



Reliability of an AI-based Conversational Agent in Pediatric Dentofacial Health Counseling

Pediyatrik Dentofasiyal Sağlık Danışmanlığında Yapay Zeka Tabanlı Konuşma Ajanının Güvenilirliği

© Ceren Yapar Gümüş¹, © Süleyman Kutalmış Büyük^{2,3}, © Hale Akdemir³

¹Department of Pediatrics, Ordu University, Faculty of Medicine, Ordu, Türkiye

²Department of Dental Biomaterials Science, Dental Research Institute, Seoul National University, Seoul, South Korea

³Department of Orthodontics, Ordu University, Faculty of Dentistry, Ordu, Türkiye

ABSTRACT

Objective: To evaluate the reliability and quality of Grok artificial intelligence's (AI) responses to commonly asked questions about pediatric dentofacial health.

Methods: Four queries related to detrimental oral habits were submitted to Grok AI in separate sessions. The generated responses were independently evaluated by orthodontists, pediatricians, and parents without a medical or dental background. Reliability was assessed using the eight-item modified DISCERN instrument, and overall information quality was evaluated using the Global Quality Scale (GQS). Group comparisons were performed using the Kruskal-Wallis test, and correlations among scores were analyzed using Spearman's rank correlation.

Results: Across evaluator groups, mean modified DISCERN scores ranged from 23.85 to 27.66, and mean GQS scores ranged from 3.61 to 4.12, corresponding, overall, to moderate-to-good reliability and quality. No statistically significant differences were observed among orthodontists, pediatricians, and parents for any query. Correlation analyses showed significant positive associations among several modified DISCERN and GQS parameters, particularly across different queries within the same evaluator groups. These findings suggested consistent scoring patterns, especially among parents and pediatricians.

Conclusion: Grok AI provided moderately reliable and generally useful responses regarding pediatric dentofacial health. However, its responses should be considered as supportive educational information

Öz

Amaç: Bu çalışmanın amacı, Grok yapay zeka (YZ) sisteminin pediyatrik dentofasiyal sağlık konusunda sık sorulan sorulara verdiği yanıtların güvenilirliğini ve kalitesini değerlendirmektir.

Yöntemler: Zararlı oral alışkanlıklarla ilişkili dört farklı soru, ayrı oturumlarda Grok YZ sistemine yöneltildi. Elde edilen yanıtlar; ortodontistler, çocuk sağlığı ve hastalıkları uzmanları ile tıp veya diş hekimliği alanında eğitim almamış ebeveynler tarafından birbirinden bağımsız olarak değerlendirildi. Yanıtların güvenilirliği, sekiz maddeden oluşan modifiye DISCERN aracı kullanılarak; genel bilgi kalitesi ise Global Quality Scale (GQS) ile değerlendirildi. Gruplar arasındaki karşılaştırmalar Kruskal-Wallis testi ile yapıldı, puanlar arasındaki ilişkiler ise Spearman sıra korelasyon analizi kullanılarak incelendi.

Bulgular: Değerlendirici gruplar genelinde ortalama modifiye DISCERN puanları 23,85 ile 27,66 arasında, ortalama GQS puanları ise 3,61 ile 4,12 arasında değişti. Bu değerler, yanıtların genel olarak orta ile iyi düzey arasında güvenilirlik ve kaliteye sahip olduğunu gösterdi. Hiçbir soru için ortodontistler, çocuk sağlığı ve hastalıkları uzmanları ve ebeveynler arasında istatistiksel olarak anlamlı bir fark saptanmadı. Korelasyon analizlerinde, çeşitli modifiye DISCERN ve GQS parametreleri arasında, özellikle aynı değerlendirici gruplarının farklı sorulara verdikleri puanlar açısından anlamlı pozitif ilişkiler belirlendi. Bu bulgular, özellikle ebeveynler ve çocuk sağlığı ve hastalıkları uzmanları arasında puanlama örüntülerinin tutarlı olduğunu düşündürdü.

