BMC Medical Informatics
and Decision Making

RESEARCH Open Access

Effectiveness of various general large language models in clinical consensus and case analysis in dental implantology: a comparative study

Yuepeng Wu[1†], Yukang Zhang[2*†], Mei Xu[3], Chen Jinzhi[4], Yican Xue[5] and Yuchen Zheng[1*]

Abstract

Background This study evaluates and compares ChatGPT-4.0, Gemini Pro 1.5(0801), Claude 3 Opus, and Qwen 2.0 72B in answering dental implant questions. The aim is to help doctors in underserved areas choose the best LLMs(Large Language Model) for their procedures, improving dental care accessibility and clinical decision-making.

Methods Two dental implant specialists with over twenty years of clinical experience evaluated the models. Questions were categorized into simple true/false, complex short-answer, and real-life case analyses. Performance was measured using precision, recall, and Bayesian inference-based evaluation metrics.

Results ChatGPT-4 exhibited the most stable and consistent performance on both simple and complex questions. Gemini Pro 1.5(0801)performed well on simple questions but was less stable on complex tasks. Qwen 2.0 72B provided high-quality answers for specific cases but showed variability. Claude 3 opus had the lowest performance across various metrics. Statistical analysis indicated significant differences between models in diagnostic performance but not in treatment planning.

Conclusions ChatGPT-4 is the most reliable model for handling medical questions, followed by Gemini Pro 1.5(0801). Qwen 2.0 72B shows potential but lacks consistency, and Claude 3 Opus performs poorly overall. Combining multiple models is recommended for comprehensive medical decision-making.

Keywords Large language models, Artificial intelligence, Dental implantology, Clinical decision-making, Case analysis

[†] Yuepeng Wu and Yukang Zhang contributed equally to this work.

*Correspondence:
Yukang Zhang
m13013319965@gmail.com
Yuchen Zheng
zhengyuchen@hmc.edu.cn
[1] Center for Plastic & Reconstructive Surgery, Department of Stomatology, Zhejiang Provincial People's Hospital, Affiliated People's Hospital, Hangzhou Medical College, Hangzhou, Zhejiang, China
[2] Xianju Traditional Chinese Medicine Hospital, Taizhou, Zhejiang, China
[3] Hangzhou Dental Hospital, West Branch, Hangzhou, Zhejiang, China
[4] College of Oceanography, HoHai University, Nanjng, Jiangsu, China
[5] Hangzhou Medical College, Hangzhou, Zhejiang, China

Introduction

Artificial Intelligence (AI) is the capability of machines or computer systems to perform tasks that typically require human intelligence [1]. It involves simulating human-like intelligence in machines through tasks such as logical reasoning, learning, problem-solving, and decision-making.In 2022, generative AI systems like ChatGPT, based on Large Language Models (LLMs), were officially launched“. These models aim to simulate human conversations, understand the meanings of words and phrases like humans, and use this understanding to generate new content based on training data [2]. LLMs have demonstrated their effectiveness in semantic parsing by converting natural language into logical forms to complete complex reasoning tasks, enhancing their effectiveness in dental medical Q&A systems and robotic planning tasks [3]. This highlights the potential and transformative capabilities of LLMs in Natural language processing(NLP) and dentistry.

During clinical consensus and case analysis, models like ChatGPT have already shown some clinical decision-making capabilities. Stephanie Cabral et al. used 20 clinical cases to evaluate the ability of LLMs to handle medical data and demonstrate clinical reasoning, comparing them with attending physicians and residents.Furthermore, the evolution of AI has significantly impacted the roles of dental healthcare professionals, with studies showing how AI-powered telehealth solutions are transforming traditional workflows and responsibilities of dental assistants and nurses in orthodontic care [4]. The results showed that LLM were superior to doctors in handling medical data and using a recognizable framework measured by the R-IDEA score for clinical diagnosis. Other clinical decision-making results showed no statistical significant difference between doctors and chatbots [5]. In providing medical information, Yazid K Ghanem et al. evaluated the content and quality of medical information on acute appendicitis generated by ChatGPT-3.5, Bard (later rebranded as Gemini), and Claude2. The results showed that various LLM scored high in quality assessment of appendicitis medical information, with overall readability far exceeding public recommendation levels [6]. These findings demonstrate that the potential of various LLMS like ChatGPT in medicine has been validated.

In the field of dentistry, Hossein Mohammad-Rahimi et al. tested the validity and reliability of ChatGPT, Bard, and Bing on endodontics-related questions, and all three chatbots achieved acceptable reliability levels (Cronbach’s alpha > 0.7) [7]. Arman Danesh et al. asked ChatGPT3.5 and ChatGPT4 questions from three different sources: INBDE Bootcamp, ITDOnline, and a board-style question list provided by the National Dental Examination Board [8]. This demonstrates the potential of various general models as tools for daily dental diagnosis and research.

The above studies invited relevant experts to evaluate the results of various LLM on specific questions using different forms of Likert scales. However, the information output by various LLM, as part of medical information, needs to meet the requirements of carrying medical text information (completeness, practicality, reliability) and the requirements of medical text attributes (clarity, simplicity, logic, authority, and adherence to relevant legal and medical ethical principles). Additionally, as an extension of the NLP field, the traditional evaluation criteria for various artificial intelligence are precision, recall, and the F1 score calculated from both. These objective indicators were not reflected in the above studies, leading to a lack of comprehensive quantitative evaluation of model assessment, which is an area that needs further exploration.A recent comprehensive review by Tomášik et al. [9] emphasized the need for more rigorous evaluation methods when assessing AI applications in orthodontics, particularly in quantitative aspects of model performance assessment.

In the field of dental implants, the auxiliary capabilities of various LLM have not been explored.Dental implants, as a branch of dentistry, combine complex knowledge from oral surgery, periodontics, and prosthetics, requiring professional doctors to have extensive medical knowledge and excellent clinical decision-making abilities [10–12]. However, global oral healthcare resources, especially for doctors capable of implant restoration, are extremely unevenly distributed.The emergence of LLM provides an opportunity to break this imbalance.Various LLM such as Gemini, Claude 3, and Qwen series, trained on different web corpora, have not yet been explored for their capabilities in the field of dental implants.The purpose of this study is to quantitatively evaluate and compare the performance of various LLM in the field of dental implants and the performance of the same LLM on questions of different difficulty levels, thereby comprehensively exploring the ability of existing LLM to transition from clinical consensus to scenario applications in the field of dental implants.

