Submitted: 23 April, 2026 Accepted: 10 June, 2026 Published: 03 November, 2026 DOI:10.22514/jocpd.2026.151

ORIGINAL RESEARCH

Journal of Clinical Pediatric Dentistry

Assessing the performance of large language models in paediatric dentistry: a study of MIH-related information quality, readability, and reliability

Emine Tatar Şatıroğlu[1,*], Hatice Kübra Başkan[1]

[1]Department of Pediatric Dentistry, Kahramanmaras Sutcu Imam University, 46100 Kahramanmaras, Türkiye

***Correspondence**
eminetatar@ksu.edu.tr
(Emine Tatar Şatıroğlu)

Abstract

Background: Artificial intelligence (AI)-based chatbots are increasingly used to obtain health-related information. However, the quality, reliability, readability, and response speed of these chatbots remain uncertain in paediatric dentistry. This study compared the performance of three large language model–based chatbots in answering frequently asked questions about molar incisor hypomineralization (MIH). **Methods**: Twenty MIH-related questions were developed from Google Frequently Asked Questions and established clinical guidelines from the European Academy of Paediatric Dentistry (EAPD) and the American Academy of Pediatric Dentistry (AAPD). Questions were submitted to three chatbots (DeepSeek V3.1, Gemini 2.5 Pro, and ChatGPT-5). Responses were assessed using the Flesch Reading Ease Score (FRES) and Flesch–Kincaid Grade Level (FKGL), quality using the Global Quality Scale (GQS), and reliability using the modified DISCERN (mDISCERN) tool, and response time. Statistical analyses included Friedman test, Wilcoxon signed-rank test with Bonferroni correction, and Spearman correlation tests ($p < 0.05$). **Results**: Although the overall FRES comparison was significant ($p = 0.025$), no significant pairwise readability differences were observed after Bonferroni correction. All responses were classified as "difficult" (18–20year reading level). DeepSeek V3.1 responded faster (2.21 s) than Gemini 2.5 Pro (5.61 s) and ChatGPT-5 (6.76 s) ($p = 0.006$). For quality, DeepSeek V3.1 scored lower than the other models ($p < 0.001$), while no difference was found between Gemini 2.5 Pro and ChatGPT-5. ChatGPT-5 achieved higher mDISCERN scores than Gemini 2.5 Pro ($p = 0.012$), with no significant differences between ChatGPT-5 and DeepSeek V3.1 or DeepSeek V3.1 and Gemini 2.5 Pro. FRES and FKGL were strongly negatively correlated ($\rho = -0.890$, $p < 0.001$). **Conclusions**: AI chatbots provide similar readability for MIH-related information, their quality, reliability, and response speed differ. All models exceeded recommended readability levels, potentially limiting accessibility. Faster responses were not associated with higher quality or reliability.

Keywords

Artificial intelligence; Chatbots; Molar incisor hypomineralization; Large language models

1. Introduction

Artificial Intelligence (AI), a field of computer science designed to emulate human intelligence, seeks to generate human-like responses using systems trained on large datasets, and was first introduced in the 1950s [1, 2]. Large language models (LLMs) represent a class of AI technologies with the potential to enhance healthcare delivery by generating human-like responses to user queries [3]. Chat Generative Pre-Trained Transformer (ChatGPT), developed by OpenAI, is a large language model designed to generate human-like responses and has undergone continuous development through successive model updates and improvements [4]. GPT-5 was introduced on 07 August 2025 as an advanced large language model with improved speed, accuracy, and instruction-following capabilities compared with earlier versions [5]. The success of ChatGPT, developed by OpenAI, has driven the emergence of various chatbot systems, including Gemini (2023) by Google AI (Mountain View, CA, USA) and DeepSeek (2023) by DeepSeek AI (Hangzhou, Zhejiang, China).

These chatbots have gained substantial user trust and contribute to improving daily life by offering services across diverse domains, including health sciences, security, education, and access control [4]. By responding to medical inquiries, these systems can increase patient knowledge, promote au-

This is an open access article under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/).

tonomy, and benefit healthcare delivery by broadening access to care and enhancing its quality [6]. Previous studies have evaluated AI chatbot performance across various dental topics, including orthodontics, prosthodontics, endodontics, and paediatric dentistry [7–9]. However, there remains a limited amount of evidence regarding chatbot-generated information specifically related to MIH. Given the increasing use of AI chatbots by parents seeking oral health information, evaluating the quality, reliability, and readability of MIH-related responses is warranted.

