J Pediatr Res
DOI: 10.4274/jpr.galenos.2026.78871

Original Article

Can a Large Language Model Effectively Answer Parents' Questions About Children's Oral Health?

İrem Nur Bitisik[1], Selcuk Savas[1], Ilgin Timarci[2], Mehmet Izgi[3], Ebru Kucukyilmaz Izgi[1]

[1]Izmir Katip Celebi University Faculty of Dentistry, Department of Pediatric Dentistry, Izmir, Türkiye
[2]Izmir Katip Celebi University Faculty of Medicine, Department of Public Health, Izmir, Türkiye
[3]Private Practice, Izmir, Türkiye

ABSTRACT

Aim: This study aimed to evaluate the readability, quality, and temporal consistency of Turkish responses generated by ChatGPT to frequently asked questions in pediatric dentistry.

Materials and Methods: Twenty open-ended questions were adapted from the frequently asked questions published on the American Academy of Pediatric Dentistry parent information page. Each question was submitted to ChatGPT (GPT-4, OpenAI, USA) in Turkish over seven consecutive days, three times per day, using a new chat session for each interaction, standardized prompts and default system settings. A total of 420 responses were analyzed. Readability was assessed using both the Ateşman and Çetinkaya-Uzun readability formulas, while response quality was evaluated using the Global Quality Score (GQS). Inter-day and intra-day comparisons were performed using repeated-measures ANOVA or Friedman tests according to the data distribution. Intraclass Correlation Coefficient (ICC) analysis was used to evaluate temporal consistency (α=0.05).

Results: A significant intra-day variability in the readability scores was observed during the first four days for the Ateşman and Çetinkaya-Uzun formulas ($p<0.05$), whereas no significant differences were identified during Days 5-7. Between-day comparisons revealed statistically significant variations in readability scores for both formulations ($p<0.001$). Ateşman's scores were found to correspond to "moderate difficulty" and "easy" readability levels, while Çetinkaya-Uzun's scores indicated an "instructional reading level". By contrast, no significant intra-day or between-day differences were observed in the GQS scores ($p=0.650$), and overall quality scores remained consistently high throughout the study period. The ICC analyses demonstrated good-to-excellent consistency for readability and GQS measurements, with overall ICC values ranging between 0.906 and 0.961.

Conclusion: The Turkish responses generated by ChatGPT to frequently asked questions in pediatric dentistry demonstrated moderate readability, high information quality, and good-to-excellent temporal consistency. However, due to the variability in readability scores across repeated assessments and over time, Large Language Models (LLMS) may serve as supportive tools for patient and parent education, however, human oversight remains necessary for the interpretation and clinical use of LLM-generated health information.

Keywords: Large language model, pediatric dentistry, readability, patient education as topic

Introduction

Large Language Models (LLM) represent a significant innovation in technology, with the objective of emulating cognitive processes which are characteristic of human intelligence through the utilization of computer systems. This development has precipitated substantial

Corresponding Author
Prof. Ebru Kucukyilmaz Izgi, Izmir Katip Celebi University Faculty of Dentistry, Department of Pediatric Dentistry, Izmir, Türkiye
E-mail: ebrukucukyilmaz@hotmail.com **ORCID:** orcid.org/0000-0002-6086-7410
Received: 08.06.2026 **Accepted:** 28.07.2026 **Epub:** 03.09.2026

Cite this article as: Bitisik IN, Savas S, Timarci I, Izgi M, Kucukyilmaz Izgi E. Can a large language model effectively answer parents' questions about children's oral health?. J Pediatr Res. [Epub Ahead of Print]

transformations across numerous domains, with particular relevance to the field of healthcare in recent years (1). The increasing data capacity, more advanced algorithms, and developments in LLMs have resulted in the wider use of AI applications in the early diagnosis of diseases, clinical decision support systems, treatment planning, and patient information processes (2). These developments provide rapid and straightforward access to information and influence the decision-making processes of patients and their relatives regarding health-related issues (2).

Pediatric dentistry is a specialty in which preventive approaches and parental education play a decisive role in treatment success (3). In this context, parents' access to accurate, understandable, and accessible information is of great importance for clinical outcomes. Today, internet-based information-seeking behaviors have changed with the widespread use of LLMs, and users have begun to prefer platforms which provide direct answers to their questions rather than relying solely on traditional search engines (4,5). However, the readability and consistency of the information provided through these platforms have become critical evaluation criteria, particularly in the field of healthcare.

In recent years, studies investigating the capacity of ChatGPT to generate health-related information have increased in the literature (6-10). Nevertheless, when the current literature is examined, it is evident that most studies have focused largely on English-language content, while research addressing Turkish-language health-related content, particularly in the field of pediatric dentistry, remains limited (11,12). Although recent Turkish studies have evaluated LLMs' performance for general parental inquiries as well as specific pediatric dental contexts such as fluoride education, sedation-related questions, pulpotomy, and molar incisor hypomineralization, these studies have primarily focused on response accuracy or single-timepoint performance (13-17). Consequently, important aspects of LLM-generated patient education, including temporal consistency across repeated interactions, changes in readability over repeated queries, and the combined evaluation of readability, information quality, and consistency, remain insufficiently investigated in Turkish-language pediatric dentistry. Considering the effect of language structure on readability, the differences in Turkish in terms of word and syllable structure necessitate a separate evaluation of the comprehensibility of texts produced in this language. Therefore, evaluating not only readability but also the informational quality and temporal consistency of LLM-generated Turkish health information has become an important research priority.

