Scientific Reports

https://doi.org/10.1038/s41598-025-28857-y

Article in Press

Exploring artificial intelligence chatbots in pediatric fluoride education: a cross-sectional study

Received: 27 June 2025

Accepted: 13 November 2025

Published online: 29 November 2025

Cite this article as: Karamüftüoğlu N., Varol E.A. & Bal C. Exploring artificial intelligence chatbots in pediatric fluoride education: a cross-sectional study. *Sci Rep* (2025). https://doi.org/10.1038/s41598-025-28857-y

Nevra Karamüftüoğlu, Ezgi Aydın Varol & Cenkhan Bal

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

ARTICLE IN PRESS

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

ARTICLE IN PRESS

EXPLORING ARTIFICIAL INTELLIGENCE CHATBOTS IN PEDIATRIC FLUORIDE EDUCATION: A CROSS-SECTIONAL STUDY

Nevra Karamüftüoğlu, DDS, PhD
Department of Pediatric Dentistry,
Gülhane Faculty of Dentistry, University of Health Sciences,
Ankara, Türkiye
Email: nvrserbest@hotmail.com
ORCID: 0000-0002-5045-5551

Ezgi Aydın Varol, DDS, PhD

Independent Researcher
Ankara, Türkiye

Cenkhan Bal, DDS, PhD

Department of Pediatric Dentistry,
Gülhane Faculty of Dentistry, University of Health Sciences,
Ankara, Türkiye

Corresponding Author

Nevra Karamüftüoğlu
Department of Pediatric Dentistry
University of Health Sciences – Gülhane Faculty of Dentistry
06010, Ankara, Türkiye
Email: nvrserbest@hotmail.com
Phone: +90 545 220 05 35

ARTICLE IN PRESS

ARTICLE IN PRESS

ABSTRACT

Background:

Large language model–based (LLM) chatbots are increasingly integrated into healthcare communication, offering accessible and interactive information. These artificial intelligence (AI) tools have the potential to influence caregiver health behaviors when tailored to user needs and literacy levels. In pediatric dentistry, fluoride remains a cornerstone of caries prevention but is also subject to public concerns and online misinformation, underscoring the need for reliable digital communication.

Objective:

This observational and exploratory study evaluated the performance of three advanced AI chatbots—ChatGPT-4.o, Google Gemini Pro, and DeepSeek V3—in providing fluoride-related information to parents and caregivers in the context of pediatric oral health.

Methods:

Twenty fluoride-related questions, derived from American Academy of Pediatric Dentistry (AAPD) guideline themes, were presented to each chatbot in standardized sessions. Responses were independently evaluated by three blinded reviewers using validated tools: EQIP, DISCERN, Global Quality Scale (GQS), Flesch Reading Ease Score (FRES), Flesch-Kincaid Reading Grade Level (FKRGL), and iThenticate similarity index. These instruments assessed quality, reliability, readability, and originality. Inter-rater reliability was confirmed with intraclass correlation coefficients (ICCs). Statistical analyses were conducted using ANOVA or Kruskal–Wallis tests with appropriate post-hoc methods.

Results:

ChatGPT-4.o achieved significantly higher EQIP (M = 4.32, SD = 0.43) and DISCERN (M = 4.20, SD = 0.48) scores than Gemini Pro and DeepSeek V3 ($p < 0.001$), indicating superior reliability and informational quality. While FRES (median = 68.5, $p = 0.12$) and Similarity Index (≤10%, $p = 0.54$) showed no significant differences, ChatGPT consistently produced more

ARTICLE IN PRESS

readable and original content. FKRGL differences were borderline (p = 0.041) but not retained after correction, and GQS outcomes were comparable. These findings suggest that ChatGPT's superior performance is not only statistically significant but also practically relevant for enhancing parental comprehension of fluoride use.

Conclusion:
Among the evaluated models, ChatGPT-4.o demonstrated the clearest and most reliable fluoride communication. Its higher EQIP and DISCERN scores highlight its potential as a supportive tool for caregiver education in pediatric dentistry. Nonetheless, these systems should be implemented cautiously, complemented with professional oversight, and continuously validated to prevent misinformation and ensure safe clinical integration.

Keywords: Artificial Intelligence; Chatbots; Fluoride; Pediatric Dentistry

ARTICLE IN PRESS

1. INTRODUCTION

Artificial intelligence (AI) and large language models (LLMs) such as ChatGPT, Gemini, and DeepSeek are increasingly being integrated into healthcare communication, offering real-time, interactive, and accessible information delivery[1,2]. Recent studies suggest that AI has the capability to influence and even modify health-related behaviors, particularly when content is tailored to user needs and literacy levels[3,4].

Despite these opportunities, pediatric dentistry faces unique challenges in digital health communication. Fluoride, one of the most effective preventive agents against dental caries, has become subject to public misperceptions and online misinformation. While its caries-preventive benefits are well documented[5], concerns about toxicity, neurodevelopmental risks, and systemic health effects have gained visibility in digital environments, often leading to parental hesitancy[6].

ARTICLE IN PRESS

These misconceptions underscore the importance of reliable, accessible communication tools for caregivers.