Molar incisor hypomineralization (MIH), a condition associated with both aesthetic concerns and rapid structural breakdown in children, has increasingly attracted parental awareness. The definition by Weerheijm *et al.* [10] (2001) of MIH as a condition characterized by hypomineralisation of systemic origin affecting one to four permanent first molars, often accompanied by involvement of the incisors, led to greater global awareness. The European Academy of Paediatric Dentistry (EAPD) diagnostic criteria for MIH include the presence of a demarcated opaque lesion affecting at least one permanent first molar and often the incisors, potentially with the presence of post-eruptive enamel breakdown, atypical restorations, and extractions due to MIH [11]. Detection of demarcated opacities in the enamel of the second primary molars indicates an increased risk of MIH [12]. While MIH has historically been diagnosed using these criteria, analysis of the etiological factors of MIH has shown that it is a multifactorial condition. Prenatal and perinatal factors, early childhood illness (especially fever), amoxicillin, and genetic factors have been found to trigger MIH etiology [13–16]. Various management options for MIH are available according to the level of enamel maturation of the tooth and the degree of hypomineralisation. Factors that complicate patient management include sensitivity, restorative marginal breakdown due to hypomineralisation, and ineffectiveness of local anaesthesia [17]. Therefore, each patient should be treated individually.

Today, it is increasingly common for individuals to seek initial information about their conditions from search engines or chatbots before consulting a dentist. Therefore, the aim of this study was to evaluate and compare the quality, readability, reliability, and response times of responses generated by ChatGPT-5, Gemini 2.5 Pro, and DeepSeek V3.1 to frequently asked questions about molar incisor hypomineralization (MIH). The null hypothesis was that there would be no significant differences between the chatbots in respect of response quality, reliability, readability, and response time.

2. Methods

This study was designed as a cross-sectional comparative study. The study dataset consisted of 20 questions developed by two paediatric dentistry specialists (ETŞ and HKB). On 06 November 2025, Google searches were performed using the terms “Molar Incisor Hypomineralisation”, “MIH”, and “cheese molar”, and questions from both the “Frequently Asked Questions” (FAQ) section and the “People Also Ask” boxes were collected. The study questions were not generated solely from online search queries but were developed through a multi-step process combining publicly available Google FAQs, common patient inquiries about MIH in routine pediatric dentistry practice, and the clinical expertise of two paediatric dentistry specialists with over 10 years of experience. The final set of questions was reviewed and refined according to the recommendations and clinical context presented in the European Academy of Pediatric Dentistry (EAPD) and the American Academy of Pediatric Dentistry (AAPD) guidelines on molar incisor hypomineralization. This approach was adopted to ensure the dataset reflected both real-world patient information needs and evidence-based clinical practice. The complete list of the 20 English questions used in this study is provided in **Supplementary material**. The questions were prepared in English and submitted to three different chatbots—DeepSeek V3.1 (DeepSeek AI, Zhejiang, China; accessed January 2026), Gemini 2.5 Pro (Google AI, Mountain View, CA, USA; accessed January 2026), and ChatGPT-5 (OpenAI, San Francisco, CA, USA; accessed January 2026) and the responses were recorded in English. Identical prompts were used for all the chatbot systems. Only the first response generated by each chatbot was evaluated, and no regenerated responses were requested. All the chatbot evaluations were performed using premium subscription versions purchased by the researchers during data collection. Access to ChatGPT-5, Gemini 2.5 Pro, and DeepSeek V3.1 were obtained through authenticated user accounts with full model functionality enabled. All the responses obtained using the chatbots’ standard modes were generated under the same conditions and during the same data collection period to minimize temporal variability associated with model updates.

All the questions were posed on the same day by a single researcher (HKB), and to minimize memory effects from previous responses, a new chat session was initiated for each question. The response times of each chatbot were recorded in seconds using a digital stopwatch. To ensure evaluator blinding, chatbot identifiers were removed and responses were coded by a third paediatric dentist who was not involved in the evaluation process prior to assessment.

All the responses generated by the AI chatbots were evaluated in respect of response time, readability, quality, and reliability. The evaluation of chatbot response quality and reliability was performed by two paediatric dentistry specialists (ETŞ, HKB), affiliated with the Department of Paediatric Dentistry, each with a minimum of 10 years of clinical experience, in line with relevant clinical guidelines [18, 19].

2.1 Evaluation of readability

All the responses were analyzed for readability using an online readability calculator accessed via https://readabilityformulas.com/readability-scoring-system.php. The readability of each response was assessed using the Flesch Reading Ease Score (FRES) and FKGL scales.

The Flesch Reading Ease Score (FRES) is a scale that measures text readability by assigning higher scores to more accessible and comprehensible content. A higher score indicates easier readability. This formula is frequently used in healthcare settings to ensure that patient-facing documents are easily understandable. The formula is:

$$a = 206.835 - 1.015\ (\text{total words/total sentences}) - 84.6$$

(total syllables/total words) [20]. The difficulty levels of the Flesch Reading Ease Score are presented in Table 1.

TABLE 1. Flesch reading ease score difficulty levels.

Flesch Reading Ease Score	Reading Level
0–29	Very Confusing
30–49	Difficult
50–59	Fairly Difficult
60–69	Standard
70–79	Fairly Easy
80–89	Easy
90–100	Very Easy

The Flesch-Kincaid Grade Level (FKGL) determines the U.S. school grade level required to understand a text. A higher score indicates lower readability and greater text complexity. The formula is:

a = 0.39 (total words/total sentences) + 11.8 (total syllables/total words) − 15.59 [21]. The age groups corresponding to grade levels in the FKGL formula are shown in Table 2 [22].