A comprehensive evaluation of LLM-generated health information requires the use of validated and language-appropriate assessment tools. For Turkish texts, the Ateşman and Çetinkaya-Uzun readability formulas are among the most widely used readability indices (18,19). While the Ateşman formula classifies texts according to their overall reading difficulty, the Çetinkaya-Uzun formula estimates the educational level required for comprehension, thereby providing complementary perspectives on textual readability. In addition to readability, the Global Quality Score (GQS) is a validated 5-point Likert-type scale widely used to evaluate the overall quality, educational value, and usefulness of patient-oriented health information, including LLM-generated content (20). Together, these complementary instruments enable a multidimensional assessment of LLM-generated health information by evaluating not only its linguistic accessibility, but also its educational quality and consistency. In order to address the aforementioned knowledge gaps, the present study adopted a repeated-measures design in order to comprehensively evaluate the readability, informational quality, and temporal consistency of ChatGPT-generated responses using validated Turkish readability formulas, the GQS, and Intraclass Correlation Coefficient (ICC) analyses. Accordingly, the present study aimed to evaluate the readability, overall quality, and consistency of ChatGPT's responses to frequently asked questions by patients and parents in the field of pediatric dentistry. The null hypotheses tested in this study were determined as follows: (i) The Ateşman readability scores of the responses generated by ChatGPT do not show statistically significant differences between different days and repetitions. (ii) The Çetinkaya-Uzun readability scores of the responses generated by ChatGPT do not show statistically significant differences between different days and repetitions. (iii) The GQS scores of the responses generated by ChatGPT do not show statistically significant differences between different days and repetitions. (iv) The responses generated by ChatGPT do not demonstrate a high level of consistency across different time periods.

Materials and Methods

Study Design

The present study was designed as an observational and descriptive investigation with the objective of evaluating the overall quality, consistency, and readability of responses generated by a LLM (ChatGPT Plus, GPT-4o, OpenAI, USA) to frequently asked questions by patients in the field of pediatric dentistry. The LLM was accessed on 10th of June,

2025 using the default system settings, with no additional prompting or parameter modification.

Question Preparation

In this study, 20 open-ended questions were prepared regarding topics frequently asked by patients and parents in the field of pediatric dentistry. Rather than constituting a novel questionnaire or psychometric instrument, the question set was designed to reflect common parental concerns regarding children's oral health. The questions employed in this study were adapted to the scope of the research based on the frequently asked questions published on the parent information page of the American Academy of Pediatric Dentistry (21). The final set of 20 questions was subsequently reviewed by two experienced pediatric dentists in order to ensure clinical relevance, clarity, and appropriateness for Turkish-speaking parents before data collection. Any differences in wording or content were resolved through discussion until consensus was reached before the data collection (Supplementary File 1). As no accepted methodology exists for determining the number of questions in LLM evaluation studies, the present sample size of 20 questions was selected in accordance with comparable pediatric dentistry studies using parent-oriented, question-based ChatGPT evaluation designs (13,14,17).

Data Collection

The finalized 20 questions were posed to ChatGPT in Turkish on three occasions per day for a duration of seven days. The timing of these data collections was as follows: in the morning between 08:00-10:00, in the afternoon between 12:00-14:00, and in the evening between 18:00-20:00. All questions were entered into the system manually in a standardized format. Responses were generated using the ChatGPT web interface under default system settings, with no custom instructions, system prompts, or parameter modifications. Each interaction was conducted in a new chat session. Before every measurement, a new conversation within the same user account was initiated in order to prevent carry-over effects from previous interactions and to ensure the independence of each generated response. The responses were evaluated by an assistant professor (10 years of clinical experience) and a professor (15 years of clinical experience), both specialized in pediatric dentistry. Any discrepancies in the assessments were resolved through discussion until a consensus was reached.

Ateşman Readability Formula

The Ateşman Readability Formula, developed by Ateşman in 1997 as an adaptation of the Flesch Reading Ease Formula into Turkish, was used to evaluate the readability level of the texts (18). According to the Ateşman readability formula, readability levels were classified as very difficult, difficult, moderately difficult, easy or very easy (18).

Çetinkaya-Uzun Readability Formula

This formula, developed to evaluate the readability of Turkish texts, is based on average word length and sentence length (19). The readability formula was calculated and readability levels were classified as frustration reading level (10^{th}-12^{th} grade), instructional reading level (8^{th}-9^{th} grade) or independent reading level (5^{th}-7^{th} grade) (19).

GQS Scale

The overall quality level of the responses was evaluated using the GQS, a 5-point Likert-type scale (20). The GQS scale levels were classified as excellent quality, low quality, moderate quality or high/excellent quality (20).

Statistical Analysis

The statistical analyses were conducted utilizing IBM SPSS Statistics Version 23.0 (IBM Corp., Armonk, NY, USA). Continuous variables are presented as mean±standard deviation or median (Q1-Q3), while the categorical variables are presented as n (%). In order to assess differences between repeated measurements, Repeated Measures ANOVA was applied when the normal distribution assumptions were met. In instances where the assumption of normality was not met, the Friedman test was employed, and if statistical significance was detected, pairwise comparisons with Bonferroni correction were conducted. For GQS scores, the assumptions for repeated-measures ANOVA were not satisfied; therefore, all within-day comparisons were performed using the Friedman test.

In order to evaluate the consistency of responses over time, the ICC was calculated. The analysis was conducted in accordance with the two-way mixed effects and absolute agreement model [ICC (3,k)] (22). Inter-rater agreement was determined by calculating Cohen's Kappa coefficient (23). In all analyses, the level of statistical significance was accepted as $p<0.05$.