Materials and methods

Research design

This study implemented a two-phase evaluation approach to assess LLM performance in dental implantology, focusing on both NLP capabilities and clinical reasoning from February 2024 to May 2024 (Fig. 1). The first phase focused on basic knowledge assessment through 20 simple questions (10 true/false and 10 numerical), derived from the International Team for Implantology (ITI) Clinical Guidebook Series. All four models (ChatGPT-4 (OpenAI, Web Version), Gemini Pro 1.5

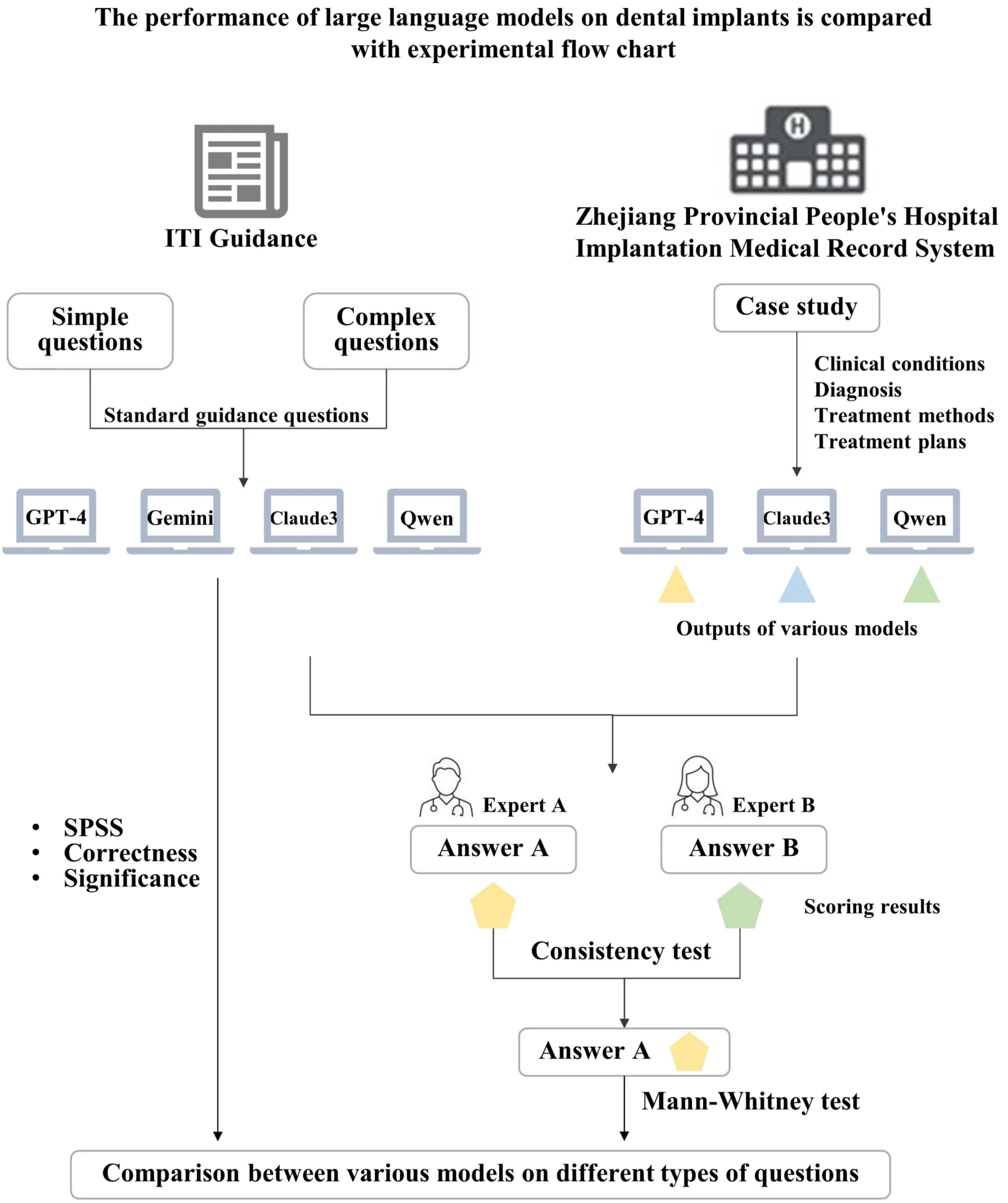
The performance of large language models on dental implants is compared with experimental flow chart
ITI Guidance
Zhejiang Provincial People's Hospital Implantation Medical Record System
Simple questions
Complex questions
Standard guidance questions
Case study
Clinical conditions
Diagnosis
Treatment methods
Treatment plans
GPT-4
Gemini
Claude3
Qwen
GPT-4
Claude3
Qwen
Outputs of various models
Expert A
Expert B
Answer A
Answer B
Scoring results
SPSS
Correctness
Significance
Consistency test
Answer A
Mann-Whitney test
Comparison between various models on different types of questions

Fig. 1 Flowchart of overall study design

(0801) (Google DeepMind, API Version), Claude 3 Opus (Anthropic, Web Version) and Qwen 2.0 72B (Alibaba Cloud, API Version)) were tested using a binary scoring system to evaluate their understanding of fundamental dental implant concepts.

The second phase examined clinical reasoning capabilities through complex short-answer questions and case analyses. Each model processed 20 complex questions covering diagnosis, treatment planning, and clinical decision-making. Additionally, six comprehensive clinical cases were analyzed, encompassing scenarios such as immediate implant placement, maxillary sinus lift procedures, and guided bone regeneration. Due to platform limitations in processing extended clinical scenarios, Gemini Pro 1.5(0801) was excluded from the case analyses portion. Responses were evaluated using a 12-point scoring system that assessed both content (completeness and reliability, 6 points) and structure (including practicality, authority, and clarity, 6 points), as illustrated in the evaluation pathway of Fig. 1.

