BMC Oral Health

RESEARCH

Open Access

A multi-dimensional performance evaluation of large language models in dental implantology: comparison of ChatGPT, DeepSeek, Grok, Gemini and Qwen across diverse clinical scenarios

Xing Wu[1,2†], Guofei Cai[3†], Bin Guo[1,2†], Leizi Ma[2,4], Siqi Shao[5], Jun Yu[6], Yuchen Zheng[2*], Linhong Wang[2*] and Fan Yang[2*]

Abstract

Background Large language models (LLMs) show promise in medicine, but their effectiveness in specialized fields like implant dentistry remains unclear. This study focuses on five recently released LLMs aiming to systematically evaluate their capabilities in clinical implantology scenarios and to investigate their respective strengths and weaknesses thoroughly to guide precise application.

Methods A comprehensive multi-dimensional evaluation was conducted using a test set of 40 professional questions (across 8 themes) and 5 complex cases. To ensure response uniformity, all queries were submitted to five LLMs (ChatGPT-o3-mini, DeepSeek-R1, Grok-3, Gemini-2.0-flash-Thinking, and Qwen2.5-max) using a pre-defined prompt. With standardized parameters to ensure a fair comparison, a single response was generated for each query without re-generation. The responses of the five LLMs were scored by three experienced senior experts from five dimensions in two rounds of double-blind. Inter-rater reliability was tested, followed by statistical analyses including Spearman'sptest, Friedman test, mixed effect model, and principal component analysis.

Results High inter-rater reliability was confirmed among the three experts (ICC for average measures ranged from 0.685 to 0.814, all $P < 0.001$). Gemini-2.0-flash-thinking achieved the highest overall performance, with a mean score of 21.9 in professional question answering and 22.2 in case analysis. This was significantly higher than ChatGPT-o3-

†Xing Wu, Guofei Cai and Bin Guo should be considered co-first author.

*Correspondence:
Yuchen Zheng
zhengyuchen@hmc.edu.cn
Linhong Wang
wanglinhong@hmc.edu.cn
Fan Yang
yangfan@hmc.edu.cn

Full list of author information is available at the end of the article

mini (mean score 19.2) in question responses and Qwen2.5-max (mean score 16.9) in case evaluations. Mixed-effects models showed Gemini-2.0-flash-thinking superiority over ChatGPT-o3-mini, while Qwen2.5-max exhibited a decline in performance. DeepSeek-R1 and Qwen2.5-max also showed positive interaction effects in specific themes (such as Theme3). The PCA results further indicate that Gemini-2.0-flash-thinking demonstrated the best comprehensive ability in both types of tasks, and reveal the existing differences in the performance of various LLMs.

Conclusion This study reveals diverse LLMs differentiated capabilities in dental implantology, recommending context-specific model selection to different clinical scenario, as Gemini-2.0-flash-Thinking demonstrates optimal performance, notably for high-level clinical support.

Trial registration The study protocol and the use of clinical case data have been approved by the Medical Ethics Committee of Zhejiang Provincial People's Hospital (Approval No. QT2025050) on March 4th, 2025. Clinical trial number is not applicable.

Keywords Large Language models, Dental implantation, Case studies, Clinical decision support, Multidimensional scaling analysis, Principal component analysis

Introduction

In recent years, the Large Language Models (LLMs), a disruptive technology within the domain of artificial intelligence, has been rapidly permeating and reshaping the landscape of medical research and clinical practice [1]. Leveraging their remarkable capabilities in natural language understanding, generation, and knowledge reasoning, LLMs are not only capable of efficiently processing and synthesizing vast amounts of medical information, but also demonstrate substantial potential for applications in critical areas such as augmenting clinical decision-making, transforming medical education paradigms, and accelerating the pace of scientific knowledge discovery [2]. Within the field of medical applications, LLMs can serve as valuable tools to assist medical students in efficiently acquiring complex medical knowledge systems [3] and in the development of intelligent medical education platforms [4–6]. Furthermore, their clinical utility is becoming increasingly evident, with applications deeply integrating with various facets of healthcare, including aiding physicians in disease diagnosis [7–9], formulating personalized treatment strategies [10, 11], and enhancing disease management and screening protocols [12, 13]. Consequently, LLMs are demonstrably propelling the advancement of medicine into a new development phase.

In dentistry, LLMs have also demonstrated considerable clinical application potential. They can enhance diagnostic and therapeutic efficiency and quality by leveraging functionalities such as automated diagnosis, multimodal analysis, personalized treatment planning, and patient education. For instance, a study by Revilla-León et al. [14] revealed that ChatGPT-4.0 scored 84% on the European Association for Osseointegration (EAO) implant dentistry certification examination, outperforming human dentists who scored 74%. Notably, ChatGPT-4.0 exhibited particular proficiency in analyzing peri-implantitis risk and complex clinical scenarios. Furthermore, Kurt Demirsoy et al. [15] validated the positive impact of ChatGPT in orthodontic patient education, where its easily understandable explanations resulted in a high patient satisfaction score of 4.27 out of 5, significantly outperforming that of traditional educational materials. Moreover, LLMs can efficiently process electronic health records and medical imaging data to enable precise diagnosis and treatment plan generation for diseases such as dental caries and periodontal disease [16]. Those studies preliminarily demonstrate the effectiveness of LLMs in diverse dental applications.

However, the practical clinical application of LLMs currently faces numerous challenges and limitations. For example, their capabilities in addressing specialized questions within dental medicine, particularly in diagnosing complex cases and formulating personalized treatment plans, remain suboptimal. In their periodontal disease assessment study, Chatzopoulos et al. [17] observed that when LLMs answer open-ended clinical inquiries, the comprehensiveness, scientific validity, logical coherence, and clarity of presentation in their responses still require improvement. This is particularly evident when dealing with intricate clinical situations characterized by multifactorial interactions and significant subjectivity, where LLMs are prone to information bias or incompleteness. Research by Pradhan et al. [18] indicated that while advanced LLMs such as ChatGPT-4o demonstrate preliminary application potential in the radiographic diagnosis of Oral Potentially Malignant Lesions, their diagnostic accuracy remains significantly lower than that of oral and maxillofacial specialists. Similarly, Tastan et al. [19] confirmed that although ChatGPT exhibits a degree of rationality in periodontitis classification, its performance in refined diagnostic aspects such as disease staging and grading is susceptible to interference from complex clinical factors, thus limiting diagnostic accuracy. Furthermore, technical challenges, including data quality heterogeneity and model "hallucinations,"

represent critical limitations hindering the broader and deeper implementation of LLMs in dental medicine [1, 16].

