Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Assessing the Quality of AI-Generated Health Information: A Comparative Study of Large Language Models in Patient Education on Dental Veneers

Dilan AYLUÇTARHAN AYÇİÇEK

dilanayluc@hotmail.com

Adıyaman University

Gamze POLAT

Van Yüzüncü Yıl University

Savaş SAĞMAK

Adıyaman University

Research Article

Keywords: Artificial Intelligence, Chat GPT, dentistry, esthetics, Gemini

Posted Date: August 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10533070/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

This study aimed to comparatively evaluate the accuracy, reliability, quality, and readability of responses generated by ChatGPT-5.3 mini, Gemini-3.5 Flash, DeepSeek-V4, and Microsoft Copilot to frequently asked questions concerning laminate veneers.

Methods

A Google search identified 266 laminate veneer FAQs from the first 30 websites; after exclusions, the 20 most frequent questions were analyzed. Responses were independently generated by four AI chatbots (ChatGPT-5.3 mini, Gemini-3.5 Flash, DeepSeek-V4, and Microsoft Copilot) and anonymized for evaluation. Answers were assessed for accuracy using a 5-point Likert scale, reliability via the modified DISCERN instrument, overall quality using the Global Quality Scale (GQS), and readability with the Flesch Reading Ease Score (FRES). Data normality was tested using the Shapiro–Wilk test. As the data were non-normally distributed, differences in accuracy, reliability (DISCERN), quality (GQS), and readability scores among the four chatbot models were analyzed using non-parametric repeated-measures tests. Inter-rater agreement was assessed using Fleiss' kappa coefficient, with statistical significance set at $p < 0.05$.

Results

Significant differences were observed among the chatbots in terms of accuracy (Likert), reliability (modified DISCERN), quality (GQS), and readability (FRES) scores (all $p < 0.05$). ChatGPT-5.3 mini achieved the highest accuracy, reliability, and quality scores, whereas DeepSeek-V4 demonstrated the lowest performance in these domains. For readability, Gemini-3.5 Flash (38.92 ± 9.26) obtained the highest FRES score, while DeepSeek-V4 (25.33 ± 11.67) showed the lowest readability. Gemini-3.5 Flash achieved the highest mean FRES score, indicating that its responses were more readable compared with those generated by the other chatbots.

Conclusion

The performance of AI chatbots in answering frequently asked questions on laminate veneers varied significantly across accuracy, reliability, quality, and readability measures. ChatGPT-5.3 mini outperformed the other models in terms of overall performance, while Gemini-3.5 Flash produced the most readable responses. The present findings indicate that AI chatbots may serve as valuable adjuncts for patient education; however, professional oversight remains is crucial for maintaining the reliability and accuracy of the information presented.

Introduction

One of the most remarkable technological advancements affecting many aspects of modern life today is artificial intelligence (AI) technology [1]. The term “artificial intelligence” was first introduced in 1956 and refers to systems designed to simulate human cognitive processes within computer-based environments [2]. These systems have been developed to carry out tasks that are generally associated with human cognitive abilities, including decision-making, problem-solving, and adapting through accumulated experience [3].

When applied in the fields of medicine and dentistry, artificial intelligence technologies can make substantial contributions to improving diagnostic accuracy and enhancing healthcare services. In contemporary dentistry, AI is utilized for various purposes, including the detection of normal and pathological structures, disease diagnosis, and the prediction of treatment outcomes. In addition, AI applications that are widely employed in dental laboratories are also becoming increasingly integrated into dental education [4].

Chatbots, which represent a subfield of artificial intelligence, are capable of understanding user queries and generating automated responses through the application of natural language processing and machine learning algorithms. Currently, numerous language models with varying characteristics and applications are available [5]. AI-powered chatbots, including ChatGPT and Gemini, DeepSeek, and Microsoft Copilot are being used increasingly widely due to their ability to generate rapid and interactive responses to complex queries [6].

Developed by OpenAI (San Francisco, CA, USA), ChatGPT is defined as an advanced large language model capable of understanding, interpreting, and generating human-like text responses. Owing to its extensive data-processing capacity in line with recent technological advancements, this system can provide large volumes of information across various fields and can be applied to a variety of language-related functions, such as translating text and generating concise summaries and text completion [7,8].

