Evaluation of the responses from different chatbots to frequently asked patient questions about impacted canines

Elif Gökçe Erkan Acar* and Başak Arslan Avan†

Ankara Medipol University Faculty of Dentistry, Department of Orthodontics, Ankara, Turkiye*

Gazi University Faculty of Dentistry, Department of Orthodontics, Ankara, Turkiye†

Background: To evaluate the responses given by ChatGPT 4.0, Google Gemini 1.5 and Claude 3.5 Sonnet chatbots to questions about impacted canines in relation to reliability, accuracy and readability.

Methods: Thirty-five questions were posed to 3 different chatbots and 105 responses were received. The answers were evaluated in relation to reliability (Modified DISCERN), accuracy (Likert scale and Accuracy of Information Index (AOI)) and readability (Flesch-Kincaid Reading Ease Score (FRES) and Flesch-Kincaid grade level (FKGL)). Statistical significance was set at $p<0.05$.

Results: Gemini had the highest modified DISCERN score (33.66 ± 2.64), followed by Claude (29.70 ± 3.08) and ChatGPT (28.13 ± 2.83). ChatGPT had the highest mean Likert score (4.76 ± 0.43), while Claude and Gemini had 4.71 ± 0.47 and 4.66 ± 0.47, respectively. For the AOI index, ChatGPT had the highest mean score (8.67 ± 0.55), which was statistically significant when compared to others (ChatGPT vs Claude: p=0.042, ChatGPT vs Gemini: p=0.036). All chatbots showed similar readability FRES and FKGL scores without any significant differences (p=0.121 and p=0.377, respectively). Claude expressed responses with significantly fewer words than the other chatbots (Claude vs ChatGPT: p=0.019, Claude vs Gemini: p=0.001) and ChatGPT was the AI service that used the most words (239.74 ± 114.21).

Conclusion: In answering questions about impacted canines, Gemini showed good, while ChatGPT and Claude provided moderate reliability. All chatbots achieved high scores for accuracy. However, the responses were difficult to understand for anyone below a college reading level. Chatbots can serve as a resource for patients seeking general information about impacted canines, potentially enhancing and expediting clinician–patient communication. However, it should be noted that the readability of chatbot-generated texts may pose challenges, thereby affecting overall comprehension. Moreover, due to patient-specific, case-based variations, the most accurate interpretation should be provided by the patient's healthcare professional. In the future, improved outcomes across all parameters may be achieved through advancements in chatbot technology and increased integration between healthcare providers.

(Aust Orthod J 2025; 41: 288 - 300. DOI: 10.2478/aoj-2025-0020)

Received for publication: March, 2025

Accepted: May, 2025.

Elif Gökçe Erkan Acar: gokceliferkan@gmail.com; Başak Arslan Avan: dt.basakarslan@gmail.com

Introduction

Chatbots are deep learning-based AI models capable of compiling information that has been introduced up to a certain time and to provide services on requested topics. The models have become widely used since data is provided in a more effortless and convenient way,[1] thereby creating the feeling of conversation with a real person[2] and being always accessible,[3] when compared to traditional knowledge acquisition. AI usage has grown across many sectors, particularly

in the areas of academic information and learning, improving student performance, marketing, and customer relations, as well as for entertainment and hobbies.[4–6] In addition, healthcare professionals and patients also use AI applications to obtain medical information and during busy workflow.[7]

According to the U.S. National Center for Health Statistics,[8] 58.5% of adults used the internet to search for health or medical information during July–December 2022 and Link et al.[9] suggested that the number of online health information seekers has been steadily increasing. An additional study found that patients who obtained medical information from the internet were more likely to make an appointment with a clinician whose subsequent information affected patient behaviour and health decision.[10] de Looper et al.[11] suggested an online health information system may positively affect a patient's contribution to a clinical consultation and provide a level of patient satisfaction. As an advantage to both patients and clinicians, Bibault et al.[12] claimed that artificial conversational agents might help inform patients regarding minor health concerns and result in clinicians focusing more on the treatment of patients rather than providing information. Therefore, the integration of large language models (LLMs) into healthcare communication has generated significant interest due to their potential to support patient education, streamline clinical workflows, and improve health literacy. However, the exact practical applicability in patient-facing settings remains underexplored.[13] It is therefore important that the information obtained from chatbots is accurate since they can affect the operation of the health system.

Various chatbot developers exist and as the capabilities of chatbots evolve, it is essential to ensure that the provided information remains contemporary and relevant. The literature has reported studies which have evaluated the accuracy of online health-related information.[12,14] The answers of chatbots to questions have been published about medical conditions related to breast cancer,[10] paediatric cardiology,[15] stroke[16] and also in the field of orthodontics regarding popular general topics,[17,18] clear aligners,[19] and interceptive orthodontic applications.[20]

While chatbots may be a useful tool for addressing health concerns, the classification of impacted canines within this framework requires further clarification. Although not life-threatening, palatally-impacted maxillary canines may lead to potential complications associated with root resorption, cyst formation, and malocclusion if left untreated.[21] Impacted canines constitute a broad clinical condition requiring simple to complex treatment approaches based on case-specific characteristics,[22] and are typically managed through a multidisciplinary collaboration of dental specialists.[23] From the patient's perspective, the condition is noted through adolescence to adulthood and the decision-making process, particularly regarding whether to extract or retain the tooth, often generates uncertainty and debate. Uncertainty may lead patients to undertake personal research in addition to a clinical consultation. To the best of current knowledge, there is no study that has evaluated the chatbot-generated answers to questions regarding impacted canine teeth. Therefore, the aim of the present study was to evaluate the responses provided by ChatGPT 4.0, Google Gemini 1.5 and Claude 3.5 Sonnet chatbots to questions about impacted canines related to information reliability, accuracy and readability. The null hypothesis of the study was that different chatbots would not show any differences regarding questioned parameters.