Currently, dental implant has become one of the primary restorative modalities for partial and complete edentulism [20]. Successful dental implant treatment and long-term prognosis stability rely not only on the surgeon's refined surgical techniques but also on meticulous preoperative planning and design, standardized postoperative maintenance, and patient compliance with long-term follow-up [21, 22]. However, the low homogeneity among implant practitioners and the uneven distribution of medical resources objectively constrain the effective promotion and widespread adoption of standardized diagnostic and treatment protocols [23]. In this context, introducing LLMs provides an opportunity to enhance the level of intelligentization in implant dentistry. Taymour et al. [24] have reported that current LLMs, including ChatGPT-3.5, ChatGPT-4, and Google Gemini, already possess considerable capabilities in answering questions related to dental implantology. In the clinical case scenario of identifying oral pathologies, Kaygisiz et al. [25] compared and evaluated the accuracy of ChatGPT-4o and Deepseek-v3 in oral pathology diagnosis and found that Deepseek-v3 performed better. Nevertheless, LLMs lack the ability to assess individual patient variations, potentially leading to generalized recommendations. This highlights that applying LLMs in clinical practice still faces numerous risks and limitations. This is because LLMs primarily rely on superficial pattern recognition based on statistical associations and vast data memorization. Consequently, their inherent limitations become particularly prominent when dealing with clinical problems requiring deep semantic understanding, complex logical reasoning, and contextualized decision-making, especially in highly specialized fields like implant dentistry that emphasize refined diagnosis and treatment [26, 27]. These limitations often preclude satisfactory outcomes in implant treatment planning scenarios that necessitate considering multimodal data, various patient-specific and surgical site factors, and even integrating systemic health conditions and laboratory indicators [16, 26]. Our preliminary research [28] has also initially confirmed that general-purpose LLMs such as ChatGPT-4.0 exhibit preliminary application potential in professional knowledge question answering within implant dentistry. However, they remain insufficient for advanced clinical decision-making tasks, such as complex case analysis and treatment plan formulation. This suggests that the in-depth application of existing general-purpose LLMs in implant dentistry still faces a bottleneck, particularly in complex clinical scenarios requiring refined, personalized treatment planning. Therefore, more advanced LLMs and systematic and in-depth performance evaluation studies are urgently needed.

The recent emergence of a new generation of LLMs, exemplified by inference models released by companies such as DeepSeek and OpenAI, signals a new paradigm in developing Artificial Intelligence (AI) technology [29, 30]. Compared to previous models, this new generation of LLMs has achieved significant advancements in model architecture, training paradigms, and inference mechanisms, demonstrating superior logical reasoning, complex problem-solving, and generalization capabilities [31, 32]. These models efficiently process larger-scale and more complex datasets and, more importantly, are capable of simulating expert-level thinking patterns, performing multi-step complex reasoning, counterfactual scenario analysis, and strategic planning. Consequently, they outperform previous models in higher-order cognitive tasks such as mathematical problem-solving, complex text logical analysis, and multi-turn dialogue interaction [29]. For example, recently Mickael et al. [33] evaluated the performance of DeepSeek-R1, ChatGPT-o1 and Llama 3.1-405B in four medical tasks: answering questions in the US Medical Licensing Examination, text-based case diagnosis and management, tumor classification and imaging report summarization through comparative experiments, and found that DeepSeek-R1 has a great advantage in reasoning ability. However, implant dentistry, characterized by its refined diagnostic and therapeutic processes, high reliance on patient education, and stringent requirements for long-term clinical efficacy [34], presents significant challenges for in-depth integration with LLMs. Therefore, this study focuses on five recently released LLMs—ChatGPT-o3-mini, DeepSeek-R1, Grok-3, Gemini-2.0-flash-Thinking, and Qwen2.5-max—aiming to systematically evaluate their capabilities in professional knowledge questions answering and complex case analysis within implant dentistry and to investigate their respective strengths and weaknesses thoroughly. The null hypothesis of this study is that there is no statistically significant difference in the performance of the five LLMs—ChatGPT-o3-mini, DeepSeek-R1, Grok-3, Gemini-2.0-flash-Thinking, and Qwen2.5-max—in professional questions answering and case analysis within implant dentistry.

Materials and methods

The design and implementation of the present study were conducted in accordance with the Declaration of Helsinki. The study protocol and the use of clinical case data have been approved by the Medical Ethics Committee of Zhejiang Provincial People's Hospital (Approval No. QT2025050). No participant personal information irrelevant to this research will be disclosed throughout

the study. Figure 1 shows a flow chart of the study implementation.

Data preparation and task design

The set of professional questions was developed based on the ITI (International Team for Implantology) Consensus Statements, EAO Clinical Practice Guidelines, and high-quality systematic reviews [35–48] in implant dentistry published in the last five years (indexed in PubMed/Cochrane). After three rounds of expert consultation and pre-scoring, 40 core questions were selected. These questions were organized into eight major themes:

Theme 1. Implant Structure Design.

Theme 2. Preoperative Assessment for Implants.

Theme 3. Implant Surgical Procedures.

Theme 4. Application of Bone Augmentation Techniques.

Theme 5. Postoperative Healing and Restoration of Implants.

Theme 6. Long-term Care and Maintenance of Implants.

Theme 7. Special Conditions and Management of Complications.

Theme 8. Pathophysiology Related to Implant Dentistry.

Each theme consisted of five open-ended questions. The case analyses were sourced from the dental implant case database of the Department of Dentistry at Zhejiang Provincial People's Hospital. From the dental implant cases conducted between January 2020 and December 2024, 20 relatively complex cases were selected by a senior dentist with 20 years of experience in implant surgery, who was not involved in the subsequent scoring process of this study. Five cases were randomly chosen for this study using a simple random sampling method. The design of the test set for evaluating the performance of LLMs is detailed in Supplementary Material 1.