DeepSeek has been developed as a large language model (LLM)-based artificial intelligence system with advanced natural language processing and clinical reasoning capabilities. Built upon a transformer architecture, this system is capable of contextually analyzing user inputs to generate coherent and meaningful text, and has attracted particular attention due to its potential applications in healthcare services and clinical decision-support processes [9].

Gemini is one of the AI-based conversational systems capable of providing up-to-date and highly accurate information through Google’s extensive data infrastructure and advanced search engine integration [10]. Microsoft Copilot, formerly known as Bing Chat, can be utilized in clinical documentation support, information management, and patient counseling services owing to its GPT-4 infrastructure and real-time internet access capabilities [11]. These continuously evolving AI technologies possess considerable potential for enhancing patient education and healthcare communication by providing accessible and reliable health-related information [12]. The increasing adoption of large language

models (LLMs) with different architectural frameworks, access models, and training strategies has further emphasized the importance of comparatively evaluating the performance of these systems [13].

An aesthetically acceptable dentition has been reported to positively influence social interactions, whereas increasing severity of dental irregularities may lead to reduced levels of individual satisfaction [14]. Both direct and indirect restorative approaches can be employed in the management of tooth discolorations and morphological defects. Although direct restorative methods enable aesthetic demands to be met in a shorter time and in a more cost-effective manner, they may present certain disadvantages in long-term use, including reduced color stability and the development of secondary caries. Conventional full-coverage crown restorations, which constitute one of the indirect restorative treatment options, can provide successful outcomes in the management of aesthetic problems; however, they require more extensive removal of healthy tooth structure. Therefore, laminate veneer restorations, which are distinguished by their minimally invasive approach, are considered an important alternative treatment option [15]. Laminate veneers are widely preferred in aesthetic dentistry because of their superior aesthetic properties, long-term clinical success, and resistance to discoloration and wear [16].

The widespread adoption of artificial intelligence technologies has substantially influenced the dissemination of health-related information, including applications in aesthetic dentistry such as laminate veneers [17]. Over the past few years, there has been a considerable an increasing number of studies comparatively evaluating the reliability, accuracy, and quality of responses generated by AI-based systems against information provided by human clinicians [18–21]. Research conducted across various medical and dental specialties has demonstrated that AI-supported chatbots possess the potential to provide health-related information; however, these studies also emphasize that the content generated by such systems should be carefully evaluated in terms of scientific accuracy and clinical reliability. The literature indicates that chatbot-generated responses may exhibit variability in content quality; may occasionally contain inaccurate, incomplete, or outdated information; may lack adequate source attribution; may incorporate algorithmic biases; and may not consistently provide the contextual depth required for clinical decision-making processes [18,19–22]. Although several studies have evaluated the quality of online information related to various dental topics, including periodontal diseases, dental implants, and prosthodontic treatments, to the authors' knowledge, no study to date has specifically focused on laminate veneers [17,23]. This highlights a significant gap in the current literature. AI-generated content has the potential to lead to misinformation and incomplete knowledge transfer regarding critical clinical issues such as material selection, clinical indications, contraindications, and application procedures, thereby contributing to the development of inaccurate patient perceptions. Thus, the present study aimed to analyze and compare the responses generated by four widely used AI-based chatbots—ChatGPT-5.3 mini, Gemini-3.5 Flash, DeepSeek-V4, and Microsoft Copilot—in response to frequently asked questions concerning laminate veneers in terms of accuracy, reliability, quality, and readability.

The present study was based on the null hypothesis that there would be no statistically significant differences among the four evaluated AI-based chatbots—ChatGPT-5.3 mini, Gemini-3.5 Flash, DeepSeek-V4, and Microsoft Copilot—with respect to the accuracy, reliability, quality, and readability of their model outputs to widely asked questions regarding laminate veneers.

Materials And Methods

The present study was carried out in accordance with the ethical standards and principles of the Declaration of Helsinki. As the study was conducted entirely in vitro and did not involve human or animal subjects, ethical approval was not required. A Google search was conducted using the keyword “frequently asked questions about laminate veneer”. The literature indicates that approximately 90 percent of users only browse the first three pages of search engine results [7]. Accordingly, as part of our study, 30 websites were examined and a total of 266 questions were recorded [21,24]. Subsequently, repetitive, similar, irrelevant and brand-related questions were excluded; from the remaining questions, the 20 most frequently asked questions were included in the study by the researchers. When identifying repetitive questions, not only questions phrased in exactly the same way but also those expressed using different words yet reflecting the same information requirement were classified as repetitive questions. (Table 1).