Sonuç: Grok YZ, pediyatrik dentofasiyal sağlıkla ilgili konularda orta düzeyde güvenilir ve genel olarak yararlı yanıtlar sundu. Bununla birlikte,

Cite this article as: Yapar Gümüş C, Kutalmış Büyük S, Akdemir H. Reliability of an AI-based conversational agent in pediatric dentofacial health counseling. Gazi Med J. 2026;37(4):504-510

Address for Correspondence/Yazışma Adresi: Ceren Yapar Gümüş, Department of Pediatrics, Ordu University, Faculty of Medicine, Ordu, Türkiye
E-mail / E-posta: cerenyapar91@gmail.com
ORCID ID: orcid.org/0000-0001-6349-2514

Received/Geliş Tarihi: 03.06.2026
Accepted/Kabul Tarihi: 13.08.2026
Publication Date/Yayınlanma Tarihi: 15.10.2026



©Copyright 2026 The Author(s). Published by Galenos Publishing House on behalf of Gazi University Faculty of Medicine. Licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND) International License.

*Telif Hakkı 2026 Yazar(lar). Gazi Üniversitesi Tıp Fakültesi adına Galenos Yayınevi tarafından yayımlanmaktadır. Creative Commons Atıf-GayriTicari-Türetilemez 4.0 (CC BY-NC-ND) Uluslararası Lisansı ile lisanslanmaktadır.

ABSTRACT

rather than a substitute for a professional pediatric or orthodontic assessment. Further validation using larger question sets, multiple AI models, and more diverse evaluator groups is needed.

Keywords: Artificial intelligence, generative artificial intelligence, natural language processing, child health education, malocclusion

ÖZ

bu yanıtlar profesyonel pediatrik veya ortodontik değerlendirmenin yerine geçen bir kaynak olarak değil, destekleyici eğitim bilgisi olarak değerlendirilmelidir. Daha geniş soru setlerinin, farklı YZ modellerinin ve daha çeşitli değerlendirici gruplarının yer aldığı ileri doğrulama çalışmalarına ihtiyaç vardır.

Anahtar Sözcükler: Yapay zeka, üretken yapay zeka, doğal dil işleme, çocuk sağlığı eğitimi, maloklüzyon

INTRODUCTION

Artificial intelligence (AI) applications, particularly chatbots powered by large language models, are being increasingly adopted in healthcare. These systems offer patients round-the-clock accessibility, user-friendliness, and rapid responses, enabling individuals to obtain information at any time of day (1,2). Owing to these features, they are regarded as a practical source of information by both the public and healthcare professionals.

Nevertheless, the accuracy and reliability of AI-based chatbots remain a matter of debate. Several studies have demonstrated that such tools may occasionally generate incorrect or irrelevant content and may even present misinformation as factual (3,4). More recent investigations have further highlighted that outputs generated by AI models can contain biases related to gender, race, or socioeconomic status (5). In addition, chatbot responses are often produced at a level exceeding recommended health literacy standards, which may hinder patients' understanding and reduce their active participation in care (6-8).

Recent studies have begun to evaluate the reliability of AI chatbots across different medical disciplines. Kurt Demirsoy et al. (9) examined ChatGPT's responses in orthodontics and found that, while the model could contribute to patient education, it remained limited in terms of factual accuracy. Similarly, Terzi et al. (10) reported that ChatGPT's responses regarding dental implants and peri-implant phenotypes were accessible but lacked sufficient depth for clinical decision-making. These findings underscore the potential utility of AI-based chatbots and emphasize the need for systematic evaluation across diverse clinical contexts. Although Grok has recently been included in comparative evaluations of large language models in dentistry, its reliability in pediatric dentofacial health counseling remains insufficiently characterized (11). This gap provided the rationale for selecting Grok for the present study.

Dentofacial health is closely linked to growth and development during childhood and adolescence. Harmful oral habits can negatively influence jaw development, disrupt muscle balance, and impair orofacial functions (12). In this context, both parents and healthcare professionals—particularly pediatricians and orthodontists—play a critical role in guiding healthy development. Preventive strategies and effective counseling, however, depend on access to trustworthy information (13). The reliability of information provided by AI-based digital advisors is therefore of great importance, not only for empowering parents but also for supporting clinicians in their educational and counseling roles.

This study aimed to assess the reliability of Grok AI, an AI-based chatbot, in providing guidance on dentofacial health. The chatbot's responses were evaluated by participants from different professional backgrounds using the modified DISCERN (mDISCERN) tool and the Global Quality Scale (GQS). By exploring the reliability of AI-based conversational agents in this field, the study seeks to contribute to both clinical practice and patient and parent education.