Model setup and task implementation

The LLMs included in this evaluation were ChatGPT-o3-mini (USA, OpenAI), DeepSeek-R1 (China, DeepSeek), Grok-3 (USA, xAI), Gemini-2.0-flash-thinking (USA, Google DeepMind), and Qwen2.5-max (China, Alibaba Cloud). These represented the most recently released generative AI chatbots available as of February 20, 2025. All LLMs' Application Programming Interfaces (APIs) were configured using Cherry Studio (China, kangfenmao). All parameters of all LLMs are required to remain consistent during the task implementation process, the temperature is set to 1.00, the max_tokens limit is not enabled, and the top_p is set to 0.80, and both the frequency_penalty and presence_penalty were set to 0. Internet search access was restricted during the questioning process. The deep thinking function was activated for LLMs possessing this capability (excluding Qwen2.5-max). The conversational context was cleared after each interaction to avoid contextual bias between queries, and all other relevant parameters were standardized across LLMs. A single author (Wu) submitted one query to each LLMs, incorporating both the question or patient case details along with a pre-defined "Prompt" to maintain response uniformity (refer to Supplementary Material 2). Each LLMs was afforded a single response opportunity, and no post-hoc modifications or re-generation of the initial response was allowed. All generated responses were documented in a digital spreadsheet (refer to Supplementary Material 3).

Comprehensive assessment of LLM performance

The LLMs performance evaluation was conducted by a panel of three experts, each possessing at least 15 years of experience in implant dentistry. To ensure inter-rater reliability and consistency in scoring, a standardization session was held prior to the scoring process based on the predefined LLMs performance evaluation standards (refer to Supplementary Material 4). This session aimed to align the scoring stringency across all panel members. For both task categories, a 5-dimension scoring system was implemented (as detailed in Table 1). This evaluation system was adapted from the methodology described by Chatzopoulos et al. [17], with specific modifications for the current study, especially in the domain of evaluating the logical reasoning capabilities of the LLMs. To maintain blinding, the LLMs-generated responses were de-identified. Independent and randomized scoring was performed, followed by a second round of scoring after a two-week interval. Subsequently, inter-rater reliability was assessed based on the results from both scoring rounds (refer to Supplementary Material 5 and 6 for detailed scoring outcomes).

Statistical methods

To descriptively summarize the evaluations of each LLMs by the raters, we initially performed descriptive statistical analyses. Given the potential violation of the normality assumption of the data, non-parametric tests were subsequently employed for further analyses. To assess the reliability of the rating process, we examined both test-retest and inter-rater reliability, quantified using the Intraclass Correlation Coefficient (ICC) for both measures. Spearman's ρ correlation coefficient was additionally calculated to validate further test-retest reliability and the correlation among the three raters' scores. Kendall's W test was used to assess inter-rater reliability. Differences in ratings within raters across the two-time points were examined using the Wilcoxon Signed-Rank Test. Furthermore, the Friedman test was applied to compare score differences among the five LLMs and across different topics, aiming

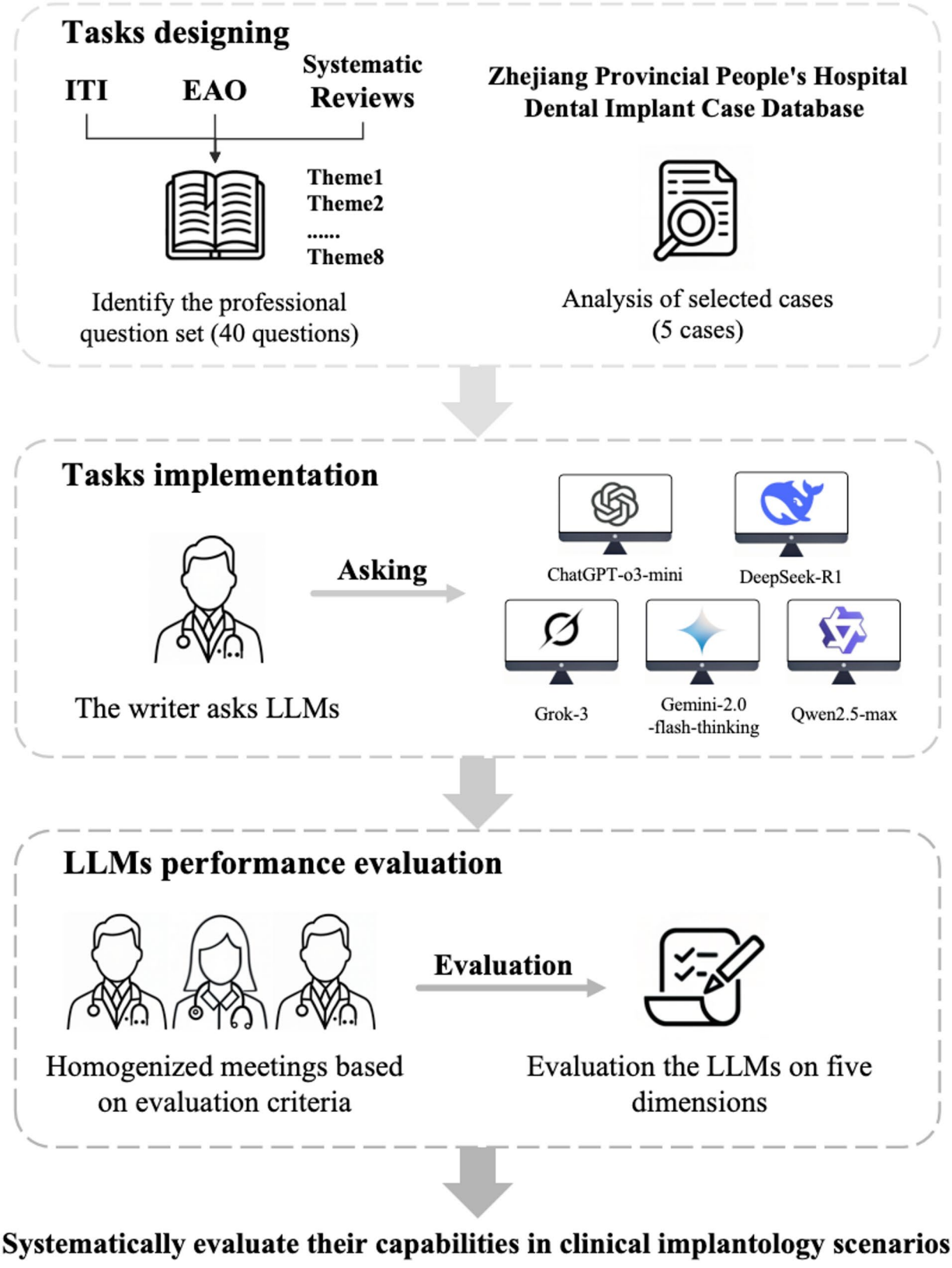

Fig. 1 Flow chart of the implementation of this study

Table 1 The evaluation dimension design of the professional questions answering and cases analysis

