher SMOG median score than the other models, whereas DeepSeek-V3.1 and ChatGPT-5 have texts that are moderately difficult for readers to understand.

Sezer et al. reported that DeepSeek achieved the highest FRES and lowest SMOG Index scores in their study, indicating relatively higher accessibility. However, even DeepSeek's responses were classified as "difficult," corresponding to a high school level. The study also found that the majority of chatbot responses did not meet the 6^{th}–8^{th} grade readability levels commonly recommended for patient education materials (31). Similarly, in our study, although we compared different question contents and LLM versions, DeepSeek-V3.1 yielded the highest FRES and lowest SMOG median values. However, in our current study, no chatbot was able to generate text at the recommended 6^{th}–8^{th} grade reading level. Nevertheless, DeepSeek-V3.1 was determined to be the model that provided results closest to these threshold values. We believe that this result may be due to DeepSeek-V3.1 being a free and publicly available platform appealing to a wider user base and therefore being trained on a much more diverse range of question contents and narrative styles.

In patient education materials, comprehensibility and applicability scores of at least 70% out of 100 are generally considered sufficient (21). In a study evaluating the performance of Güven et al. 's artificial intelligence chatbots in responding to patient questions about traumatic dental injuries, the median PEMAT-P comprehensibility score for all chatbots was above the 70% threshold level (32). Similarly, in our study, all chatbots had a PEMAT-P comprehensibility median score above 70%. This demonstrates that LLMs can provide a similar level of comprehensibility across different patient questions and subject areas. However, high comprehensibility scores alone are not sufficient to ensure the accuracy of the information provided or its suitability for individual patient characteristics. Therefore, the information provided by chatbots must be evaluated under the supervision of clinicians and corrected when necessary. Furthermore, testing such tools in different languages and cultures is important for accessibility and linguistic adaptation.

The PEMAT does not directly measure readability; therefore, it is recommended that readability assessments be conducted alongside the PEMAT. Furthermore, it is known that the difficulty of reading texts is consistently associated with lower comprehension levels. Therefore, it is important to analyze readability quickly and objectively (33). Comparing PEMAT readability scores with readability assessments obtained from an online platform is an effective method for increasing the text validity. Statistical analysis revealed no statistically significant difference between the PEMAT readability scores and the readability assessments obtained from the online platform.

Conclusion

Our study comprehensively evaluated the accuracy, comprehensibility, and readability of responses to

patient questions regarding endodontic treatment using three different LLMs (ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.1).

The findings revealed that DeepSeek-V3.1 demonstrated superior performance compared with the other models, particularly in terms of accuracy. In terms of comprehensibility, all models achieved similar scores above the threshold considered sufficient for patient education. However, in terms of readability, significant differences were observed across the models in terms of various metrics. Nevertheless, none of the models reached the recommended 6th–8th grade level for patient information materials.

The results obtained show that LLM-based systems can be a valuable resource in patient communication. However, responses must be evaluated under clinician supervision in terms of accuracy and compatibility with individual patient characteristics.

In this context, it was concluded that LLMs can be used as complementary tools in clinical practice and support patient education and communication, but should not be used in a decision-making position on their own.

Disclosures

Peer-review: Externally peer-reviewed.

Author Contributions: Conception – M.T.; Design – M.T., Ö.A.; Supervision – Ö.A., M.G.; Materials – M.T., Ö.A.; Data Collection and/or Processing – M.G., H.O.; Analysis and/or Interpretation – H.O.; Literature Review – M.G., M.G.K.; Writer – M.T.; Critical Review – Ö.A., M.G., M.G.K.

Conflict of Interest: No conflict of interest was declared by the authors.

Funding: The authors declared that this study has received no financial support.