MATERIALS AND METHODS

In this study, frequently asked questions about detrimental oral habits were submitted to Grok AI, an artificial- intelligence-based conversational agent. To minimize the risk that the system's adaptive learning would influence subsequent responses, each question was entered separately into new sessions. The generated answers were then evaluated by three distinct participant groups through an online survey. Each group reviewed the same set of responses, thereby allowing a comprehensive assessment of Grok AI's accuracy, clarity, and relevance to the topic.

The following questions were directed to the chatbot:

1. What are the effects of mouth breathing on the dentofacial region?
2. What are the impacts of prolonged pacifier and bottle use on the head and facial region?
3. How do habits such as nail-biting, thumb-sucking, and biting foreign objects (e.g., pencils) affect the dentofacial region?
4. What are the effects of frenulum and tongue-tie on the dentofacial region?

The responses generated by Grok AI were assessed by three groups, including two expert panels, to capture perspectives across different demographics and educational backgrounds. The participant groups were defined as follows:

1. Pediatricians (medical specialists in child health),
2. Orthodontists (dental specialists in orthodontics),
3. Parents: individuals with at least an undergraduate degree, who have children but no professional background in medicine or dentistry.

Inclusion Criteria

- Having completed specialty training in pediatrics or orthodontics and voluntarily consenting to participate in the survey.
- For inclusion in the parent group: having at least an undergraduate degree, being a parent, and not belonging to a health-related profession.

Exclusion Criteria

- Being a resident in pediatrics or orthodontics.
- Declining to provide informed consent for participation.
- Parents who, despite meeting the above criteria, were affiliated with medical or dental professions.

The sample size was calculated using the G*Power program (version 3.1.9.2; Axel Buchner, University of Düsseldorf, Düsseldorf, Germany) to detect a medium effect size with 90% power. The required sample size was determined to be 66 participants (effect size: 0.4382637).

Evaluation of Responses

The chatbot’s answers were evaluated using two validated instruments: the mDISCERN questionnaire and the GQS. Assessments were performed independently by three groups: orthodontists, pediatricians, and parents with no professional background in healthcare.

Evaluation Scales

The DISCERN instrument is a widely used three-part tool designed to evaluate the reliability and quality of online health information (3,14). In this study, only the first section consisting of eight items was applied to examine the reliability of responses generated by the AI-based chatbot (14). The eight criteria assessed in the mDISCERN scale were as follows: (i) clarity of aims, (ii) achievement of stated aims, (iii) relevance of content to the topic, (iv) citation of information sources, (v) currency of information, (vi) impartiality and balance, (vii) provision of additional sources of support, and (viii) acknowledgment of areas of uncertainty. Each item was rated on a scale from 1 (no) to 5 (yes). The total score was categorized as poor (8-15), moderate (16-31), and good (32-40) (15).

The GQS is a five-point instrument used to assess the quality of medical information sources (2). The scale is defined as follows:

- 1 point: poor quality, lacking information, not useful for patients,
- 2 points: poor quality, many important details missing, limited usefulness,
- 3 points: moderate quality, some important information adequately addressed, others insufficiently covered,

- 4 points: good quality, most key information adequate and useful for patients,
- 5 points: excellent quality, fluent, comprehensive, and highly useful.

Accordingly, GQS scores of 1-2 indicate low quality, scores of 3 indicate moderate quality, and scores of 4-5 indicate high quality.

This study was approved by the Ordu University Non-Interventional Scientific Research Ethics Committee (approval no: 2025/246, date: 18.07.2025). Electronic informed consent was obtained from all participants before participation in the online survey.

Statistical Analysis

Statistical analyses were performed using IBM SPSS Statistics for Windows, version 20.0 (IBM Corp., Armonk, NY, USA). The normality of mDISCERN and GQS score distributions was assessed using the Shapiro-Wilk test. Descriptive statistics were presented as mean ± standard deviation and minimum-maximum values. Because the score distributions did not meet normality assumptions and the GQS represents an ordinal rating scale, nonparametric methods were used. The Kruskal-Wallis test was applied to compare mDISCERN and GQS scores among the three evaluator groups. Spearman’s rank correlation analysis was used to evaluate associations between mDISCERN and GQS scores within each evaluator group. All statistical tests were two-tailed, and a p-value of < 0.05 was considered statistically significant.