Category	Dimension	Description
Problem Answering Ability	Accuracy	Assesses the consistency of the LLMs' responses to dental implantology questions with authoritative knowledge, and the extent of knowledge point coverage.
	Integrity	Measures whether the LLMs' answer encompasses all essential elements of the question, and if the logic is clear and coherent.
	Relevance	Examines the degree of close association between the LLMs' responses and cited sources with the core issues in dental implantology.
	Diversity	Evaluates the LLMs' capacity to present multiple lines of thought, methodologies, and perspectives when addressing questions that can be answered from several aspects.
	Interpretability	Verifies whether the LLMs' explanations and reasoning processes for its answers are clear, logical, and readily understandable.
Case Analysis Ability	Accuracy of disease diagnosis	Determines the LLMs' ability to make correct diagnoses based on case information and to accurately articulate the rationale behind the diagnosis.
	Rationality of treatment plan	Evaluates whether the implant treatment plan formulated by the LLMs, based on the case, is scientifically sound, comprehensive, and aligned with practical considerations.
	Comprehensiveness of risk assessment	Examines the LLMs' ability to comprehensively identify risks associated with implant surgery and post-operative phases, and to propose effective mitigation strategies.
	Knowledge application ability	Measures the accuracy of the LLMs' application of multidisciplinary knowledge in case analysis and the reliability of its citations to relevant literature.
	Logical reasoning ability	Verifies the LLMs' ability to engage in rigorous reasoning based on case information to formulate sound diagnoses, treatment plans, and risk management strategies.

to identify statistically significant variations. A linear mixed-effects model was utilized to explore the interaction effects of LLMs and topics on the ratings further. Finally, to comprehensively analyze the overall performance of LLMs across multiple evaluation dimensions, Principal Component Analysis (PCA) was conducted. All statistical analyses were conducted using IBM SPSS Statistics (Version 27.0, IBM Corp., Armonk, NY, USA) and the statistical software R (version 4.4.2, R Foundation for Statistical Computing).

Results

Results of rater reliability testing

Tables 2 and 3 present the descriptive statistics of ratings provided by the three raters for LLMs in the professional question-answering and case analysis sections, respectively, across two time points. The heat map of the scoring results (Fig. 2) shows the two scores of the three raters, and the higher the score in the figure, the more concentrated the red.

To evaluate the reliability of rater assessments of LLMs performance, this study initially conducted tests for both test-retest and inter-rater reliability. As shown in Table 4, the ICC single measures and average measures for test-retest reliability for Rater 1, Rater 2, and Rater 3 were all relatively high. Spearman's ρ correlation coefficient values were all above 0.641 and demonstrated significant statistical significance ($p < 0.001$), indicating good consistency and stability in the raters' scores across different time points. Furthermore, the results of the difference test (Wilcoxon Signed-Rank Test) presented in Table 5 revealed that, across the two-time points, the rating differences for the same LLMs by the same rater were not statistically significant ($p > 0.05$), further validating the rating stability of the raters.

Table 6 presents the results of the inter-rater reliability analysis of the average scores from the three raters. The results showed that the inter-rater reliability reached a level above moderate. Specifically, the range of ICC for average measures of the five LLMs was from 0.685 to 0.814, and the ICC values of all models were statistically significant ($p < 0.001$). Furthermore, Kendall's W test also revealed significant results ($p < 0.001$). These findings suggest that there is an acceptable level of inter-rater reliability among the raters and for the overall rating outcomes. Figure 3 displays the Spearman's ρ correlation coefficient matrix for the average scores among the three raters. The results indicate a high degree of correlation in scores assigned by different raters for the same LLMs, suggesting a considerable level of agreement among raters in their evaluation of individual LLMs, thus implying strong reliability of the assessment results. Conversely, the correlation among different LLMs was lower, and even negative correlations were observed. This indicates that the actual performance of each LLMs varied distinctly and that raters were effectively able to differentiate between models with varying levels of capability.

Performance of LLMs in professional question answering and case analysis

Tables 7 and 8 present the descriptive statistics of the final scores for the five LLMs in the professional question-answering and case analysis sections, respectively. In the professional question-answering section, Gemini-2.0-flash-thinking demonstrated the highest performance (mean score 21.9), followed by DeepSeek-R1 (mean score

Table 2 Descriptive statistics of the scores given by the three raters to evaluate the ability of the various LLMs to the professional questions answering at two different time points

	ChatGPT						DeepSeek						Grok						Gemini						Qwen					
	Rater1		Rater2		Rater3		Rater1		Rater2		Rater3		Rater1		Rater2		Rater3		Rater1		Rater2		Rater3		Rater1		Rater2		Rater3	
Stage of scoring	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
Mean	18.4	18.2	19.1	19.0	20.3	20.1	19.0	18.8	19.3	19.2	20.2	20.4	18.1	17.9	19.0	18.8	19.4	19.3	20.9	21.0	21.8	22.1	22.9	22.7	18.8	18.6	19.1	19.2	20.0	19.7
SE of the mean	0.2	0.2	0.2	0.2	0.3	0.3	0.1	0.1	0.1	0.1	0.2	0.1	0.2	0.3	0.2	0.3	0.3	0.2	0.3	0.2	0.2	0.2	0.2	0.2	0.1	0.1	0.2	0.2	0.2	0.2
Median	19.0	18.5	19.0	19.0	20.0	21.0	19.0	19.0	19.0	19.0	20.0	21.0	18.0	18.0	19.0	19.0	20.0	20.0	21.0	21.0	22.0	22.0	23.5	23.0	19.0	19.0	19.0	19.0	20.0	20.0
Mode	19.0	19.0	19.0	20.0	20.0	21.0	19.0	19.0	19.0	19.0	20.0*	21.0	18.0	17.0*	19.0	18.0	20.0	20.0	20.0*	20.0	22.0	21.0	24.0	23.0	19.0	19.0	18.0	19.0	20.0	20.0
Minimum	16.0	15.0	16.0	16.0	15.0	16.0	17.0	17.0	18.0	18.0	17.0	18.0	16.0	15.0	16.0	15.0	15.0	16.0	17.0	18.0	19.0	19.0	19.0	20.0	17.0	17.0	18.0	17.0	17.0	17.0
Maximum	21.0	21.0	21.0	23.0	23.0	23.0	20.0	20.0	21.0	21.0	22.0	22.0	21.0	23.0	23.0	24.0	23.0	23.0	24.0	23.0	24.0	25.0	24.0	25.0	20.0	20.0	21.0	21.0	22.0	22.0
Sum up	737.0	729.0	765.0	761.0	810.0	805.0	758.0	753.0	773.0	767.0	809.0	816.0	722.0	715.0	760.0	750.0	776.0	773.0	836.0	838.0	873.0	884.0	915.0	907.0	751.0	745.0	763.0	767.0	798.0	788.0
standard deviation	1.1	1.3	1.1	1.5	1.9	1.7	0.8	0.9	0.9	0.9	1.5	0.9	1.2	1.6	1.3	1.6	1.8	1.4	1.7	1.3	1.2	1.5	1.4	1.1	0.8	0.8	1.0	1.0	1.4	1.2
variance	1.1	1.8	1.1	2.1	3.6	2.9	0.7	0.8	0.8	0.9	2.3	0.9	1.3	2.6	1.7	2.5	3.2	2.1	2.8	1.7	1.5	2.4	2.1	1.3	0.6	0.7	1.0	1.0	1.9	1.5