Table 1
Questions submitted to the AI chatbots(ChatGPT-5.3 mini, Gemini,-3.5 Flash, Microsoft Copilot, DeepSeek-V3)

1	**What is a Laminate Veneer?**
2	What are the steps and duration of the laminate veneer procedure?
3	What are the clinical indications for laminate veneer treatment?
4	What are the contraindications for laminate veneer treatment?
5	What are the advantages of laminate veneer treatment?
6	What are the disadvantages of laminate veneer treatment?
7	What are the types of laminate veneers?
8	What are the types of laminate veneers according to their application technology, and what are the main differences between these methods?
9	What are the recommended maintenance and care practices for laminate veneers?
10	What is the anticipated clinical longevity of laminate veneers?
11	Is the laminate veneer procedure associated with pain or discomfort?
12	Is local anesthesia necessary during laminate veneer procedures?
13	Is debonding or detachment a possible complication of laminate veneers?
14	Do laminate veneers provide a natural esthetic appearance?
15	Do laminate veneers undergo color changes over time?
16	Do laminate veneers have any detrimental effects on dental hard tissues?
17	How durable are laminate veneers?
18	How do laminate veneers compare with composite veneers in terms of clinical outcomes and esthetic performance?
19	How do laminate veneers compare with porcelain crowns in terms of clinical indications, esthetic outcomes, and tooth preservation?
20	How do laminate veneers compare with zirconia crowns in terms of clinical performance, esthetic outcomes, and conservative tooth preparation?

The performance of four AI-based chatbots, namely ChatGPT-5.3 mini (https://openai.com/chatgpt), Gemini-3.5 Flash (https://ai.google.dev/gemini-api), DeepSeek-V4 (https://www.deepseek.com), and Microsoft Copilot (https://copilot.microsoft.com), was evaluated. All AI tools were accessed via their publicly available free-tier web interfaces. It has been reported that ChatGPT can provide different and faster responses when the same question is asked again or at different points in time [25]. For this reason, all questions were asked only once. Similarly, Gemini generates three different responses (drafts) for each query [26]. In this study, the first draft response was used as the basis for evaluation.

Microsoft Copilot, on the other hand, has different conversation style modes; in this context, the 'balanced' option—the standard mode—was selected for the analysis. For DeepSeek, the first and only response was included in the evaluation. To minimise potential variability in response generation, queries were posed only once. In line with previous studies, separate user accounts and independent browser sessions were used for each chatbot in order to minimise potential biases and cross-platform effects [27, 28]. Furthermore, prior to data collection, browser data including search history and cookies was cleared to reduce the effects of personalisation and algorithmic memory. To ensure reproducibility and methodological standardization, all four models were accessed between April 10 and April 14, 2026, using the most up-to-date stable versions available at the time of access. No restrictions were imposed on the length or content of the responses. All questions were submitted using a standardized prompt format without any additional instructions, role assignments, or contextual explanations (e.g., directives such as "respond as a dental expert"). This approach was adopted to ensure that the generated responses reflected the default response-generation behavior of each model, independent of prompt engineering effects.

A fourth independent researcher anonymized all chatbot responses to ensure objectivity and assigned randomized letter codes (A–D) to each output; all platform identifiers and any identifying information were excluded from the analysis. The de-identified question–answer sets were subsequently merged into a single PDF document and provided to the evaluators for assessment. All responses were independently assessed by three researchers with expertise in the field, in line with the current literature and clinical practice. This procedure was implemented to maintain complete blinding of the evaluators to chatbot identities throughout the analyses of accuracy, reliability, quality, and readability, thereby minimizing potential assessment bias.

Inter-rater agreement among expert was assessed by the evaluators using Cohen's kappa coefficient, an appropriate statistical measure for multiple raters (kappa values of < 0.20 indicate poor agreement; 0.21–0.40 fair agreement; 0.41–0.60 moderate agreement; 0.61–0.80 good agreement; and 0.81–1.00 very good agreement) [29]. Expert evaluations were based on a primary reference source, such as Esthetic Rehabilitation in Fixed Prosthodontics, along with up-to-date, evidence-based scientific literature.