RESULTS

The quality and reliability of Grok AI’s responses were evaluated using the mDISCERN and GQS instruments across three participant groups: orthodontists, pediatricians, and parents. As shown in Table 1, the mean mDISCERN scores for the four queries did not differ significantly among the groups. The mean mDISCERN scores ranged from 23.85 to 24.90 in the orthodontist group, from 25.95 to 27.66 in the pediatrician group, and from 25.52 to 27.09 in the parent group. Similarly, GQS scores showed no statistically significant differences among the three groups, with mean values ranging from 3.61 to 4.12. These findings indicate that Grok AI’s responses were rated at comparable levels by orthodontists, pediatricians, and parents, corresponding overall to moderate-to-good levels of reliability and quality.

Table 1. Comparison of Grok AI responses by mDISCERN and GQS scores across groups.

	Parameters	Orthodontists (n = 21)		Pediatricians (n = 24)		Parents (n = 21)		p*
		Mean (SD)	Min-max	Mean (SD)	Min-max	Mean (SD)	Min-max	
Query 1	mDISCERN	23.85 (3.38)	18-32	25.95 (4.17)	21-39	26.19 (5.06)	18-40	0.120
	GQS	3.71 (0.64)	2-5	4.04 (0.62)	3-5	4.09 (0.70)	3-5	0.166
Query 2	mDISCERN	24.90 (3.68)	20-37	27.66 (5.97)	19-40	27.09 (5.44)	18-40	0.175
	GQS	3.85 (0.72)	2-5	4.04 (0.75)	3-5	4.09 (0.83)	3-5	0.626
Query 3	mDISCERN	24.76 (3.78)	15-32	26.54 (5.83)	18-40	26.66 (5.62)	18-40	0.588
	GQS	3.71 (0.84)	2-5	3.95 (0.75)	3-5	4.04 (0.74)	3-5	0.470
Query 4	mDISCERN	24.85 (3.81)	18-36	27.33 (5.53)	18-40	25.52 (4.63)	18-40	0.248
	GQS	3.61 (0.80)	2-5	4.12 (0.67)	3-5	4.00 (0.77)	3-5	0.112

*: Results of Kruskal-Wallis test.

GQS: Global Quality Scale, SD: Standard deviation, Max: Maximum, Min: Minimum, mDISCERN: Modified DISCERN.

In the orthodontist group (Table 2), significant positive correlations were observed primarily among scores from the same assessment instrument across different items. Among GQS scores, the strongest correlations were found between Query 2 and Query 3 ($r = 0.855$, $p < 0.001$), Query 2 and Query 4 ($r = 0.821$, $p < 0.001$), and Query 3 and Query 4 ($r = 0.792$, $p < 0.001$). Similarly, mDISCERN scores showed significant positive correlations across different queries, including Query 2 and Query 3 ($r = 0.756$, $p < 0.001$), Query 2 and

Query 4 ($r = 0.747$, $p < 0.001$), and Query 3 and Query 4 ($r = 0.719$, $p < 0.001$). These findings suggest that orthodontists exhibit a consistent scoring pattern when evaluating Grok AI's responses across different dentofacial health topics.

In the pediatrician group (Table 3), significant positive correlations were also observed among GQS scores across different queries. These correlations ranged from $r = 0.616$ to $r = 0.865$, with the strongest association observed between GQS scores for Query 3 and

Table 2. Correlation of GQS and mDISCERN scores in orthodontist (n = 21).

		Query 1 GQS	Query 1 mDISCERN	Query 2 GQS	Query 2 mDISCERN	Query 3 GQS	Query 3 mDISCERN	Query 4 GQS
Query 1 mDISCERN	r	0.053						
	p	0.820						
Query 2 GQS	r	0.551*	0.081					
	p	0.010	0.727					
Query 2 mDISCERN	r	-0.289	0.568*	0.091				
	p	0.205	0.007	0.696				
Query 3 GQS	r	0.414	0.107	0.855*	0.142			
	p	0.062	0.646	< 0.001	0.540			
Query 3 mDISCERN	r	-0.077	0.728*	0.034	0.756*	0.141		
	p	0.739	< 0.001	0.882	< 0.001	0.542		
Query 4 GQS	r	0.634*	0.091	0.821*	0.103	0.792*	-0.054	
	p	0.002	0.696	< 0.001	0.656	< 0.001	0.817	
Query 4 mDISCERN	r	0.064	0.598*	0.330	0.747*	0.383	0.719*	0.372
	p	0.782	0.004	0.144	< 0.001	0.087	< 0.001	0.097

Bold values are statistically significant. *: Correlation is significant at the 0.01 level (2-tailed).

