



Accuracy, completeness and readability of large language model chatbots answering dental patient questions: a systematic review

Precisión, exhaustividad y legibilidad de chatbots de modelos de lenguaje en preguntas de pacientes odontológicos: revisión sistemática

Ernesto Learsys Campello Trujillo¹  

Lourdes Hernandez Cuetara¹  

Yivelin Fernandez Queija¹  

¹ Universidad de Ciencias Médicas de la Habana, La Habana, Cuba.

Submitted: 02-01-2026 | Revised: 01-03-2026 | Accepted: 07-05-2026 | Published: 12-05-2026

Corresponding Author: Ernesto Learsys Campello Trujillo 

ABSTRACT

Introduction: Large language model chatbots are increasingly used by patients to seek answers about oral health, but the quality of the information they provide remains uncertain. **Methods:** A systematic review was conducted following PRISMA 2020. Scopus, PubMed and Crossref were searched up to 2026 for studies that evaluated chatbots answering patient questions about oral health against a reference standard, with no language restriction. Screening was assisted by artificial intelligence and verified by the authors. **Results:** Twenty-six studies published in 2025 and 2026 were included. ChatGPT was the most frequently evaluated model (21 studies), followed by Gemini (16) and DeepSeek (12). Accuracy was generally moderate to high but varied by model and clinical domain, with no model consistently superior. Readability was a recurrent weakness, with most responses exceeding the recommended sixth-grade reading level. Completeness was higher for chatbot responses than for clinician responses in one comparative study. **Conclusions:** Large language model chatbots provide largely accurate but frequently hard-to-read answers to dental patient questions. Their use should be adjunctive to professional care, and readability of outputs must be improved before they can safely support patient education.

Keywords: artificial intelligence; large language models; dentistry; patient education; health literacy; systematic review.

RESUMEN

Introducción: Los chatbots basados en modelos de lenguaje se usan cada vez más para que los pacientes resuelvan dudas sobre salud bucal, pero la calidad de la información que ofrecen sigue siendo incierta. **Métodos:** Se realizó una revisión sistemática según PRISMA 2020. Se buscó en Scopus, PubMed y Crossref hasta 2026 estudios que evaluaran chatbots que responden preguntas de pacientes sobre salud bucal frente a un patrón de referencia, sin restricción de idioma. El cribado fue asistido por inteligencia artificial y verificado por los autores. **Resultados:** Se incluyeron veintiséis estudios publicados en 2025 y 2026. ChatGPT fue el modelo más evaluado (21 estudios), seguido de Gemini (16) y DeepSeek (12). La precisión fue en general de moderada a alta, pero varió según el modelo y el ámbito clínico, sin que ningún modelo fuera superior de forma constante. La legibilidad fue la debilidad recurrente, con la mayoría de las respuestas por encima del nivel de lectura recomendado de sexto grado. La exhaustividad fue mayor en las respuestas de los chatbots que en las de los clínicos en un estudio comparativo. **Conclusiones:** Los chatbots basados en modelos de lenguaje ofrecen respuestas en gran medida precisas, pero con frecuencia difíciles de leer, a las preguntas de los pacientes odontológicos. Su uso debe ser complementario de la atención profesional y es necesario mejorar la legibilidad antes de que puedan apoyar de forma segura la educación de los pacientes.

Palabras clave: inteligencia artificial; modelos de lenguaje; odontología; educación del paciente; alfabetización en salud; revisión sistemática.

INTRODUCTION

Dental patients increasingly obtain health information from digital sources before and after clinical consultation, and the volume and accessibility of that information shape how they understand and act on oral health advice. Oral diseases remain among the most prevalent conditions worldwide, and a large share of that burden is preventable through behaviours that depend on accurate, comprehensible patient education. When patients search for answers about pain, treatment options or postoperative care, the quality of what they read can determine whether they seek timely care, adhere to instructions, or defer treatment altogether.⁽¹⁾ Written patient information has long been criticised for exceeding the reading ability of the average adult, and the arrival of conversational artificial intelligence has added a new, largely unregulated source to the information patients consult.⁽²⁾

Large language models (LLMs) have changed how patients interact with health information. Unlike static web pages, chatbots such as ChatGPT, Gemini, DeepSeek and Claude generate free-text, conversational answers to questions phrased in ordinary language, and patients have adopted them rapidly because the responses are immediate, personal in tone and free of the navigation effort that conventional websites require.⁽³⁾ In medicine, early evaluations reported that chatbot answers to patient questions were rated more empathetic and, in many cases, at least as accurate as physician-written responses, which accelerated interest in their clinical and educational use.⁽³⁾ The models evaluated in those studies have since been superseded by successive releases, and the performance of the newest models in the dental domain has not been summarised.

The dental literature has grown quickly but unevenly. Narrative and systematic reviews have described the applications of LLMs in dental diagnosis, education and decision support, and a scoping review of implant dentistry catalogued their expanding uses without a quantitative synthesis of patient-facing performance.^(1,4,5) Existing reviews that did address patient education were either confined to a single specialty, such as orthodontics, or synthesised chatbot responses to patient questions across medicine without dental-specific detail.^(6,7) As a result, it remains unclear whether chatbots can be recommended as a source of oral health information, and precisely where their strengths and weaknesses lie.

The rapid release cycle of these models compounds the problem. The evidence base is a moving target: a study that evaluated one version of a chatbot may say little about the version a patient uses a year later, and the proprietary models change without public documentation of what improved. This does not make the evidence unusable, but it means that the stable conclusions are more important than the transient rankings of particular versions. A synthesis that separates the durable findings, such as the readability deficit, from the version-specific rankings is therefore more useful to clinicians and patients than a leaderboard that will be outdated by publication.

The question matters clinically because patient education is a core dental intervention: preoperative instructions, postoperative care and self-management of conditions such as gingivitis all depend on the patient having understood what they were told. Oral health literacy is itself a recognised determinant of outcomes, since adults with low oral health literacy attend less for preventive care and present with more advanced disease. A tool that returns accurate but unreadable answers may be of little help to the patient who asks it a question in plain language, and may be more likely to be trusted and acted upon because of its confident, conversational register.