Two dental implant specialists with over 15 years of clinical experience independently evaluated all responses. Intraclass Correlation Coefficient (ICC) analysis confirmed inter-rater reliability (complex questions: 0.965; case analyses: 0.797), with regular calibration sessions ensuring consistency in scoring. All clinical cases were anonymized and selected from real patient scenarios, maintaining compliance with ethical guidelines while representing diverse clinical challenges in implant dentistry.

Example revision

“In this study, we evaluated various LLMs for their ability to answer dental implant-related questions. The models used in this study were as follows:

1. ChatGPT-4: A state-of-the-art conversational AI model developed by OpenAI. We utilized the web-based version of ChatGPT-4 to simulate conversational interactions in dental implantology. This version is capable of processing both text and complex medical queries in natural language through an interactive web interface.
2. Gemini Pro 1.5 (0801): This version of Gemini is used via its API and offers capabilities in contextual understanding and multi-step reasoning. The Pro version was selected for its enhanced handling of complex medical queries, such as those related to dental implantology.
3. Claude 3 Opus: A conversational agent developed by Anthropic, Claude-3 is designed to produce high-quality text responses with an emphasis on clarity and reliability in medical contexts.
4. Qwen 2.0 72B: This model, designed for general-purpose language understanding, was accessed through Alibaba’s API to provide insights into implant-related medical queries based on a wide range of medical and dental training data.”

Question selection and case sampling methodology

The study material was systematically sourced from two primary channels: the International Team for Implantology (ITI) Clinical Guidebook Series and the hospital’s clinical database. From the ITI guidelines, questions were extracted using a structured screening protocol that assessed their relevance to current clinical practice and alignment with consensus statements. Two independent reviewers conducted the initial screening, with a third reviewer resolving any disagreements. Questions were selected to ensure comprehensive coverage of implant dentistry’s core domains while maintaining balanced representation across different complexity levels.

Clinical cases were identified through a stratified random sampling of the hospital’s database (2020–2023), with stratification based on the SAC Classification system (straightforward, advanced, complex). Each selected case required complete documentation including CBCT scans, clinical photographs, and minimum one-year follow-up data. To ensure representativeness, cases were stratified by anatomical location (anterior/posterior, maxilla/mandible) and implant placement timing.

The validation process employed a modified Delphi (A structured approach to achieving consensus through multiple rounds of anonymous expert assessments - to determine evaluation criteria. The panel, made up of X dental professionals with more than Y years of experience, determined the final criteria after a Z-round evaluation.)technique with two rounds of independent review by external implantologists. This systematic approach assessed content validity while minimizing selection bias. A pilot phase ($n = 8$) verified the clarity and applicability of both questions and cases, with refinements made based on quantitative and qualitative feedback from pilot participants. Cases with incomplete documentation or unclear treatment outcomes were excluded from the final selection.

Question design and evaluation criteria

Simple questions

Simple questions are divided into true/false questions and numerical fill-in-the-blank questions. These questions aim to test the model’s understanding and accuracy of clear, single answers. For true/false questions, given their clear judgment points, the study used binary (yes/no) questions with specific prompt formats (only yes or no answers) for the models to answer. For numerical answers or numerical range questions, due to the rigor of medical issues, the model's response must match the answer in the guidebook or fall within the range of values provided in the guidebook. For single-element fill-in-the-blank questions, the model must provide an answer conceptually identical to the intended meaning to be considered correct.

Complex questions and case analyses

Complex questions and case analyses require multifaceted answers. To evaluate these responses comprehensively, we developed a scoring system based on established medical text evaluation principles and the EQIP (Expanded Quality Information Patient) scale [21]. Our evaluation framework assessed both content quality through completeness [13, 14] and reliability [14, 15], as well as structural elements including clarity [14, 16], simplicity [14, 17], logic [14, 18], authority [14, 19], and adherence to relevant legal and medical ethical principles. For objective assessment of text credibility, we incorporated precision and recall metrics based on Bayesian inference, which are commonly used in evaluating AI-generated texts [20].

Our evaluation methodology involved expert evaluators identifying key clinical concepts from ITI guidelines as reference standards. Each model's response was analyzed by comparing it to these standards, with each key point representing a distinct clinical concept, diagnostic criterion, or treatment consideration [14, 15]. We tracked three elements in this analysis: correct information present in the model's answer (True Positives), incorrect or irrelevant information added by the model (False Positives), and important information from guidelines omitted by the model (False Negatives). Completeness scores were calculated by measuring the proportion of correct information included in the model's answer relative to all information that should have been included (multiplied by 3 points). Reliability scores were calculated by measuring the proportion of correct information provided by the model relative to all information in the model's answer (multiplied by 3 points) [20]. This mathematical approach ensured objective evaluation of the models' responses while adhering to established medical text assessment principles [13, 14].

Table 1 Scoring rules for medical texts output by LLM

Evaluation criteria	Full score
Text Content	
Completeness (Should provide comprehensive information)	3 points
Reliability (Whether the answers given are reliable)	3 points
Text Structure	
Practicality (Whether the answer provides detailed explanations for practical situations)	1 point
Authority (Whether authoritative citations are provided when quoting answers)	1 point
Compliance with Laws and Ethical Principles (Whether relevant experts should be consulted/Whether medical ethics should be consulted when dealing with complex issues)	1 point
Clarity (Whether there are any ambiguities or contradictions)	1 point
Logic	1 point
Simplicity	1 point
Total Score	

Based on the EQIP scale [21], the structural evaluation encompasses six criteria, each worth one point: practicality (detailed explanations for practical situations), authority (presence of authoritative citations), compliance with legal and ethical principles (recognition of expert consultation needs), clarity (absence of ambiguities), logic (clear reasoning), and simplicity (accessible presentation). A set of scoring criteria for medical texts generated by LLMs was developed based on these considerations, as shown in Table 1.