Results

The results are presented according to the predefined outcome measures, including readability, overall quality, and temporal consistency of the ChatGPT-generated responses.

The findings obtained using the Ateşman readability formula are presented in Table I. A significant difference was observed among repeated assessments within Days

1-4 (p=0.017, p=0.016, p=0.003, and p=0.018, respectively), whereas no significant intra-day differences were detected for Days 5-7 (p=0.937, p=0.754, and p=0.222, respectively). Furthermore, between-day comparisons revealed statistically significant differences in Ateşman readability scores (p<0.001). The highest mean readability score was observed on Day 2, Repetition 1 (79.2±15.1), whereas the lowest mean score was recorded on Day 7, Repetition 2 (60.3 ±15.6). Despite these statistically significant differences, the readability scores remained within the same readability category throughout the study period, indicating that the observed variations were not associated with a meaningful change in the practical readability of the generated responses.

The results obtained using the Çetinkaya-Uzun readability formula are presented in Table II. A significant intra-day variability was observed among repeated assessments on Days 1-4 (p=0.022, p=0.020, p=0.003, and p=0.009, respectively), while no significant differences were detected on Days 5-7 (p=0.887, p=0.525, and p=0.399, respectively). Furthermore, between-day comparisons revealed statistically significant differences in readability scores (p<0.001). Similarly, although statistically significant differences were observed, these variations did not result in meaningful changes in the corresponding educational reading levels, suggesting that the practical comprehensibility of the responses remained largely stable over time.

Table I. Comparison of Ateşman readability scores between days and repeated assessments within each day

Day	Mean ± SD	Median (Q1-Q3)	Repetition	Mean ± SD	Median (Q1-Q3)	p value
Day 1	69.4±16.2^{AB}	72.3 (58.4-81.6)	Repetition 1	64.4±14.9^{a}	62.8 (56.3-78.3)	p=0.017$^{£}$
			Repetition 2	75.7±14.8^{b}	72.1(68.7-81.7)	
			Repetition 3	68.1±17.2^{ab}	73.6 (64.8-79.2)	
Day 2	73.7±18.2^{BC}	75.1 (61.7-86.8)	Repetition 1	79.2±15.1^{a}	77.8 (71.3-85)	p=0.016$^{¥}$
			Repetition 2	69.4±20^{b}	69.9 (57.2-83.2)	
			Repetition 3	72.4±18.5^{ab}	72.4 (64.6-79.7)	
Day 3	68.9±16.8^{AB}	70.4 (57.1-80.3)	Repetition 1	71.3±12.2^{a}	71.5 (61.4-81.4)	p=0.003$^{¥}$
			Repetition 2	61.1±15.5^{b}	62.7 (49.1-70.4)	
			Repetition 3	74.3±19.6^{ab}	77.9 (61.9-84.9)	
Day 4	72.3±14.6^{BC}	73.5 (61.2-82.4)	Repetition 1	69.8±15^{a}	68.2 (59-81.4)	p=0.018$^{¥}$
			Repetition 2	76.2±16.9^{b}	73.8 (63.8-94.3)	
			Repetition 3	70.7±11.1^{ab}	70.1 (63.2-76.2)	
Day 5	71.1±14.2^{B}	72.1 (60.8-80.1)	Repetition 1	71.6±16.3^{a}	76.5 (62.7-81.5)	p=0.937$^{¥}$
			Repetition 2	70.8±10.9^{a}	71.6 (62.7-79.8)	
			Repetition 3	70.6±15.4^{a}	69.9 (59-78.6)	
Day 6	75.3±14.2^{C}	77.2 (65.1-87.9)	Repetition 1	76.1±12^{a}	77.6 (69.6-83.6)	p=0.754$^{¥}$
			Repetition 2	75.8±17.7^{a}	75.9 (60.5-88.1)	
			Repetition 3	73.9±12.7^{a}	74 (65.2-82.3)	
Day 7	63.3±15.4^{A}	64.8 (51.6-73.9)	Repetition 1	65.9±14.4^{a}	64.8 (55.6-74)	p=0.222$^{¥}$
			Repetition 2	60.3±15.6^{a}	57.3 (47.4-70.2)	
			Repetition 3	63.6±16.2^{a}	65.1 (56-72.9)	
Between-day comparison p value	<0.001*$^{£}$		Between 21 repeated assessments p value	<0.001*$^{£}$		

Data are presented as mean ± SD and median (Q1-Q3)
$^{£}$: Friedman test, $^{¥}$: Repeated measures ANOVA
*Comparisons between repetitions within each day were performed using repeated-measures ANOVA for normally distributed data and the Friedman test for non-normally distributed data
Different lowercase superscript letters ($^{a, b}$) indicate statistically significant differences between repeated assessments within the same day (p<0.05)
Different uppercase superscript letters ($^{A, B, C}$) indicate statistically significant differences between days (p<0.05)
SD: Standard deviation; Q1: First quartile; Q3: Third quartile

The results of the GQS evaluation are presented in Table III. No statistically significant differences were found between repetitions within the same day for any of the days analyzed (respectively: p=0.497; p=0.595; p=0.368; p=0.692; p=0.311; p=0.174; p=0.054). In the analysis in which all repetitions were evaluated together (1st-21st repetitions), no significant difference was detected (p=0.650).

The ICC analysis results for the Ateşman readability scores are presented in Table IV. Reliability ranging from moderate to good was observed in the within-day repeated measurements (ICC=0.588-0.883). In the analysis evaluating all repetitions together, the ICC value was 0.958 [95% confidence interval (CI): 0.925-0.980]. All ICC values were found to be statistically significant (p<0.001).