This study is the first to comparatively evaluate three advanced AI chatbots—ChatGPT-4.o (OpenAI, San Francisco, USA), Google Gemini Pro (Google DeepMind), and DeepSeek V3 (Hangzhou DeepSeek AI Co., Ltd.)—in delivering fluoride-related information. The selection of these models was based on their widespread availability, frequent use in health-related queries, and representativeness of current state-of-the-art LLMs.

Fluoride-related questions were developed by reviewing the most recent American Academy of Pediatric Dentistry (AAPD) guideline on fluoride[7]. Additionally, the 2024–2025 AAPD Reference Manual provides the most current policy framework on pediatric oral health recommendations, forming the foundation for question development[8]. Guideline themes were adapted into caregiver-focused questions to provide a standardized foundation for evaluation. The absence of direct parental input represents a limitation that future studies should address.

The aim of this study was to assess the reliability, quality, readability, and originality of fluoride-related responses generated by three AI chatbots in the context of pediatric dentistry. By doing so, this work addresses a critical gap in the literature, as no prior comparative evaluation has systematically examined multiple advanced AI models for caregiver education in pediatric oral health, nor explored their potential in clarifying fluoride-related misconceptions.

2. MATERIALS AND METHODS

2.1. Development and Presentation of Fluoride-Related Questions

This study followed a multi-step methodology to formulate, distribute, and analyze fluoride-related queries presented to the selected chatbot models (Figure 1). Questions regarding fluoride were developed based on the most recent guideline on fluoride published by AAPD[7]. Key preventive themes

ARTICLE IN PRESS

were identified and reformulated into caregiver-focused questions. Specifically, practitioner-oriented recommendations were rephrased into plain-language items framed around common parental concerns and decision-making scenarios**.** While this expert-driven process provided standardization, it must be noted that direct parental input was not included, which represents a limitation of the study. Future research should integrate parental perspectives in question development.

All questions were presented to each AI model in independent conversation sessions (Figure 2). An initial pool of items was pilot-tested on 10 questions to ensure clarity and relevance before full-scale evaluation. All finalized questions (n = 20) were presented identically across platforms. The chatbots were accessed between March 15–20, 2025, in independent conversation windows. To maintain standardization, only the first response was analyzed, and no follow-up prompts were allowed.

ARTICLE IN PRESS

2.2. Evaluation Criteria

The chatbot-generated responses were analyzed using several validated instruments, including EQIP, DISCERN, GQS, FRES, FKRGL, and the Similarity Index. EQIP measures completeness and structural clarity, whereas DISCERN evaluates reliability and balance of information, thereby providing complementary perspectives on content quality. GQS, originally developed for educational video content, was adapted to text evaluation in this study. It rates the overall quality and usefulness of content on a 5-point scale (1 = poor, 5 = excellent), reflecting global comprehensibility, consistency, and benefit for caregivers. Scoring ranges for all tools are summarized in Table 1, and the formulae for the readability indices.

All responses were independently evaluated by three blinded reviewers. Inter-rater reliability was verified using intraclass correlation coefficients (ICCs), with both intra- and inter-rater values exceeding 0.70, indicating acceptable agreement. For the quality scales, a score of 1 reflected low-quality content with insufficient information and minimal value to users,

ARTICLE IN PRESS

while a score of 5 denoted highly consistent, comprehensive, and clinically valuable content (Figure 3). In addition, FRES and FKRGL were used to assess the readability and comprehensibility of the chatbot responses.

2.3. Statistical Analysis

All analyses were performed using IBM SPSS Statistics, version 25.0 (IBM Corp., Armonk, NY, USA), with a significance level set at $p < 0.05$.

Normality was tested with Shapiro–Wilk and homogeneity with Levene's test. One-way ANOVA was used for normally distributed data, with Tukey's HSD post-hoc test. For non-parametric distributions, the Kruskal–Wallis test was applied, followed by Bonferroni-adjusted Mann–Whitney U tests. ANOVA was chosen for continuous parametric data (e.g., EQIP, FRES), while Kruskal–Wallis was used for ordinal or skewed distributions (e.g., DISCERN, GQS, Similarity Index). Post-hoc tests were applied to control for Type I error. Table 2 presents the tests applied to each evaluation metric.

3. RESULTS

3.1. Descriptive Analysis

Descriptive statistics for the chatbot responses were examined across two domains: general knowledge questions and clinical application questions. The analyses included mean, standard deviation (SD), median, and interquartile ranges (Q1–Q3). ChatGPT-4.o consistently achieved the highest mean values in both domains, outperforming Gemini Pro and DeepSeek V3. For general questions, ChatGPT-4.o obtained the highest score ($M = 4.32$, $SD = 0.43$), followed by Gemini Pro ($M = 3.74$, $SD = 0.66$) and DeepSeek V3 ($M = 3.48$, $SD = 0.57$). A similar trend was observed for clinical questions, with ChatGPT-4.o again achieving the top performance ($M = 4.20$, $SD = 0.48$), while Gemini Pro and DeepSeek produced comparatively lower scores.

3.2. Comparative Evaluation Across Quality Metrics

ARTICLE IN PRESS

Comparative results are illustrated in Figures 4–9, showing EQIP, DISCERN, FRES, FKRGL, and Similarity Index distributions. ChatGPT-4.o significantly outperformed Gemini and DeepSeek in EQIP (M = 4.32, SD = 0.43) and DISCERN (M = 4.20, SD = 0.48) scores ($p < 0.001$), reflecting higher reliability and informational quality. GQS outcomes were comparable across the three models, with no statistically significant differences, indicating that while overall global quality was adequate, it did not distinguish performance between systems.