References

1. Bujnowska-Fedak MM, Waligóra J, Mastalerz-Migas A. The internet as a source of health information and services. In: Advancements and innovations in health sciences. Springer; 2019:1-16. https://doi.org/10.1007/5584_2019_396
2. Johnson SB, King AJ, Warner EL, Aneja S, Kann BH, Bylund CL. Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. JNCI Cancer Spectr. 2023;7(2):pkad015. https://doi.org/10.1093/jncics/pkad015
3. Yeung AWK, Tosevska A, Klager E, Eibensteiner F, Tsagkaris C, Parvanov ED, et al. Medical and health-related misinformation on social media: bibliometric study of the scientific literature. J Med Internet Res. 2022;24(1):e28152. https://doi.org/10.2196/28152
4. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595. https://doi.org/10.3389/frai.2023.1169595
5. Dhamani N, Engler M. Introduction to generative AI. Manning; 2024.
6. OpenAI. Introducing GPT-5. OpenAI. https://openai.com/index/introducing-gpt-5/ Access date: September 5, 2025.
7. Rao A, Kim J, Kamineni M, Pang M, Lie W, Dreyer KJ, et al. Evaluating GPT as an adjunct for radiologic decision making: GPT-4 versus GPT-3.5 in a breast imaging pilot. J Am Coll Radiol. 2023;20(10):990-997. https://doi.org/10.1016/j.jacr.2023.05.003
8. Campbell DJ, Estephan LE. ChatGPT for patient education: an evolving investigation. J Clin Sleep Med. 2023;19(12):2135-2136. https://doi.org/10.5664/jcsm.10728
9. Alhur A. Redefining healthcare with artificial intelligence (AI): the contributions of ChatGPT, Gemini, and Co-pilot. Cureus. 2024;16(4):e57795. https://doi.org/10.7759/cureus.57795
10. Google. Gemini 2.5: our most intelligent AI model. Google Blog. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/#gemini-2-5-thinking Access date: September 5, 2025.
11. Huang C-Y, Lee Y-P, Sun A, Chiang C-P. Performance of ChatGPT-4, Gemini, and DeepSeek-V3 on answering the multiple choice questions from Taiwan national dental technician licensing examinations and their self-learning abilities over a three-week period. J Dent Sci. 2025. https://doi.org/10.1016/j.jds.2025.07.011
12. DeepSeek-V3.1 release. https://api-docs.deepseek.com/news/news250821. Access date: September 5, 2025.
13. Abbott PV. Endodontics—current and future. J Conserv Dent Endod. 2012;15(3):202-205. https://doi.org/10.4103/0972-0707.97935
14. Leong DJX, Yap AUJ. Quality of life of patients with endodontically treated teeth: a systematic review. Aust Endod J. 2020;46(1):130-139. https://doi.org/10.1111/aej.12372
15. Find the questions people also ask. AlsoAsked. https://alsoasked.com/ Access date: September 5, 2025.
16. See what people are searching everywhere. AnswerThePublic. https://answerthepublic.com/ Access date: September 5, 2025.
17. Worrall AP, Connolly MJ, O'Neill A, O'Doherty M, Thornton KP, McNally C, et al. Readability of online COVID-19 health information: a comparison between four English speaking countries. BMC Public Health. 2020;20(1):1635. https://doi.org/10.1186/s12889-020-09710-5
18. Cheng C, Dunn M. Health literacy and the Internet: a study on the readability of Australian online health information. Aust N Z J Public Health. 2015;39(4):309-314. https://doi.org/10.1111/1753-6405.12341
19. Uzunçıbuk H, Marrapodi MM, Ronsivalle V, Cicciu M, Minervini G. Lessons to be learned when designing comprehensible patient-oriented online information about temporomandibular disorders. J Oral Rehabil. 2025;52(2):222-229. https://doi.org/10.1111/joor.13798
20. Readability formulas. https://readable.com/readability/flesch-reading-ease-flesch-kincaid-grade-level/ Access date: September 5, 2025.

21. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns. 2014;96(3):395-403. https://doi.org/10.1016/j.pec.2014.05.027
22. Shoemaker SJ, Wolf MS, Brach C. The patient education materials assessment tool (PEMAT) and user's guide. Rockville, MD: Agency for Healthcare Research and Quality; 2013. https://doi.org/10.1037/t37641-000
23. Shoemaker SJ, Wolf MS, Brach C. The patient education materials assessment tool (PEMAT) and user's guide: an instrument to assess the understandability and actionability of print and audiovisual patient education materials (version 1.0). Rockville, MD: Agency for Healthcare Research and Quality; 2013. AHRQ Publication No. 14-0002-EF.
24. Maqbool M, Tirmazi SSM, Shakoor A, Akram Z, Nazir R, Chohan AN, et al. Perception of dental house officers regarding endodontic file separation during endodontic treatment. Biomed Res Int. 2023;2023:1044541. https://doi.org/10.1155/2023/1044541
25. Huynh R, Peters CI, Zafar S, Peters OA. Evaluating the stress of root canal treatment in patients and dentists compared to other dental treatments: a systematic review and meta-analysis. Eur J Oral Sci. 2023;131(4):e12941. https://doi.org/10.1111/eos.12941
26. Bashah A, Salem A, Al-Waqeerah A, Ghaleb E, Wahan N, Awad A, et al. Evaluation of DeepSeek, Gemini, ChatGPT-4o, and perplexity in responding to salivary gland cancer. BMC Oral Health. 2025;25(1):1-15. https://doi.org/10.1186/s12903-025-06726-4
27. Kaygisiz ÖF, Teke MT. Can DeepSeek and ChatGPT be used in the diagnosis of oral pathologies? BMC Oral Health. 2025;25(1):638. https://doi.org/10.1186/s12903-025-06034-x
28. Singal A, Goyal S. Comparative evaluation of AI platforms "Google Gemini 2.5 Flash, Google Gemini 2.0 Flash, DeepSeek V3 and ChatGPT 4o" in solving multiple-choice questions from different subtopics of anatomy. Surg Radiol Anat. 2025;47(1):1-8. https://doi.org/10.1007/s00276-025-03707-8
29. Özcivelek T, Özcan B. Comparative evaluation of responses from DeepSeek-R1, ChatGPT-o1, ChatGPT-4, and dental GPT chatbots to patient inquiries about dental and maxillofacial prostheses. BMC Oral Health. 2025;25(1):871. https://doi.org/10.1186/s12903-025-06267-w
30. Readable. https://app.readable.com/text/ Access date: September 5, 2025.
31. Sezer B, Aydoğdu T. Performance of advanced artificial intelligence models in traumatic dental injuries in primary dentition: a comparative evaluation of ChatGPT-4 Omni, DeepSeek, Gemini Advanced, and Claude 3.7 in terms of accuracy, completeness, response time, and readability. Appl Sci. 2025;15(14):7778. https://doi.org/10.3390/app15147778
32. Guven Y, Ozdemir OT, Kavan MY. Performance of artificial intelligence chatbots in responding to patient queries related to traumatic dental injuries: a comparative study. Dent Traumatol. 2025;41(3):338-347. https://doi.org/10.1111/edt.13020
33. Davis TC, Wolf MS, Bass PF, Middlebrooks M, Kennen E, Baker DW, et al. Low literacy impairs comprehension of prescription drug warning labels. J Gen Intern Med. 2006;21(8):847-851. https://doi.org/10.1111/j.1525-1497.2006.00529.x