* indicates the presence of multiple modes and shows the smallest value

19.5), and ChatGPT-o3-mini and Qwen2.5-max (both with a mean score of 19.2). Grok-3 exhibited a comparatively lower performance (mean score 18.7). The results for the case analysis section revealed a similar trend, with Gemini-2.0-flash-thinking again leading (mean score 22.2), while Qwen2.5-max obtained the lowest score (mean score 16.9).

Figure 4 illustrates the distribution and variability of average scores for each LLMs in both the professional question answering (A) and case analysis (B) sections. The results indicate that Gemini-2.0-flash-thinking achieved significantly higher scores in the professional question answering task compared to ChatGPT-o3-mini ($p < 0.001$) and DeepSeek-R1 ($p < 0.001$). Furthermore, a statistically significant difference was also observed between DeepSeek-R1 and Grok-3 ($p < 0.01$). In the case analysis task, Gemini-2.0-flash-thinking also significantly outperformed Qwen2.5-max ($p < 0.01$), while the extent of score differences among the other models varied.

Variations in difficulty for topic-specific professional questions and cases, and results of mixed-effects model analysis

Figure 5 compares the difficulty levels of professional questions and cases across different themes and cumulatively displays the scores for each LLMs in the professional question answering and case analysis sections, categorized by theme. The results indicate that Theme 1 and Theme 6 were relatively more straightforward, with all LLMs achieving higher scores in these themes. Conversely, Theme 2, Theme 3, Theme 4, Theme 5, Theme 8, and Case presented comparatively more incredible difficulty.

Table 9 presents the detailed results of the mixed-effects model analysis of LLMs performance. The model intercept was 20.73 ($p < 0.001$), indicating that the baseline model, ChatGPT-o3-mini, had an estimated mean score of approximately 20.73 points under the baseline theme, Theme 1. The fixed effects analysis revealed that, in comparison to the baseline model ChatGPT-o3-mini, Gemini-2.0-flash-thinking achieved significantly higher scores (Estimate = 1.7, $p < 0.001$), while Qwen2.5-max obtained significantly lower scores (Estimate = -1, $p = 0.040$). DeepSeek-R1 and Grok-3 did not show significant differences from ChatGPT-o3-mini. Regarding theme effects, compared to Theme 1, Theme 2 (Estimate = -1.133, $p = 0.058$) showed a slightly lower score, but this difference was not statistically significant. Themes 3 (Estimate = -2.8, $p < 0.001$), 4 (Estimate = -1.7, $p = 0.005$), 5 (Estimate = -2.367, $p < 0.001$), 7 (Estimate = -1.8, $p = 0.003$), and 8 (Estimate = -1.8, $p = 0.003$) all exhibited significantly reduced scores. Theme 6 (Estimate = -0.7, $p = 0.240$) showed a non-significant score reduction. The effect of case difficulty on scores was not significant (Estimate = -0.133, $p = 0.823$). Regarding significant interaction effects, Gemini-2.0-flash-thinking showed significant positive interaction effects compared to ChatGPT-o3-mini in Theme 2 (Estimate = 1.567, $p = 0.023$), Theme 3 (Estimate = 1.6, $p = 0.020$), and Theme 5 (Estimate = 1.4, $p = 0.042$). DeepSeek-R1 also showed significant positive interaction effects in Theme 3 (Estimate = 1.72, $p = 0.013$) and Theme 7 (Estimate = 1.7, $p = 0.014$). Similarly, Qwen2.5-max exhibited significant positive interaction effects in Theme 3 (Estimate = 2.133, $p = 0.002$) and Theme 5 (Estimate = 1.6, $p = 0.020$) compared to ChatGPT-o3-mini. Qwen2.5-max demonstrated a significant negative interaction effect in Case analysis (Estimate = -2.689, $p < 0.001$).

Performance of LLMs across different evaluation dimensions and PCA results

Tables 10 and 11, respectively, provide descriptive statistics for the five evaluation dimensions (Dimensions 1–5) for LLMs performance in professional question answering and case analysis. Radar charts in Fig. 6 visually present the performance of each LLMs across these different dimensions. In professional question answering (Fig. 6A), Gemini-2.0-flash-thinking exhibited higher mean scores across all dimensions compared to other LLMs. All LLMs demonstrated relatively balanced performance across dimensions without apparent weaknesses in any specific dimension. In case analysis (Fig. 6B), Gemini-2.0-flash-thinking particularly excelled in Dimension 2 (Reasonableness of Treatment Plan), Dimension 4 (Knowledge Application Ability), and Dimension 5 (Logical Reasoning Ability), whereas Qwen2.5-max consistently achieved relatively lower scores across all dimensions.