Materials and methods

Since no materials derived from human or animal subjects were used, ethical committee approval was not required. Based on the approach used by Arslan et al.,[24] a statistical power analysis was conducted using G*Power software (Heinrich Heine Universität, Dusseldorf, Germany, version 3.1.9.7), which indicated that a minimum of 35 questions was required to achieve 80% statistical power.

To construct the question set, all prior browser history and cookies were cleared to avoid algorithmic bias.[24] The keywords "impacted teeth" and "impacted canine" were entered into the Google search engine (Google LLC, Mountain View, CA, USA) and the first 10 pages were reviewed. Patient-oriented websites such as clinician and clinic pages, health information blogs, and Q&A fora were included in the analysis. Frequently recurring themes and questions were identified and a total of 35 questions, which reflected the concerns of patients seeking information about impacted teeth, were formulated (Table I). The questions were intended to simulate natural, layperson inquiries and were designed with an emphasis on simplicity and clarity.

Table I. Questions used in the study

	Questions
1	What is impacted tooth?
2	What are the most common impacted tooth in humans?
3	What is the frequency of canine impaction?
4	What is the most common cause of canine impaction?
5	What are the signs of impacted canine tooth?
6	Do impacted canines have to be removed?
7	What age do you treat impacted canine?
8	Can an impacted canine come down on its own?
9	What happens if impacted canine tooth is left untreated ?
10	If impacted canine is left untreated, what would be the side effects?
11	Who does the treatment for impacted canine teeth?
12	How is the decision made whether to extract or maintain an impacted tooth?
13	Should I have orthodontic treatment to treat impacted canine?
14	How does the orthodontic treatment for impacted canine is done?
15	What problems can occur during the treatment of impacted canine?
16	Does treatment of impacted canine can be done with braces?
17	Does treatment of impacted canine can be done with aligners?
18	Can impacted canine tooth erupt on its own?
19	Can impacted canine tooth erupt with jaw expansion?
20	Can all impacted canine teeth be saved with orthodontic treatment?
21	How long does it take to erupt an impacted canine?
22	Does orthodontic treatment of impacted canine tooth cause pain?
23	Is x-ray necessary for impacted canine tooth treatment?
24	How is impacted canine tooth exposure surgery done?
25	Who does the canine tooth exposure surgery?
26	What is closed technique for impacted canine?
27	What is gold chain in impacted canine treatment?
28	How long does it take to recover from impacted canine exposure surgery?
29	How is oral care performed after the canine exposure surgery?
30	Does impacted tooth exposure surgery cause pain?
31	When can I return to work and school after impacted canine exposure surgery?
32	Does orthodontic treatment to bring an impacted canine into position cause damage to the tooth?
33	Does another tooth regrow after treatment of impacted tooth?
34	Will my teeth shift after the orthodontic treatment of an impacted canine tooth?
35	How is retention done after impacted canine tooth treatment?

The initial draft of the question set was prepared by one researcher (E.G.E.A) and subsequently reviewed by both authors to ensure that the questions were both clinically relevant and accessible to individuals without specialised dental knowledge.

Questions were subsequently posed to three different artificial intelligence chatbots: ChatGPT 4.0 (ChatGPT), Claude 3.5 Sonnet (Claude), and Google Gemini 1.5 (Gemini). The questions were asked once using the same internet connection and laptop (MacBook Pro, 2,3 GHz Dual-Core Intel Core i5). During this process, the search history of the web browser and chatbots was closed and a new tab was opened for each question. All responses were obtained on the same day, and so the paid version of chatbots was used to overcome question limitation. The responses to questions were copied to Microsoft Word files created for each chatbot. Any references, visual images and word counts of responses were also recorded to the same files. An evaluation of responses was conducted by two orthodontists in separate rooms under similar conditions and using four different assessment tools to evaluate response reliability (modified DISCERN), accuracy (Likert scale and Accuracy of Information Index (AOI)) and readability (Flesch-Kincaid Reading Ease Score (FRES) and Flesch-Kincaid grade level (FKGL)).