GQS: Global Quality Scale, mDISCERN: Modified DISCERN.

Table 3. Correlation of GQS and mDISCERN scores in pediatricians (n = 24).

		Query 1 GQS	Query 1 mDISCERN	Query 2 GQS	Query 2 mDISCERN	Query 3 GQS	Query 3 mDISCERN	Query 4 GQS
Query 1 mDISCERN	r	0.122						
	p	0.571						
Query 2 GQS	r	0.657**	0.394					
	p	< 0.001	0.057					
Query 2 mDISCERN	r	0.165	0.835**	0.588**				
	p	0.441	< 0.001	0.003				
Query 3 GQS	r	0.653**	0.264	0.698**	0.417*			
	p	0.001	0.213	< 0.001	0.043			
Query 3 mDISCERN	r	0.297	0.797**	0.521**	0.792**	0.529**		
	p	0.158	< 0.001	0.009	< 0.001	0.008		
Query 4 GQS	r	0.629**	0.202	0.616**	0.399	0.865**	0.415*	
	p	0.001	0.343	0.001	0.053	< 0.001	0.044	
Query 4 mDISCERN	r	0.149	0.862**	0.446*	0.867**	0.373	0.843**	0.357
	p	0.488	< 0.001	0.029	< 0.001	0.073	< 0.001	0.087

*: Correlation is significant at the 0.05 level (2-tailed). Bold values are statistically significant. **: Correlation is significant at the 0.01 level (2-tailed).

GQS: Global Quality Scale, mDISCERN: Modified DISCERN.

Query 4 ($r = 0.865$, $p < 0.001$). Strong positive correlations were also detected among mDISCERN scores across different queries, ranging from $r = 0.792$ to $r = 0.867$. In addition, significant correlations were found between GQS and mDISCERN scores for Query 2 ($r = 0.588$, $p = 0.003$) and for Query 3 ($r = 0.529$, $p = 0.008$). These findings indicate that pediatricians demonstrate a consistent evaluation pattern, particularly in terms of response quality and reliability across the four queries.

Among parents (Table 4), GQS scores across different queries showed very strong positive correlations. The strongest correlations were detected between Query 2 and Query 4 ($r = 0.928$, $p < 0.001$), Query 2 and Query 3 ($r = 0.884$, $p < 0.001$), and Query 3 and Query 4 ($r = 0.871$, $p < 0.001$). Similarly, mDISCERN scores showed strong positive correlations across different queries, including Query 2 and Query 3 ($r = 0.892$, $p < 0.001$); Query 3 and Query 4 ($r = 0.868$, $p < 0.001$); Query 1 and Query 4 ($r = 0.852$, $p < 0.001$); Query 1 and Query 3 ($r = 0.831$, $p < 0.001$); Query 2 and Query 4 ($r = 0.780$, $p < 0.001$); and Query 1 and Query 2 ($r = 0.677$, $p = 0.001$). A significant positive correlation was also found between Query 3 GQS and Query 3 mDISCERN scores ($r = 0.447$, $p = 0.042$). These results suggest that parents tended to evaluate the quality and reliability of Grok AI's responses consistently across different questions.

Grok AI's responses were rated similarly by all three evaluator groups, with no significant between-group differences in mDISCERN or GQS scores. Correlation analyses revealed significant positive associations among several GQS and mDISCERN parameters, particularly across different queries evaluated by the same evaluator groups. These findings suggest consistency in evaluators' scoring patterns, particularly among parents and pediatricians.

DISCUSSION

In this study, the reliability and quality of responses generated by Grok AI, an AI-based conversational agent, were evaluated in

the context of pediatric dentofacial health, based on assessments by orthodontists, pediatricians, and parents. The findings demonstrated that both mDISCERN and GQS scores exhibited overall moderate-to-good reliability and quality. Mean mDISCERN scores ranged between 23 and 27, while GQS scores ranged from 3.6 to 4.1, with no statistically significant differences observed among the three groups. These results suggest that Grok AI's responses were perceived as broadly comparable by both professional and non-professional evaluators.