The ICC analysis results for the Çetinkaya-Uzun readability scores are presented in Table V. Moderate, good, and excellent levels of reliability were obtained in the within-day repeated measurements (ICC=0.650-0.919). In the analysis evaluating all measurements together, the ICC value was 0.961 (95% CI: 0.932-0.982). All ICC values obtained in the analyses were statistically significant (p<0.001).

The ICC analysis results for the GQS scores are presented in Table VI. In the analysis evaluating all measurements together, the ICC value was 0.919 (95% CI: 0.919-0.979). All ICC analyses were found to be statistically significant (p<0.001).

Table II. Comparison of Çetinkaya-Uzun readability scores between days and repeated assessments within each day

Day	Mean ± SD	Median (Q1-Q3)	Repetition	Mean ± SD	Median (Q1-Q3)	p value
Day 1	41.6±9.4^{AB}	43.1 (35.2-47.9)	Repetition 1	39.2±8.1^{a}	39.7 (32.5-46.6)	p=0.022$^{£}$
			Repetition 2	45±8.5^{b}	42.4 (40.1-48.4)	
			Repetition 3	40.6±10.7^{ab}	43.9 (38.6-47.2)	
Day 2	44.1±11.1^{BC}	45.2 (37.4-50.8)	Repetition 1	47.9±9.8^{a}	45.4 (42.5-50.7)	p=0.020$^{¥}$
			Repetition 2	47.1±12.3^{b}	42.8 (32.8-48.1)	
			Repetition 3	43.6±10.8^{ab}	43.9 (37.4-47.8)	
Day 3	41.1±9.8^{AB}	41.8 (34.7-47.1)	Repetition 1	42.1±7.1^{a}	41.5 (38.9-47.7)	p=0.003$^{¥}$
			Repetition 2	37.3±9.4^{b}	36.2 (31.3-42.8)	
			Repetition 3	43.9±11.6^{ab}	45 (35.8-49.6)	
Day 4	43.1±9.2^{BC}	43.9 (36.9-48.5)	Repetition 1	42.2±9.2^{a}	40.5 (34.3-50.4)	p=0.009$^{¥}$
			Repetition 2	45.4±10.6^{b}	42.7 (38.1-55.3)	
			Repetition 3	41.6±7.4^{a}	39.8 (35.6-46.6)	
Day 5	42.6±8.2^{B}	42.7 (36.4-47.6)	Repetition 1	42.9±8.7^{a}	43.7 (37.2-48.6)	p=0.887$^{¥}$
			Repetition 2	42.1±6^{a}	41.4 (37.2-48.4)	
			Repetition 3	42.8±9.7^{a}	40.7 (36.3-46.4)	
Day 6	44.8±8.8^{C}	46.3 (39.1-51.4)	Repetition 1	45.3±7.6^{a}	45.3 (41.3-49.3)	p=0.525$^{¥}$
			Repetition 2	45.6±10.8^{a}	44.6 (38.4-53.6)	
			Repetition 3	43.4±7.9^{a}	42.7 (37.7-48.1)	
Day 7	38.1±9.4^{A}	38.5 (31.7-43.8)	Repetition 1	39±9.1^{a}	37.2 (31-44.3)	p=0.399$^{¥}$
			Repetition 2	36.6±9.5^{a}	36 (31.2-40.8)	
			Repetition 3	38.5±9.7^{a}	39.2 (34.7-43.8)	
Between-day comparison p value	<0.001*$^{£}$		Between 21 repeated assessments p value	<0.001*$^{£}$		

Data are presented as mean ± SD and median (Q1-Q3)
$^{£}$: Friedman test, $^{¥}$: Repeated measures ANOVA
*Comparisons between repetitions within each day were performed using repeated-measures ANOVA for normally distributed data and the Friedman test for non-normally distributed data
Different lowercase superscript letters ($^{a, b}$) indicate statistically significant differences between repeated assessments within the same day (p<0.05)
Different uppercase superscript letters ($^{A, B, C}$) indicate statistically significant differences between days (p<0.05)
SD: Standard deviation; Q1: First quartile; Q3: Third quartile

Table III. Comparison of Global Quality Score (GQS) scores according to days and repeated assessments within each day

Day	Repetition	Mean ± SD	Median (Q1-Q3)	p value
Day 1	Repetition 1	4.4±0.59	4 (4-5)	p=0.497
	Repetition 2	4.2±0.77	4 (4-5)	
	Repetition 3	4.3±0.55	4 (4-5)	
Day 2	Repetition 1	4.3±0.64	4 (4-5)	p=0.595
	Repetition 2	4.2±0.89	4 (4-5)	
	Repetition 3	4.1±0.97	4 (4-5)	
Day 3	Repetition 1	4.2±0.67	4 (4-5)	p=0.368
	Repetition 2	4.3±0.57	4 (4-5)	
	Repetition 3	4.3±0.57	4 (4-5)	
Day 4	Repetition 1	4.2±0.99	4 (4-5)	p=0.692
	Repetition 2	4.3±0.57	4 (4-5)	
	Repetition 3	4.2±0.70	4 (4-5)	
Day 5	Repetition 1	4.5±0.51	4 (4-5)	p=0.311
	Repetition 2	4.3±0.66	4 (4-5)	
	Repetition 3	4.3±0.57	4 (4-5)	
Day 6	Repetition 1	4.3±0.47	4 (4-5)	p=0.174
	Repetition 2	4.2±0.88	4 (4-5)	
	Repetition 3	4.4±0.50	4 (4-5)	
Day 7	Repetition 1	4.0±1.05	4 (4-5)	p=0.054
	Repetition 2	4.3±0.79	4 (4-5)	
	Repetition 3	4.2±0.75	4 (4-5)	
Between-day comparison	-	-	-	p=0.650