This evaluation form is divided into two parts: text content and text structure. Evaluators need to first annotate and break down the key points of the standard answers from the guidelines and the responses from various LLMs. Then, they calculate True Positives (TP) = factors in the model's answers that match the standard answers; False Positives (FP) = new factors added by the model that are not mentioned in the standard answers; False Negatives (FN) = factors mentioned in the standard answers but omitted in the model's answers. This expresses the comprehensiveness of the model's responses. Completeness corresponds to the model's recall rate, which is the number of FN (cases where ChatGPT did not provide correct medical advice but was actually needed). Precision (with a maximum score of 3) is calculated as TP (fully matched information) / (TP + FP) * 3, representing the accuracy of the model's provided medical advice or answers that match professional guidelines or are considered accurate. Reliability is expressed by precision (with a maximum score of 3), calculated as TP / (TP + FN) * 3, indicating how much of the model's advice aligns with the guidelines.

The study compares the key points of complex questions in the guidelines and case(In the actual medical records, the personal information of patients was anonymized, retaining only the chief complaint, present illness history, past medical history, family history, physical examination, and auxiliary examination parts that reflect the patient's condition. These were input into various large-scale models, requesting them to output diagnoses, treatment plans, and optimal treatment strategies.) analyses with the diagnoses, treatment plans, and suitable treatment plans provided for the patients, matching the key points of the model's answers with those in the guidelines.

Text completeness indicates the elements mentioned in the standard guidelines but not omitted by the model. Reliability represents the proportion of factors mentioned by the model that are consistent with the guidelines. Practicality evaluates whether the model can provide valuable answers based on specific analysis of the current issue, such as whether the model can analyze specific problems when providing follow-up analysis. Authoritativeness checks whether the model can provide relevant authoritative citations when adding answers. Legal and ethical principles assess whether the model acknowledges its limitations as an AI and reminds users to consult relevant medical professionals. Clarity evaluates whether the output text is unambiguous, with appropriate sentence and word meaning. Logic assesses whether the model's response is well-organized. Accessibility checks whether the expressions used are free from obscure terms and long, complex sentences.The scoring proportions for each criterion in this study refer to the EQIP expansion scale, which balances the scores for medical content and text structure at 50% each.

The present study invited two dental implant experts to use the above rules to annotate the key points in the guidelines and compare them with the key points extracted from the anonymized responses of the models to determine precision and recall. They further reviewed the model's responses to complex questions and scored them based on their clinical experience and the patient treatment records, considering the diagnosis, chief complaint, and the most suitable treatment plan. The results were saved. Responses from models that completely misunderstood or hallucinated were scored as 0.

Through this design, this study ensures the scientific rigor and accuracy in evaluating the performance of various LLMs in the field of dental implantology.

Statistical methods

We implemented a systematic statistical framework to evaluate LLM performance in dental implantology:

Sample size calculation

To evaluate the adequacy of sample sizes, we conducted preliminary studies with various models. For simple questions, with an expected average score of 9/12 points (SD = 0.1), using the general formula [22] with 95% confidence level (Z = 1.96) and margin of error of 0.1, the minimum sample size was calculated to be 4. For complex questions and case analyses (margin of error = 0.1, SD = 0.2), the minimum sample size was 16. To improve reliability and account for question diversity, we uniformly set the sample size at 20 for each group Tables (2 and 3).

Table 2 Presentation of simple and complex questions posed to different general LLM

Question ID	Question content
1	What are the patient-identifying factors required for aesthetic implant restoration of anterior teeth?
2	How to obtain long-term soft tissue stability in aesthetic implant of anterior teeth?
3	What are the biological types of gum in the implant treatment area?
4	What are the requirements for transitional dentures during the healing period?
5	What are the types of barrier membranes used for bone defects around implants? What are the advantages of each?
6	How many millimeters from the adjacent tooth is required for the ideal placement of the implant in the proximal and distal aesthetic areas of the tooth?
7	How long after tooth loss is delayed dental implant placement typically performed?
8	What is the average depth of the alveolar crest square soft tissue (including gingival groove depth) around the implant?
9	What is the torque range required for immediate implantation to sustain the initial stability of a single tooth?"
10	Does insufficient keratinized gingival mucosa (KAM) play a decisive role in maintaining peri-implant soft tissue health while ensuring good oral hygiene?

Table 3 Presentation of clinical cases posed to different general LLM

Case ID	Case focus
1	Immediate implant placement: Diagnosis, treatment planning, therapeutic strategies
2	Maxillary sinus lift implant: Diagnosis, treatment planning, therapeutic strategies
3	GBR surgery: Diagnosis, treatment planning, therapeutic strategies
4	Immediate loading of implants: Diagnosis, treatment planning, therapeutic strategies
5	Early loading implant surgery: Diagnosis, treatment planning, therapeutic strategies
6	All-on–6 procedure for advanced periodontal disease: Diagnosis, treatment planning, therapeutic strategies

Data collection and reliability

To evaluate the scientific validity of the scoring method, this study first had two experts score the aforementioned complex questions and case analyses based on the evaluation criteria. The Intraclass Correlation Coefficient (ICC) test was used to compare the consistency of their scores across different types of questions.

Statistical analysis methods

For model response evaluation, different statistical approaches were applied to analyze simple questions and complex questions/case analyses. For simple questions, we calculated the exact match ratio and its 95% confidence intervals (Wilson score interval) and used t-tests to analyze between-group significance. For complex questions and case analyses, all answers were stored in Excel spreadsheets and analyzed using Mann-Whitney U tests for pairwise comparisons between models to determine significant differences in response quality. Data visualization and graphing were performed using GraphPad Prism 8.0.1 software.

Results

Consistency

The consistency of the ratings (shown in Table 4) was assessed using the Intraclass Correlation Coefficient (ICC). The average consistency of the two dentists' scores on complex problems and case analyses was found to be very high, with ICC values of 0.965 and 0.797, respectively. This indicates that the average agreement among multiple raters was strong. For both complex questions and case analyses, the dentists, who had similar expertise, showed high consistency in their ratings.