Correlation analyses provided additional insight into evaluator scoring patterns. Significant positive correlations were observed among several GQS and mDISCERN parameters, particularly across different queries evaluated by the same evaluator groups. In the orthodontist group, correlations were primarily observed within GQS scores and within mDISCERN scores across different queries. In the pediatrician group, within-scale correlations and some GQS–mDISCERN correlations were significant. Among parents, very strong positive correlations were observed across GQS scores and mDISCERN scores for different queries. These findings suggest that participants consistently evaluated Grok AI's responses across dentofacial health topics. However, these results should be interpreted as evidence of consistent scoring patterns rather than complete agreement between the two instruments.

Similar findings have been reported in previous studies. Kurt Demirsoy et al. (9) examined ChatGPT's responses in the field of orthodontics and reported that, although the content was accessible, limitations remained in terms of factual accuracy. Similarly, Terzi et al. (10) evaluated AI-generated responses regarding dental implants and peri-implant phenotypes and concluded that, while the information may be useful for patient education, it lacked sufficient depth to support clinical decision-making. In line with these reports, our findings suggest that Grok AI may serve as a supportive educational resource to provide general information for parents. However, it

Table 4. Correlation of GQS and mDISCERN scores in parents (n = 21).

		Query 1 GQS	Query 1 mDISCERN	Query 2 GQS	Query 2 mDISCERN	Query 3 GQS	Query 3 mDISCERN	Query 4 GQS
Query 1 mDISCERN	r	0.331						
	p	0.142						
Query 2 GQS	r	0.747**	0.405					
	p	< 0.001	0.069					
Query 2 mDISCERN	r	0.211	0.677**	0.113				
	p	0.358	0.001	0.627				
Query 3 GQS	r	0.750**	0.424	0.884**	0.255			
	p	< 0.001	0.055	< 0.001	0.265			
Query 3 mDISCERN	r	0.308	0.831**	0.347	0.892**	0.447*		
	p	0.174	< 0.001	0.124	< 0.001	0.042		
Query 4 GQS	r	0.829**	0.325	0.928**	0.078	0.871**	0.263	
	p	< 0.001	0.151	< 0.001	0.735	< 0.001	0.249	
Query 4 mDISCERN	r	0.259	0.852**	0.199	0.780**	0.342	0.868**	0.189
	p	0.256	< 0.001	0.386	< 0.001	0.129	< 0.001	0.412

*: Correlation is significant at the 0.05 level (2-tailed). Bold values are statistically significant. **: Correlation is significant at the 0.01 level (2-tailed). GQS: Global Quality Scale, mDISCERN: Modified DISCERN.

should not be considered a substitute for professional pediatric or orthodontic assessment.

The general medical literature also supports similar conclusions. Liu et al. (16) demonstrated that, although ChatGPT offers practical benefits in generating medical content, issues of accuracy and currency remain. Giannakopoulos et al. (17) reported that large language models provide only limited reliability in the context of evidence-based dentistry. While Hwang et al. (18) highlighted the contribution of deep learning applications in dentistry, Yu et al. (19) emphasized that, despite the growing role of AI in clinical decision support systems, concerns regarding reliability and bias persist. Likewise, Topol (20) highlighted the transformative potential of AI in healthcare delivery but underscored the need to address ethical and transparency-related challenges. Taken together with our findings in pediatric dentofacial health, these observations support a cautious interpretation of AI-based conversational agents as adjunctive educational tools rather than as independent clinical decision-making resources.

One of the factors to consider when interpreting information generated by large language models is "hallucination." These models can occasionally generate incorrect or unsubstantiated information in a convincing manner; this can pose a problem, particularly when answering questions related to medicine or dentistry (21,22). Users may not always recognize such errors, especially when the provided response appears clear and authoritative. Although our study did not specifically evaluate hallucinations, the variability in reliability scores indicates that health information generated by AI should be treated with caution and, when necessary, verified by healthcare professionals.