Data are presented as mean **±** SD and median (Q1-Q3)
Within-day comparisons were performed using the Friedman test
SD: Standard deviation, Q1: First quartile, Q3: Third quartile

Table IV. ICC reliability analysis of Ateşman readability scores

Repetitions	ICC	95% CI	Interpretation	p value
A1-A2-A3	0.736	0.448-0.887	Moderate	p<0.001
A4-A5-A6	0.836	0.649-0.930	Good	p<0.001
A7-A8-A9	0.733	0.421-0.887	Moderate	p<0.001
A10-A11-A12	0.883	0.746-0.950	Good	p<0.001
A13-A14-A15	0.812	0.600-0.920	Good	p<0.001
A16-A17-A18	0.588	0.117-0.826	Moderate	p<0.001
A19-A20-A21	0.802	0.589-0.915	Good	p<0.001
A1-A21	0.958	0.925-0.980	Excellent	p<0.001

ICC values were calculated using a two-way mixed-effects model with absolute agreement [ICC (3,k)]
A1-A21 represent sequential Ateşman readability measurements across all repetitions and days
ICC: Intraclass Correlation Coefficient, CI: Confidence interval

Table V. ICC reliability analysis of Çetinkaya-Uzun readability scores

Repetitions	ICC	95% CI	Interpretation	p value
C1-C2-C3	0.699	0.386-0.869	Moderate	p<0.001
C4-C5-C6	0.870	0.722-0.945	Good	p<0.001
C7-C8-C9	0.797	0.555-0.915	Good	p<0.001
C10-C11-C12	0.919	0.818-0.966	Excellent	p<0.001
C13-C14-C15	0.799	0.574-0.915	Good	p<0.001
C16-C17-C18	0.650	0.261-0.851	Moderate	p<0.001
C19-C20-C21	0.822	0.629-0.924	Good	p<0.001
C1-C21	0.961	0.932-0.982	Excellent	p<0.001

ICC values were calculated using a two-way mixed-effects model with absolute agreement [ICC (3,k)]
C1-C21 represent sequential Çetinkaya-Uzun readability score measurements across all days and repetitions
ICC: Intraclass Correlation Coefficient, CI: Confidence interval

Table VI. ICC reliability analysis of GQS scores

Repetitions	ICC	95% CI	Interpretation	p value
G1-G2-G3	0.839	0.663-0.931	Good	p<0.001
G4-G5-G6	0.596	0.137-0.829	Moderate	p<0.001
G7-G8-G9	0.867	0.722-0.943	Good	p<0.001
G10-G11-G12	0.783	0.543-0.908	Good	p<0.001
G13-G14-G15	0.791	0.564-0.911	Good	p<0.001
G16-G17-G18	0.767	0.517-0.900	Good	p<0.001
G19-G20-G21	0.906	0.803-0.960	Excellent	p<0.001
G1-G21	0.919	0.919-0.979	Excellent	p<0.001

ICC values were calculated using a two-way mixed-effects model with absolute agreement [ICC (3,k)]
G1-G21 represent sequential GQS measurements across all days and repetitions
ICC: Intraclass Correlation Coefficient, CI: Confidence interval, GQS: Global Quality Scale

Discussion

The findings of the present study demonstrated that ChatGPT-generated responses to frequently asked question in pediatric dentistry generally exhibited a moderate level of readability, high quality scores, and overall good consistency between repetitions. However, variability in readability scores was observed both within the same day and across different days. Accordingly, the first and second null hypotheses, proposing that the Ateşman and Çetinkaya-Uzun readability scores did not differ significantly across different days and repetitions, were rejected. The third null hypothesis, which proposed that GQS scores did not demonstrate significant differences over time, was accepted, whereas the fourth null hypothesis, which proposed that ChatGPT responses did not demonstrate a high level of consistency over time, was rejected. Collectively, these results suggest that although ChatGPT maintains a relatively stable level of informational quality over time, the linguistic complexity of its responses may fluctuate depending on when the same prompt is submitted.

According to the findings obtained in the present study, the readability results demonstrated that the Ateşman readability scores were generally within the range of 60-80, indicating that the responses were between the "moderately difficult" and "easy" reading levels. Similarly, the Çetinkaya-Uzun scores ranging between 35 and 47 suggested that the contents could be classified within the '"instructional reading level" category. These results indicate that the responses are understandable for a certain user population. However, the "instructional reading level" identified by the Çetinkaya-Uzun formula approximately corresponds to an eighth- to ninth-grade educational level, suggesting that individuals with limited educational attainment or lower health literacy may still experience difficulties in fully comprehending or applying the information provided. Consequently, readability should be considered as an important determinant of the practical usefulness of LLM-

generated patient education materials, in addition to their factual accuracy and informational quality.