DISCERN[25] is a tool which tests the quality of health information related to a treatment plan and consists of 2 parts, the first of which asks 8 questions regarding reliability, while the other asks 7 questions which provide information about treatment. Content evaluation was assessed by a modified DISCERN tool of which only the first 8 questions were utilised[19] (Table II). In the modified DISCERN, the questioning of health content is achieved according to parameters related to clarity, the source of information, reference to uncertainty areas and finally, a total score for reliability is obtained by summing the scores of all parameters (Table II). As the total score increases, the reliability of information also increases according to the categorisation of poor (8–15 points), moderate (16–31 points), or good (32–40 points).[26]

The second assessment criteria regarding accuracy used a 5-point Likert scale [19,20] and an Accuracy of Information Index (AOI) recommended by Daraqel *et al*. [18] (Table II). A Likert scale evaluates the extent of true and false statements within a given text using a 5-point scale, of which number 1 indicates that answers were completely false and number 5 indicates completely true (Table II). AOI included five key components of factual accuracy, corroboration, consistency, clarity and specificity, and relevance of the response scaling each component from 0 to 2, representing poor to excellent. Subsequently, a total score is obtained by summing the scores of the five components. Therefore, as the score increases, an improvement in the quality of that response is indicated.

As the 3rd evaluation step, the responses were analysed according to readability using the Flesch-Kincaid Reading Ease Score and the Flesch-Kincaid Grade Level by the Microsoft Word Flesch-Kincaid calculation feature (version 16.66.1 [22101101]; Microsoft, Redmond, WA, USA).[27] FRES and FKGL are calculated by syllable, word and sentence counts of an English text using the formulas 206.835-1.015x(total words/total sentences)-84.6x(total syllables/total words) and 0.39x(total words/total sentences)+11.8x(total syllables/total words)-15.59, respectively. In this scoring system, a calculation yields an objective score between 0 and 100, which identifies the reading level as easy or difficult. As the score increases to 100, the document becomes easier to read, while the grade section indicates the educational level of individuals who can comprehensively read the document (Table II). A flowchart of the study is shown in Fig. 1.

Statistical analysis

All data obtained were analysed using the SPSS 22 for Windows (SPSS Inc., Chicago, IL, USA) for statistical analysis. Descriptive statistics were presented as mean, standard deviation and median (minimum to maximum). For normally distributed data, the one-way ANOVA test was used to compare groups, while the Kruskal-Wallis H test was applied for non-normally distributed data. The post-hoc Tukey test was used to determine which pairs showed significant differences between the groups. A significance level of 0.05 was set, therefore a p-value less than 0.05 indicated a statistically significant difference.

Results

Thirty-five questions were asked of 3 different chatbots and 105 responses were obtained. In addition to texts, Gemini provided 30 images with captions for 13 of the 35 questions. After an evaluation was

Table II. The Modified DISCERN, Likert Scoring, AOI Index, FRES and FKGL

MODIFIED DİSCERN		
Questions	Scoring (Minimum-maximum)	
1. Are the aims clear?	1-5	
2. Does it achieve its aims?	1-5	
3. Is it relevant?	1-5	
4. Is it clear what sources of information were used to compile the publication?	1-5	
5. Is it clear when the information used or reported in the publication was produced?	1-5	
6. Is it balanced and unbiased?	1-5	
7. Does it provide details of additional sources of support and information?	1-5	
8. Does it refer to areas of uncertainty?	1-5	
Total Modified DİSCERN Score	8-40 (Poor (8–15 point), moderate (16–31 point), or good (32–40 point))	
LIKERT Scale		
Parameter	Score	
The chatbot's answers are completely incorrect	1	
The chatbot's answers contain more incorrect items than correct items	2	
The chatbot's answers contain an equal balance of correct and incorrect items	3	
The chatbot's answers contain more correct items than incorrect items	4	
The chatbot's answers are completely correct	5	
AOI INDEX		
Item	Definition	Score (minimum-maximum)
Factual accuracy	The response aligns with known facts, data, or established knowledge on the subject	0-2
Corroboration	The response is based on evidence from textbooks, studies, or guidelines	0-2
Consistency	The response is internally consistent and does not contain contradictory 2 statements	0-2

Table II. Continued

AOI INDEX		
Item	Definition	Score (minimum-maximum)
Clarity and specificity	The response is clear and specific, avoiding vague or ambiguous language	0-2
Relevance	The response directly addresses and adheres to the question or topic posed	0-2
Total AOI score	The sum of all scores	0-10
FLESCH KINCAID READABILITY		
Flesh Reading Ease Score	Grade Level	Reading level
90-100	5	Very Easy
80-90	6	Easy
70-80	7	Fairly Easy
60-70	8-9	Standard and/or Plain
50-60	10-12	Fairly Difficult
30-50	College	Difficult
0-30	College Grad	Very Difficult

completed, the scores were analysed for intraclass correlation coefficients (ICC) which were found to range between 0.869 to 0.996, thereby showing high reliability between examiners. Therefore, the average result of the 2 evaluators was calculated and consequently, mean scores were obtained for the Modified DISCERN, Likert and the AIO index. Flesch Kincaid Readability and word count calculations were used without any modifications since they were objective parameters. The results are shown in Tables III and IV.

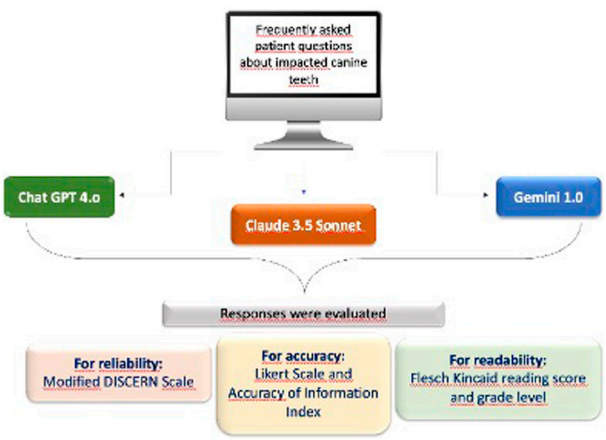

Figure 1. Flowchart of the study.