Single Measure ICC (A, 1): Evaluates agreement between a single rater's judgment and the true value (lower values here suggest less agreement).

Average Measure ICC (A, K): Evaluates agreement when considering the average of multiple raters' judgments (higher values suggest stronger reliability).

Table 4 Inter-rater reliability analysis of dentists' ratings (ICC analysis)

Analysis type / Type	ICC	95% CI
Complex Problems		
- Single Measure ICC (A,1)	0.396	0.162–0.904
- Average Measure ICC (A, K)	0.965	0.890–0.997
Case Analysis		
- Single Measure ICC (A,1)	0.136	0.065–0.289
- Average Measure ICC (A, K)	0.797	0.633–0.911

Simple questions

Table 5 shows the performance of each model on simple questions, including the significance of the comparisons between models. Gemini Pro 1.5(0801) achieved the highest accuracy rate (0.80), while Qwen 2.0 72B had the lowest accuracy rate (0.60). ChatGPT-4 and Claude 3 Opus had accuracy rates of 0.74 and 0.72, respectively. The standard deviations for all models were similar, ranging from 0.05 to 0.07.

When comparing the models, there were no significant differences between ChatGPT-4 and Gemini Pro 1.5(0801), or between ChatGPT-4 and Claude 3 Opus. However, significant differences were observed between ChatGPT-4 and Qwen 2.0 72B, and between Qwen 2.0 72B and Gemini Pro 1.5(0801). The comparison between Qwen 2.0 72B and Claude 3 Opus showed a difference close to significance.

Table 5 Significance testing of different models on simple questions

Group comparison	*p*-value
ChatGPT-4 vs. Qwen 2.0 72B	0.035
ChatGPT-4 vs. Claude 3 Opus	0.752
ChatGPT-4 vs. Gemini Pro 1.5(0801)	0.316
Qwen 2.0 72B vs. Claude 3 Opus	0.074
Qwen 2.0 72B vs. Gemini Pro 1.5(0801)	0.002
Claude 3 Opus vs. Gemini Pro 1.5(0801)	0.187

Table 6 Accuracy of different models on simple questions

Model	Accuracy	CL95% (Wilson Score Interval)
ChatGPT-4	0.74	0.604 to 0.841
Qwen 2.0 72B	0.6	0.462 to 0.724
Claude 3 Opus	0.72	0.583 to 0.825
Gemini Pro 1.5(0801)	0.8	0.670 to 0.888

Complex questions

We compared the performance of four LLMs (ChatGPT-4, Qwen 2.0 72B, Claude 3 Opus, and Gemini Pro 1.5) on complex dental implantology questions. The evaluation included comprehensive statistical analysis with metrics such as average score, standard deviation, and percentile distributions. Table 6 summarizes these performance metrics, while Fig. 2 illustrates the significant differences between models. Our analysis revealed that ChatGPT-4 significantly outperformed Claude 3 Opus ($p = 0.001$), and Gemini Pro 1.5(0801) performed significantly better than Claude 3 Opus ($p = 0.033$). No significant differences were found between other model pairs (all $p > 0.05$) (Tables 7A and 7B).

SASDPLLM: Statistical Analysis of Scores in Diagnostic Performance of LLM.

SASTRLLM: Statistical Analysis of Scores in Therapeutic Regimen of LLM.

SASTSLLM: Statistical Analysis of Scores in Therapeutic Schedule of LLM.

SASCPMLM: Statistical Analysis of Scores for Complex Problems of LLM.

The specific case analysis results are shown in Fig. 3. In the diagnostic module, across 120 trials, Claude 3 Opus had an average score of 9.88 with a standard deviation of 1.56, while ChatGPT-4 had an average score of 9.83 with a standard deviation of 2.10, showing greater variability than Claude 3 Opus. The Qwen algorithm showed the highest average score of 10.90, combined with the smallest standard deviation of 0.68, indicating higher consistancy in its scores. Statistical analysis of significance showed that Qwen 2.0 72B had significant differences compared to Claude 3 Opus ($p = 0.001$) and ChatGPT-4 ($p = 0.002$), while the p-values for comparisons with other groups were all greater than 0.05.