Previous studies have similarly reported that the readability levels of health-related content generated by LLMs is generally acceptable, although substantial variation exists depending on the clinical topic, language, and assessment method employed (24-27). In a study conducted by Aypar Akbağ (26), the mean Ateşman readability score of content related to gestational diabetes generated by ChatGPT and Gemini was reported as 77.8. The researchers stated that these contents had a high level of readability and were considered understandable for users (26). Similarly, in a study conducted by Boztaş Demir and Görgülü (27), the Turkish responses generated by ChatGPT and Gemini to patient questions in the field of orthodontics were reported to be generally adequate in terms of accuracy, comprehensiveness, and readability. Erdat et al. (25) also reported that ChatGPT's responses to questions related to colorectal cancer demonstrated moderate readability and good quality characteristics. Overall, the findings of the present study are consistent with this growing body of evidence indicating that LLM-generated health information is generally understandable and of acceptable educational quality. Nevertheless, direct comparison across studies should be interpreted cautiously as the evaluated clinical topics, response languages, and readability formulas differ considerably.

When the literature specific to pediatric dentistry is considered, an interesting pattern emerges. Although direct numerical comparison is limited by differences in language and readability indices, Kocaoğlu et al. (15) and Aydın Varol et al. (16) similarly reported that ChatGPT-generated responses addressing procedural or guideline-based pediatric dental topics in English generally required university-level reading ability. By contrast, the parental frequently asked questions evaluated in the present study demonstrated moderate-to-easy readability. This discrepancy is likely attributable to the inherently greater linguistic complexity of procedure-oriented subjects, such as pediatric dental sedation and pulpotomy, compared with preventive oral health information intended for parents, rather than reflecting a true language-related difference. These observations further emphasize that the readability of LLM-generated responses depends not only on the characteristics of the language model itself, but also on the complexity and intended audience of the clinical topic being addressed.

One of the notable findings of the present study was that readability scores varied significantly despite the use of identical prompts. Significant differences were observed both between repeated assessments performed on the same day and between different days, indicating that the linguistic characteristics of ChatGPT-generated responses are not entirely stable over time. This situation may be explained by the probabilistic nature of LLMs. LLMs do not always generate the same output for the same input and may demonstrate variability in characteristics such as sentence structure, word choice, and text length. As readability indices are inherently sensitive to these textual characteristics, fluctuations in readability scores are expected even when the underlying informational content remains largely unchanged. These findings suggest that readability assessments performed at a single time point may not adequately represent the linguistic variability of LLM-generated health information. Importantly, the temporal fluctuations observed in the present study provide support for concerns regarding potential model drift and response variability which have previously been discussed in the pediatric dentistry literature. Kocaoğlu et al. (15), for example, identified their cross-sectional design as an important limitation and recommended repeated longitudinal evaluations in order to determine whether chatbot performance changes over time. Similarly, Aydın Varol et al. (16) acknowledged that the temporal consistency of individual chatbots could not be assessed within their study design, while Uçar Gündoğar and Sarıoğlu (17) limited repeated evaluations to a single day, precluding an assessment of between-day variability. By employing a seven-day, three-times-daily repeated-measures protocol, the present study directly addresses this methodological gap and demonstrated that although the overall informational quality of ChatGPT responses remains remarkably stable, their linguistic complexity may fluctuate across repeated interactions. These findings indicate that temporal reproducibility constitutes an independent dimension of LLM performance which cannot be adequately captured by conventional single-timepoint evaluations.

When the quality evaluation was examined, ChatGPT responses demonstrated a consistent performance with high GQS scores across all time periods, and no significant differences were observed between repetitions. Similarly, previous studies have reported that the responses of LLMs in the healthcare field are generally of moderate-to-high quality (28,29). In a systematic review published in 2024, Li et al. (29) reported that ChatGPT mostly demonstrated "passing" or '"moderate" performance across different healthcare applications, although it was not yet considered

fully reliable for clinical use. In another systematic review, ChatGPT was reported to have a broad range of applications in healthcare education, research, and clinical practice, offering significant advantages in areas such as improving writing quality, data analysis, and patient education (28). That same study also emphasized that despite high efficiency and explanatory capacity in content generation, risks such as inaccuracies, bias, and "hallucinations" remain important concerns. The phenomenon of "hallucinations", defined as the ability of LLMs to generate different or inaccurate responses to the same input, has been highlighted as a potential risk for users, particularly in the healthcare field (28,29). Therefore, despite the high-quality scores observed, such systems should not be regarded as definitive standalone sources of information. Recent discussions regarding the design philosophy of LLMs further reinforce these concerns. Contemporary LLMs are generally optimized to generate a response to user prompts rather than explicitly indicating uncertainty or declining to answer. Consequently, inaccurate or unsupported information may occasionally be presented in a fluent and confident manner, making it difficult for users to distinguish between reliable and erroneous content. In the context of pediatric dentistry, where parents may rely on online information to make decisions regarding their children's oral health, this behavior highlights the importance of professional verification and reinforces that LLM-generated information should complement, but not replace, clinical judgment.

The stability of GQS scores observed in the present study parallels findings reported in recent pediatric dentistry investigations. Both Karamüftüoğlu et al. (14) and Aydın Varol et al. (16) similarly observed no significant differences in GQS values, although their comparisons were performed between different chatbot platforms rather than across time. Considered together with the present findings, these studies suggest that GQS may represent a relatively robust measure of the overall educational quality of LLM-generated content, remaining comparatively stable regardless of whether the source of variation is model selection or repeated prompting over time. Furthermore, Kocaoğlu et al. (15) demonstrated no significant association between DISCERN quality scores and FKGL readability scores across multiple chatbots, supporting the concept that readability and informational quality constitute distinct dimensions of LLM-generated health information. The present study further strengthens this interpretation by demonstrating that readability may fluctuate substantially over time while overall quality remains stable, indicating that linguistic accessibility and informational quality should be evaluated independently rather than being considered interchangeable indicators of content performance.