For accuracy, a five-point Likert scale was employed. According to this scale; Score 1; the chatbot's answers are completely incorrect, Score 2; the chatbot's answers contain more incorrect items than correct items, Score 3; the chatbot's answers contain an equal balance of correct and incorrect items, Score 4; the chatbot's answers contain more correct items than incorrect items and Score 5; the chatbot's answers are completely correct (21). In this context, the chatbot-generated responses were distributed to three restorative dentistry specialists remained blinded to the identity of the chatbot responsible for each response.

Readability was evaluated using the Flesch Reading Ease (FRES) score, a reliable and widely used tool in the literature for this purpose [30]. Readability analysis was conducted using an online readability

assessment tool, Character Calculator (charactercalculator.com), which automatically computes the Flesch Reading Ease Score (FRES) based on syllable, word, and sentence counts. In the FRES scale, scores vary from 0 to 100, where lower scores represent greater reading difficulty [31]. For instance, scores between 0 and 10 represent extremely difficult readability, typically comprehensible only to American university graduates, whereas scores between 90 and 100 indicate very easy readability that can be understood even by elementary school-level readers. The FRE score was calculated using the standard formula: 206.835 − 1.015 × (total words / total sentences) − 84.6 × (total syllables / total words) [31,32].

Three restorative dentistry specialists (D.A.A., G.P., and S.S.) independently evaluated the quality of the generated responses was assessed with the five-point Modified Global Quality Score (GQS) [33]. The scoring criteria were defined as follows: a score of 1 indicated low-quality responses containing incomplete information with no clinical or practical value; a score of 2 represented generally poor-quality responses with limited usefulness; a score of 3 indicated moderate-quality responses with partially sufficient information; a score of 4 indicated good-quality responses containing most of the relevant information; and a score of 5 was considered to represent excellent-quality responses with comprehensive and highly clinically useful information.

The reliability of the information was evaluated using the modified DISCERN instrument. The assessment comprised seven domains: informational accuracy, clinical relevance, completeness, clarity and comprehensibility, structural coherence, absence of misleading content, and scientific reliability. Each item was scored as 1 (No), 2–4 (Partially), or 5 (Yes) [34].

Statistical Analysis

All statistical analyses were conducted using IBM SPSS Statistics (version 26.0; IBM Corp., Armonk, NY, USA). Normality of the data was assessed using the Shapiro–Wilk test. As the data did not satisfy the assumptions of normality and each question was evaluated by all four chatbot models, non-parametric repeated-measures analyses were performed. Differences in accuracy, reliability (DISCERN), quality (GQS), and readability scores among ChatGPT-5.3 mini, Gemini-3.5 Flash, Microsoft Copilot, and DeepSeek-V4 were analyzed using the Friedman test. When statistically significant differences were identified, pairwise comparisons were performed using the Wilcoxon signed-rank test with Bonferroni correction. Inter-rater agreement between the expert evaluators was assessed using Fleiss' kappa coefficient. Statistical significance was defined as a two-sided p-value < 0.05.

Results

Inter-rater reliability among the three independent expert evaluators was assessed using Fleiss’ kappa coefficient, which is appropriate for multiple raters. The level of agreement was found to be fair-to-moderate across all evaluated domains, including accuracy, mDISCERN, and GQS scores. (Table 2).

Table 2
Inter-rater agreement coefficients for the Likert, mDISCERN and GQS scales.

Scale	Fleiss' Kappa	Interpretation
Likert	0.337	Fair / Moderate
mDISCERN	0.639	Fair / Moderate
GQS	0.309	Fair / Moderate

When the analysis results regarding the Likert scale were examined, a statistically significant difference was identified among the chatbots ($p < 0.05$). According to the Likert scale analysis results, ChatGPT-5.3 mini (4.400 ± 0.298) achieved the highest accuracy score, whereas DeepSeek-V4 (4.150 ± 0.333) obtained the lowest score (Table 3).

When the results regarding the modified DISCERN variable were examined, a highly statistically significant difference was found among the chatbots ($p < 0.05$). According to DISCERN analysis results, ChatGPT-5.3 mini (4.517 ± 0.425) achieved the highest reliability score, whereas DeepSeek-V4 (3.517 ± 0.465) obtained the lowest score (Table 3).

When the analysis results regarding the GQS scores, a statistically significant difference was identified among the chatbots ($p < 0.05$). According to GQS analysis results, ChatGPT-5.3 mini (4.400 ± 0.317) achieved the highest quality score, whereas DeepSeek-V4 (4.033 ± 0.373) obtained the lowest score (Table 3).