To further investigate the comprehensive performance patterns of LLMs across the five evaluation dimensions, we conducted a PCA on the dimension scores for both professional question-answering and case analysis sections. The results are shown in Fig. 7. The PCA results for professional question answering (Fig. 7A) indicate that the first two principal components (PC1 and PC2) explained a substantial portion of the variance in the dimension scores (66.4%). As shown in the loadings plot of Fig. 7A, PC1 primarily reflects the composite ability of LLMs in answering questions, particularly in Dimension 3 (Relevance), Dimension 4 (Diversity), and Dimension 5 (Interpretability). In the loadings plot, the vectors for these dimensions point in the positive direction of PC1 and have relatively long lengths, indicating that these dimensions contribute significantly to PC1. PC2 is primarily positively correlated with Dimension 2 (Completeness), although the correlation is weaker than that of the dimensions in PC1 and negatively correlated with Dimension 1 (Accuracy) and Dimension 5 (Interpretability). In the PCA scatter plot, the centroid of the ellipse for

Table 3 Descriptive statistics of the scores given by the three raters when evaluating the ability of the various LLMs to the cases analysis at two different time points

	ChatGPT						DeepSeek						Grok						Gemini						Qwen					
	Rater1		Rater2		Rater3		Rater1		Rater2		Rater3		Rater1		Rater2		Rater3		Rater1		Rater2		Rater3		Rater1		Rater2		Rater3	
Stage of scoring	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
Mean	19.8	19.6	21.2	21.0	21.2	20.8	18.8	18.6	18.4	18.8	19.4	19.2	19.2	19.0	20.0	20.4	18.4	18.0	21.4	21.4	22.6	23.0	22.6	22.2	17.2	16.8	17.4	17.2	16.4	16.4
SE of the mean	0.5	0.4	0.4	0.3	0.4	0.5	0.5	0.4	0.5	0.4	0.4	0.4	0.7	0.8	0.3	0.4	0.5	0.5	0.7	0.5	0.4	0.4	0.6	0.4	1.0	0.7	0.7	0.5	0.5	0.5
Median	20.0	19.6	21.3	21.0	21.3	21.0	19.0	18.6	18.3	18.8	19.5	19.3	19.3	19.5	20.0	20.5	18.3	17.8	21.0	21.3	22.5	23.0	22.5	22.3	17.0	16.7	17.7	17.0	16.3	16.3
Mode	20.0	20.0	21.0*	21.0	21.0*	21.0	19.0	19.0	18.0	18.0*	20.0	19.0*	19.0	20.0	20.0	21.0	18.0	17.0*	20.0*	21.0	22.0	22.0*	22.0*	22.0*	16.0	17.0	18.0	17.0	16.0	16.0
Minimum	18.0	18.0	20.0	20.0	20.0	19.0	17.0	17.0	17.0	18.0	18.0	18.0	17.0	16.0	19.0	19.0	17.0	17.0	20.0	20.0	22.0	22.0	21.0	21.0	15.0	15.0	15.0	16.0	15.0	15.0
Maximum	21.0	20.0	22.0	22.0	22.0	22.0	20.0	19.0	20.0	20.0	20.0	20.0	21.0	20.0	21.0	21.0	20.0	20.0	23.0	23.0	24.0	24.0	24.0	23.0	20.0	19.0	19.0	19.0	18.0	18.0
Sum up	99.0	98.0	106.0	105.0	106.0	104.0	94.0	93.0	92.0	94.0	97.0	96.0	96.0	95.0	100.0	102.0	92.0	90.0	107.0	107.0	113.0	115.0	113.0	111.0	86.0	84.0	87.0	86.0	82.0	82.0
standard deviation	1.1	0.9	0.8	0.7	0.8	1.1	1.1	0.9	1.1	0.8	0.9	0.8	1.5	1.7	0.7	0.9	1.1	1.2	1.5	1.1	0.9	1.0	1.3	0.8	2.2	1.5	1.5	1.1	1.1	1.1
variance	1.2	0.8	0.7	0.5	0.7	1.2	1.2	0.8	1.3	0.7	0.8	0.7	2.2	3.0	0.5	0.8	1.3	1.5	2.3	1.3	0.8	1.0	1.8	0.7	4.7	2.2	2.3	1.2	1.3	1.3