Pediatrics is an important field for the evaluation of digital advisors, particularly because parental education and child health literacy may influence healthcare-seeking behavior and decision-making (23). Kubba and Foran (23), as well as Walsh et al. (24), demonstrated that parents frequently use online health information, which can significantly influence their decision-making processes. Baumann et al. (25) reported that parents seek child health information through multiple sources, including digital media, print media, and personal contacts, and that this behavior may vary according to whether the concern involves general child health and development or acute illness. Similarly, Min et al. (26), in a pediatric medical chatbot project developed in Singapore, emphasized that such tools may provide accurate, empathetic, and accessible information in urgent care contexts, although improvements are needed to better accommodate health literacy. DeWitt et al. (27) reviewed mobile health applications and noted that, although content was generally user-friendly, enhancements in readability and structure were required. Likewise, Klapow et al. (28) found that chatbot initiatives targeting parents were valuable for educational purposes but suggested that improvements in literacy adaptation and accuracy could further enhance user experience. Collectively, these findings underscore the growing role of digital tools in pediatric counseling. By directly including parents as evaluators, this study adds a parental perspective to assessing AI-generated pediatric dentofacial health information, although further research is needed to confirm and expand these findings.

Study Limitations

This study has several limitations. First, because the analysis was based on a limited number of queries, the findings may not reflect Grok AI's performance across the full spectrum of pediatric dentofacial health topics. Second, only one AI-based conversational agent was evaluated, and the lack of transparency regarding Grok AI's training data and update frequency may have influenced the accuracy and reproducibility of its responses. Third, the evaluations were subjective, and participants' varying levels of knowledge and experience may have affected the scoring. Moreover, inter-rater reliability was not assessed, which may limit interpretation of consistency across evaluators. The single-center design of the study may also limit the generalizability of the findings. Although the mDISCERN and GQS instruments are commonly used to assess the reliability and quality of online health information, both remain partially dependent on evaluators' interpretation. As the study was cross-sectional and based on hypothetical queries, it may not fully capture real-life interactions between parents, clinicians, and AI-based tools. In addition, each query was evaluated using a single generated response, and variations in Grok AI's responses across repeated sessions were not assessed. The evolving nature of large language models also means that responses may change over time, which could further limit reproducibility.

Future research should include comparative analyses of different AI models, larger and more diverse international samples, and direct assessments of patient experiences. In particular, the role of AI-based tools in parental education for pediatric populations warrants further investigation. Such efforts may clarify the potential contribution of AI-supported counseling to preventive care in addressing dentofacial health issues during childhood.

CONCLUSION

This study demonstrated that Grok AI provided responses with moderate-to-good reliability in the context of pediatric dentofacial health, as evaluated by orthodontists, pediatricians, and parents. These findings suggest that AI-based conversational agents may serve as supportive tools in parental education and preventive pediatrics. However, they remain insufficient for clinical decision-making, and further validation in larger and more diverse settings is warranted.

Ethics

Ethics Committee Approval: This study was approved by the Ordu University Non-Interventional Scientific Research Ethics Committee (approval no: 2025/246, date: 18.07.2025).

Informed Consent: Electronic informed consent was obtained from all participants before participation in the online survey.

Footnotes

Authorship Contributions

Concept: C.Y.G., S.K.B., Design: C.Y.G., S.K.B., Data Collection or Processing: C.Y.G., S.K.B., H.A., Analysis or Interpretation: C.Y.G., S.K.B., Literature Search: C.Y.G., S.K.B., H.A., Writing: C.Y.G., S.K.B., H.A.

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

Financial Disclosure: The authors declared that this study received no financial support.