The clinical context adds urgency to the question. Dental advice is not risk-free when it is wrong or misunderstood: a patient who misreads an instruction about medication after surgery, or about when to return for review of a suspicious lesion, may be harmed in a way that a patient seeking general wellness advice would not. Unlike regulated medical devices, consumer chatbots are not cleared for clinical use, and their answers carry no audit trail or accountability for the advice given. A synthesis of what these tools actually achieve when asked dental questions is therefore a prerequisite for any sensible recommendation about whether patients should use them and how clinicians should respond to the answers they bring to the consultation. The three outcomes chosen here, accuracy, completeness and readability, correspond to the properties that a patient would need an answer to have before it could substitute for, or even usefully complement, professional advice. The objective of this systematic review was to synthesize the evidence on the accuracy, completeness and readability of large language model chatbot responses to patient questions about oral health.

METHODS

Protocol and registration

This systematic review was planned and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) statement.⁽⁸⁾ The protocol was not registered in a public repository, which is acknowledged as a limitation of the review.

Eligibility criteria

Studies were eligible if they evaluated at least one large language model or conversational artificial intelligence chatbot that generated answers to patient-oriented questions about oral health, and if they reported at least one outcome among accuracy, completeness, quality, reliability or readability, assessed against a reference standard such as expert judgement, clinical guidelines or a validated appraisal instrument. Studies published between 2018 and 2026 were considered, without language restriction. Excluded were studies that evaluated chatbots on examination-style or multiple-choice questions intended for clinicians or students, studies of image-based diagnosis without a patient-question component, reviews, editorials, conference abstracts without primary data, and studies outside the oral health domain.

The decision to exclude examination-style questions was deliberate and important to the question of the review. A substantial body of dental work has tested chatbots against multiple-choice and licensing examination questions, which measures factual recall by clinicians rather than the accuracy of answers given to patients in ordinary language. Those studies answer a different question and were therefore set aside so that the synthesis reflected patient-facing use.

In the terminology of the PICO framework, the population was patient-oriented questions about oral health; the intervention or exposure was a response generated by a large language model or conversational chatbot; the comparator was expert judgement, clinical guidelines or a validated appraisal instrument; and the outcomes were accuracy, completeness, quality, reliability and readability. Studies were eligible regardless of the language of the questions or the language of the responses, provided the evaluation was reported sufficiently to be extracted.

Information sources and search strategy

Scopus, PubMed and Crossref were searched. The searches combined terms for large language models and chatbots with terms for dentistry and oral health and terms for patient questions, education, accuracy and readability. In Scopus the query used field codes, for example TITLE-ABS-KEY("large language model" OR "ChatGPT" OR "GPT" OR "generative artificial intelligence") AND TITLE-ABS-KEY("dental" OR "dentistry" OR "oral health") AND TITLE-ABS-KEY("patient"). A second Scopus query added terms for answering, quality, accuracy and readability to increase specificity. Equivalent free-text queries were used in PubMed and Crossref. The two Scopus queries returned four hundred records, PubMed returned two hundred and Crossref returned one hundred, and after deduplication the three sources yielded the set of unique records that entered screening. The searches were run in 2026, and the last search was performed on the date of analysis. A record of each query, its source and the number of records returned was retained for reproducibility, and the full search strings are available from the corresponding author on request.

Selection process

Records retrieved from the three databases were deduplicated by digital object identifier and title. Title and abstract screening was performed with the assistance of an artificial intelligence model, and each decision was subsequently verified by the authors; uncertain records were carried forward to full-text screening rather than excluded. Full texts were retrieved through legitimate open-access routes, and studies whose full text could not be obtained but whose abstract reported sufficient data were retained with abstract-level extraction, as indicated in the results. The screening process was conducted by a single reviewer assisted by the model, without independent duplicate screening, which is reported as a limitation. This assisted screening is consistent with emerging guidance on the use of artificial intelligence in systematic review workflows, while acknowledging that it does not replace human judgement.⁽⁸⁾

Data extraction

For each included study the following items were extracted: first author and year, country, the chatbot models evaluated, the number and source of the questions, the reference standard

or appraisal instrument used, and the reported findings on accuracy, completeness, quality and readability. Where a study reported a numeric score, the value and its scale were recorded. Where a study reported readability, the instrument was recorded, most often the Flesch Reading Ease score, the Flesch-Kincaid grade level or the Ateşman index. Where a study reported inter-rater reliability, the statistic and its value were recorded. Extraction was performed by a single reviewer and checked against the source text, which is reported as a limitation.

Risk of bias and quality appraisal

Because the included studies were predominantly cross-sectional comparative evaluations rather than diagnostic accuracy or intervention studies, no single standard risk-of-bias tool applied uniformly. Methodological quality was therefore appraised across four domains adapted from established instruments for evaluating patient information: whether the question set reflected real patient questions, the number of independent evaluators, whether evaluation was blinded, and whether inter-rater reliability was reported.^(2,9) The appraisal of individual studies on these domains is summarised in the results, and no study was excluded on the basis of quality.

Synthesis

Findings were synthesised narratively by outcome domain. A meta-analysis was not appropriate because the studies used different models, question sets, scales and appraisal instruments. Counts and proportions were used to describe the body of evidence, and representative quantitative values from individual studies are reported where available. Several included studies measured quality with the Global Quality Score, which anchors a five-point ordinal rating of information quality, and several others applied DISCERN or the Patient Education Materials Assessment Tool to score reliability, understandability and actionability.^(9,10) Where studies used these shared instruments their results are directly comparable; where they used bespoke Likert scales the scores are reported as presented.

For each outcome domain, the synthesis distinguishes the direction of the finding from its magnitude. Because the scales were ordinal and heterogeneous, the review reports whether a model was rated highest, lowest or statistically indistinguishable rather than pooling numeric means, and it reports exact values only where a study provided them. This conservative approach prioritises the direction and consistency of the evidence over spuriously precise pooled estimates, and it is the reason the results are presented as ranges and rankings rather than as a single summary statistic.

RESULTS

Study selection and characteristics

The searches identified 700 records, of which 113 were duplicates. Of the 587 unique records screened, 561 were excluded at title and abstract, and 26 reports were sought for retrieval. All 26 reports were assessed for eligibility and all were included, yielding 26 studies in the review (Figure 1). The complete flow of records through each stage is shown in the PRISMA diagram, with the reasons for exclusion recorded at the full-text stage. The 26 included studies were all published in 2025 or 2026; 19 were published in 2026 and 7 in 2025 (Table 1).