* indicates the presence of multiple modes and shows the smallest value

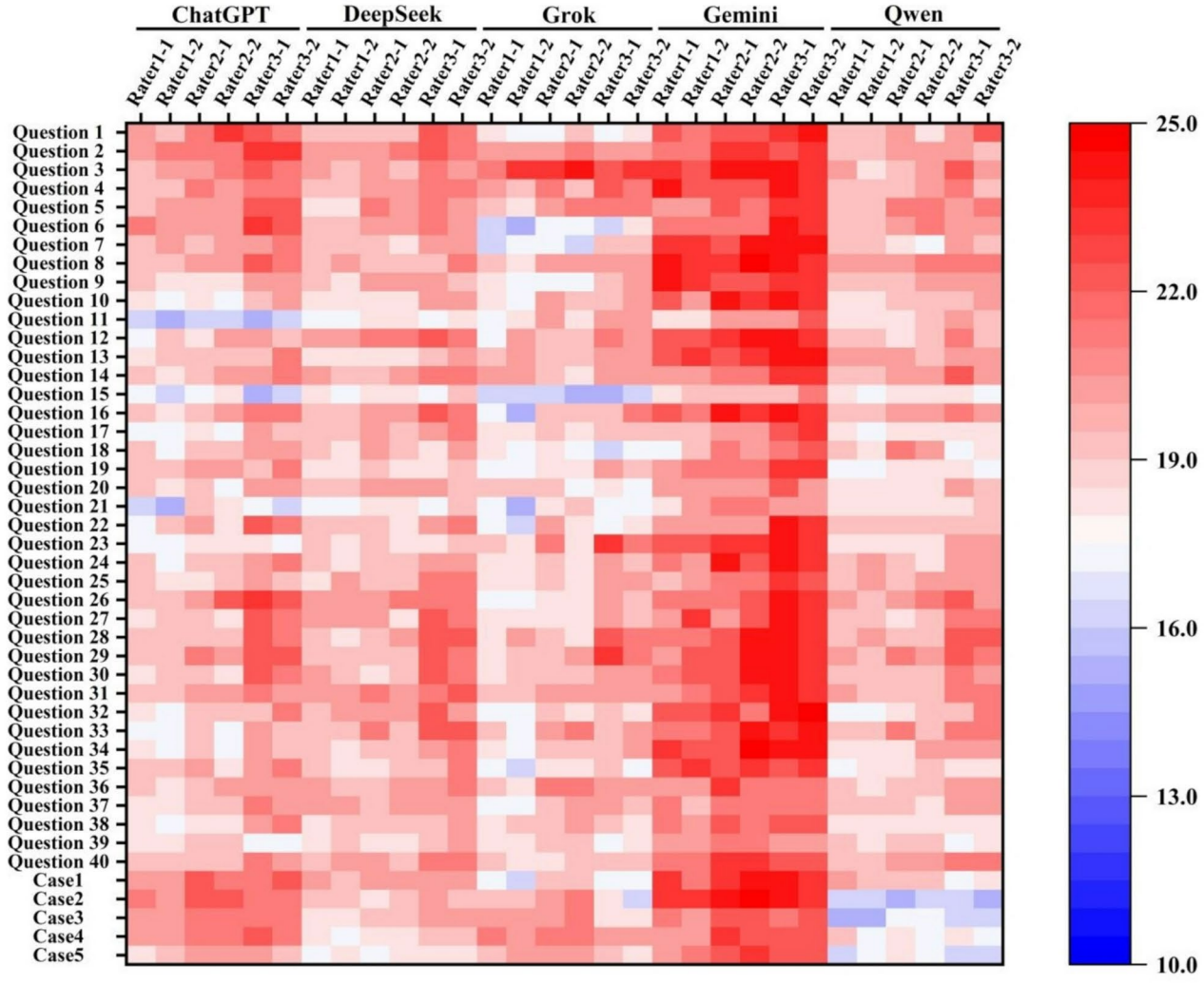

Fig. 2 Heat maps of the ratings given by the three raters at two different time points to evaluate the ability of the LLMs

Table 4 Test-retest reliability tests of the three raters' ratings of each LLMs at two different time points

	Rater1			Rater2			Rater3		
	ICC [95%CI]		Spearman's ρ	ICC [95%CI]		Spearman's ρ	ICC [95%CI]		Spearman's ρ
	Single	Average		Single	Average		Single	Average	
ChatGPT	0.717 [0.541, 0.834]	0.835 [0.702, 0.909]	0.679 ($P<0.001$)	0.727 [0.553, 0.84]	0.842 [0.712, 0.913]	0.736 ($P<0.001$)	0.845 [0.736, 0.912]	0.916 [0.848, 0.954]	0.715 ($P<0.001$)
DeepSeek	0.660 [0.459, 0.797]	0.795 [0.63, 0.887]	0.641 ($P<0.001$)	0.652 [0.447, 0.792]	0.789 [0.617, 0.884]	0.661 ($P<0.001$)	0.664 [0.464, 0.8]	0.798 [0.634, 0.889]	0.689 ($P<0.001$)
Grok	0.725 [0.552, 0.838]	0.841 [0.711, 0.912]	0.735 ($P<0.001$)	0.739 [0.573, 0.847]	0.850 [0.729, 0.917]	0.659 ($P<0.001$)	0.809 [0.678, 0.89]	0.894 [0.808, 0.942]	0.817 ($P<0.001$)
Gemini	0.723 [0.546, 0.838]	0.839 [0.707, 0.912]	0.719 ($P<0.001$)	0.648 [0.442, 0.789]	0.786 [0.613, 0.882]	0.665 ($P<0.001$)	0.726 [0.553, 0.839]	0.841 [0.713, 0.913]	0.735 ($P<0.001$)
Qwen	0.798 [0.66, 0.884]	0.888 [0.795, 0.938]	0.741 ($P<0.001$)	0.735 [0.565, 0.845]	0.847 [0.722, 0.916]	0.710 ($P<0.001$)	0.817 [0.69, 0.895]	0.899 [0.817, 0.945]	0.749 ($P<0.001$)

Gemini-2.0-flash-thinking is clearly separated from other LLMs and located on the right side of the plot, suggesting its superior performance in the composite ability represented by PC1. The centroid of the ellipse for DeepSeek-R1 is positioned between Gemini-2.0-flash-thinking and other models. The centroids of the ellipses for ChatGPT-o3-mini and Qwen2.5-max are relatively close in the plot, indicating a degree of similarity in their dimensional performance patterns. In contrast, the centroid of the ellipse for Grok-3 is located furthest to the left.

Table 5 Difference tests of the three raters' ratings of each LLMs at two different time points

	Rater1	Rater2	Rater3
ChatGPT	-1.235 (*P* = 0.217)	-0.743 (*P* = 0.458)	-0.888 (*P* = 0.375)
DeepSeek	-1.225 (*P* = 0.221)	-0.769 (*P* = 0.442)	0.934 (*P* = 0.350)
Grok	-1.210 (*P* = 0.226)	-1.122 (*P* = 0.262)	-0.897 (*P* = 0.370)
Gemini	-0.07 (*P* = 0.944)	1.634 (*P* = 0.102)	-1.414 (*P* = 0.157)
Qwen	-1.713 (*P* = 0.087)	0.498 (*P* = 0.618)	-1.616 (*P* = 0.106)

Table 6 Reliability analysis of the mean scores of the three raters for each LLMs

	ICC [95%CI]		Kendall's W
	Single	Average	
ChatGPT	0.520 [0.121, 0.754]	0.764 [0.293, 0.902]	0.623 (*P* < 0.001)
DeepSeek	0.420 [0.065, 0.676]	0.685 [0.172, 0.862]	0.648 (*P* < 0.001)
Grok	0.490 [0.260, 0.674]	0.742 [0.513, 0.861]	0.298 (*P* < 0.001)
Gemini	0.443 [0.083, 0.692]	0.705 [0.214, 0.871]	0.590 (*P* < 0.001)
Qwen	0.593 [0.361, 0.755]	0.814 [0.629, 0.903]	0.287 (*P* < 0.001)

The PCA results for case analysis (Fig. 7B) indicate that the first two principal components (PC1 and PC2) explained a substantial portion of the variance in the dimension scores (68.3%). In the dimension loadings plot, PC1 for case analysis also primarily reflects the composite ability of LLMs in Dimension 1 (Accuracy of Disease Diagnosis), Dimension 2 (Reasonableness of Treatment