Of the mean modified DISCERN scores, Gemini 1.5 had the highest (33.66 ± 2.64), followed by Claude (29.70 ± 3.08) and ChatGPT (28.13 ± 2.83). The difference between Gemini 1.5 and the other chatbots was found to be statistically significant (ChatGPT vs Gemini: p=0.028, Claude vs Gemini: p=0.011). ChatGPT had the highest mean Likert score of 4.76 ± 0.43, while Claude and Gemini had 4.71 ± 0.47 and 4.66 ± 0.47, respectively. No statistical difference was found between the chatbots (p=0.61). Following the order of Likert scores, ChatGPT had the highest mean AOI index score (8.67 ± 0.55), as statistically significant when compared to the others (ChatGPT vs Claude: p=0.042, ChatGPT vs Gemini: p=0.036). Claude had 8.37 ± 0.78 and Gemini had 8.11 ± 0.84 as the lowest.

Gemini (40.19 ± 9.85) had the highest FRES score followed by ChatGPT (37.89 ± 8.08) and Claude (35.46 ± 10.47), and ChatGPT (12.99 ± 1.70) had the highest FKGL score very close to other chatbots (Claude: 12.64 ± 1.39; Gemini: 12.44 ± 1.83). FRES and the grade levels of chatbots did not show any

Table III. Modified DISCERN, Likert and AOI index scores of ChatGPT, Claude and Gemini

	ChatGPT		Claude		Gemini				
	Mean ± SD	Median (Min-Max)	Mean ± SD	Median (Min-Max)	Mean ± SD	Median (Min-Max)	*P* value		
Modified Discern Score	28.13 ± 2.83	29.00 (20.00-32.00)	29.70 ± 3.08	31.00 (23.00-33.00)	33.66 ± 2.64	34.00 (28.00-38.00)	ChatGPT vs Claude	ChatGPT vs Gemini	Claude vs Gemini
							0.385	0.028*	0.011*
Likert Score	4.76 ± 0.43	5.00 (4.00-5.00)	4.71 ± 0.47	5.00 (3.50-5.00)	4.66 ± 0.47	5.00 (4.00-5.00)		0.61	
AOI Index	8.67 ± 0.55	9.00 (7.00-9.00)	8.37 ± 0.78	9.00 (6.00-9.00)	8.11 ± 0.84	8.00 (6.00-10.00)	ChatGPT vs Claude	ChatGPT vs Gemini	Claude vs Gemini
							0.042*	0.036*	0.360

*P*values for groups were calculated using the one way ANOVA test and if significant difference vas observed, posthoc Tukey test was used to determine which pairs show significant differences.

*P < 0.05.

Table IV. Flesch-Kincaid Readilibity Scores, Flesch-Kincaid Grade Levels and word counts of ChatGPT, Claude and Gemini

	ChatGPT		Claude		Gemini		*P* value		
	Mean ± SD	Median (Min-Max)	Mean ± SD	Median (Min-Max)	Mean ± SD	Median (Min-Max)			
Flesch-Kincaid Reading Ease Score	37.89 ± 8.08	36.90 (14.80-51.00)	35.46 ± 10.47	36.30 (5.70-51.10)	40.19 ± 9.85	38.40 (18.30-62.40)	0.121		
Flesch-Kincaid Grade Level	12.99 ± 1.70	13.30 (9.70-17.50)	12.64 ± 1.39	12.40 (10.40-15.90)	12.44 ± 1.83	12.60 (8.40-15.90)	0.377		
Word Count	239.74 ± 114.21	237.00 (64.00-490.00)	185.89 ± 49.91	181.00 (95.00-285.00)	219.29 ± 96.54	187.00 (65.00-430.00)	ChatGPT vs Claude	Claude vs. Gemini	ChatGPT vs. Gemini
							0.019*	0.001*	0.041*

P values for groups were calculated using the Kruskal-Wallis H test and if significant difference vas observed, posthoc Tukey test was used to determine which pairs show significant differences.*p<0.05.

statistically significant differences between each other (p=0.121 and 0.377, respectively). The FRES of all chatbots ranged between a 30-50 score interval corresponding to responses that would be difficult to understand by anyone below a college reading level as indicated in Table II. Claude expressed responses related to word count with significantly fewer words than the other chatbots (Claude vs ChatGPT: p=0.019, Claude vs Gemini: p=0.001). A statistical difference was also found between ChatGPT and Gemini (p=0.041). ChatGPT was the chatbot that used the most words with an average of 239 words and Claude the least with an average of 185.