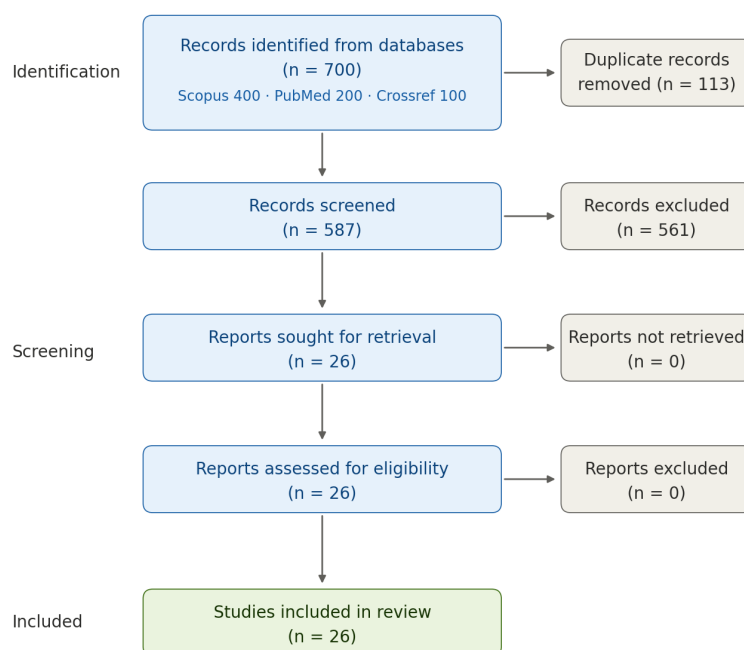
Seventeen of the 26 studies originated in Türkiye, with single studies from Poland, Saudi Arabia, Jordan, Iran and China; four studies did not state the country of origin in the available text. ChatGPT was evaluated in 21 studies, Gemini in 16, DeepSeek in 12, Claude in 7, Microsoft Copilot in 4, Grok in 3, and Perplexity and a domain-specific dental model in 2 studies each. The clinical domains spanned the breadth of dentistry, including implant dentistry, restorative dentistry, paediatric dentistry, endodontics, periodontology, oral surgery, orthodontics and oral medicine. The number of questions posed to each chatbot ranged from 10 to 126, and most studies used between 20 and 60 questions derived from patient-facing question-and-answer lists, guideline documents or online patient forums.

The reference standards and appraisal instruments varied. Expert judgement on Likert-type scales was the dominant approach, but several studies also applied the Global Quality Score, the DISCERN instrument and the Patient Education Materials Assessment Tool, which measure related but distinct constructs of quality, reliability and understandability. Readability was quantified with the Flesch Reading Ease score, the Flesch-Kincaid grade level and, in Turkish-language studies, the Ateşman index. A minority of studies tested a single model, whereas most compared three or more models on the same question set, which permitted within-study comparison of performance.

Table 1. Characteristics of the 26 included studies.

Study (year)	Country	LLMs evaluated	Clinical domain	Questions (n)
Özüdoğru (2026)	Türkiye	ChatGPT-3.5, ChatGPT-4o, Gemini 2.0 Flash, DeepSeek-R1	Anesthesia in dentistry	60
Cangül (2026)	Türkiye	ChatGPT-5, Gemini 2.5 Pro, DeepSeek-V3.2, Claude Sonnet-4.5	Restorative dentistry	25
Kocaoğlu (2026)	Türkiye	Multiple LLMs	Pediatric dental sedation	NA
Mysior (2026)	Poland	Six LLMs (Polish and English)	Six dental disciplines	30
Sözen (2026)	Türkiye	Multiple LLMs	Dental trauma	NA
Alkan Aygör (2026)	Türkiye	AI chatbots (multilingual)	Tooth bleaching	NA
(JORMAS) (2026)	NA	GPT-4o, Claude Sonnet 5, Gemini 2.5 Pro	Periodontal/peri-implant	40
(BMC Oral Health) (2026)	NA	ChatGPT 5.5, Claude 4.6, DeepSeek 3.2, Gemini 3.1 Pro, Grok 4.20, PerioGPT	Periodontal-orthodontic	30
Işık (2026)	Türkiye	ChatGPT-4o, Claude 3.7 Sonnet, Copilot, DeepSeek-V2, Gemini 2.0, Grok-3	Prosthodontics	10
(JORMAS) (2026)	NA	ChatGPT-4.5, DeepSeek, Claude Sonnet 3.7	Dental implants	27
(BMC Oral Health) (2026)	Türkiye	ChatGPT-5, ChatGPT-4o, Gemini-2.5 Flash, DeepSeek-V3, Copilot	Zirconia crowns	20
Bitişik (2026)	Türkiye	ChatGPT (GPT-4)	Pediatric dentistry	20
(JOMI) (2026)	NA	ChatGPT-3.5, Gemini, Copilot	Implant dentistry	20
(JERD) (2026)	Türkiye	ChatGPT, Gemini, DeepSeek	Dental caries	59
Bayram Sonmez (2026)	Türkiye	ChatGPT	Temporomandibular disorders	NA
Arpaci (2025)	Türkiye	ChatGPT-4, Google Gemini	General oral health	NA
Aboalshamat (2025)	Saudi Arabia	ChatGPT-4o mini	Pregnancy oral health	NA
Tuzlalı (2025)	Türkiye	ChatGPT-o1, Deepseek-R1, Gemini, Claude-3.5-Sonnet, Perplexity-Pro	Dental implants	45
Alnsour (2025)	Jordan	ChatGPT 3.5	Crown restorations	126
Esmailpour (2025)	Iran	Three AI chatbots	Prosthodontics	31
Demir Cicek (2025)	Türkiye	DeepSeek V3, GPT 5, Gemini 2.5 Flash	Root canal treatment	37
Ispir (2026)	Türkiye	ChatGPT 5.2, Gemini, DeepSeek V3, Grok 3	General dental questions	33
Altan (2026)	Türkiye	ChatGPT 5.2, Gemini 3, DeepSeek-V3.2	Laminate veneers	25
Zhang (2026)	China	ChatGPT-5, Gemini 2.5, Perplexity Pro, DeepSeek v3.2	Oral cancer	22
Akdeniz (2026)	Türkiye	ChatGPT GPT-5.1, Gemini 2.5 Flash, Claude Sonnet 4.5	Periodontology	20
Özdemir (2025)	Türkiye	ChatGPT-4o, Gemini, Copilot	Deleterious oral habits	43

Note: LLM = large language model; n = number of questions posed to each chatbot; NA = not reported.