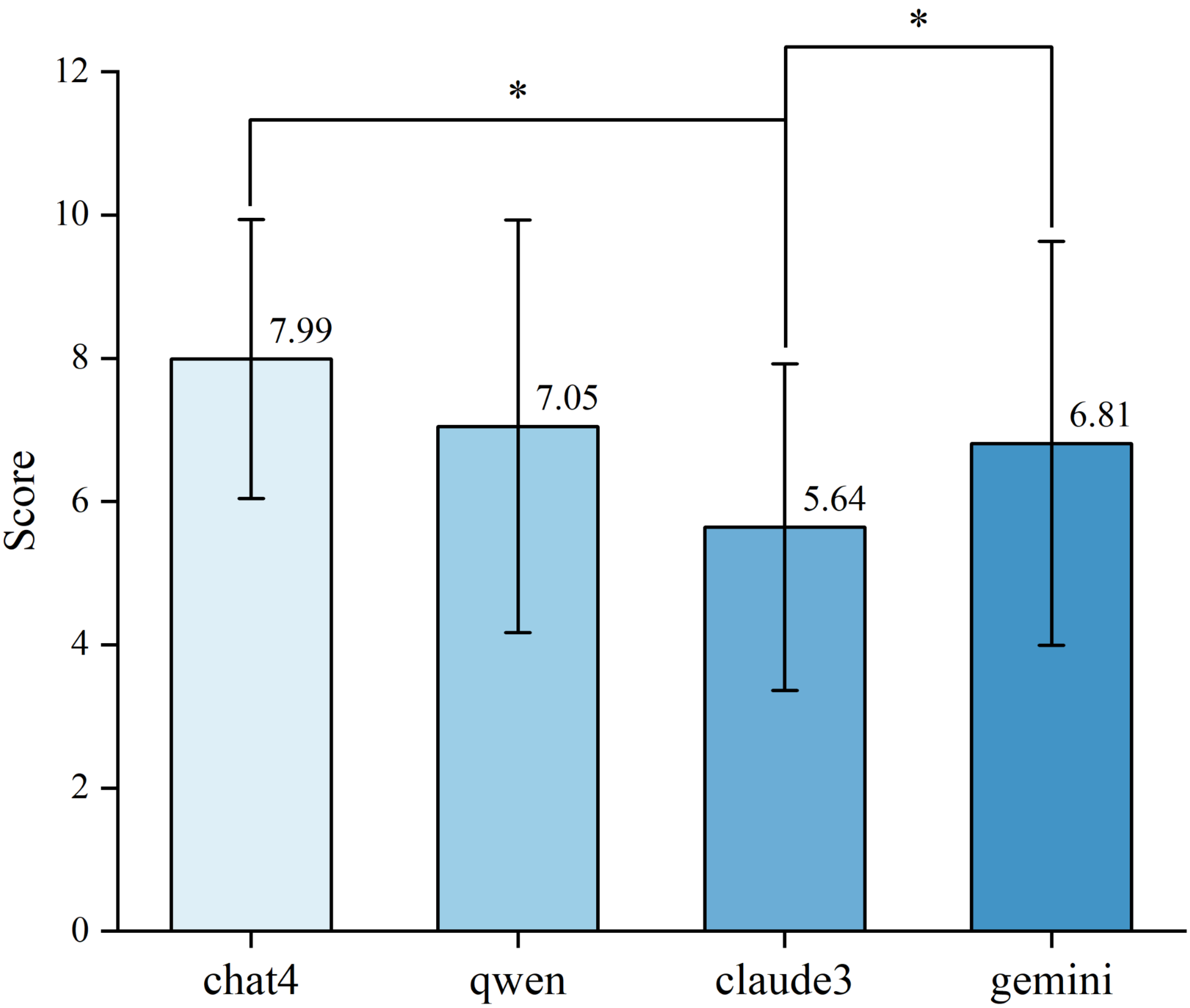

Fig. 2 The performance and significance of different models in diagnostic scores

Table 7A A statistical analysis of LLM performance in diagnostic and treatment planning

Model	Metric	Mean	Standard deviation	Minimum	Maximum
ChatGPT4	SASDPLLM	9.83	2.10	3.00	12
Qwen 2.0 72B	SASDPLLM	10.9	0.68	9.30	12
Claude 3 Opus	SASDPLLM	9.88	1.56	3.00	12
ChatGPT4	SASTRLLM	9.28	2.41	3.00	12
Qwen 2.0 72B	SASTRLLM	9.07	2.24	5.00	12
Claude 3 Opus	SASTRLLM	9.24	2.41	3.00	12

Table 7B Statistical analysis of LLM performance in treatment scheduling and complex problem solving

Model	Metric	Mean	Standard deviation	Minimum	Maximum
ChatGPT4	SASTSLLM	10.02	1.23	4.5	12.00
Qwen 2.0 72B	SASTSLLM	7.89	4.46	0	12.00
Claude 3 Opus	SASTSLLM	9.83	2.31	3.00	12.00
ChatGPT4	SASCPMLM	7.99	1.95	5.00	10.00
Qwen 2.0 72B	SASCPMLM	7.05	2.88	0	11.40
Claude 3 Opus	SASCPMLM	5.64	2.28	2.00	12.00
Gemini Pro 1.5(0801)	SASCPMLM	6.81	2.82	2.00	12..00

Regarding treatment plans and planning modules: Claude 3 Opus had average scores of 9.41 and 9.83, with standard deviations of 2.43 and 2.31, respectively. ChatGPT-4 had average scores of 9.28 and 10.02, with