TABLE 2. FKGL grade level and age groups.

Flesch-Kincaid Grade Level Score	Estimated Reading Grade Level	Estimated Age
6	6th grade	11–12
7	7th grade	12–13
8	8th grade	13–14
9	High school freshman	15–16
10	High school sophomore	16–17
11	High school junior	17–18
12	High school senior	18–19
13	College freshman	19–20
14	College sophomore	20–21
15	College junior	21–22
16	College senior	22–23
17	College graduate	24+

2.2 Evaluation of response time

The operational processing time of each chatbot, from the completion of data input to the generation of the final response, was measured digitally in seconds and recorded as the system's response speed. Timing started immediately after prompt submission and ended when the final response was fully generated.

When measuring the response times of the chatbots, all procedures were conducted using a computer equipped with an Apple M4 processor and 16 GB random access memory (RAM) to minimize hardware-related variability. A stable 100 Mbps fibre-optic internet connection was used to reduce fluctuations in data transmission speed.

2.3 Quality and reliability assessment

The quality assessment was conducted using the Global Quality Scale, a 5-point Likert scale which is shown in Table 3 [23]. When the chatbot responses were evaluated using the GQS, scores were recorded for each question on a scale ranging from a minimum of 1 to a maximum of 5 points.

TABLE 3. Global quality scale.

Point	Explanation
1	Poor quality, poor flow of the site, most information missing, not at all useful for patients
2	Generally poor quality and poor flow, some information listed but many important topics missing, of very limited use to patients
3	Moderate quality, suboptimal flow, some important information is adequately discussed but other topics are poorly discussed, somewhat useful for patients
4	Good quality and generally good flow, most of the relevant information is listed, but some topics not covered, useful for patients
5	Excellent quality and excellent flow, very useful for patients

Designed in 1998 to help healthcare consumers and information providers evaluate the quality and reliability of written information about treatment options for a health problem, DISCERN consists of 15 questions [24]. It was later modified in 2012 by Singh *et al.* [25] as a 5-question scale, which is presented in Table 4. In this scoring system, each item is assigned 1 point for "yes" and 0 points for "no", yielding a total score between 0 and 5. The modified DISCERN (mDISCERN) scale was used in this study to evaluate and record the total score of each chatbot response for every question.

TABLE 4. mDISCERN scale.

No.	Reliability questions
1	Are the aims clear and achieved?
2	Are reliable sources of information used?
3	Is the information presented balanced and unbiased?
4	Are additional sources of information listed for patient reference?
5	Are areas of uncertainty mentioned?

Two researchers, each unaware of the other's responses, evaluated the answers to the questions using the GQS and mDISCERN scales. These results were then combined into a single Excel file and saved as researcher 1 and researcher 2. The inter-rater reliability between the two evaluators was evaluated for each scoring system using Cohen's kappa. Substantial agreement was observed for the GQS evaluations (κ = 0.811; $p < 0.001$), while almost perfect agreement was found for mDISCERN evaluations (κ = 0.871, $p < 0.001$).

IBM® SPSS Statistics 31 software (IBM Corp., Armonk, NY, USA) was used for all statistical analyses. Since the same set of 20 questions was evaluated across all three chatbots, comparisons were performed using statistical methods appropriate for related samples. Differences among chatbots in FKGL, FRES, GQS, mDISCERN, and response time scores were assessed using the Friedman test. When a significant overall difference was detected, pairwise comparisons were conducted using the Wilcoxon signed-rank test with Bonferroni correction. Quantitative data were summarized using mean and standard deviation (SD) values. Correlations between study variables were evaluated using Spearman's rank correlation coefficient. A value of $p < 0.05$ was accepted as the level of statistical significance.

3. Results

The readability of chatbot responses was evaluated using FKGL and FRES indices. Friedman analyses revealed no statistically significant differences among the chatbots in terms of FKGL scores ($\chi^2(2) = 1.200$, $p = 0.549$). Although no statistically significant pairwise differences were detected in intergroup comparisons, Gemini 2.5 Pro was found to provide more readable responses compared to the others based on both the FKGL and FRES assessments. Similarly, no statistically significant pairwise differences were identified for FRES scores after Bonferroni correction, despite a significant overall Friedman test result ($\chi^2(2) = 7.351$, $p = 0.025$). In all three chatbots, readability was classified as difficult according to the FRES test. The FKGL results showed that Gemini 2.5 Pro corresponded to a 12th-grade reading level (18–19 years), while DeepSeek V3.1 and ChatGPT-5 corresponded to a 13th-grade reading level (19–20 years).

The mean response time for answering the questions was 2.21 seconds for DeepSeek V3.1, 5.61 seconds for Gemini 2.5 Pro, and 6.76 seconds for ChatGPT-5. Friedman analysis showed that DeepSeek V3.1 provided significantly faster responses than both Gemini 2.5 Pro and ChatGPT-5, whereas no statistically significant difference was observed between Gemini 2.5 Pro and ChatGPT-5.