Discussion

Access to information can be universally achieved which has led to changes in how people search for knowledge.[28] The internet has become a significant source of online public health information, as search engines become the primary access point for patients.[29] Recently, chatbots have emerged in association with the development of artificial intelligence-based large language models which can communicate in written and spoken forms. There is an increasing popularity due to their ease of use and quick answers to everyday questions.[1] In addition to their various applications in different fields, chatbots are being used by patients and clinicians in medicine and dentistry for both diagnostic purposes and information acquisition.[30,31]

The responses from chatbots to questions regarding clear aligner treatment,[19] interceptive treatment,[20] temporary anchorage devices,[32] orthognathic surgery[31] and general orthodontic topics[17,18,33] have been evaluated. However, to the best of current knowledge, no studies were found regarding impacted canines. The treatment of impacted canines is a challenging process about which patients are curious. The decision between extraction and forced eruption can be controversial and determined by a number of clinical factors.[34] For this reason, frequently asked patient questions about impacted canines was compiled, and chatbot responses were evaluated. Previous studies on chatbots have been conducted using ChatGPT 4.0, Google Bard, Copilot AI, and Microsoft Bing applications. In the present research, ChatGPT 4.0, Google Gemini 1.5, and Claude 3.5 Sonnet, which had not been previously applied in orthodontics, were chosen because of their widespread use. Chatbots are essentially applications that have the potential to present information introduced up to a certain time and provide responses on requested topics. However, there can be noted variations in chatbots produced by different developers.[35] As identified from in-application information, ChatGPT 4.0 and Claude 3.5 Sonnet were 'trained' using data up to the end of 2023, and April 2024, respectively. For Gemini 1.5, clear information about the timeframe of its data could not be obtained, since continuous updates have occurred and commercial secrets protected. This situation is considered to be a main reason for the differences in responses provided by artificial intelligence models to posed questions.

Chatbots may provide different answers to a repeated question or when asked the same question following a change in question structure.[36] To avoid confusion in word count and the readability parameters of answers, each question was asked only once and the web browser and chatbot application's tracking history were turned off, so that a new tab was required and opened for each question.[17,19] While ChatGPT and Claude provided similar and plain texts related to flow and integrity, Gemini provided references at the end of sentences and added images and image captions in addition to text for some answers (13 out of 35 answers), as noted by Aziz et al.[37] It is known that images combined with the texts strengthen the narration[38] and the retention of knowledge,[39] and so Gemini might be advantageous in this respect. However, it was also detected that the images in some answers were irrelevant (14 out of 30 images). For this reason, a careful examination is recommended when evaluating responses. While Gemini indicated the information source was an academic article or a website URL, it was noted that ChatGPT and Claude did not directly provide any data. When a second question was asked regarding the source of the answer, ChatGPT stated that it obtained the information from the general dental literature, relevant academic sources and guides, and gave the names of well-known textbooks for guidance. Claude based its information on general medical and dental information introduced to it up until April 2024. The sources Gemini provided were mostly clinician or university-related informational websites and academic publications. In this regard, it was considered beneficial for all chatbots to take their sources from websites which have the HONcode seal, which is an indicator that health websites have reliable and understandable information, as well as

from popular textbooks and peer-reviewed academic articles in the relevant field.[16]

According to the average results of the modified DISCERN tool, Gemini showed good reliability at the lower limit with a statistically significant difference (ChatGPT vs Gemini: p=0.028, Claude vs Gemini: p=0.011), while ChatGPT and Claude showed moderate reliability. That Gemini scored higher than ChatGPT is consistent with the studies by Dursun et al.[19] and Taymour et al.[40] which evaluated chatbot responses to questions about clear aligners and dental implantology, respectively. Gemini's score differed from Johnson's study which assessed the reliability of responses to dental trauma questions.[41] The moderate reliability of ChatGPT is similar to the results of Onder et al.[26] and Kılınç et al.[17] who evaluated the responses to questions about hypothyroidism during pregnancy and frequently asked general questions about orthodontics. ChatGPT showed the highest and Gemini showed the lowest Likert scores regarding the accuracy of responses; however, the results were not statistically different (p=0.61). Because the Likert scores were greater than 4 and close to the maximum score of 5, chatbots generally provided correct answers to the questions. In another accuracy index, AOI, ChatGPT received the highest score similar to Likert, followed by Claude and Gemini (ChatGPT vs Claude: p=0.042, ChatGPT vs Gemini: p=0.036). The high accuracy demonstrated by ChatGPT was consistent with the findings of recent comparative medical studies on emergency responses to avulsion injuries,[42] on paediatric radiology,[43] and on answers related to retinal detachment.[44] While Gemini received a significantly higher score in the modified DISCERN system compared to the others, its low score in accuracy indices may have developed due to the diversity of queried parameters and that the scoring was conducted over a wider range of 0 to 5 and the total score was finally used. From a mathematical perspective, the wider scoring range of the DISCERN tool could have contributed to a more distinct representation of intergroup differences. Furthermore, items assessing the presence of the references or source transparency may have negatively affected the scores of ChatGPT and Claude in the modified DISCERN, as these models typically do not provide explicit source citations.