REFERENCES

1. Abu Arqub S, Al-Moghrabi D, Allareddy V, Upadhyay M, Vaid N, Yadav S. Content analysis of AI-generated (ChatGPT) responses concerning orthodontic clear aligners. *Angle Orthod.* 2024; 94: 263-72.
2. Acar AH. Can natural language processing serve as a consultant in oral surgery? *J Stomatol Oral Maxillofac Surg.* 2024; 125: 101724.
3. Kılınç DD, Mansız D. Examination of the reliability and readability of Chatbot Generative Pretrained Transformer's (ChatGPT) responses to questions about orthodontics and the evolution of these responses in an updated version. *Am J Orthod Dentofacial Orthop.* 2024; 165: 546-55.
4. 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.
5. Kim J, Cai ZR, Chen ML, Simard JF, Linos E. Assessing biases in medical decisions via clinician and AI chatbot responses to patient vignettes. *JAMA Netw Open.* 2023; 6: e2338050.
6. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of artificial intelligence Chatbot responses to top searched queries about cancer. *JAMA Oncol.* 2023; 9: 1437-40.
7. Sriwastwa A, Ravi P, Emmert A, Chokshi S, Kondor S, Dhal K, et al. Generative AI for medical 3D printing: a comparison of ChatGPT outputs to reference standard education. *3D Print Med.* 2023; 9: 21.
8. Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol.* 2023; 29: 721-32.
9. Kurt Demirsoy K, Buyuk SK, Bicer T. How reliable is the artificial intelligence product large language model ChatGPT in orthodontics? *Angle Orthod.* 2024; 94: 602-7.
10. Terzi M, Yavuz MC, Bicer T, Buyuk SK. Evaluation of artificial intelligence robot's knowledge and reliability on dental implants and peri-implant phenotype. *Sci Rep.* 2025; 15: 9519.
11. Dave M, Tattar R, Alafaleg R, Barry S, Ariyaratnam S, Roudsari RV, et al. Performance of large language models (ChatGPT4-0, Grok2 and Gemini) in UK dentistry and dental hygiene and therapy assessments. *Br Dent J.* 2025.
12. Giannini L, Galbiati G, Cressoni P, Esposito L. Bad oral habits: a review of the literature. *J Biol Regul Homeost Agents.* 2021; 35: 403-6.
13. Hood CA, Hunter ML, Kingdon A. Demographic characteristics, oral health knowledge and practices of mothers of children aged 5 years and under referred for extraction of teeth under general anaesthesia. *Int J Paediatr Dent.* 1998; 8: 131-6.
14. Meade MJ, Dreyer CW. Web-based information on orthodontic clear aligners: a qualitative and readability assessment. *Aust Dent J.* 2020; 65: 225-32.
15. Onder CE, Koc G, Gokbulut P, Taskaldiran I, Kuskonmaz SM. Evaluation of the reliability and readability of ChatGPT-4 responses regarding hypothyroidism during pregnancy. *Sci Rep.* 2024; 14: 243.
16. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. *J Med Internet Res.* 2023; 25: e48568.
17. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing Chat in supporting evidence-based dentistry: comparative mixed methods study. *J Med Internet Res.* 2023; 25: e51580.
18. Hwang JJ, Jung YH, Cho BH, Heo MS. An overview of deep learning in the field of dentistry. *Imaging Sci Dent.* 2019; 49: 1-7.
19. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng.* 2018; 2: 719-31.
20. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019; 25: 44-56.
21. Gravel J, D'Amours-Gravel M, Osmanliu E. Learning to fake it: limited responses and fabricated references provided by ChatGPT for medical questions. *Mayo Clin Proc Digit Health.* 2023; 1: 226-34.
22. Templin T, Perez MW, Sylvia S, Leek J, Sinnott-Armstrong N. Addressing 6 challenges in generative AI for digital health: a scoping review. *PLOS Digit Health.* 2024; 3: e0000503.
23. Kubbe C, Foran HM. Online health information seeking by parents for their children: systematic review and agenda for further research. *J Med Internet Res.* 2020; 22: e19985.
24. Walsh AM, Hyde MK, Hamilton K, White KM. Predictive modelling: parents' decision making to use online child health information to increase their understanding and/or diagnose or treat their child's health. *BMC Med Inform Decis Mak.* 2012; 12: 144.
25. Baumann I, Jaks R, Robin D, Juvalta S, Dratva J. Parents' health information seeking behaviour - does the child's health status play a role? *BMC Fam Pract.* 2020; 21: 266.
26. Min C, Lim RX, Tan SW, Ganapathy S. Experience developing a pediatric medical chatbot in Singapore: a digital innovation for improved emergency care. *Front Digit Health.* 2025; 7: 1557804.
27. DeWitt A, Kientz J, Liljenquist K. Quality of mobile apps for child development support: search in app stores and content analysis. *JMIR Pediatr Parent.* 2022; 5: e38793.
28. Klapow MC, Rosenblatt A, Lachman J, Gardner F. The feasibility and acceptability of using a digital conversational agent(chatbot)for delivering parenting interventions: systematic review. *JMIR Pediatr Parent.* 2024; 7: e55726.