Friedman analysis revealed a statistically significant difference among the chatbots in terms of GQS scores ($\chi^2(2) = 20.118$, $p < 0.001$). *Post-hoc* Wilcoxon signed-rank tests with Bonferroni correction showed that DeepSeek V3.1 had significantly lower GQS scores than both Gemini 2.5 Pro and ChatGPT-5, whereas no statistically significant difference was observed between Gemini 2.5 Pro and ChatGPT-5.

For mDISCERN, Friedman analysis indicated a statistically significant difference among the chatbots ($\chi^2(2) = 9.864$, $p = 0.007$). *Post-hoc* Wilcoxon signed-rank tests with Bonferroni correction showed that ChatGPT-5 had significantly higher mDISCERN scores than Gemini 2.5 Pro. No statistically significant differences were observed between DeepSeek V3.1 and Gemini 2.5 Pro or between DeepSeek V3.1 and ChatGPT-5.

Table 5 presents the comparison of FKGL, FRES, response time, GQS, and mDISCERN scores among the three chatbots. No statistically significant differences were observed in FKGL scores among the chatbots. Although the overall Friedman test indicated a significant difference for FRES scores, *post-hoc* pairwise comparisons did not reveal any statistically significant differences after Bonferroni correction. Significant differences were observed for response time, GQS, and mDISCERN scores. DeepSeek V3.1 generated significantly faster responses than both Gemini 2.5 Pro and ChatGPT-5, while no significant difference was found between Gemini 2.5 Pro and ChatGPT-5. For GQS, DeepSeek V3.1 scored significantly lower than both Gemini 2.5 Pro and ChatGPT-5, whereas no significant difference was observed between Gemini 2.5 Pro and ChatGPT-5. For mDISCERN, ChatGPT-5 achieved significantly higher scores than Gemini 2.5 Pro ($p = 0.012$), while no significant differences were found between DeepSeek V3.1 and the other chatbots.

In the Spearman correlation analysis presented in Table 6, statistically significant negative correlations were found between the FRES and FKGL scores and between FRES and response time. In the FRES, lower scores indicate increased difficulty, whereas in the FKGL, higher scores indicate increased difficulty. The significant negative correlation between these two readability measures is consistent with their inverse scoring systems. In addition, the negative correlation between FRES and response time suggests that faster responses may sometimes be associated with more complex or longer sentence structures.

4. Discussion

The provision of health-related information and advice by artificial intelligence systems has created a need to evaluate the reliability, quality, and readability of the information they provide. In this study, the quality, reliability, and readability of chatbot responses to frequently asked questions about molar incisor hypomineralization, an important challenge in paediatric dentistry, were assessed.

Readability refers to the numerical values obtained through systematic formulas that measure the reading level required to understand written documents. Various formulas are used to assess the readability of texts [26]. In the FRES formula, which is commonly used for assessing the readability of medical documents, higher scores indicate greater readability, whereas in the FKGL formula, higher scores indicate increased reading difficulty [20, 21]. In this context, a negative correlation is expected between the FRES and FKGL readability measures. The results of the readability evaluations of the chatbot responses in this study showed similar readability levels for all three chatbots, which supports the null hypothesis. According to the FRES readability scoring, all the chatbots were classified at a "Difficult" level, and the FKGL results indicated a reading level corresponding to individuals aged 18–20 years. In the literature, it is generally recommended that patient education materials be written at approximately 6th- to 8th-grade reading level [27]. An increase in the difficulty level of text readability limits the accessibility of chatbot-generated information for users with low health literacy. Consistent with this study, there are also reports in the dental literature demonstrating that chatbot responses are generally classified at a "Difficult" readability level [28–30]. The current study findings suggest that the medical information provided by current chatbots may

TABLE 5. Comparison of readability, quality, reliability, and response time among chatbots.

Variable	Chatbot	Mean ± SD	Mean Rank	Friedman Test
FKGL				
	DeepSeek V3.1	13.47 ± 2.35	2.10	
	Gemini 2.5 Pro	12.52 ± 2.41	1.80	$\chi^2(2) = 1.200$, $p = 0.549$
	ChatGPT-5	13.27 ± 2.50	2.10	
FRES				
	DeepSeek V3.1	31.05 ± 12.16	1.85	
	Gemini 2.5 Pro	34.80 ± 14.72	2.48	$\chi^2(2) = 7.351$, $p = 0.025$
	ChatGPT-5	29.45 ± 13.99	1.68	
Response Time (s)				
	DeepSeek V3.1^{a}	2.21 ± 0.68	1.45	
	Gemini 2.5 Pro^{b}	5.61 ± 3.00	2.10	$\chi^2(2) = 10.300$, $p = 0.006$
	ChatGPT-5^{b}	6.76 ± 3.50	2.45	
GQS				
	DeepSeek V3.1^{a}	3.80 ± 0.95	1.40	
	Gemini 2.5 Pro^{b}	4.80 ± 0.41	2.53	$\chi^2(2) = 20.118$, $p < 0.001$
	ChatGPT-5^{b}	4.45 ± 0.51	2.08	
mDISCERN				
	DeepSeek V3.1^{a,b}	2.80 ± 1.01	2.13	
	Gemini 2.5 Pro^{a}	2.45 ± 0.60	1.53	$\chi^2(2) = 9.864$, $p = 0.007$
	ChatGPT-5^{b}	3.00 ± 0.46	2.35	

FKGL: Flesch–Kincaid Grade Level; FRES: Flesch Reading Ease Score; GQS: Global Quality Scale; mDISCERN: modified DISCERN; SD: standard deviation. ^{a,b}*: Values with different superscript letters indicate statistically significant differences according to the Bonferroni-corrected Wilcoxon signed-rank post hoc test ($p < 0.0167$). Values sharing the same superscript letter are not significantly different.*

TABLE 6. Correlation matrix of readability measures, response time, GQS, and mDISCERN scores.