No significant differences were observed for FRES ($p = 0.12$) or Similarity Index ($p = 0.54$). However, ChatGPT demonstrated higher median readability (FRES = 68.5) and lower similarity values (≤10%), suggesting clearer and more original responses. FKRGL showed borderline significance ($p = 0.041$), but this did not remain significant after Bonferroni correction.

From a practical perspective, these findings indicate that ChatGPT's superior performance may enhance caregivers' understanding of fluoride use and support informed decision-making in pediatric dentistry.

ARTICLE IN PRESS

3.3. Summary of Pairwise Differences

Post-hoc analyses confirmed ChatGPT's superiority over Gemini and DeepSeek in EQIP and DISCERN ($p < 0.05$). No significant differences were found between Gemini and DeepSeek. The implication is that parents using ChatGPT are more likely to receive accurate, reliable, and comprehensible fluoride education than with the other platforms.

3.4. Graphical Representation

Figure 4 and Figure 5 illustrate the distribution of chatbot performance across general and clinical domains. Figure 6 illustrates the distribution of EQIP (Ensuring Quality Information for Patients) scores across the three chatbot models—ChatGPT, DeepSeek, and Gemini. The box plot displays the median values, interquartile ranges, and potential outliers for each

ARTICLE IN PRESS

model, allowing a comparative evaluation of the quality of patient-directed health information provided by these AI systems. Figure 7 demonstrates the comparative distribution of DISCERN scores among ChatGPT, DeepSeek, and Gemini. The results indicate that ChatGPT achieved substantially higher median and overall scores, suggesting superior reliability and quality of health information related to treatment choices. In contrast, DeepSeek and Gemini performed similarly, but with lower scores and narrower ranges. Figure 8 presents the Flesch-Kincaid Reading Grade Level (FKRGL) scores of the chatbot-generated responses. Among the models, DeepSeek and Gemini exhibited higher median FKRGL scores, indicating that their outputs were written at a more advanced reading level. ChatGPT's responses, on the other hand, were more accessible, requiring a lower grade-level comprehension. While DeepSeek showed greater score variability, Gemini maintained a narrower and more consistent range. Figure 9 illustrates the Flesch Reading Ease Score (FRES) distributions across the chatbot models. ChatGPT achieved the highest median FRES values, indicating that its responses were the most readable. DeepSeek showed moderate readability scores with a wider range, while Gemini’s lower FRES values suggest that its responses were more difficult to comprehend. These findings imply that ChatGPT may be more suitable for delivering easily understandable health information. The clinical relevance of these evaluation metrics in pediatric dentistry is summarized in Table 3, highlighting how higher EQIP and DISCERN scores translate into improved caregiver comprehension, trust, and adherence to preventive guidance.

ARTICLE IN PRESS

4. DISCUSSION

This study provides a focused, comparative evaluation of three advanced AI-driven chatbots—ChatGPT-4.o, Google Gemini Pro, and DeepSeek V3—for communicating fluoride-related information to caregivers in pediatric dentistry. ChatGPT-4.o consistently achieved higher EQIP and DISCERN scores than the comparators, indicating stronger reliability, structure, and informational quality for caregiver-facing content. These findings align

ARTICLE IN PRESS

with recent evidence suggesting that AI-based conversational agents can generate patient-directed materials of acceptable quality when developed and evaluated under transparent, standardized methods and when readability levels are appropriately considered (1–3, 9, 10). Consistent with prior audits, our readability analysis revealed higher median FRES and lower FKRGL scores for ChatGPT-4.o, supporting the premise that readability is a key determinant of comprehension and behavior change in digital health education (9, 10).

These results are in agreement with Li et al. (11), who demonstrated that chatbot interventions can enhance mental well-being and engagement in self-care routines. Synthesizing these studies suggests that AI tools hold considerable promise for shaping positive health behaviors—provided they are accurate, context-sensitive, and guided by evidence-based frameworks. In meta-analyses and systematic reviews, chatbots have demonstrated efficacy in improving lifestyle behaviors such as physical activity, healthy eating, and sleep, supporting the potential for similar benefits in oral-health literacy and preventive behavior when fluoride messages are both accurate and accessible (4, 9).

Fluoride remains a cornerstone of caries prevention in childhood, and authoritative bodies consistently affirm that appropriately used topical and systemic fluoride are safe and effective (5, 7, 12). AAPD provides comprehensive, periodically updated best-practice recommendations that can be embedded into chatbot outputs to reduce misinformation and guide caregivers toward sound decisions (7, 12). Considering the increasing prevalence of fluoride-related misinformation on social media, integrating AAPD-based evidence and clearly signposted professional sources within chatbot interfaces is crucial (13–15).

Quality appraisal frameworks such as DISCERN and EQIP—originally designed for written patient materials—translate well to chatbot outputs because they assess reliability, balance, completeness, and presentation (9, 16, 17). In this study, these instruments were used to benchmark

ARTICLE IN PRESS

information quality across models, while the Global Quality Score (GQS) provided a general usefulness rating. Notably, recent scholarship cautions that the GQS is sometimes inconsistently cited in the literature, emphasizing the importance of accurate methodological reporting (18). Readability indices (FRES/FKRGL) remain pragmatic proxies for accessibility and are widely used in audits of patient-education materials; nevertheless, future work should include user-testing with caregivers to capture subtleties that readability formulas may miss (19).