ChatGPT was found to provide more accurate information than the other chatbots. This is consistent with the study by Daraqel et al.,[18] who compared Chat GPT 3.5 and Google Bard, and Hatia et al.,[20] who examined the response accuracy of ChatGPT 4.0 to questions about interceptive orthodontics. Despite the high level of accuracy, it was still apparent that there was no chatbot that correctly answered all the questions.[20]

In addition to reliability and accuracy, the readability and understanding of information provided by chatbots are critical factors that enhance their overall value. A previous study demonstrated that many public health materials were written above the recommended 6th to 8th grade reading levels, thereby reducing their accessibility for the general population.[45] Consequently, the readability of health-related content plays a vital role in promoting effective patient education and engagement. To assess this issue, various linguistic parameters can be utilised, such as sentence length and the frequency of complex vocabulary. The Flesch-Kincaid Reading Ease Score and Flesch-Kincaid Grade Level used in the present study evaluated an English written text by examining how many words, sentences, and contained syllables and indicate an approximate educational level that a person requires to read the text easily.[17] A high FRES score indicates better readability, while a score between 60 and 70 represents standard English. In the present study, all chatbots remained below an acceptable limit and were found to be difficult to read at college level. The finding appears as an important disadvantage, considering that these applications are used for the purpose of gathering medical information and providing a social benefit for patients. Although it was not statistically significant, Gemini received a higher FRES (40.19 ± 9.85) and lower FKGL (12.44 ± 1.83) when compared to the other chatbots, in concordance with a previous study in which ChatGPT 3.5, ChatGPT 4.0, Gemini, and Copilot were examined.[19] Earlier studies conducted on ChatGPT also found that the understanding of the responses was low.[17,46] Yurdakurban et al.[47] found that chatbot-generated answers about orthognathic surgery required at least a college-level education for readability when evaluated by the Simple Measure of Gobbledygook index (https://readabilityformulas.com/the-smog-readability-formula). Contrary to the present findings, an endodontic study found that GPT-3.5, Google Bard, and Bing's responses were at an acceptable level of reliability.[48] Different results may occur since FRES does not evaluate

texts related to differences in sentence structure, the use of professional terminology, and jargon of the writing style which may also be factors affecting content readability. However, a literature review has indicated that the majority of existing studies has predominately focused on accuracy and reliability parameters, while also highlighting the need for further research addressing readability.[20,31–33]

From an ethical perspective, inaccurate medical information obtained through chatbots could result in harm to patients and privacy risks associated with patient data.[49] It has been previously demonstrated that AI chatbots providing incorrect or misleading medication-related guidance may pose risks to patient safety.[50] Recent research by Zada et al.[51] found that several popular LLMs generated persuasive but sometimes inaccurate content about harmful health myths often including fake references and patient stories. These outputs can easily mislead users seeking reliable advice. The findings underscore the potential harm to patients who rely on AI tools for medical guidance without proper oversight. It has been noted that AI developers include disclaimers or warnings about potential errors within these applications to address the attendant risks. In contrast, during an examination with a real clinician, the person is held accountable for their statements to patients and may even face legal consequences when necessary. Although companies legally protect themselves in the case of chatbots, artificial intelligence developers must also exercise sensitivity in this regard and refrain from providing responses to controversial topics or implement applications under the guidance of expert clinicians. It would therefore be beneficial for patients who use chatbots for informational purposes to be aware of these issues and act selectively.

A limitation of the present study is that both the questions and answers were evaluated in English. The content of responses, word counts, and readability metrics may vary when examined in other languages. Another limitation arises from the evolving nature of chatbot technology. These applications are continuously updated and may provide different answers to the same question at different times. This variability can introduce biases in the evaluation process, as answers may differ depending on the version of the chatbot interrogated. Additionally, measurement errors may occur due to factors related to the subjective interpretation of responses, particularly in the assessment of reliability and accuracy. Although objective scoring criteria were used, evaluator bias cannot be completely eliminated. However, it is also important to note that the study involved two evaluators, and statistical analysis revealed a high level of agreement, which helps to mitigate potential bias. The question number of 35, while based on the power analysis, may still limit the generalisability of the findings. Variability in chatbot behaviour due to model updates or phrasing differences in user input may also introduce a level of measurement error. These factors should be carefully considered when interpreting the study's findings. Future research should aim to minimise such biases and measurement errors, while also monitoring the continuous advancements in chatbot technology. Despite these limitations, the data obtained from the present study may still provide valuable insights for clinicians, patients, plus chatbot developers, and may further contribute to the development of these technologies.

Conclusion

(1) The null hypothesis was rejected. In replying to questions about impacted canines, Gemini showed good reliability, while ChatGPT and Claude provided moderate reliability. All chatbots achieved high scores for accuracy and provided college level texts which were difficult to read.
(2) ChatGPT and Claude did not provide sources and images with their answers, unlike Gemini. However, when evaluating visuals, compatibility with the text should also be considered.
(3) The answers provided by chatbots to questions about impacted canines may be used by patients to acquire basic information prior to a dental visit or during the treatment process. However, since health-related issues can vary from person to person, it is considered that the most accurate source of information is the clinician.
(4) Chatbots are applications that frequently receive updates and are continuously improved. Therefore, future studies may be useful in evaluating and monitoring knowledge development.