	FRES	FKGL	GQS	mDISCERN	Response time
FRES					
Spearman Correlation	1.000	−0.890***	0.053	0.170	−0.330
Sig. (2-tailed)		$p < 0.001$	0.686	0.195	0.010
FKGL					
Spearman Correlation	−0.890***	1.000	0.053	−0.200	0.245
Sig. (2-tailed)	$p < 0.001$		0.686	0.125	0.060
GQS					
Spearman Correlation	0.053	−0.094	1.000	0.122	0.182
Sig. (2-tailed)	0.686	0.476		0.351	0.163
mDISCERN					
Spearman Correlation	0.170	−0.200	0.122	1.000	−0.230
Sig. (2-tailed)	0.195	0.125	0.351		0.078
Response time					
Spearman Correlation	−0.330	0.245	0.182	−0.230	1.000
Sig. (2-tailed)	0.010	0.060	0.163	0.078	

****: Correlation at 0.001 (2-tailed).*
FKGL: Flesch–Kincaid Grade Level; FRES: Flesch Reading Ease Score; GQS: Global Quality Scale; mDISCERN: modified DISCERN.

remain complex for parents with low health literacy, indicating that such information should be simplified by clinicians in clinical practice.

Although no statistically significant difference was found in readability among the chatbots in the present study, previous studies in the literature have reported that DeepSeek provides more readable outputs compared to ChatGPT and Gemini [31, 32]. In a 2026 study, Prasad *et al.* [31] stated that DeepSeek produced significantly more readable output compared to ChatGPT and Gemini. In a 2026 study, Acar *et al.* [32] assessed readability using the FRES scale and demonstrated that DeepSeek and ChatGPT provided more readable responses than Gemini. Although statistically significant differences were observed among chatbot responses, both studies reported that all three chatbots produced output at a "fairly difficult" readability level. Taken together with the findings of the present study, these results suggest that adequate health literacy is necessary for users to comprehend chatbot-generated medical information.

One of the main reasons AI chatbots are currently widely preferred by users is that they provide rapid responses to questions, thereby reducing the time spent searching for information on websites. In the current technological era, where the perception of time feels increasingly compressed, quick access to information is highly valuable. In this study, DeepSeek V3.1 provided significantly faster responses compared to Gemini 2.5 Pro and ChatGPT-5. The correlation analysis showed no positive relationship between response speed and quality (GQS) or reliability (mDISCERN). Moreover, the negative correlation between FRES and response time ($\rho = -0.330$; $p = 0.01$) suggests that faster responses may sometimes be associated with more complex or longer sentence structures. Response time results should be interpreted with caution, as chatbot latency may be influenced by factors such as server load and network conditions despite efforts to standardize the testing environment.

Studies in the literature evaluating chatbot response times have also shown that faster responses may conflict with quality, indicating that the fastest system does not always provide the best answer. Sezer and Aydogdu evaluated chatbot performance for questions on "Pulp Therapy for Immature Permanent Teeth", and found that Gemini Advanced produced the fastest responses, but these were not the most accurate answers [28]. Similarly, Li *et al.* [7] reported that although ChatGPT-4 required more time to respond compared to ChatGPT-3.5, it provided more accurate answers.

The Global Quality Score (GQS) is a 5-point scale used to assess information quality and usefulness for patients. It ranges from 1 (poor quality, poorly organized, and not useful) to 5 (excellent quality, well-structured, and highly useful), with intermediate scores indicating progressively better quality, completeness, and usability [23]. When the response quality of the chatbots in the current study was evaluated, DeepSeek V3.1 was found to have significantly lower scores compared to Gemini 2.5 Pro and ChatGPT-5, while no significant difference was observed between Gemini 2.5 Pro and ChatGPT-5. This finding leads to the rejection of the null hypothesis. In addition, several studies in the literature have reported that chatbot responses are generally of similar quality levels [8, 33]. In one of those studies, Acar *et al.* [32] compared chatbot responses across prosthodontics and paediatric dentistry domains. No significant differences were determined between the chatbots for paediatric dentistry questions, but DeepSeek was found to generate higher-quality responses for paediatric dentistry questions compared to prosthodontic questions [32]. Conversely, in a study of AI responses to patient questions about bone graft materials in dentistry, Yıldız *et al.* [34] reported that ChatGPT provided higher-quality responses compared to DeepSeek. The differences observed in chatbot systems may be related to variations in model architecture, training datasets, optimization strategies, and response-generation mechanisms. Some models may prioritize concise and rapidly generated responses, whereas others may produce more comprehensive explanations that improve perceived quality and reliability. Differences in the way models handle uncertainty, source attribution, and medical reasoning may also contribute to variability in GQS and mDISCERN scores.