From a trust and design standpoint, Jin et al. (20) reported that gendered cues and personalization in chatbot design increase acceptance, particularly among female caregivers—a finding relevant to pediatric oral-health promotion. Aligning chatbot responses with plain-language and health-literacy standards can also prevent miscommunication in vulnerable populations (21), and emerging NLP approaches such as neural machine translation models have shown promise for improving linguistic accessibility and health-literacy alignment in patient-education materials (22). Bhattacharya and Pissurlenkar (23) further underscored that chatbots must undergo rigorous validation to avoid perpetuating misinformation, especially on sensitive topics such as fluoride use where public perceptions are polarized. Similarly, Buldur and Sezer (24) found that ChatGPT-4 generated answers comparable in accuracy to official FDA guidance on dental amalgam, though occasional deviations from regulatory phrasing highlighted the importance of expert review. Gugnani et al. (25) also showed that while ChatGPT responses to parental questions were generally coherent and clear, gaps in clinical nuance and preventive advice remained. Together, these findings reinforce that ChatGPT-4 may serve as a supplementary educational tool—but never a substitute—for professional consultation.

Despite the demonstrated potential of AI chatbots in health education, caution is warranted. The main risks involve misinformation due to absent source attribution, outdated content, or algorithmic bias reflecting users' pre-existing beliefs. For example, a caregiver skeptical about fluoride

ARTICLE IN PRESS

safety might receive responses unintentionally reinforcing these concerns. Although not directly assessed here, such effects warrant investigation in future studies as possible risk factors influencing information trust and behavioral intent. The similarity analysis also revealed occasional content overlaps with existing online material, suggesting a need for stronger originality safeguards in future AI models.

Trust, disclosure, and clinical oversight are thus central to the safe deployment of AI tools in pediatric dentistry. Evidence indicates that healthcare users generally view AI-assisted communication positively when human supervision is explicit and limitations are transparently disclosed (2, 26-30). Accordingly, any clinical application of chatbots should: (i) disclose AI involvement and dataset constraints; (ii) offer clear escalation pathways to clinicians; (iii) provide direct links to authoritative resources such as AAPD fluoride guidelines; and (iv) undergo periodic expert validation. These safeguards are particularly vital for topics prone to controversy or misinformation amplification (12–14).

The strengths of this study include its standardized prompting, multi-metric evaluation (quality, readability, originality), and blinded, multi-rater scoring with satisfactory inter-rater agreement. Nonetheless, limitations must be acknowledged. The study focused solely on English-language responses and fluoride-specific queries, without incorporating caregiver-generated questions or real-world multi-turn interactions. The absence of parental input in question development limits ecological validity and should be addressed in subsequent research. Future studies should incorporate multilingual evaluations, longitudinal monitoring, and outcome-based trials assessing whether chatbot interactions influence fluoride acceptance, brushing frequency, or recall adherence.

We also encourage alignment with emerging reporting guidance for chatbot studies in healthcare to enhance transparency and reproducibility (2, 26–28). Overall, our findings support the cautious, supervised use of advanced AI chatbots as adjuncts—rather than replacements—for pediatric

ARTICLE IN PRESS

dental counseling and oral-health literacy initiatives. Embedding AAPD-based clinical content, ensuring readability and transparency, and maintaining expert oversight can help translate the technological advantages of AI chatbots into meaningful parental understanding and safer fluoride-related decision-making.

5. CONCLUSION

This study offers a novel and methodologically robust evaluation of how artificial intelligence chatbots convey fluoride-related information to pediatric patients and caregivers. Among the models assessed, ChatGPT-4.o consistently demonstrated superior performance across multiple metrics of quality, reliability, and readability. Its higher EQIP and DISCERN scores suggest that it is more capable of delivering accurate, evidence-based, and user-friendly health information than Gemini Pro or DeepSeek V3.

However, while ChatGPT-4.o shows potential as a digital health education tool, our findings also highlight the limitations and risks associated with relying on AI-generated content in clinical contexts. Variability in response quality, lack of clinical nuance, and moderate reading complexity underscore the need for cautious implementation and professional validation.

ARTICLE IN PRESS

The central contribution of this study is to show that AI chatbots vary substantially in their ability to communicate fluoride-related information, and only carefully validated systems can meaningfully support caregiver education. Based on these findings, developers should prioritize improved fine-tuning for pediatric vocabulary, inclusion of verified references, and alignment with health literacy standards. Clinicians should adopt AI chatbots as supportive—not substitutive—tools to reinforce preventive guidance, while public health professionals and policymakers may explore their integration into oral health literacy campaigns, particularly in underserved settings.

ARTICLE IN PRESS

Ultimately, AI chatbots may serve as valuable adjuncts to professional guidance in promoting pediatric oral health literacy, but their safe and effective use will depend on transparent validation, user-centered design, and the irreplaceable role of human oversight.

DECLERATIONS

Ethics Approval and Consent to Participate

This study did not involve human participants, personal data, or biological materials. It was based solely on the evaluation of publicly available AI-generated content. Therefore, ethics committee approval or Institutional Review Board (IRB) review was not required.