Nevertheless, consistently high GQS scores should not be interpreted as evidence of complete factual reliability. Recent pediatric dentistry studies have demonstrated that even LLM-generated responses rated highly for quality may contain fabricated or inaccurate references. Uçar Gündoğar and Sarıoğlu (17) reported that although DeepSeek R1 demonstrated greater overall consistency and accuracy than ChatGPT-4o when generating information related to molar-incisor hypomineralization, both models exhibited substantial rates of fabricated citations. Similarly, Sağlam et al. (13) found that the accuracy of cited references was consistently lower than the accuracy of the explanatory content itself across all evaluated LLMs. Taken together with the present findings, these observations indicate that fluent language, stable quality scores, and coherent explanations should not be interpreted as guarantees of source reliability. Consequently, LLM-generated health information, particularly when intended for patient education, should continue to undergo professional verification before being incorporated into clinical communication or educational resources.

Upon examination of the consistency analyses evaluated in the present study, it was seen that the responses generated by ChatGPT generally exhibited a satisfactory level of consistency, as evidenced by the high ICC values obtained. In particular, the high ICC values observed in the analyses evaluating all repetitions collectively suggest that the model is generally capable of producing stable and predictable responses to similar prompts. However, the presence of significant differences in readability scores both within the same day and across different days, despite the absence of significant differences in GQS scores over time, suggests that LLMs may exhibit variability in linguistic complexity while maintaining a more stable level of overall information quality. The findings, when considered as a whole, suggest that the evaluation of health-related content generated by LLMs should consider not only a single response, but also temporal variability and response consistency over time.

From a clinical perspective, when the findings obtained from this study are evaluated collectively, LLMs appear to be potential tools which may support rapid access to information for parents in the field of pediatric dentistry. However, the results also demonstrate that these systems should not be considered completely reliable and standardized sources of information. In particular, the

fact that readability levels may not be equally suitable for all user groups and that responses may exhibit a certain degree of variability over time suggests that LLMs should be regarded as supportive tools in clinical practice. Although the Cetinkaya-Uzun readability scores corresponded to the instructional reading level (approximately the 8th-9th grade), this level may still exceed the reading and health literacy skills of a substantial proportion of caregivers. Consequently, despite the consistently high quality of the information provided, some parents may experience difficulty in fully understanding or applying the generated recommendations. Furthermore, the possibility that semi-technical or clinical expressions may be misinterpreted by individuals with low health literacy should also be taken into consideration. These findings highlight the importance of optimizing the readability of LLM-generated patient education materials in addition to ensuring their informational quality particularly as such tools are intended for use by diverse populations with varying levels of health literacy. Therefore, it is recommended that health-related information provided by LLMs should, when necessary, be supported by expert evaluation and interpreted with caution.

From a methodological perspective, the present study possesses several important strengths. Unlike previous investigations which primarily adopted cross-sectional or single-timepoint designs, the present study incorporated a longitudinal repeated-measures protocol consisting of three independent assessments per day over seven consecutive days. This design enabled simultaneous evaluations of readability, information quality, and temporal reproducibility while minimizing the likelihood that the findings reflected a single stochastic response. Furthermore, the combined use of two validated Turkish readability formulas, the GQS, and intraclass correlation analyses provided a multidimensional assessment framework, allowing for the complementary evaluation of linguistic accessibility, educational quality, and response stability. Such an approach offers a more comprehensive characterization of LLM-generated patient education materials in comparison to studies relying on a single evaluation metric.

Nevertheless, several limitations should be considered when interpreting the findings. First, the present study evaluated only ChatGPT; therefore, the results cannot be generalized to other LLMs, which may demonstrate different linguistic characteristics, quality profiles, or temporal stability. This deliberate methodological choice allowed for a rigorous repeated-measures design focused on the temporal reproducibility of a single model but precluded direct comparisons between competing LLMs. Future studies incorporating multiple LLMs within longitudinal repeated-measures protocols would help determine whether the temporal variability observed in the present study is model-specific or represents a more general characteristic of contemporary LLMs. Second, the evaluation period was limited to seven consecutive days. Given that commercial LLMs undergo continuous algorithmic refinement and periodic updates, longer follow-up periods may reveal additional patterns of temporal variation which could not be captured within the present study. Finally, although readability, information quality, and temporal consistency were comprehensively assessed, this study did not evaluate the factual accuracy, reference validity, or clinical correctness of the generated responses. These dimensions remain essential components of the performance of LLMs and should be incorporated into future investigations in order to provide a more holistic assessment of LLM-assisted patient education.

Conclusion

This study demonstrated that the Turkish responses generated by ChatGPT to frequently asked questions in the field of pediatric dentistry generally exhibited moderate readability, high quality, and good consistency. However, variability in readability scores was observed both within the same day and across different days. The findings obtained indicate that LLMs may be used as supportive tools in patient and parent information processes; however, as the generated content may vary in terms of readability and consistency, these systems should not be regarded as completely reliable sources of information in clinical settings without human supervision and expert evaluation.

Ethics

Ethics Committee Approval: This study was conducted using responses generated by ChatGPT and did not involve human participants, patient data, personal information, or biological materials. Therefore, ethics committee approval was not required for this study.

Informed Consent: This study was conducted using responses generated by ChatGPT and did not involve human participants, patient data, personal information, or biological materials. Therefore, informed consent was not required for this study.