Although DISCERN was originally designed to assess the reliability of health information in written texts, it is also currently used to evaluate health-related information generated by AI tools and shared on social media platforms [29]. In the present study, when the reliability of the chatbot responses was evaluated using mDISCERN, statistically significant differences were observed among all the chatbots. *Post-hoc* analyses showed that ChatGPT-5 scored significantly higher than Gemini 2.5 Pro. However, no statistically significant differences were observed between ChatGPT-5 and DeepSeek V3.1 or between DeepSeek V3.1 and Gemini 2.5 Pro. This finding is consistent with the study by Bostan *et al.* [35] on AI-generated responses to acne-related questions, in which ChatGPT-4.5 was reported to provide significantly more reliable answers than DeepSeek, thereby supporting the results of the current study. In a study by Gumus *et al.* [36] evaluating chatbot knowledge regarding nocturnal enuresis, no significant differences between DeepSeek, Gemini, and ChatGPT were reported, and all three chatbots were found to demonstrate insufficient performance based on mDISCERN assessments. In addition, Aydınbelge-Dizdar and Dizdar evaluated chatbot responses to positron emission tomography/computed tomography (PET/CT) scans and reported findings contrary to those of the current study, showing that Gemini 2.5 Pro provided significantly higher-quality responses than ChatGPT. The authors attributed this difference to the fact that Gemini 2.5 Pro included references in its responses [37]. In the present study, the finding that Gemini 2.5 Pro achieved a lower mean mDISCERN score than DeepSeek despite a higher GQS score suggests deficiencies in reliability-related criteria such as source citation and the reporting of uncertainty. This inconsistency is thought to arise from differences in the scope and content of the medical topics examined across studies. It may also be considered an indication of the lack of standardization in the training of large language models (LLMs). Despite their potential value as educational tools, chatbot-generated responses should not be considered a substitute for professional dental consultation. Inaccurate, incomplete, or misleading information may contribute to misunderstandings regarding disease severity, delayed professional care, or inappropriate treatment expectations. Therefore, AI-generated information

should be interpreted cautiously and verified by qualified dental professionals, particularly when used for paediatric dental conditions such as MIH.

Although several studies have compared AI chatbots across different dental specialties, evidence regarding MIH-specific information remains limited. MIH is a common developmental dental condition that frequently prompts parental information-seeking behaviour due to its complex diagnosis, aesthetic implications, and treatment requirements. By focusing specifically on MIH-related frequently asked questions derived from real-world search behaviour, the present study provides topic-specific evidence regarding the quality, reliability, readability, and response characteristics of contemporary AI chatbots in a clinically relevant area of paediatric dentistry.

The findings of this study should be interpreted in the light of several limitations. First, the questions used may not fully represent the diversity of real patient inquiries or the full range of clinical scenarios, which may limit the generalizability of the results. A further limitation is that no formal sample size calculation was performed, as the study was exploratory in nature and based on a predefined set of MIH-related frequently asked questions. In addition, all questions were posed in English, and the responses were analyzed in the same language; therefore, the findings may not fully reflect chatbot performance in other languages. Artificial intelligence models (ChatGPT-5, Gemini 2.5 Pro, DeepSeek V3.1) are continuously updated and evolving systems. The data collection reflects the performance of the models at a specific point in time, and future updates to their algorithms may lead to different readability and quality outcomes. Although standardization was attempted by asking all the questions in the same format and initiating a new chat session for each question, random variability inherent to chatbot systems could not be completely eliminated. The GQS and mDISCERN scales used to assess quality and reliability involve a certain degree of subjective interpretation. To minimize bias, evaluations were performed independently by two researchers, and high inter-rater agreement was sought to reduce potential subjectivity. In addition, the present study did not specifically assess the presence of unsafe or potentially misleading clinical recommendations in chatbot responses. Since the study focused on frequently asked MIH-related questions and evaluated readability, quality, reliability, and response time was beyond its scope. Future studies should further investigate the safety and clinical accuracy of AI-generated paediatric dental information.

5. Conclusions

The results of this study demonstrated that AI-based chatbots provide generally similar readability levels in response to frequently asked questions about MIH, although there are notable differences in response quality, reliability, and speed. Importantly, all the evaluated systems generated content with readability levels above those recommended for patient education, which may pose a significant barrier to accessibility for individuals with low health literacy. Notably, faster response generation was not associated with higher quality or greater reliability, highlighting a critical balance between speed and information integrity. The performance variability observed across the models indicates that large language models are not yet standardized for patient-oriented medical information delivery. These findings suggest that despite the growing role of chatbots in disseminating health information, they should not be considered standalone reliable sources, particularly for specific paediatric dental conditions such as MIH. Therefore, the information provided should be interpreted with caution and supported by professional clinical guidance whenever possible. Future research should focus on developing standardized evaluation frameworks, integrating evidence-based clinical guidelines, and optimizing AI-generated medical content in terms of both accuracy and readability.