Figure 1. PRISMA 2020 flow diagram of study selection.

Accuracy of chatbot responses

Reported accuracy was moderate to high but inconsistent across models and domains. In the studies that used expert ratings on Likert-type scales, mean accuracy or quality scores typically ranged from approximately 3.5 to 4.9 on a five-point scale, with the highest scores concentrated in a small set of models. Gemini achieved the highest accuracy, usefulness and global quality scores in studies of dental caries, zirconia crowns and general dental patient questions.(11-13)

A domain-specific periodontal model and Gemini reached the highest accuracy in a six-model comparison of periodontal and orthodontic questions, with mean scores of 4.93 and 4.90 respectively.(14)

When chatbots were compared directly with clinicians, no significant difference in scientific accuracy was observed between three chatbots and expert periodontologists, although the chatbots produced more complete and more empathetic responses.(15) A study of implant-related questions likewise found no statistically significant difference in accuracy among three chatbots, with total accuracy scores of 28.8, 30.9 and 29.1 out of 40.(16)

In a three-model comparison for periodontal and peri-implant questions, ChatGPT-4o was rated lowest, whereas Gemini received the highest accuracy and clinical safety ratings.(17) No statistically significant difference in reliability or usefulness was found among three chatbots answering dental implant questions, with the only difference emerging in the avoidance of medical jargon.(18)

Several studies evaluated single models or narrower question sets. ChatGPT responses to temporomandibular disorder questions and to parents' questions about children's oral health were assessed for accuracy and temporal consistency, and a large cross-sectional evaluation of six models across six dental disciplines generated 2,700 answers that two dentists independently rated for safety.(19-21)

Accuracy of responses to questions on paediatric dental sedation, dental local and general anaesthesia, and dental trauma were each reported separately, with the models compared against guideline-derived reference answers.(22-24) Answers to questions on the World Dental Federation patient page and on restorative dentistry topics were evaluated against expert responses, with models scoring within the acceptable range in most domains.(25,26) The lowest accuracy ratings were recorded for models asked to handle clinically complex or safety-critical content, and no single model was consistently superior across the full set of studies.

The designs of the remaining studies reinforce the same picture. A study that posed sixty questions on local and general anaesthesia in dentistry, split between patient and professional questions drawn from American guidelines, compared four chatbots on accuracy, reliability and readability and reported that performance depended on the difficulty of the item.(23) A six-model

evaluation of prosthodontic patient education used only ten standardised questions but blinded raters and still detected differences in accuracy and clarity among the models.(27)

A study of paediatric dental sedation questions and one of dental trauma emergencies each assessed guideline concordance and reported that the models were informative but incomplete on specific clinical points.(22,24) A single-model evaluation of ChatGPT for parents' questions about children's oral health tested the same twenty questions repeatedly over seven days and reported temporal variability in readability that settled after the first days, a finding that bears on the stability of chatbot advice.(20)

Evaluations of implant, prosthodontic, denture and root canal questions similarly reported acceptable mean accuracy with domain-dependent gaps, and a study of pregnancy-related oral health combined a survey of three hundred women with expert ratings of the model's responses.(28-31)

Taken together, the accuracy estimates describe a body of evidence in which no model failed outright and none dominated. The between-model differences were statistically significant in several comparisons but small in absolute terms, with the top two models separated by fractions of a scale point in most studies, and the practical importance of these differences is uncertain. The more consistent finding across the accuracy studies is not which model won, but that the models performed better on well-defined, guideline-based questions and worse on questions that required clinical nuance, personalisation or safety judgement. This gradient, rather than any single ranking, is the stable feature of the accuracy evidence and is the one most relevant to how a patient or clinician should interpret a chatbot answer.

Completeness and readability of chatbot responses

Completeness was assessed in a smaller number of studies than accuracy and its operational definition varied. Some studies treated completeness as whether the answer addressed every component of a multi-part question, others as the depth of the explanation offered, and still others as the absence of important omissions relative to a guideline. This inconsistency in definition is itself a finding of the review, because it means that completeness scores across studies cannot be pooled.

Where completeness was reported, the models generally performed well: the comparative study against periodontologists recorded higher completeness for chatbot responses than for expert responses, a statistically significant difference, and the periodontal comparison found completeness above four on a five-point scale for the strongest models.(15,17) Response completeness was highest for Claude in that comparison, with a mean score of 4.77, while Gemini scored highest for clinical safety and overall quality.(17)Completeness and comprehensibility were also rated in studies of implant, prosthodontic and root canal questions, generally alongside accuracy.(16,27,30)

The completeness advantage, however, was not accompanied by better readability, which suggests that the models compensate for their readability deficit by writing longer and more thorough answers rather than clearer ones.

Readability was the most consistent weakness identified across the included studies. In an evaluation of oral cancer patient education, the mean Flesch-Kincaid grade level ranged from 12.65 for ChatGPT-5 to 15.65 for Perplexity Pro, and all models produced text above the recommended sixth-grade reading level.(32)The same pattern was observed in a periodontal and orthodontic comparison, where the most accurate domain-specific model produced the least readable output, with a Flesch Reading Ease score of 12.7 and a Flesch-Kincaid grade level of 15.1.(14)

In a study of dental crown restorations, approximately 53% of ChatGPT responses were classified as difficult to read even though most were rated useful, with a mean Global Quality Score of 4.70 out of 5 and a mean CLEAR reliability score of 24.37 out of 25.(33)

In the Turkish-language evaluations, ChatGPT produced the most readable responses, whereas Gemini achieved higher accuracy but lower readability.(11,13) For questions about deleterious oral habits, ChatGPT-4o answered 93% of questions correctly and 76.7% of its responses were rated as excellent quality, but its responses were significantly less readable than those of Gemini and Copilot.(34)

Readability, quality and consistency were also assessed for questions on tooth bleaching, laminate veneers, dental implant and denture patient questions, and pregnancy-related oral health.(28,29,31,35,36)Across these studies, the recurring result was that the models that produced the most accurate content were not the ones that produced the most readable text, and the two qualities coincided in the same model in only a minority of comparisons.