The results showed the Flesch Reading Ease Score (FRES) variable also revealed a statistically significant difference among the chatbots ($p < 0.05$). According to FRES results, Gemini-3.5 Flash (38.917 ± 9.262) achieved the highest readability score, whereas DeepSeek-V4 (25.328 ± 11.671) obtained the lowest score (Table 3).

Table 3
A comparison of the Likert, mDISCERN, GQS and FRES scores of four different chatbots.

	Chatbot	Mean	SD	p value
Likert	ChatGPT	4.400	0.298	**0.039**
	Gemini	4.367	0.403	
	Microsoft Copilot	4.383	0.271	
	Deepseek	4.150	0.333	
Modified Discern	ChatGPT	4.517	0.425	**0.000**
	Gemini	4.467	0.365	
	Microsoft Copilot	4.133	0.274	
	Deepseek	3.517	0.465	
Global Quality Scale	ChatGPT	4.400	0.317	**0.001**
	Gemini	4.350	0.366	
	Microsoft Copilot	4.267	0.256	
	Deepseek	4.033	0.373	
Flesch Reading Ease Score	ChatGPT	36.255	15.376	**0.004**
	Gemini	38.917	9.262	
	Microsoft Copilot	29.747	12.356	
	Deepseek	25.328	11.671	

Discussion

In line with rising aesthetic expectations and the ever-increasing influence of social media, the field of cosmetic dentistry has undergone significant development and transformation. Aesthetic requirements, particularly in the anterior region, mean that patient expectations are no longer limited to purely aesthetic and functional aspects, but must also be met from a psychosocial perspective [35]. In this context, patients generally prefer laminate veneers in line with their aesthetic expectations; consequently, given the high expectations of patients, it is essential to provide comprehensive information regarding the treatment to be carried out [36]. On the other hand, it is reported that patients initially turn to the internet to seek information about treatment options; during this process, they can easily access both posts by specialist doctors and social media content from individuals sharing their treatment experiences, and can shape their choices by evaluating this information [37].

The quality of health information regarding laminate veneers is of great importance due to the high level of interest shown by patients. Artificial intelligence chatbots are considered a potential resource for patients to access information about laminate veneers. In this context, the aim of this study is to evaluate the accuracy (Likert scale), reliability (Modified DISCERN) and quality (GQS) of the responses generated by four different chatbot models (ChatGPT-5. 3 mini, Gemini 3.5 Flash, DeepSeek-V4 and Microsoft Copilot) in response to frequently asked questions about laminate veneers. The findings revealed statistically significant differences between chatbot models in terms of accuracy, reliability, quality and readability; consequently, the null hypothesis of the study was rejected. These results have shown that the models evaluated performed at varying levels, particularly in the context of patient education, and that the quality of the content can vary depending on the artificial intelligence model used.

In the literature, various indices have been used to assess the accuracy of artificial intelligence chatbot responses; in this study, however, a five-point Likert scale was chosen [38,39]. The study found that all artificial intelligence chatbot models generally produced answers of an acceptable standard to the questions most frequently asked by patients regarding laminate veneers. All three evaluators awarded the highest accuracy score to the responses from ChatGPT-5.3 mini, followed by Microsoft Copilot, Gemini 3.5 Flash and Deep Seek-V4, in that order. In a study conducted by Topdağı et al. [21], the responses provided by ChatGPT-3.5, ChatGPT-4, Gemini and Copilot to questions regarding digital smile design were evaluated comparatively in terms of accuracy, reliability, quality and readability. The results of the study revealed that ChatGPT-4 demonstrated the highest performance among the models examined, with a median accuracy score of 5 (4–5). Furthermore, it has been reported that there are statistically significant differences in the accuracy scores of artificial intelligence models. The study conducted by Ozdemir et al. [40] aimed to evaluate the responses provided by various artificial intelligence programmes to questions relating to the field of restorative dentistry in terms of accuracy, reliability, consistency and readability. In the evaluation of knowledge-based questions, ChatGPT-4o demonstrated the highest performance with an accuracy rate of 93.3 percent, whereas Copilot showed the lowest performance with an accuracy rate of 68.9 percent. However, in a study by Özcivelek et al. [6], in which the responses of various artificial intelligence programmes were evaluated in terms of accuracy, quality, readability, comprehensibility and applicability, it was reported that Dental GPT achieved the highest accuracy score, whereas ChatGPT-4 demonstrated the lowest accuracy performance.