Conflict of interest

The authors declare that there is no conflict of interest.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Corresponding author

Başak Arslan Avan, Gazi University Faculty of Dentistry, Department of Orthodontics, Bişkek Road. 1st Street No:4, Emek, Ankara, Turkiye, Postal Code: 06490, Telephone: +905073464970. Email: dt.basakarslan@gmail.com

References

[1] Farid H. Chatbots in Digital Marketing. In: Chintalapati & Pandey, eds. Chapter 3: Contemporary Approaches in Digital Marketing and the Role of Machine Intelligence. Hershey, (PA): IGI Global; 2023:46–67.
[2] Skjuve M, Følstad A, Fostervold KI, Brandtzaeg PB. My chatbot companion-a study of human-chatbot relationships. Int J Hum Comput Stud 2021;149:102601.
[3] Dobbala MK, Lingolu MSS. Conversational ai and chatbots: Enhancing user experience on websites. AJCSIT 2024;7:11.
[4] Ait Baha T, El Hajji M, Es-Saady Y, Fadili H. The impact of educational chatbot on student learning experience. Educ Inf Technol 2024;29:10153–76.
[5] Cheng Y, Jiang H. Customer–brand relationship in the era of artificial intelligence: understanding the role of chatbot marketing efforts. J Prod Brand Manag 2022;31:252–64.
[6] García-Méndez S, De Arriba-Perez F, González-Castaño FJ, Regueiro-Janeiro JA, Gil-Castiñeira F. Entertainment chatbot for the digital inclusion of elderly people without abstraction capabilities. IEEE Access 2021;9:75878–91.
[7] 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 2024;26:e56930.
[8] Wang X, Cohen RA. Health Information Technology Use Among Adults: United States, July–December 2022. US Department of Health and Human Services, Centers for Disease Control and Prevention 2023.
[9] Bachl M, Link E, Mangold F, Stier S. Search engine use for health-related purposes: Behavioral data on online health information-seeking in Germany. Health Commun 2024;39:1651–64.
[10] Li H, Li D, Zhai M, Lin L, Cao Z. Associations Among Online Health Information Seeking Behavior, Online Health Information Perception, and Health Service Utilization: Cross-Sectional Study. J Med Internet Res 2025;27:e66683.
[11] De Looper M, van Weert JC, Schouten BC, Bolle S, Belgers EH, Eddes EH, et al. The influence of online health information seeking before a consultation on anxiety, satisfaction, and information recall, mediated by patient participation: field study. J J Med Internet Res 2021;23:e23670.
[12] Bibault J-E, Chaix B, Guillemassé A, Cousin S, Escande A, Perrin M, et al. A chatbot versus physicians to provide information for patients with breast cancer: blind, randomized controlled noninferiority trial. J Med Internet Res 2019;21:e15787.
[13] Topol E. Deep medicine: how artificial intelligence can make healthcare human again. Hachette UK 2019
[14] Alpaydın MT, Buyuk SK, Bavbek NC. Information on the Internet about clear aligner treatment—an assessment of content, quality, and readability. J Orofac Orthop 2021;83(Suppl 1):1.
[15] Gritti MN, AlTurki H, Farid P, Morgan CT. Progression of an artificial intelligence chatbot (ChatGPT) for pediatric cardiology educational knowledge assessment. Pediatr Cardiol 2024;45:309–13.
[16] Lam WY, Au SCL. Stroke care in the ChatGPT era: Potential use in early symptom recognition. J Acute Dis 2023;12:129–30.
[17] 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.
[18] Daraqel B, Wafaie K, Mohammed H, Cao L, Mheissen S, Liu Y, et al. The performance of artificial intelligence models in generating responses to general orthodontic questions: ChatGPT vs Google Bard. Am J Orthod Dentofacial Orthop 2024;165:652–62.
[19] Dursun D, Bilici Geçer R. Can artificial intelligence models serve as patient information consultants in orthodontics? BMC Med Inform Decis Mak 2024;24:211.
[20] Hatia A, Doldo T, Parrini S, Chisci E, Cipriani L, Montagna L, et al. Accuracy and completeness of ChatGPT-Generated information on interceptive orthodontics: a Multicenter Collaborative Study. J Clin Med 2024;13:735.
[21] Mavreas D, Athanasiou AE. Factors affecting the duration of orthodontic treatment: A systematic review. Eur J of Orthod 2008;30:386–95.
[22] Grisar K, Luyten J, Preda F, Martin C, Hoppenreijs T, Politis C, et al. Interventions for impacted maxillary canines: A systematic review of the relationship between initial canine position and treatment outcome. Orthod Craniofac Res 2021;24:180–93.
[23] Mancini A, Chirico F, Colella G, Piras F, Colonna V, Marotti P, et al. Evaluating the success rates and effectiveness of surgical and orthodontic interventions for impacted canines: a systematic review of surgical and orthodontic interventions and a case series. BMC Oral Health 2025;25:295.
[24] Arslan C, Kahya, K, Cesur E, Cakan DG. An evaluation of orthodontic information quality regarding artificial intelligence (AI) chatbot technologies: A comparison of ChatGPT and google BARD. Australas Orthod J 2024;40:149–57.
[25] 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 1999;53:105–11.
[26] Onder C, Koc G, Gokbulut P, Taskaldiran I, Kuskonmaz S. Evaluation of the reliability and readability of ChatGPT-4 responses regarding hypothyroidism during pregnancy. Sci Rep 2024;14:243.
[27] Kincaid JP, Fishburne Jr RP, Rogers RL, Chissom BS. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel (Research Branch Report 8-75). Naval Technical Training Command 1975.
[28] Chu JT, Wang MP, Shen C, Viswanath K, Lam TH, Chan SSC. How, when and why people seek health information online: qualitative study in Hong Kong. Interact J Med Res 2017;6:e7000.
[29] Bachl M, Link E, Mangold F, Stier S. Search engine use for health-related purposes: Behavioral data on online health information-seeking in Germany. Health Commun 2024;39:1–14.
[30] Davis RJ, Ayo-Ajibola O, Lin ME, Swanson MS, Chambers TN, Kwon DI, et al. Evaluation of oropharyngeal cancer information from revolutionary artificial intelligence chatbot. The Laryngoscope 2024;134:2252–7.
[31] Balel Y. Can ChatGPT be used in oral and maxillofacial surgery? J Stomatol Oral Maxillofac Surg 2023;124:101471.