Language added a further dimension of variability. In the studies conducted in Turkish, the Ateşman index was used to grade readability, and the models that performed best in English did not always translate that advantage into Turkish; ChatGPT produced the most readable Turkish output in two evaluations, while the model with the highest accuracy in one of them was not the most readable.(11,13) A multilingual study of tooth bleaching answers reported that response quality and readability shifted when the same question was posed in a different language.(35) These observations indicate that the accuracy-readability trade-off is not uniform across languages, and that a model ranked highest in one language may not hold that rank in another.

The instrument-based quality scores corroborated the expert ratings. Where the Global Quality Score was applied, mean scores clustered in the upper half of the five-point scale; in the oral cancer evaluation, Perplexity Pro reached a mean score of 4.05 and in the crown restoration study ChatGPT reached 4.70.(32,33) DISCERN-based reliability was similarly favourable for the best-performing model in the oral cancer study, with a mean of 46.36, but this did not extend to readability.(32)

Reliability assessed with the CLEAR tool in the crown study reached 24.37 out of 25, indicating that the content was trustworthy even when its presentation was not accessible.(33)The consistency of these findings across instruments and across expert raters strengthens confidence that the observed accuracy-readability gap is real rather than an artefact of a single measurement approach.

Methodological quality and risk of bias

Methodological quality was heterogeneous (Table 2). Most studies used expert-generated or guideline-derived question sets rather than questions sampled from real patients, which limits ecological validity; a minority drew questions from online patient forums or question-and-answer lists published for patients. The number of independent evaluators ranged from one to nine, and several studies did not report whether evaluation was blinded. Inter-rater reliability was reported in a subset of studies and was generally acceptable, with intraclass correlation coefficients above 0.75 where stated.(11,13)

Table 2. Methodological quality of the included studies across four appraisal domains.

Appraisal domain	Studies reporting (n)
Question set derived from real patient questions or patient-facing FAQs	18
Two or more independent evaluators	21
Blinded evaluation of responses	16
Inter-rater reliability reported	14

The absence of a shared question bank, shared appraisal scales and shared reference standards precluded quantitative pooling, and the resulting heterogeneity is a source of uncertainty in the synthesis. Only one study directly compared chatbot responses with those of clinicians on the same questions, which limits the strength of the claim that chatbots perform at expert level.(15)

The distribution of these quality characteristics shaped how much weight each study could carry. Studies that used patient-derived questions and multiple blinded evaluators provided the most credible estimates, whereas those that used a single evaluator and expert-authored questions risked inflating apparent accuracy, because expert-authored questions tend to be well-defined and therefore easier for a model to answer. The studies that reported inter-rater reliability were generally the same studies that reported the largest question sets, and their estimates are the most defensible in the synthesis. The absence of a formal risk-of-bias instrument tailored to this study type meant that the appraisal remained descriptive rather than a scored summary, which is reported transparently rather than concealed.

DISCUSSION

The main finding of this review is that large language model chatbots produce answers to dental patient questions that are broadly accurate but frequently written at a reading level too high for the intended audience. Across 26 studies published in 2025 and 2026, accuracy and quality scores were moderate to high, yet readability was the most recurrent failure, with most models exceeding the sixth-grade level that patient education materials are recommended to meet.

Completeness was less consistently measured but, where it was assessed, chatbot responses

were at least as complete as expert responses and occasionally more so. The picture that emerges is of a capable but poorly calibrated communicator: the content is present, but the delivery is pitched at the wrong audience. These results place chatbots in a familiar position within patient education: they are reliable in content and deficient in accessibility.

The accuracy findings are consistent with the broader literature on chatbot-generated health information. A systematic cross-sectional synthesis of chatbot responses to patient questions across medicine similarly reported high but variable accuracy alongside poor readability, and concluded that no model could be recommended unconditionally.(7)

The general pattern in which Gemini and more recent model versions outperformed earlier releases has also been described in non-dental specialties, including patient education for mitral valve surgery and frozen shoulder, where accuracy was high but readability remained inadequate. (37,38)

Within dentistry, the present results agree with a systematic review of chatbots in orthodontic patient education, which reported acceptable accuracy but suboptimal readability.(6)The convergence across specialties suggests that the readability deficit is a property of current model behaviour rather than a dental-specific artefact.

The finding that chatbots can match or exceed clinicians on completeness and empathy, while no accuracy difference was observed against periodontologists, mirrors the landmark physician-versus-chatbot comparison, in which chatbot responses were rated more empathetic than physician responses.(3,15)

This comparison should be interpreted cautiously: the dental studies used structured expert questions and blinded raters, conditions that favour the evaluation of content rather than of clinical judgement under uncertainty. The superiority of chatbots on completeness does not establish that they are safer or more appropriate for real patients, because completeness of text is not equivalent to clinical appropriateness for a specific individual.

The readability failure has a plausible mechanism. Large language models are trained predominantly on professional and technical text, and their default register reproduces the vocabulary and sentence structure of that training corpus. When asked a plain-language question, the model often answers at a college or postgraduate reading level unless explicitly instructed to simplify, which explains why grade-level estimates clustered between 12 and 15 rather than near the sixth-grade target. This is a solvable problem in principle, since the same studies that measured readability also showed that prompting and post-processing can lower grade level, but it has not been solved by default in any of the models evaluated.(7,39)

A further consequence of the model-ranking heterogeneity is that no single chatbot can be endorsed as the default for patient use. The highest accuracy scores were split among Gemini, a domain-specific periodontal model and Perplexity depending on the question domain, while the best readability was variously attributed to ChatGPT, Claude and Copilot.