These findings demonstrate that current large language models are capable of generating answers to patient-centred questions about laminate veneers with a high degree of accuracy. However, the presence of statistically significant differences between the models suggests that the accuracy of the information provided by artificial intelligence systems may vary depending on the model used. For this reason, though AI-based chatbots may be regarded as a useful resource in patient education and information-seeking processes, it is important that the information provided is verified by dentists before it is applied in clinical practice.

Nowadays, the internet has become a widely used source for patients seeking health-related information. However, as the accuracy and reliability of information provided on online platforms cannot always be guaranteed, it is of the utmost importance to assess the reliability of health information. The fact that AI-based chatbots can occasionally generate incorrect, incomplete or misleading information, and that users tend to accept this information as accurate, raises significant concerns regarding the use of these systems in the healthcare sector [41]. In this context, factors such as the citation of information sources, the provision of references and citations, and the absence of misleading information are among the key criteria for assessing the reliability of health information. In this study, the reliability assessment was carried out using the Modified DISCERN scale.

The findings of our study have shown that all the AI chatbots evaluated generally produced responses of a high standard of reliability and quality to questions frequently asked by patients regarding laminate veneer treatments. However, statistically significant differences were observed between the chatbots in terms of their Modified DISCERN and GQS scores. The highest reliability and quality scores were achieved by ChatGPT-5.3 mini, whilst it was found that Deep Seek-V4's reliability and quality scores were lower than those of the other three models.

A study conducted by Gassem et al. [42] reported that ChatGPT 3.5, ChatGPT 4, Gemini and Copilot were compared in terms of the accuracy, reliability, quality and readability of the information they generated regarding orthodontic retention. The research findings revealed that ChatGPT 3.5 achieved the highest score in terms of accuracy, whereas ChatGPT 4 outperformed the other models in terms of reliability. Furthermore, it was reported that both ChatGPT models achieved better results than Gemini and Copilot in quality assessments. Similarly, in another study on MIH, ChatGPT-4 and ChatGPT-3.5 achieved the highest reliability scores [43]; however, a study by Shrestha et al. [44] reported that ChatGPT could exhibit lower accuracy, reliability and quality in certain scenarios. In a study in the field of paediatric dentistry published in the literature, it was reported that ChatGPT-4o achieved the highest MGQS score, whilst DeepSeek-V3 performed the worst, demonstrating significant differences between the models [45].

The findings of the present study are consistent with numerous studies in the literature in that ChatGPT-5.3 mini outperforms other models in terms of reliability and quality. This may be related to ChatGPT-5.3 mini's capacity to generate more comprehensive, balanced and structured responses to patient-centred healthcare queries. However, the fact that some studies report higher performance from different models may be attributable to differences in the scope of the question sets used, the dental discipline assessed, the evaluation methods employed and the versions of the artificial intelligence systems examined.

However, it is thought that the lower reliability and quality scores observed in some models, such as DeepSeek, may be linked to more limited validation mechanisms in a clinical context, a lack of source attribution, and responses containing more general or superficial information. Furthermore, differences in some models' capacity for real-time information integration and reference generation may lead to variability in performance, particularly on reliability-based scales such as DISCERN and GQS.

Readability is a concept that refers to the degree to which a written text is comprehensible to the reader, and it determines the text's cognitive processability. The level of readability is influenced not only by individual factors such as the reader's cognitive capacity, educational level, socio-cultural background, areas of interest and purpose of reading, but also by text-specific characteristics such as the text's content structure, conceptual density, terminology used, linguistic style and formal organisation [17]. According to the findings of our study, the Gemini-3.5 Flash model was found to have the highest readability score, whilst the DeepSeek-V4 model had the lowest.

A study published in the literature evaluated the performance of three AI chatbots in responding to questions frequently asked by patients regarding dental prostheses. The study reported that Gemini achieved the highest Flesch Reading Ease (FRES) score [18]. Similarly, in another study, responses generated by ChatGPT-3.5, ChatGPT-4, Gemini and Copilot were comparatively assessed in terms of accuracy, reliability, quality and readability; this study also reported that Gemini achieved the highest FRES score [7].