Footnotes

Authorship Contributions

Concept: Irem Nur Bitisik, Ebru Kucukyilmaz Izgi, Design: Ebru Kucukyilmaz Izgi, Irem Nur Bitisik, Data Collection: Selcuk Savas, Mehmet Izgi, Analysis or Interpretation: Ilgin Timarci, Literature Search: Mehmet Izgi, Selcuk Savas, Writing: Irem Nur Bitisik, Selcuk Savas.

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

Financial Disclosure: The authors received no financial support for the conduct, authorship, or publication of this study.

References

1. Michalski RS, Carbonell JG, Mitchell TM, editors. Machine learning: an artificial intelligence approach. 1st ed. Berlin, Germany: Springer-Verlag; 1983.
2. Agrawal P, Nikhade P. Artificial intelligence in dentistry: past, present, and future. Cureus. 2022; 14:e27405.
3. Schwendicke F, Samek W, Krois J. Artificial intelligence in dentistry: chances and challenges. J Dent Res. 2020; 99:769-74.
4. Sezer B, Okutan AE. Evaluation of ChatGPT-4's performance on pediatric dentistry questions: accuracy and completeness analysis. BMC Oral Health. 2025; 25:1427.
5. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. J Esthet Restor Dent. 2023; 35:1098-102.
6. Shahsavar Y, Choudhury A. User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. JMIR Hum Factors. 2023; 10:e47564.
7. Issa J, Sohrabniya F, Brinz J, Dyszkiewińska-Konwińska M, Chaurasia A. The role of large language models in dental diagnosis, treatment planning, and prognosis. J Stomatol. 2025; 78:298-303.
8. Islam MR, Urmi TJ, Mosharrafa RA, Rahman MS, Kadir MF. Role of ChatGPT in health science and research: a correspondence addressing potential application. Health Sci Rep. 2023; 6:e1625
9. Alhaidry HM, Fatani B, Alrayes JO, Almana AM, Alfhaed NK. ChatGPT in dentistry: a comprehensive review. Cureus. 2023; 15:e38317
10. Fatani B. ChatGPT for future medical and dental research. Cureus. 2023; 15:e37285.
11. Gülçin Çetin S, Karadağ GG, Çetin SG. Comparative evaluation of artificial intelligence models in answering pediatric dentistry questions of the Turkish Dental Specialization Exam (DUS). Dicle Dent J. 2025; 26:31-7.
12. Aşık A, Kuru E. Analysis of ChatGPT's answers to pedodontics questions asked in the dentistry specialization training entrance exam: cross-sectional study. Turkiye Klinikleri J Dent Sci. 2025; 31:401-6.
13. Sağlam C, Aşık A, Kuru E, et al. Evaluating the efficacy of large language models in providing information for parental inquiries regarding primary care of pediatric oral and dental health. J Clin Pediatr Dent. 2026; 50:132-41.
14. Karamüftüoğlu N, Varol EA, Bal C. Exploring artificial intelligence chatbots in pediatric fluoride education: a cross-sectional study. Sci Rep. 2025; 16:182.
15. Kocaoğlu MH, Demirel A, Kaya İ. Accuracy, quality, and readability analyses of responses from large language models to questions on pediatric dental sedation. BMC Oral Health. 2026; 26:492.
16. Aydın Varol E, Öztürk Z, Bal C, Karamüftüoğlu N. Comparison of two different artificial intelligence chatbots that provide information to patients and parents about primary tooth pulpotomy treatments. BMC Oral Health. 2026; 26:685.
17. Uçar Gündoğar Z, Sarıoğlu D. Comparative assessment of quality, consistency, and reference accuracy of MIH-related clinical information generated by ChatGPT-4o and DeepSeek R1. BMC Oral Health. 2026; 26.
18. Ateşman E. Türkçede okunabilirliğin ölçülmesi. AÜ Tömer Dil Dergisi. 1997; 58:171-4.
19. Çetinkaya G, Uzun L. Türkçe ders kitaplarındaki metinlerin okunabilirlik özellikleri. In: Ülper H, editor. Türkçe Ders Kitabı Çözümlemeleri. Ankara: Pegem Akademi; 2010. p. 141-55.
20. Bernard A, Langille M, Hughes S, Rose C, Leddin D, Veldhuyzen van Zanten S. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. Am J Gastroenterol. 2007; 102:2070-7.
21. American Academy of Pediatric Dentistry. Frequently asked questions for parents. Available from: https://www.aapd.org/resources/parent/faq/. Accessed: June 2026.
22. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York: Wiley; 1981. p. 212-25.
23. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33:159-74.
24. Koç U, Güneş YC, Çolakoğlu MN, et al. ChatGPT-generated informed consent forms in interventional radiology: a randomized controlled evaluation of comprehension, attitudinal responses, and readability. Eur J Radiol. 2026; 195:112624.
25. Erdat EC, Yalçıner M, Utkan G. Readability and quality analysis of ChatGPT o1's responses on colorectal cancer: a study of an AI language model. Acta Haematol Oncol Turc. 2025; 58:74-80.
26. Aypar Akbağ NN. Assessing artificial intelligence-generated patient educational material on gestational diabetes mellitus. J Perinat Neonatal Nurs. 2025; 39:210-7.
27. Boztaş Demir G, Görgülü S. The Turkish proficiency of ChatGPT and Gemini in orthodontics: an evaluation of the accuracy, completeness, and readability of responses to patient questions. Yeditepe J Dent. 2025; 21:151-8.
28. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel). 2023; 11.
29. Li J, Dada A, Puladi B, Kleesiek J, Egger J. ChatGPT in healthcare: a taxonomy and systematic review. Comput Methods Programs Biomed. 2024; 245:108013.