This instability means that a recommendation to use one model for implant questions would not transfer to oral cancer or paediatric questions, and that rankings are sensitive to the version of the model tested, which changed rapidly over the review window. Patients are therefore best served by advice that does not depend on any particular model, and by tools that cross-check answers rather than by loyalty to a single platform. The same instability was reported in evaluations of chatbots in orthodontic and temporomandibular disorder counselling, where prompt wording changed the answer as much as the choice of model.(39,40)

Several sources of heterogeneity limit the conclusions that can be drawn. The studies differed in the models tested, the number and provenance of questions, the scales used to score responses and the choice of reference standard, so the numeric scores are not directly comparable across studies. The dominance of a single country, Türkiye, in 17 of 26 studies means that the evidence reflects a particular set of models, question styles and appraisal conventions, and the results may not generalise to other languages or health systems. The same geographic concentration has been noted in reviews of chatbot evaluation in other health fields, and it cautions against treating the pooled picture as globally representative.(41,42)

Four studies reported data only at abstract level because full text was not retrievable, and their quantitative detail was therefore incomplete.

The findings can be set against the wider evaluation of chatbots in specialised dental education and counselling. Studies of orthodontic consultation, bruxism information, orthodontic information accuracy and prompt sensitivity in orthodontic counselling all reported the same general result: acceptable factual content that depends on how the question is phrased, alongside readability that falls short of patient needs.(39,43-45)A comparative study in temporomandibular disorder education likewise described informative but inconsistently readable answers.(40)

Evaluations in pregnancy-related health and in paediatric chronic conditions outside dentistry have reached the same conclusion about the accuracy-readability trade-off, which supports the external validity of the dental findings.^(46,47) These adjacent findings reinforce the conclusion that the accuracy-readability trade-off is systematic across the dental domain rather than an artefact of any single instrument or model generation.

In clinical practice the most realistic scenario is not that a chatbot replaces the dentist, but that a patient arrives at the consultation having already read a chatbot answer and formed expectations from it. The evidence reviewed here suggests that the clinician can usually assume the answer is broadly correct in content but should be prepared to correct its register and its emphasis. Where the chatbot answer conflicts with the clinician's advice, the clinician should explain the source of the discrepancy, because the patient will have experienced the chatbot as confident and may trust it over a cautious verbal explanation.

None of the studies reviewed tested this downstream effect on trust or adherence, and it remains an important unanswered question, but the accuracy-readability gap identified here provides a concrete reason to anticipate it. Addressing the patient's own reading level, rather than simply restating the correct answer, is the intervention most directly supported by these findings.

The non-dental comparisons are instructive because they isolate the properties of the models from the properties of the dental content. Evaluations of chatbot responses to questions about paediatric health, attention-deficit hyperactivity disorder, pregnancy-related concerns and a paediatric neuromuscular condition each reported that the responses were accurate but written above the recommended reading level, and that model rankings differed by outcome.^(41,42,46,47) The repetition of the same accuracy-readability gap across ophthalmology, cardiology, orthopaedics and now dentistry indicates that the problem is not the domain knowledge of the models, which is adequate, but the register in which they choose to write. This is a design characteristic of current systems rather than a gap in their medical or dental training, and it is the single most tractable target for improvement.

The strengths of this review should be weighed against those limitations. The search spanned three large databases with Scopus as the primary source, the included evidence is recent and homogeneous in its focus on patient-facing questions, and the synthesis distinguishes accuracy, completeness and readability rather than collapsing them into a single score.

The explicit decision to exclude examination-style questions means that the conclusions speak directly to the patient-facing use case, which is the context in which most of the public actually encounters these tools. The transparent reporting of the appraisal domains, and of the studies that could only be extracted at abstract level, allows the reader to judge the weight of the synthesis rather than taking a summary statement on faith.

The practical implications follow from these limitations. Patients who use chatbots for oral health questions are likely to receive answers that are factually sound but written above their reading level, which could reduce comprehension and, in turn, adherence. Clinicians who recommend or tolerate chatbot use should direct patients toward outputs that are rewritten in plain language, or should verify the specific advice before it is acted upon. For developers, the consistent readability failure indicates that prompt-level or post-processing safeguards that lower grade level without sacrificing accuracy should be a priority. For researchers, the field would benefit from a shared, patient-derived question bank and a common appraisal protocol so that future studies can be pooled, and from routine reporting of inter-rater reliability and blinding. These recommendations are the concrete remedy for each of the limitations identified, rather than a generic call for further research.

A further property that bears on real-world use is consistency over time. The included studies that repeated the same question across days reported that responses varied between and within days, particularly for readability, and one study recorded higher response consistency for one model than for its rivals even though the models were similarly accurate.^(16,20) This matters because a patient or clinician who receives a different answer to the same question on different days may reasonably lose confidence in the tool, and because temporal instability complicates any attempt to audit or regulate the advice given. The observation that accuracy was more stable than readability suggests that the volatile component is stylistic rather than factual, which is reassuring for content but still problematic for a tool intended to educate consistently.

The safety dimension deserves particular emphasis. Accuracy and readability are not the only properties that matter when a patient acts on advice, and the included studies that measured clinical safety found that the highest-accuracy models were not always the safest. In the periodontal comparison, Gemini achieved the highest accuracy and safety ratings while a different model scored highest on completeness, which shows that a model can be accurate without being safe, or complete without being appropriate.⁽¹⁷⁾

For safety-critical topics such as dental trauma, sedation and oral cancer, the consequences of a confidently delivered but wrong or unreadable answer can be serious, and the evidence reviewed here does not support relying on any chatbot for these situations without professional verification. This is consistent with the broader medical literature, which has repeatedly cautioned that chatbot answers to patient questions, however fluent, should not be treated as medical advice.(3,7)

CONCLUSIONS

Large language model chatbots provide largely accurate answers to patient questions about oral health, but the completeness of their responses varies and their readability is frequently too low for patient audiences. Readability, rather than factual accuracy, is the principal barrier to their use in patient education. The methodological quality of the underlying studies is uneven, with patient-derived questions, blinding and reliability reporting each absent from a substantial share of the evidence, which tempers the confidence that can be placed in any single estimate.

Until readability is systematically improved and the evidence base is strengthened with shared question banks and common appraisal protocols, chatbots should serve as a supplement to, not a substitute for, professional dental advice.