ARTICLE IN PRESS

Funding Declaration

This research received no external funding.

Human Ethics and Consent to Participate Declarations

Human Ethics and Consent to Participate declarations: not applicable

Author Contributions

All authors contributed significantly to the development of this manuscript.

- **Nevra Karamüftüoğlu**: Conceptualization, methodology, data analysis, writing—original draft preparation.
- **Ezgi Aydın Varol**: Data collection, software implementation, review and editing.
- **Cenkhan Bal**: Evaluation framework design, statistical validation, visualization.

  All authors read and approved the final version of the manuscript.

Competing Interests

ARTICLE IN PRESS

The authors declare that they have no financial or non-financial competing interests related to the content of this manuscript.

Data availibility

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

REFERENCES

1. Subramanian, C. R., Yang, D. A. & Khanna, R. Enhancing health care communication with large language models—the role, challenges, and future directions. *JAMA Netw. Open* **7**, e240347 (2024).
2. Rahman, A. et al. Comparative analysis based on DeepSeek, ChatGPT, and Google Gemini: Features, techniques, performance, future prospects. Preprint at https://arxiv.org/abs/2503.04783 (2025).
3. Aggarwal, A., Tam, C. C., Wu, D., Li, X. & Qiao, S. Artificial intelligence–based chatbots for promoting health behavioral changes: Systematic review. *J. Med. Internet Res.* **25**, e40789 (2023).
4. Younis, H. A. et al. A systematic review and meta-analysis of artificial intelligence tools in medicine and healthcare: applications, considerations, limitations, motivation and challenges. *Diagnostics* **14**, 109 (2024).
5. Weyant, R. J. et al. Topical fluoride for caries prevention: executive summary of the updated clinical recommendations and supporting systematic review. *J. Am. Dent. Assoc.* **144**, 1279–1291 (2013).
6. Samaranayake, L., Porntaveetus, T., Tsoi, J. & Tuygunov, N. Facts and fallacies of the fluoride controversy: a contemporary perspective. *Int. Dent. J.* **75**, 100833 (2025).

ARTICLE IN PRESS

7. American Academy of Pediatric Dentistry. American Academy of Pediatric Dentistry Fluoride Therapy. *The Reference Manual of Pediatric Dentistry; American Academy of Pediatric Dentistry: Chicago, IL, USA*, 288-291(2020).
8. American Academy of Pediatric Dentistry. *American Academy of Pediatric Dentistry Reference Manual*. American Academy of Pediatric Dentistry (2008).
9. Singh, B. et al. Systematic review and meta-analysis of the effectiveness of chatbots on lifestyle behaviours. *npj Digit. Med.* **6**, 118 (2023).
10. Moult, B., Franck, L. S. & Brady, H. Ensuring quality information for patients: development and preliminary validation of a new instrument to improve the quality of written health care information. *Health Expect.* **7**, 165–175 (2004).
11. Li, H., Zhang, R., Lee, Y.-C., Kraut, R. E. & Mohr, D. C. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *npj Digit. Med.* **6**, 236 (2023).
12. American Academy on Pediatric Dentistry Liaison with Other Groups Committee, & American Academy on Pediatric Dentistry Council on Clinical Affairs. Policy on use of fluoride. *Pediatr. Dent*. **30**, 34-35 (2008).
13. Esmaeilzadeh, F., Movahhed, T., Hasani Yaghooti, M. R., Hoseinzadeh, M. & Babazadeh, S. Content analysis of fluoride-related posts on Instagram: prevalence of misinformation. *BMC Oral Health* **24**, 1179 (2024).
14. American Dental Association. Community water fluoridation at optimal levels is safe and effective (press release). https://www.ada.org/about/press-releases/community-water-fluoridation-at-optimal-levels-is-safe-and-effective (2024).
15. Keleş, Z. H. Is fluoride opposition a marketing tool on social media? *Int. Dent. J.* **74**, 93 (2024).

ARTICLE IN PRESS

16. Charnock, D., Shepperd, S., Needham, G. & Gann, R. DISCERN: An instrument for judging the quality of written consumer health information on treatment choices. *J. Epidemiol. Community Health* **53**, 105–111 (1999).
17. Yeung, A. W. K. Evaluation of content quality of online health information by Global Quality Score: A case study of researchers misnaming it and citing secondary sources. *Publications* **13**, 23 (2025).
18. Laymouna, M., Ma, Y., Lessard, D., Schuster, T., Engler, K. & Lebouché, B. Roles, users, benefits, and limitations of chatbots in health care: Rapid review. *J. Med. Internet Res.* **26**, e56930 (2024).
19. Jindal, P. & MacDermid, J. C. Assessing reading levels of health information: uses and limitations of Flesch formula. *Educ. Health* **30**, 84–88 (2017).
20. Jin, E., Ryoo, Y., Kim, W. J. & Song, Y. G. Bridging the health literacy gap through AI chatbot design: The impact of gender and doctor cues on chatbot trust and acceptance. *Internet Res.* **34**, 1299-1329 (2025).
21. Diviani, N., van den Putte, B., Giani, S. & van Weert, J. C. Low health literacy and evaluation of online health information: a systematic review of the literature. *J. Med. Internet Res.* **17**, e112 (2015).
22. Hindelang, M., Sitaru, S. & Zink, A. Transforming health care through chatbots for medical history-taking and future directions: comprehensive systematic review. *JMIR Med. Inform.* **12**, e56628 (2024).
23. Bhattacharya, B. S. & Pissurlenkar, V. S. Assistive chatbots for healthcare: a succinct review. Preprint at https://arxiv.org/abs/2308.04178 (2023).
24. Buldur, M. & Sezer, B. Evaluating the accuracy of Chat Generative Pre-trained Transformer version 4 (ChatGPT-4) responses to United States Food and Drug Administration (FDA)

ARTICLE IN PRESS

ARTICLE IN PRESS

frequently asked questions about dental amalgam. *BMC Oral Health* **24**, 605 (2024).