AVAILABILITY OF DATA AND MATERIALS

The datasets generated and/or analyzed during this study are available from the corresponding author upon reasonable request.

AUTHOR CONTRIBUTIONS

ETŞ and HKB—conception; review and editing the manuscript. HKB—data collection. ETŞ—statistical analysis and manuscript writing. Both authors contributed to editorial changes in the manuscript. Both authors read and approved the final manuscript.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

This study did not require Ethics Committee approval as it did not involve any experiments on human or animal subjects and was conducted using publicly available digital data (chatbot outputs).

ACKNOWLEDGMENT

The authors would like to thank Kahramanmaras Sutcu Imam University for technical contributions to this study.

FUNDING

This research received no external funding.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

SUPPLEMENTARY MATERIAL

Supplementary material associated with this article can be found, in the online version, at https://oss.jocpd.com/files/article/2117523627672780800/attachment/Supplementary%20material.docx.

REFERENCES

[1] Schwendicke F, Samek W, Krois J. Artificial intelligence in dentistry: chances and challenges. Journal of Dental Research. 2020; 99: 769–774.

[2] Dufey-Portilla N, Frisman AB, Robles MG, Peña-Bengoa F, Ávila CC, Nagendrababu V, *et al.* Assessing the validity of ChatGPT-4o and Google Gemini Advanced when responding to frequently asked questions in endodontics. Journal of Applied Oral Science. 2025; 33: e20250321.

[3] Ahmed WM, Azhari AA, Alfaraj A, Alhamadani A, Zhang M, Lu CT. The quality of AI-generated dental caries multiple choice questions: a comparative analysis of ChatGPT and Google Bard language models. Heliyon. 2024; 10: e28198.

[4] Guleria A, Krishan K, Sharma V, Kanchan T. ChatGPT: ethical concerns and challenges in academics and research. The Journal of Infection in Developing Countries. 2023; 17: 1292–1299.

[5] Taşyürek M, Adıgüzel Ö, Ortaç H. Comparative evaluation of responses from ChatGPT-5, Gemini 2.5 Flash, Grok 4, and Claude Sonnet-4 chatbots to questions about endodontic iatrogenic events. Healthcare. 2025; 13: 2615.

[6] Alhaidry HM, Fatani B, Alrayes JO, Almana AM, Alfhaed NK. ChatGPT in dentistry: a comprehensive review. Cureus. 2023; 15: e38317.

[7] Li C, Zhang J, Abdul-Masih J, Zhang S, Yang J. Performance of ChatGPT and dental students on concepts of periodontal surgery. European Journal of Dental Education. 2025; 29: 36–43.

[8] Balel Y. Comparative study of technical and patient-related question answering quality of DeepSeek-R1 and ChatGPT-4o in the field of oral and maxillofacial surgery. Oral and Maxillofacial Surgery. 2025; 29: 163.

[9] Park H, Lee S, Lee KC. ChatGPT performance in orthodontics: assessment of accuracy and repeatability in patient instruction and management using Global Quality Score. Journal of Orofacial Orthopedics. 2026. PMID: 41615450.

[10] Weerheijm KL, Jälevik B, Alaluusua S. Molar-incisor hypomineralisation. Caries Research. 2001; 35: 390–391.

[11] Weerheijm KL, Duggal M, Mejàre I, Papagiannoulis L, Koch G, Martens LC, *et al.* Judgement criteria for molar incisor hypomineralisation (MIH) in epidemiologic studies: a summary of the European meeting on MIH held in Athens, 2003. European Journal of Paediatric Dentistry. 2003; 4: 110–113.

[12] Garot E, Denis A, Delbos Y, Manton D, Silva M, Rouas P. Are hypomineralised lesions on second primary molars (HSPM) a predictive sign of molar incisor hypomineralisation (MIH)? A systematic review and a meta-analysis. Journal of Dentistry. 2018; 72: 8–13.

[13] Crombie F, Manton D, Kilpatrick N. Aetiology of molar-incisor hypomineralization: a critical review. International Journal of Paediatric Dentistry. 2009; 19: 73–83.

[14] Alaluusua S. Aetiology of molar-incisor hypomineralisation: a systematic review. European Archives of Paediatric Dentistry. 2010; 11: 53–58.

[15] Jeremias F, de Souza JF, Silva CM, Cordeiro Rde C, Zuanon AC, Santos-Pinto L. Dental caries experience and molar-incisor hypomineralization. Acta Odontologica Scandinavica. 2013; 71: 870–876.

[16] Silva MJ, Scurrah KJ, Craig JM, Manton DJ, Kilpatrick N. Etiology of molar incisor hypomineralization—a systematic review. Community Dentistry and Oral Epidemiology. 2016; 44: 342–353.

[17] William V, Messer LB, Burrow MF. Molar incisor hypomineralization: review and recommendations for clinical management. Pediatric Dentistry. 2006; 28: 224–232.