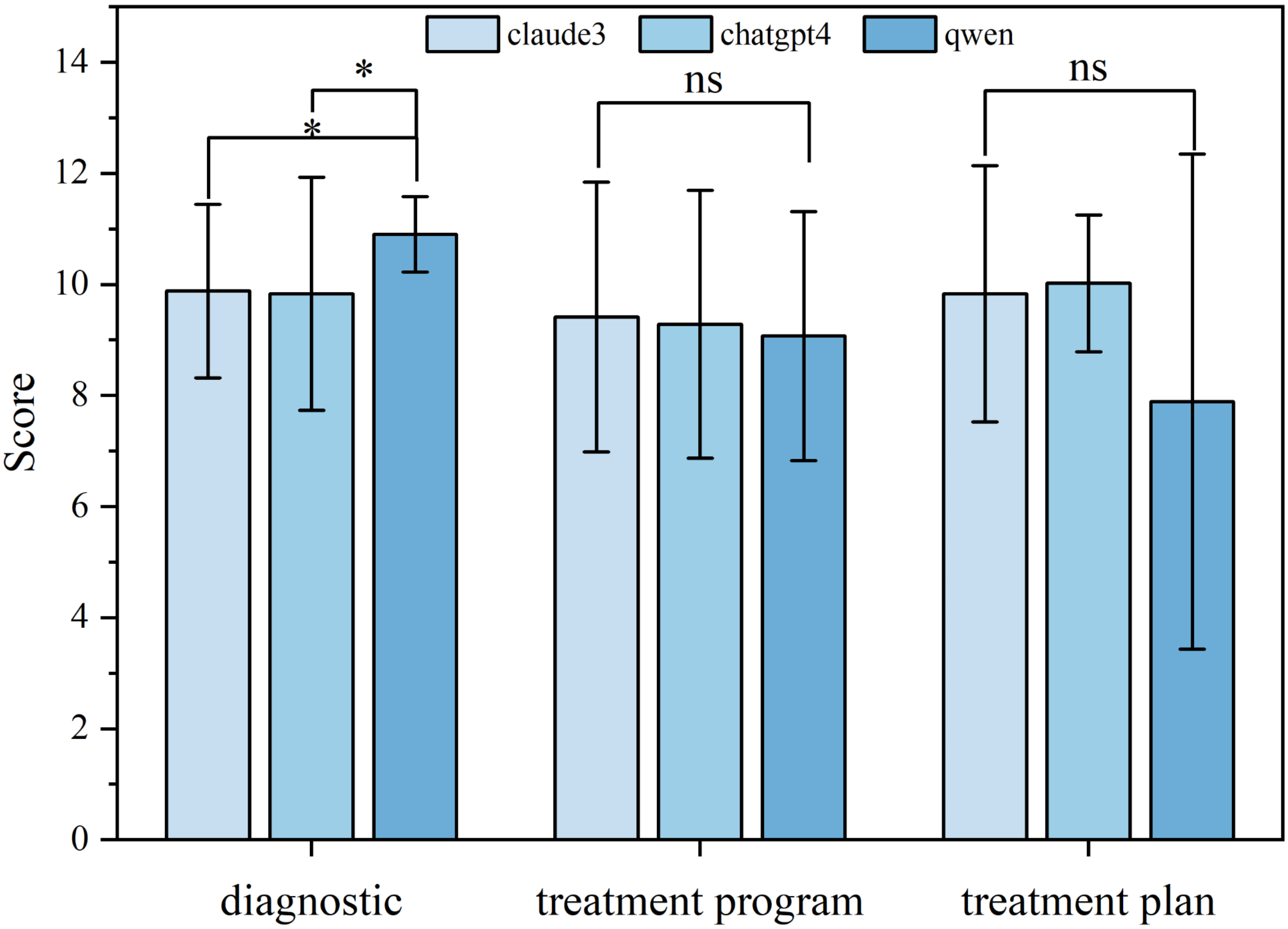

Fig. 3 Comparison of scores across LLMs in diagnostic, treatment program, and treatment plan domains

standard deviations of 2.41 and 1.23, respectively. Qwen 2.0 72B had average scores of 9.07 and 7.89, with standard deviations of 2.24 and 4.46, respectively. In terms of p-values, the scores given by different models for treatment plans and planning modules did not show significant differences.

Discussion

This study proposes a novel evaluation method that objectively assesses AI-generated medical text while incorporating traditional AI metrics. Our findings reveal distinct performance patterns among the tested models across different question types.

For simple medical questions, Gemini Pro 1.5(0801) achieved the highest accuracy (0.80) with minimal variation (SE = 0.057). This performance aligns with findings from Masalkhi et al. in ophthalmology [23] and Mihalache et al. in general medicine [24, 25]. ChatGPT-4 and Claude 3 Opus showed comparable performance (accuracy: 0.74 and 0.72 respectively), while Qwen 2.0 72B demonstrated lower accuracy (0.60).

In complex question analysis, ChatGPT-4 achieved the highest mean score (7.99 ± 1.95), demonstrating consistent performance. This reliability mirrors Walker et al.'s findings in hepatopancreatobiliary medicine [26]. Qwen 2.0 72B showed promise with the highest median score (8.5) but greater variability, suggesting potential for specific applications. Claude 3 Opus performed significantly lower than both ChatGPT-4 ($p = 0.001$) and Gemini Pro 1.5(0801) ($p = 0.033$), possibly due to its Constitutional AI architecture emphasizing safety over specialized medical reasoning [27] and limited professional domain adaptation [28].

Differences in diagnostic and treatment planning capabilities and their clinical implications

Diagnostic capabilities analysis revealed that Claude 3 Opus and ChatGPT-4 maintained consistent performance (9.88 ± 1.56 and 9.83 ± 2.10 respectively), though Claude's strict evaluation criteria may affect its clinical decision flexibility (Sonoda et al., 2024). Qwen 2.0 72B showed the highest average score (10.9 ± 0.68) but with notable fluctuations in treatment planning (7.89 ± 4.46). This variability highlights important boundaries for AI systems in clinical applications: while current models excel at identifying and analyzing symptoms, their reliability decreases when required to develop individualized treatment plans integrating multiple factors.

This imbalance between diagnostic and treatment planning capabilities has direct clinical implications—in dental implantology, models may accurately determine patient indications but lack necessary clinical judgment when selecting specific implant systems or surgical approaches. This suggests dental practitioners can consider using AI tools to assist with initial assessments, but final treatment decisions should remain professional-led. This disparity becomes particularly prominent in complex cases involving insufficient bone volume or occlusal disorders.

Statistical analyses revealed significant differences in diagnostic performance (Qwen 2.0 72B vs. Claude 3 Opus: $p = 0.001$; vs. ChatGPT-4: $p = 0.002$) but not in treatment planning ($p > 0.05$), indicating similar treatment planning capabilities across models with room for improvement in all.

Study limitations include the focused scope within dental implantology and the reliance on ITI guidelines for evaluation standards. The models' limited explanation of technical aspects may reflect constraints in their training data regarding specialized implant knowledge.Beyond current clinical applications, AI's potential in implant dentistry extends to biomaterials development. Recent work by Thurzo and Varga demonstrates the integration of AI with advanced 4D biomaterials for personalized implant scaffolds, highlighting opportunities for interdisciplinary innovation [29]. Future research should explore such cross-disciplinary approaches to advance patient-specific implant treatments, particularly in complex cases involving insufficient bone volume or occlusal disorders.

Conclusion

This study introduces a novel evaluation method, providing an objective assessment of LLMs in dental implant research. ChatGPT-4 demonstrates the most consistent performance, while Gemini Pro 1.5(0801) and Qwen 2.0 72B show variable results, and Claude 3 Opus performs poorly. The findings highlight significant differences between models, recommending their use as auxiliary tools in medical decision-making, with a focus on integrated results from multiple models.

Abbreviations

AI	Artificial Intelligence
LLM	Large Language Model
NLP	Natural Language Processing
ITI	International Team for Implantology
TP	True Positives
FP	False Positives
FN	False Negatives
EQIP	Expanded Quality Information Patient

Acknowledgements

I would like to extend my sincere gratitude to the dental experts at Zhejiang Provincial People's Hospital for their invaluable assistance in providing data evaluation and scoring. Their expertise and meticulous efforts have significantly contributed to the quality and reliability of this research.