25. Gugnani, N., Pandit, I. K., Gupta, M., Gugnani, S. & Kathuria, S. Parental concerns about oral health of children: Is ChatGPT helpful in finding appropriate answers? *J. Indian Soc. Pedod. Prev. Dent.* **42**, 104–111 (2024).
26. Pupong, K., Hunsrisakhun, J., Pithpornchaiyakul, S. & Naorungroj, S. Development of chatbot-based oral health care for young children and evaluation of its effectiveness, usability, and acceptability: mixed methods study. *JMIR Pediatr. Parent.* **5**, e62738 (2025).
27. Or, A., Sukumar, S., Ma, A., Ang, D., Liu, M., Ritchie, H. E. & Sarrafpour, B. Enhancing dental students' history-taking skills with a generative artificial intelligence chatbot. *J. Dent. Educ.* e13952 (2025).
28. Maruska, E. E., Elgreatly, A., Madaio, W., Razoky, K., Bay, C. & Mahrous, A. Comparing dentist and chatbot answers to dental questions for quality and empathy. *JADA Found. Sci.* **4**, 100044 (2025).
29. Collaborative, T. C. Reporting guideline for chatbot health advice studies: The Chatbot Assessment Reporting Tool (CHART) statement. *BMJ Med.* **4**, e001632 (2025).
30. Zhou, Y., Oniani, D., Sreekumar, S., DeAlmeida, R. & Wang, Y. Toward improving health literacy in patient education materials with neural machine translation models. Preprint at https://arxiv.org/abs/2209.06723 (2022).
31. Liu, X. et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *The Lancet Digit. Health* **2**, e537–e548 (2020).
32. Rivera, S. C. et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. *The Lancet Digit. Health* **2**, 549–560 (2020).

ARTICLE IN PRESS

ARTICLE IN PRESS

33. Vasey, B. et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ* **377**, e070904 (2022).
34. Collins, G. S. et al. TRIPOD+ AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e078378 (2024).
35. Tejani, A. S. et al. Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. *Radiol. Artif. Intell.* **6**, e240300 (2024).

ARTICLE IN PRESS

ARTICLE IN PRESS

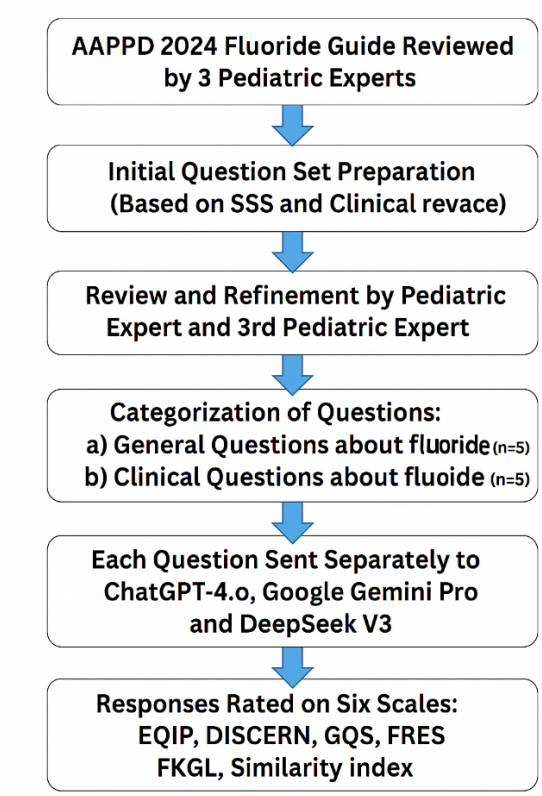


Figure 1. Flowchart of the study

ARTICLE IN PRESS

General Questions	Clinical Questions
• What is the relationship between fluoride and teeth?	• How to apply fluoride treatment on teeth?
• Is fluoride good or bad for children?	• What are the methods of fluoride treatment?
• When can my child use fluoridated toothpaste?	• Until what age is fluoridation applied?
• Are there side effects of fluoride?	• How often is fluoride treatment done?
• Are there any countries where fluoride is banned?	• Is fluoride excreted from the body?