[32] Tanaka OM, Gasparello GG, Hartmann GC, Casagrande FA, Pithon MM. Assessing the reliability of ChatGPT: a content analysis of self-generated and self-answered questions on clear aligners, TADs and digital imaging. Dental Press J Orthod 2023;28: e2323183.
[33] Perez-Pino A, Yadav S, Upadhyay M, Cardarelli L, Tadinada A. The accuracy of artificial intelligence–based virtual assistants in responding to routinely asked questions about orthodontics. Angle Orthod 2023;93:427–32.
[34] Tanaka OM, Weissheimer A, Pithon MM, Gasparello GG, Araújo EA. Focus on leveling the hidden: managing impacted maxillary canines. Dental Press J Orthod 2024;29:e24spe5.
[35] Smutny P, Bojko M. Comparative Analysis of Chatbots Using Large Language Models for Web Development Tasks. Appl Sci 2024;14:10048.
[36] Vaishya R, Misra A, Vaish A. ChatGPT: Is this version good for healthcare and research? Diabetes Metab Syndr 2023;17: 102744.
[37] Aziz AAA, Abdelrahman HH, Hassan MG. The use of ChatGPT and Google Gemini in responding to orthognathic surgery-related questions: A comparative study. J World Fed Orthod 2025;14: 20–6.
[38] Carifio J, Perla RJ. A critique of the theoretical and empirical literature of the use of diagrams, graphs, and other visual aids in the learning of scientific-technical content from expository texts and instruction. Interchange 2009;40:403–36.
[39] Alhazmi K. The Effect of Multimedia on Vocabulary Learning and Retention. World J Engl Lang 2024;14:390.
[40] Taymour N, Fouda SM, Abdelrahaman HH, Hassan MG. Performance of the ChatGPT-3.5, ChatGPT-4, and Google Gemini large language models in responding to dental implantology inquiries. J Prosthet Dent 2025. [Epub ahead of print]
[41] Johnson AJ, Singh TK, Gupta A, Sankar H, Gill I, Shalini M, et al. Evaluation of validity and reliability of AI Chatbots as public sources of information on dental trauma. Dent Traumatol 2024;41 41:187–193.
[42] Mustuloğlu Ş, Deniz BP. Evaluation of Chatbots in the Emergency Management of Avulsion Injuries. Dent Traumatol 2025:1–8.
[43] Sami MA, Samad MA, Parekh K, Suthar PP. Comparative accuracy of ChatGPT 4.0 and Google Gemini in answering pediatric radiology text-based questions. Cureus 2024;16: e70897.
[44] Strzalkowski P, Strzalkowska A, Chhablani J, Pfau K, Errera MH, Roth M, et al. Evaluation of the accuracy and readability of ChatGPT-4 and Google Gemini in providing information on retinal detachment: A multicenter expert comparative study. Int J Retina Vitreous 2024;10:61.
[45] Mishra V, Dexter JP. Comparison of readability of official public health information about COVID-19 on websites of international agencies and the governments of 15 countries. JAMA Netw Open 2020;3:e2018033–e2018033.
[46] Abou-Abdallah M, Dar T, Mahmudzade Y, Michaels J, Talwar R, Tornari C. The quality and readability of patient information provided by ChatGPT: can AI reliably explain common ENT operations? Eur Arch Otorhinolaryngol 2024;281:6147–6153.
[47] Yurdakurban E, Topsakal KG, Duran GS. A comparative analysis of AI-based chatbots: Assessing data quality in orthognathic surgery related patient information. J Stomol Oral Maxillofac Surg 2024;125:101757.
[48] Mohammad-Rahimi H, Ourang SA, Pourhoseingholi MA, Dianat O, Dummer PMH, Nosrat A. Validity and reliability of artificial intelligence chatbots as public sources of information on endodontics. Int Endod J 2024;57:305–14.
[49] Lima NGM, Costa L, Santos PB. ChatGPT in orthodontics: Limitations and possibilities. Australas Orthod J 2024;40:19–21.
[50] Andrikyan W, Sametinger SM, Kosfeld F, Jung-Poppe L, Fromm MF, Maas R, et al. Artificial intelligence-powered chatbots in search engines: a cross-sectional study on the quality and risks of drug information for patients. BMJ Qual Saf 2025;34:100–9.
[51] Zada T, Tam N, Barnard F, Van Sittert M, Bhat V, Rambhatla S. Medical Misinformation in AI-Assisted Self-Diagnosis: Development of a Method (EvalPrompt) for Analyzing Large Language Models. JMIR Form Res 2025;9:e66207.