Author contributions

YPW, as the first author, collected the clinical data and supervised all the procedures. YKZ designed the study and performed the statistical analysis. Yuchen Zheng, as a corresponding author, prepared the manuscript and interpreted the data. All authors critically discussed the results, revised the manuscript, and approved the final version.

Funding

Not applicable.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

The research protocol received evaluation and authorization from the Ethics Committee of the Zhejiang Provincial People's Hospital (No. QT2023249). All patient medical records used in this study were included with the informed consent of the patients. This study was conducted in accordance with the principles of the Declaration of Helsinki.

Consent for publication

All authors have read and approved the final manuscript and consent to its publication.

Relevant guidelines and regulations

The study was conducted following the International Team for Implantology (ITI) Clinical Guidebook Series standards and relevant clinical practice guidelines for dental implantology.

Competing interests

The authors declare no competing interests.

Received: 20 June 2024 / Accepted: 12 March 2025
Published online: 26 March 2025

References

1. Morandín-Ahuerma F. What is artificial intelligence? Int J Res Publ Rev. 2022;3(12):1947–51.
2. Abd-alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large Language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ. 2023;9(1):e48291.
3. Huang H, Zheng O, Wang D, Yin J, Wang Z, Ding S, et al. ChatGPT for shaping the future of dentistry: the potential of multi-modal large Language model. Int J Oral Sci. 2023;15(1):29.
4. Surovková J, Haluzová S, Strunga M, Urban R, Lifková M, Thurzo A. The new role of the dental assistant and nurse in the age of advanced artificial intelligence in telehealth orthodontic care with dental monitoring: preliminary report. Appl Sci. 2023;13(8):5212.
5. Cabral S, Restrepo D, Kanjee Z, Wilson P, Crowe B, Abdulnour RE, et al. Clinical reasoning of a generative artificial intelligence model compared with physicians. JAMA Intern Med. 2024;184(5):581.
6. Ghanem YK, Rouhi AD, Al-Houssan A, Saleh Z, Moccia MC, Joshi H, et al. Dr. Google to dr. ChatGPT: assessing the content and quality of artificial intelligence-generated medical information on appendicitis. Surg Endosc. 2024;38(5):2887–93.
7. 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(3):305–14.

8. Danesh A, Pazouki H, Danesh K, Danesh F, Danesh A. The performance of artificial intelligence Language models in board-style dental knowledge assessment. J Am Dent Assoc. 2023;154(11):970–4.
9. Tomášik J, Zsoldos M, Oravcová Ľ, Lifková M, Pavleová G, Strunga M, Thurzo A. AI and face-driven orthodontics: a scoping review of digital advances in diagnosis and treatment planning. AI. 2024;5(1):158–76.
10. Greenstein G, Cavallaro J, Tarnow D. Dental implantology: numbers clinicians need to know. Compend Contin Educ Dent. 2019;40:e1–26.
11. Shah KK, Sivaswamy V. Assessment of knowledge on implant abutment and platform switching among dental specialists practicing implantology. J long Term Eff Med Implants. 2023;33(1):31–7.
12. Fatani B, Almutairi ES, Almalky HA, Mubarki MI, Al-Safadi A. A comparison of knowledge and skills related to up-to-date implant techniques among prosthodontists, periodontists, and oral surgeons: a cross-sectional study. Cureus. 2022;14(10):e30259.
13. Macrina FL. Scientific integrity: an introductory text with cases. Washington DC: ASM; 1995.
14. Sharma S. How to become a competent medical writer? Perspect Clin Res. 2010;1(1):33–7.
15. Adams S. Under construction: reviewing and producing information reliability on the web. Rotterdam: Erasmus University; 2006.
16. Collier R. A call for clarity and quality in medical writing. Can Med Assoc J. 2017;189(46):E1407.
17. Plavén-Sigray P, Matheson GJ, Schiffler BC, Thompson WH. The readability of scientific texts is decreasing over time. eLife. 2017;6:e27725.
18. Stuyt PMJ. Why don't medical textbooks teach? The lack of logic in the differential diagnosis. Ned Tijdschr Geneeskd. 2003;61(11):e1–5.
19. Roger A, Aïmeur E, Rish I. Towards ethical multimodal systems. arXiv [Preprint]. 2023 [cited 2024 May 23]. Available from: arxiv:2304.13765.
20. van Rijsbergen CJ. Information retrieval. 2nd ed. London: Butterworth; 1979.
21. Charvet-Berard AI, Chopard P, Perneger TV. Measuring quality of patient information documents with an expanded EQIP scale. Patient Educ Couns. 2008;70(3):407–11.
22. Sullivan LM. Essentials of biostatistics for public health. 4th ed. Burlington: Jones & Bartlett Learning; 2022.
23. Masalkhi M, Ong J, Waisberg E, Lee AG. Google DeepMind's gemini AI versus ChatGPT: a comparative analysis in ophthalmology. Eye. 2024. https://doi.org/10.1038/s41433-024-02958-w.
24. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. JAMA Ophthalmol. 2023;141(6):589–95.
25. Mihalache A, Grad J, Patil NS, Huang RS, Popovic MM, Mallipatna A, et al. Google gemini and bard artificial intelligence chatbot performance in ophthalmology knowledge assessment. Eye. 2024. https://doi.org/10.1038/s41433-024-03067-4.
26. Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Müller BP, Raptis DA, et al. Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. J Med Internet Res. 2023;25:e47479.
27. Aydın Ö, Karaarslan E. Is ChatGPT leading generative AI? What is beyond expectations? APJESS. 2023;11(3):118–34. https://doi.org/10.21541/apjess.1293702.
28. Bahir D, Zur O, Attal L, et al. Gemini AI vs. ChatGPT: A comprehensive examination alongside ophthalmology residents in medical knowledge. Graefes Arch Clin Exp Ophthalmol. 2025;263:527–36. https://doi.org/10.1007/s00417-024-06625-4.
29. Thurzo A, Varga I. Advances in 4D shape-memory resins for AI-aided personalized scaffold bioengineering. Bratisl Med J. 2025. https://doi.org/10.1007/s44411-025-00043-6.

Publisher's note