[18] Lygidakis NA, Garot E, Somani C, Taylor GD, Rouas P, Wong FSL. Best clinical practice guidance for clinicians dealing with children presenting with molar-incisor-hypomineralisation (MIH): an updated European Academy of Paediatric Dentistry policy document. European Archives of Paediatric Dentistry. 2022; 23: 3–21.

[19] American Academy of Pediatric Dentistry. Molar-incisor hypomineralization. 2025. Available at: https://www.aapd.org/research/oral-health-policies--recommendations/molar-incisor-hypomineralization/ (Accessed: 07 January 2026).

[20] Flesch R. A new readability yardstick. Journal of Applied Psychology. 1948; 32: 221–233.

[21] Kincaid JP, Fishburne III RP, Rogers RL, Chissom BS. Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for navy enlisted personnel. 1975. Available at: https://apps.dtic.mil/sti/citations/tr/ADA006655 (Accessed: 06 February 2026).

[22] Restrepo E, Ko N, Warner ET. An evaluation of readability and understandability of online education materials for breast cancer survivors. Journal of Cancer Survivorship. 2024; 18: 457–465.

[23] Gudapati JD, Franco AJ, Tamang S, Mikhael A, Hadi MA, Roy V, *et al.* A study of global quality scale and reliability scores for chest pain: an Instagram-post analysis. Cureus. 2023; 15: e45629.

[24] Charnock D. The DISCERN handbook: quality criteria for consumer health information on treatment choices. Radcliffe Medical Press: Oxford. 1998.

[25] Singh AG, Singh S, Singh PP. YouTube for information on rheumatoid arthritis—a wakeup call? The Journal of Rheumatology. 2012; 39: 899–903.

[26] Büker M, Mercan G. Readability, accuracy and appropriateness and quality of AI chatbot responses as a patient information source on root canal retreatment: a comparative assessment. International Journal of Medical Informatics. 2025; 201: 105948.

[27] Özduran E, Hanci V. Evaluating the readability, quality and reliability of online information on Behçet's disease. Reumatismo. 2022; 74: 49–60.

[28] Sezer B, Aydoğdu T. Performance of advanced artificial intelligence models in pulp therapy for immature permanent teeth: a comparison of ChatGPT-4 Omni, DeepSeek, and Gemini advanced in accuracy, completeness, response time, and readability. Journal of Endodontics. 2025; 51: 1675–1684.

[29] Halboub E, Hakami RAM, Khalufi KNA, Hakami SAH, Alhajj MN. Assessment of quality, understandability, actionability, and readability of responses of selected chatbots to the top searched queries about oral cancer. Digital Dentistry Journal. 2025; 1: 100008.

[30] Karamüftüoğlu N, Varol EA, Bal C. Exploring artificial intelligence chatbots in pediatric fluoride education: a cross-sectional study. Scientific Reports. 2026; 16: 182.

[31] Prasad S, Koseoglu M, Antonopoulou S, Sears LM, Launardo V, Ariani N, *et al.* Readability and performance of AI chatbot responses to frequently asked questions in maxillofacial prosthodontics. The Journal of Prosthetic Dentistry. 2026; 135: 195.e1–195.e9.

[32] Acar NK, Sengul F, Bardakci E, Celikel P. Assessing the accuracy, reliability, quality, and readability of artificial intelligence chatbots in patient education: insights from zirconia crowns. BMC Oral Health. 2026; 26: 502.

[33] Ceylan Şen S, Çardakci Bahar Ş, Saraç Atagün Ö, Ustaoğlu G, Yildiz ZH, Toker H. Quality and reliability of AI information on dental implant failure: a comparative multi-model analysis. Journal of Craniofacial Surgery. 2026; 37: 675–680.

[34] Yildiz ZH. Assessing the usefulness and quality of artificial intelligence-generated content on bone graft materials in dentistry based on patient questions. Journal of Dental Health and Oral Research. 2025; 6: 1–10.

[35] Bostan E, Uçar MT, Dönmez E. Evaluation of the accuracy, reliability, quality, and readability of artificial intelligence chatbots-generated responses to acne-related questions. Turkish Journal of Dermatology. 2025; 19: 235–243.

[36] Gumus K, Yilmaz AB, Cinbek AF, Ozdemir HB, Kutman KG. Evaluating AI chatbots in enuresis nocturna information: a comparative analysis of readability, reliability and quality. The French Journal of Urology. 2026; 36: 103062.

[37] Aydinbelge-Dizdar N, Dizdar K. Reliability and readability evaluation of chatbots responses as a patient information resource for the most common PET/CT scans. Revista Española de Medicina Nuclear e Imagen Molecular. 2025; 44: 500065.

How to cite this article: Emine Tatar Şatıroğlu, Hatice Kübra Başkan. Assessing the performance of large language models in paediatric dentistry: a study of MIH-related information quality, readability, and reliability. Journal of Clinical Pediatric Dentistry. 2026; 50(6): 117-124. doi: 10.22514/jocpd.2026.151.