REFERENCES

1. KAPLAN TT, AKYOL GE. The role of large language models in dental diagnosis, decision-making, and communication: A systematic review. *Japanese Dental Science Review*. 2026;62:57-67. doi: 10.1016/j.jdsr.2026.01.002
2. 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(2):105-111. doi: 10.1136/jech.53.2.105
3. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med*. 2023;183(6):589. doi: 10.1001/jamainternmed.2023.1838
4. Eyüboğlu TF, Özcan M. Large Language Models in Dentistry: A Narrative Review of Reliability, Clinical Utility, and Safety Considerations. *Curr Oral Health Rep*. 2026;13(1):19. doi: 10.1007/s40496-026-00439-5
5. Kong HJ, Kim YL. Large language models in implant dentistry: a scoping review of applications, performance, and limitations. *BMC Oral Health*. 2026;26(1):1492. doi: 10.1186/s12903-026-08747-z
6. Alkadhi OH. Large language models and chatbots in orthodontic patient education: A systematic review and meta-analysis. *International Orthodontics*. 2026;24(4):101212. doi: 10.1016/j.ortho.2026.101212
7. Whittaker P, Sun M. Quality and readability of chatbot responses to patient questions: A systematic cross-sectional meta-synthesis. *Health Informatics J*. 2025;31(4):14604582251388879. doi: 10.1177/14604582251388879
8. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;n71. doi: 10.1136/bmj.n71
9. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): A new measure of understandability and actionability for print and audiovisual patient information. *Patient Education and Counseling*. 2014;96(3):395-403. doi: 10.1016/j.pec.2014.05.027
10. Bernard A, Langille M, Hughes S, Rose C, Leddin D, Veldhuyzen van Zanten S. A Systematic Review of Patient Inflammatory Bowel Disease Information Resources on the World Wide Web. *Am J Gastroenterology*. 2007;102(9):2070-2077. doi: 10.1111/j.1572-0241.2007.01325.x
11. Erdal B, Duman FM, Altinkılıç MR, Horoz M. Artificial Intelligence as a Digital Consultant: Evaluating Large Language Models for Patient Inquiries on Dental Caries. *J Esthet Restor Dent*. 2026;jerd.70284. doi: 10.1111/jerd.70284
12. Acar NK, Sengul F, Bardakci E, Celikel P. Assessing the accuracy, reliability, quality, and readability of artificial intelligence chatbots in patient education: insights from zirconia crowns. *BMC Oral Health*. 2026;26(1):502. doi: 10.1186/s12903-026-07884-9
13. İspir NG, Aslan G, Akay G. Quality, Readability, Understandability, and Actionability of AI Chatbots' Responses to Dental Patient Questions: A Comparative Evaluation. *Diagnostics*. 2026;16(20):3430. doi: 10.3390/diagnostics16203430
14. Çolak R, Çolak MK, Cicek O. Comparative evaluation of AI-driven large language models for periodontal–orthodontic patient questions: accuracy, comprehensiveness, readability, and temporal consistency. *BMC Oral Health*. 2026;26(1):2039. doi: 10.1186/s12903-026-09382-4
15. Akdeniz C, Doğru AG. Comparison of artificial intelligence–based chatbots and expert periodontists in responding to patient questions: a multi-dimensional analysis. *BMC Oral Health*. 2026;26(1):1278. doi: 10.1186/s12903-026-09315-1
16. Yagci F, Eraslan R, Albayrak H, İpekten F. Accuracy and Reliability of Artificial Intelligence Chatbots as Public Information Sources in Implant Dentistry. *Int J Oral Maxillofac Implants*. 2025;41(2):1-23. doi: 10.11607/jomi.11280
17. Ağırbaş R, Yüksekaya V. Comparative evaluation of large language models in answering patient questions following periodontal and peri-implant examination: an expert-based study. *Journal of Stomatology Oral and Maxillofacial Surgery*. 2026;127(6):102956. doi: 10.1016/j.jormas.2026.102956
18. Mavi E, Çevik Sönmez T. Comparing the performance of conversational artificial intelligence models in answering patient questions about dental implants. *Journal of Stomatology Oral and Maxillofacial Surgery*. 2026;127(6):102893. doi: 10.1016/j.jormas.2026.102893
19. Unuvar BS, Unuvar EEO. Evaluating chat generative pre-trained transformer responses to common patient questions on temporomandibular disorders. *Rev Assoc Med Bras*. 2026;72(3):e20251888. doi: 10.1590/1806-9282.20251888
20. Bitisik İN, Savas S, Timarci I, İzgi M, Kucukyılmaz İzgi E. Can a Large Language Model Effectively Answer Parents' Questions About Children's Oral Health?. *Jpr*. 2026;13(3):303-313. doi: 10.4274/jpr.galenos.2026.78871
21. Mysior M, Mysior MP, Maslowski P, Skośkiewicz-Malinowska K. A cross-sectional evaluation of Large Language Model answers to dental questions. *BMC Oral Health*. 2026;26(1):1937. doi: 10.1186/s12903-026-09223-4
22. Kocaoğlu MH, Demirel A, Kaya İ. Accuracy, quality, and readability analyses of responses from large language models to questions on pediatric dental sedation. *BMC Oral Health*. 2026;26(1):492. doi: 10.1186/s12903-026-08026-x
23. Özüdoğru S, Tokgöz Kaplan T. Evaluation of the accuracy, quality, and readability of large language models on local anesthesia and general anesthesia-sedation in dentistry. *BMC Anesthesiol*. 2026;26(1):368. doi: 10.1186/s12871-026-03934-5
24. Sözen E, Akpınar H, Yaman S. Comparative performance of large language models for patient-oriented support in dental trauma emergencies. *BMC Oral Health*. 2026;26(1):665. doi: 10.1186/s12903-026-08037-8
25. Arpacı A, Oztürk AU, Okur I, Sadry S. Evaluation of the accuracy of ChatGPT-4 and Gemini's responses to the World Dental Federation's frequently asked questions on oral health. *BMC Oral Health*. 2025;25(1):1293. doi: 10.1186/s12903-025-06624-9