These findings suggest that differences in the models' readability performance may stem not only from their content generation capacity, but also from variations in language modelling strategies, response length, sentence structure and terminology usage. In particular, the tendency of some models, such as Gemini, to produce shorter, simpler and more user-focused responses may have contributed to higher Flesch Reading Ease (FRES) scores. Conversely, in models that generate more technical and detailed responses, increased conceptual density and the use of medical terminology may lead to lower readability scores.

Furthermore, the fact that the FRES scores for all models were generally found to be low suggests that text generated by artificial intelligence in specialist fields such as dentistry may have limitations in terms of comprehensibility at the patient level. This highlights the need to consider not only accuracy and reliability but also readability as a critical parameter, particularly in patient information processes.

Taken as a whole, the findings of this study indicate that the performance of artificial intelligence chatbots must be assessed from multiple perspectives. The imbalance between accuracy, reliability and readability highlights that high accuracy does not always equate to high comprehensibility or clinical reliability. This underscores the importance of adopting a holistic evaluation approach when considering AI outputs in clinical practice.

From a clinical perspective, the increasing use of artificial intelligence-based systems in patient information processes carries the risk that incorrect or incomplete information may directly influence the patient's decision-making process. Particularly in fields such as cosmetic dentistry, where patient expectations are high, information obtained from online sources may influence treatment choices. This situation can lead to misguided treatment requests, unrealistic expectations and deviations in clinical decision-making processes. Therefore, the integration of human expert oversight and verification mechanisms into the clinical use of AI outputs is of great importance.

Several limitations should be acknowledged. First, the fact that large language models have a constantly updated and rapidly evolving structure. In this study, the ChatGPT-5.3 mini, Gemini-3.5 Flash, DeepSeek-V4 and Microsoft Copilot versions available during the data collection period were evaluated. However, the performance of these models may change over time as new versions are released. Consequently, the findings obtained reflect only the performance of the evaluated model versions at the time of data collection and cannot be directly generalised to subsequent versions. To support the reproducibility of the study, the versions of the artificial intelligence models used have been clearly reported. It is recommended that future studies re-evaluate the findings using up-to-date model versions and conduct a comparative analysis of performance changes over time. Second, only four AI chatbots were evaluated. Third, readability was assessed using a single readability index, and patient perceptions were not directly investigated. Fourthly, each question was posed to each chatbot only once. This meant that it was not possible to assess the consistency of the responses, given the possibility that the models might generate different responses to the same question at different times. Consequently, the findings obtained reflect the models' performance within the context of a single interaction.

Future research that examines the long-term performance of AI chatbots, evaluates different large language models, and focuses on patient-centred outcomes—such as the impact of information provided by AI on patients' level of understanding, trust and treatment decision-making processes—will help to demonstrate the potential of these technologies in clinical practice more comprehensively.

Conclusion

This study has revealed that AI-generated responses to frequently asked questions about laminate veneers showed significant differences in terms of accuracy, reliability, quality and readability, depending on the chatbot used. ChatGPT-5.3 mini demonstrated the highest performance in terms of accuracy, reliability and overall quality assessments, whilst Gemini-3.5 Flash achieved the highest scores for readability. In contrast, DeepSeek-V4 performed the worst across most evaluation parameters. These findings indicate that AI-based chatbots have significant potential for supporting patient education and access to information regarding laminate veneers; however, the quality of the information provided may vary depending on the model used. Consequently, whilst these systems may be considered a useful supplementary resource for patient information, they should not replace clinical advice, and it should be borne in mind that the information provided must be verified by dentists. Future artificial intelligence models should focus not only on the accuracy of information but also on reliability, content quality and patient-friendly readability; this could help to improve patient communication and digital health literacy.

Abbreviations

AI: Artificial Intelligence

LLM: Large Language Model

FAQ: Frequently Asked Question

GQS: Global Quality Scale

FRES: Flesch Reading Ease Score

DISCERN: Quality Criteria for Consumer Health Information

RDS: Restorative Dentistry Specialist

SPSS: Statistical package for the social sciences

Declarations

Acknowledgements

Not applicable.

Authors' contributions

D. A. A.: Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review and editing

G. P.: Data curation, Investigation, Methodology, Writing – original draft, Writing – review and editing, Supervision

S.S.:Data curation, Investigation, Methodology, Writing – original draft, Writing – review and editing

Funding

Not applicable.

Data availability

The datasets analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable.

Competing Interests

The authors declare that they have no competing interests.