Figure 2. General and clinical questions about flouride

ARTICLE IN PRESS

Score	Global Quality Score (GQS) Description
1	Poor quality, poor flow of video, 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 others poorly discussed, somewhat useful for patients
4	Good quality and generally flow. Most of the revelant information is listed, but some topics not covered, useful for patients
5	Excellent quality and flow, very useful for patients

Figure 3. Global Quality Score (GQS) Description (18)

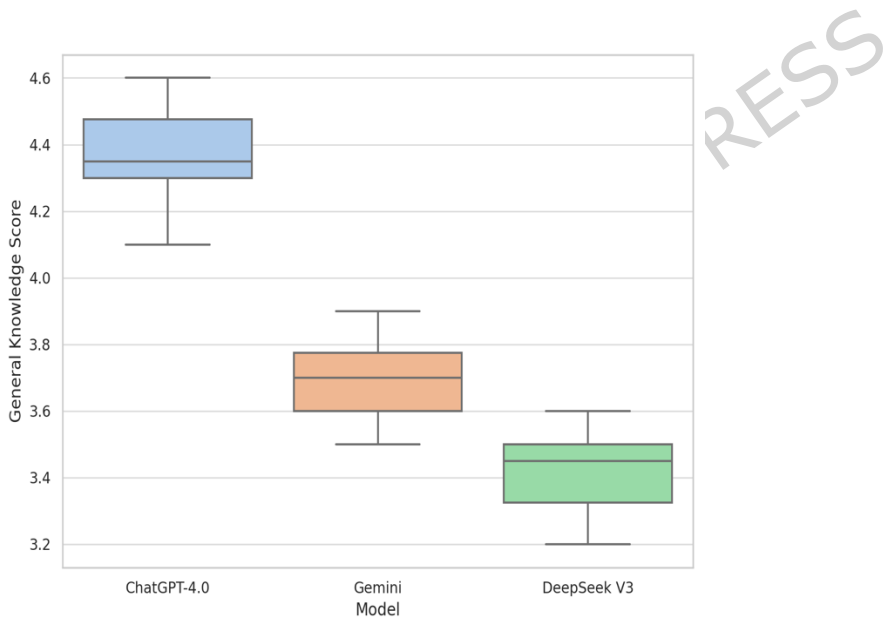


Figure 4. Distribution of general knowledge scores across ChatGPT-4.o, Gemini, and DeepSeek V3

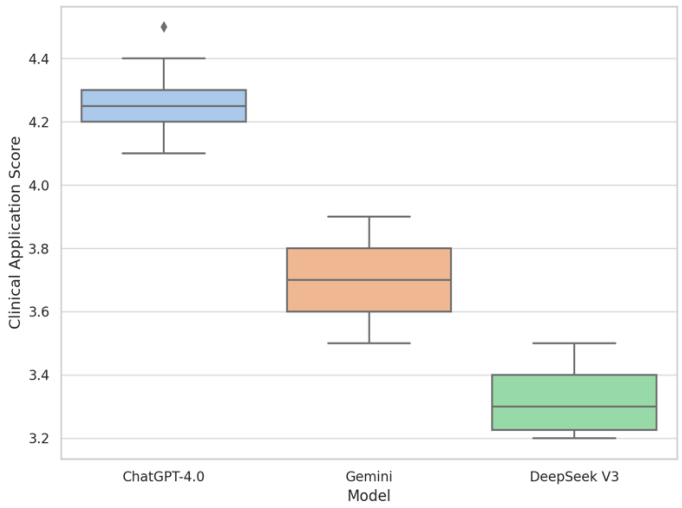


Figure 5. Distribution of clinical application scores across ChatGPT-4.o, Gemini, and DeepSeek V3