26. Cangül S, Taşyürek M, Tunç T, Adıgüzel Ö, Ortaç H. From algorithms to empathy: can large language models effectively answer patients' questions in restorative dentistry?. *BMC Oral Health*. 2026;26(1):934. doi: 10.1186/s12903-026-08255-0
27. Işık AD, Yerliyurt K. Comparative efficacy of six contemporary large language models in prosthodontic patient education: a multi-parameter blinded assessment. *BMC Oral Health*. 2026;26(1):1442. doi: 10.1186/s12903-026-08705-9
28. Tuzlalı M, Baki N, Aral K, Aral CA, Bahçe E. Evaluating the performance of AI chatbots in responding to dental implant FAQs: A comparative study. *BMC Oral Health*. 2025;25(1):1548. doi: 10.1186/s12903-025-06863-w
29. Esmailpour H, Rasaie V, Babaei Hemmati Y, Falahchai M. Performance of artificial intelligence chatbots in responding to the frequently asked questions of patients regarding dental prostheses. *BMC Oral Health*. 2025;25(1):574. doi: 10.1186/s12903-025-05965-9
30. Demir Cicek B, Cicek O. Evaluating the Response of AI-Based Large Language Models to Common Patient Concerns About Endodontic Root Canal Treatment: A Comparative Performance Analysis. *JCM*. 2025;14(21):7482. doi: 10.3390/jcm14217482
31. Aboalshamat KT, Alnafeı JH, Alkhatabi LA, Alhawsawi GY, Alahmadi SM, Almalki SO, et al. Can ChatGPT be trusted? Evaluating AI responses to oral health questions among pregnant Arabic-speaking women. *BMC Oral Health*. 2025;25(1):1597. doi: 10.1186/s12903-025-06909-z
32. Zhang B, Shi W, Zhang Y. Performance of Large Language Models in Oral Cancer Patient Education: An Evaluation of Reliability, Readability, and Patient Communication Quality. *Oral Health Prev Dent*. 2026;24:659-667. doi: 10.3290/j.ohpd.c.2786
33. Alnsour MM, Al-Ghannam W, Alraheam IA, Sabrah AHA, Oweis Y, AL-Omiri MK. Assessing the suitability of ChatGPT in responding to public inquiries about dental crown restorations. *BMC Oral Health*. 2025;25(1):1949. doi: 10.1186/s12903-025-07281-8
34. Özdemir ÖT, Kavan MY, Güven Y. Evaluation of the readability, quality, and accuracy of AI chatbot responses to questions about deleterious oral habits. *BMC Oral Health*. 2025;25(1):1812. doi: 10.1186/s12903-025-07298-z
35. Alkan Aygör F, Bahadır HS, Bulut AC, Öztürk C. Performance of AI chatbots in answering questions on tooth bleaching: a multilingual comparative study. *BMC Oral Health*. 2026;26(1):1590. doi: 10.1186/s12903-026-08852-z
36. Altan B, Çınar Ş. Performance of AI chatbots in responding to patients' frequently asked questions about laminate veneer restorations. *BMC Oral Health*. 2026;26(1):2085. doi: 10.1186/s12903-026-09438-5
37. Akdag BB, Bademci MS, Peker I, Karaca GO, Kandemir C, Özcecm B, et al. Performance of large language models in mitral valve surgery patient education: a comparative analysis. *BMC Med Inform Decis Mak*. 2026;26(1):352. doi: 10.1186/s12911-026-03662-3
38. Liang J, Gui W, Li H. Performance of five large language models in frozen shoulder patient education: a multidimensional evaluation of readability, quality, clinical alignment, and safety. *Sci Rep*. 2026;16(1):25296. doi: 10.1038/s41598-026-66627-6
39. Yıldırım E, Tunalı E, Karaçay Ş. Prompt sensitivity of large language models in orthodontic patient counseling: a scenario-based experimental study. *BMC Oral Health*. 2026;26(1):1599. doi: 10.1186/s12903-026-09058-z
40. Özdal Zincir Ö, Hatipoğlu Ş, Çifçi Özkan E. Large Language model (LLM) in temporomandibular disorder education: a comparative study. *BMC Oral Health*. 2025;26(1):112. doi: 10.1186/s12903-025-07468-z
41. Lahut D, Bilgin DD, Başbüyük M, Büke Ö, Karabayır N. Reliability and readability of AI platforms for pediatric health advice: a comparative analysis. *Sci Rep*. 2026;16(1):24131. doi: 10.1038/s41598-026-54812-6
42. Bilgiç B, Turan S, Avcil S, Pekcanlar AA. How do large language models answer ADHD-related questions? A comparative study of ChatGPT, Gemini, and DeepSeek. *BMC Psychiatry*. 2026;26(1):653. doi: 10.1186/s12888-026-08280-x
43. Li J, Ni J, Zheng T, Zuo Z, Wang Y. Evaluating large language models for orthodontic consultation in patients with periodontitis: a study of reliability, quality, and readability. *BMC Oral Health*. 2026;26(1):1160. doi: 10.1186/s12903-026-08448-7
44. Altuntaş MC, Erdinç Akyol G. A bilingual evaluation of chatbot performance in bruxism-related information: accuracy and readability across two models. *BMC Oral Health*. 2026;26(1):1397. doi: 10.1186/s12903-026-08581-3
45. Mughni M, Ilyas MR, Gilani SMA, Fida M, Sukhia RH. Evaluating the informational accuracy of large language models in patient-directed orthodontic retainer guidance: a cross-sectional comparison of Chat GPT-4.1, Gemini 2.5, Microsoft Copilot GPT-4.1 and DeepSeek-V3. *BMC Oral Health*. 2026;26(1):1961. doi: 10.1186/s12903-026-08960-w
46. Keyif B, Yurtçu E, Başbuğ A, Goynumer FG. Evaluation of AI language models in answering pregnancy-related questions assessed by obstetrics specialists. *Sci Rep*. 2026;16(1):9322. doi: 10.1038/s41598-026-40609-0
47. Çelik G, Hekim HH, Filiz MB, Toraman NF, Yapar D. Can ChatGPT provide reliable guidance to families of children with Duchenne Muscular Dystrophy? A dual perspective of parents and experts. *BMC Pediatr*. 2026;26(1):862. doi: 10.1186/s12887-026-07126-5

FUNDING

No financing

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTIONS

Drafting – original draft: Ernesto Learsys Campello Trujillo, Lourdes Hernandez Cuetara, Yivelin Fernandez Queija.

Writing–review and editing: Ernesto Learsys Campello Trujillo, Lourdes Hernandez Cuetara, Yivelin Fernandez Queija.