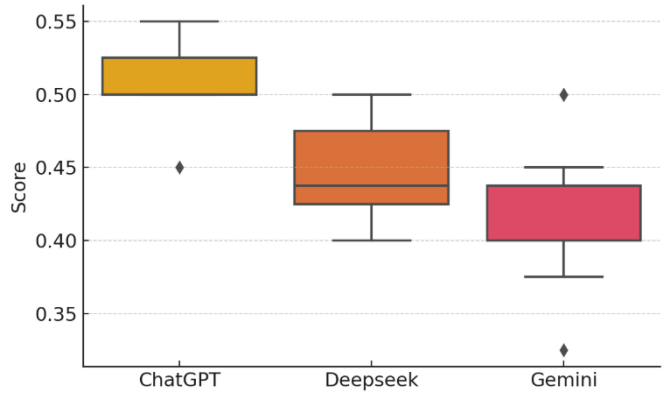


Figure 6. Comparison of EQIP scores among the chatbot models

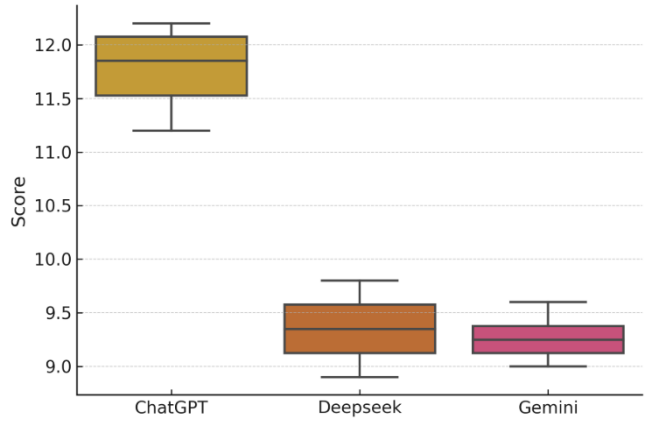


Figure 7. Comparison of DISCERN scores across chatbot models

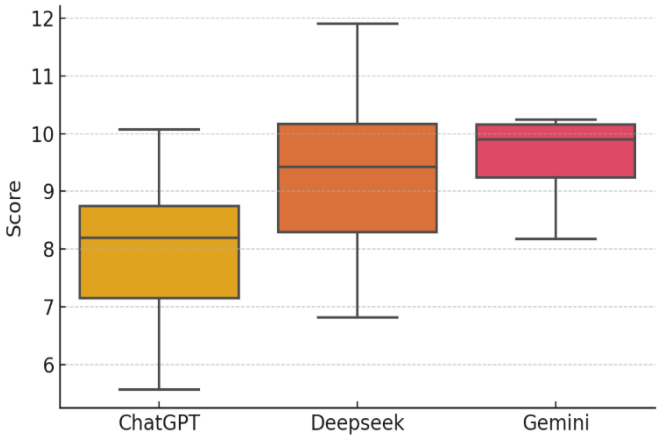


Figure 8. Comparison of FKRGL (Flesch-Kincaid Reading Grade Level) scores across chatbot responses.

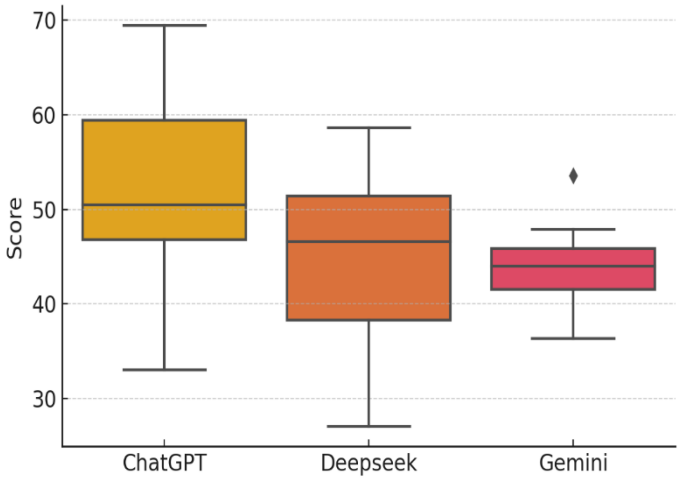


Figure 9. Comparison of FRES (Flesch Reading Ease Score) among chatbot responses.

ARTICLE IN PRESS

Table 1. Scoring ranges and evaluation criteria for tools used in the study

Tool	Purpose	Formula / Method	Scoring Range	Interpretation
EQIP	Completeness and structural clarity of content	Structured checklist	1–5	1 = low-quality, insufficient information; 5 = highly consistent, comprehensive, and valuable
DISCERN	Reliability and balance of health information	Standardized questionnaire	1–5	1 = unreliable, biased content; 5 = reliable, balanced, high-quality content
GQS	Overall quality and flow of information	Global Quality Scale	1–5	1 = poor quality; 5 = excellent quality, useful and easy-to-follow
FRES	Readability (ease of understanding)	**Formula:** 206.835 − 1.015 × (total words ÷ total sentences) − 84.6 × (total syllables ÷ total words)	0–100	90–100 = very easy (5th grade), 60–70 = standard (8th–9th grade), 30–50 = college level, 10–0 = very difficult
FKRGL	Reading grade level required (U.S. system)	**Formula:** 0.39 × (total words ÷ total sentences) + 11.8 × (total syllables ÷ total words) − 15.59	0–18+	0–6 = basic, 6–12 = intermediate, 12–18 = advanced
Similarity Index (iThenticate)	Originality of text; overlap with existing sources	Percentage overlap with databases	0–100%	0–10% = highly original; 10–20% = acceptable; 20–40% = high similarity; 40–100% = very high similarity

Table 2. Statistical tests applied to each evaluation metric

Metric	Primary Test	Post-hoc Test	Purpose
FRES	ANOVA	–	Readability (ease of understanding)
EQIP	ANOVA	Tukey HSD	Completeness and structural clarity
FKRGL	Kruskal–Wallis	Mann–Whitney U with Bonferroni correction	Reading grade level (text complexity)
DISCERN	Kruskal–Wallis	Mann–Whitney U with Bonferroni correction	Reliability and balance of information

ARTICLE IN PRESS

Table 3. Clinical implications of evaluation metrics

Metric	Primary Focus	Clinical Implication in Pediatric Dentistry
EQIP	Completeness and structural clarity of information	High scores indicate that parents receive comprehensive, clear, and easy-to-follow guidance on fluoride use. Low scores may lead to misinterpretations and reduced adherence to preventive recommendations.
DISCERN	Reliability and balance of information	High DISCERN scores reflect trustworthy and unbiased information, which can enhance parental confidence in clinical decision-making and strengthen trust in dental professionals.
GQS	Overall quality and usefulness of content	High GQS values suggest that the information is comprehensive and practically applicable, supporting parents in everyday oral health–related decisions.
FRES	Readability (ease of understanding)	Higher scores indicate that information is easier for parents to comprehend, facilitating more effective communication of fluoride-related health guidance.
FKRGL	Reading level required for comprehension	Lower grade levels ensure accessibility to a wider audience, whereas higher levels may create barriers to understanding among parents with limited health literacy.
Similarity Index	Originality and risk of plagiarism	Low values (<10%) indicate originality and reliability of content. Higher similarity ratios may signal redundancy or reduced credibility of the provided information.

ARTICLE IN PRESS

ARTICLE IN PRESS

ARTICLE IN PRESS