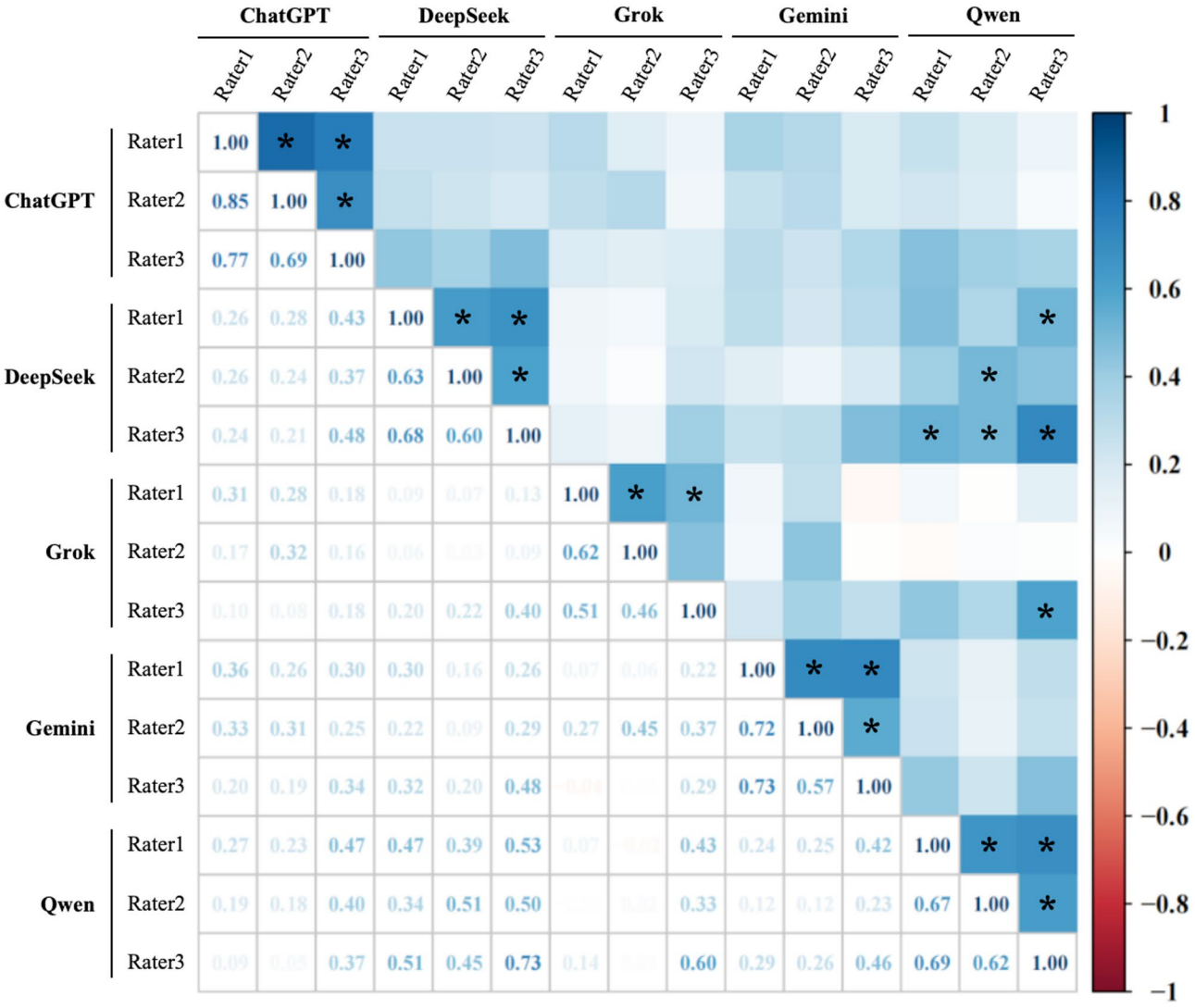

Fig. 3 Matrix plot of correlation analysis of the mean scores of the three raters for each LLMs. Significance codes: * $p < 0.001$

Table 7 Final descriptive statistics of the ratings given by the three raters when evaluating the competence of each LLMs for the professional questions answering section

Stage of scoring	ChatGPT	DeepSeek	Grok	Gemini	Qwen
Mean	19.2	19.5	18.7	21.9	19.2
SE of the mean	0.2	0.1	0.2	0.2	0.1
Median	19.4	19.7	18.7	22.2	19.4
Mode	19.5	19.8	18.17*	22.8	19.50*
Minimum	15.7	17.5	15.7	19.2	17.5
Maximum	21.5	20.8	22.7	23.7	20.5
Sum up	767.8	779.3	749.3	875.5	768.7
standard deviation	1.2	0.8	1.2	1.1	0.8
variance	1.5	0.6	1.5	1.3	0.7

* indicates the presence of multiple modes and shows the smallest value

Table 8 Final descriptive statistics of the ratings given by the three raters when evaluating the competence of each LLMs for the cases analysis section

Stage of scoring	ChatGPT	DeepSeek	Grok	Gemini	Qwen
Mean	20.6	18.7	19.1	22.2	16.9
SE of the mean	0.3	0.4	0.5	0.5	0.6
Median	20.8	19.0	19.3	21.7	16.7
Mode	19.3*	19.0	17.5*	21.3*	15.7*
Minimum	19.3	17.7	17.5	21.3	15.7
Maximum	21.2	19.8	20.5	23.7	18.7
Sum up	103.0	93.7	95.7	111.2	84.7
standard deviation	0.7	0.8	1.1	1.0	1.2
variance	0.5	0.7	1.2	1.1	1.5

* indicates the presence of multiple modes and shows the smallest value

Plan), Dimension 4 (Knowledge Application Ability), and Dimension 5 (Logical Reasoning Ability). PC2 for case analysis is primarily positively correlated with Dimension 3 (Comprehensiveness of Risk Assessment) and Dimension 4 (Knowledge Application Ability) to a certain extent. However, the strength is less than that of the dimensions in PC1. In the PCA scatter plot, the centroid of the ellipse for Gemini-2.0-flash-thinking is again clearly distinguished from other LLMs, further demonstrating its superiority in case analysis, particularly in the composite ability represented by PC1. Compared to professional question answering, the PCA plot for case analysis shows a higher degree of differentiation among LLMs. ChatGPT-o3-mini is slightly inferior to Gemini-2.0-flash-thinking, while the centroid of the ellipse for Qwen2.5-max is located further away from other models in the PCA plot for case analysis, further confirming its relatively weaker performance in case analysis. DeepSeek-R1 and Grok-3, relatively speaking, are more prominent in Knowledge Application Ability and Comprehensiveness of Risk Assessment, respectively.

Discussion

This study aimed to comprehensively analyze the performance disparities of five mainstream LLMs—ChatGPT-o3-mini, DeepSeek-R1, Grok-3, Gemini-2.0-flash-Thinking, and Qwen2.5-max—in the professional domain of dental implantology by constructing a multi-task, multi-dimensional evaluation system. The findings definitively rejected the null hypothesis, demonstrating significant differentiation in performance among the LLMs across professional dental implantology tasks. This underscores the objective variability in LLMs capabilities within specialized medical applications and provides an evidence-based rationale for the precise and targeted application of LLMs in dental implantology in the future.

This study's primary outcome reveals the exceptional performance of the Gemini-2.0-flash-Thinking model

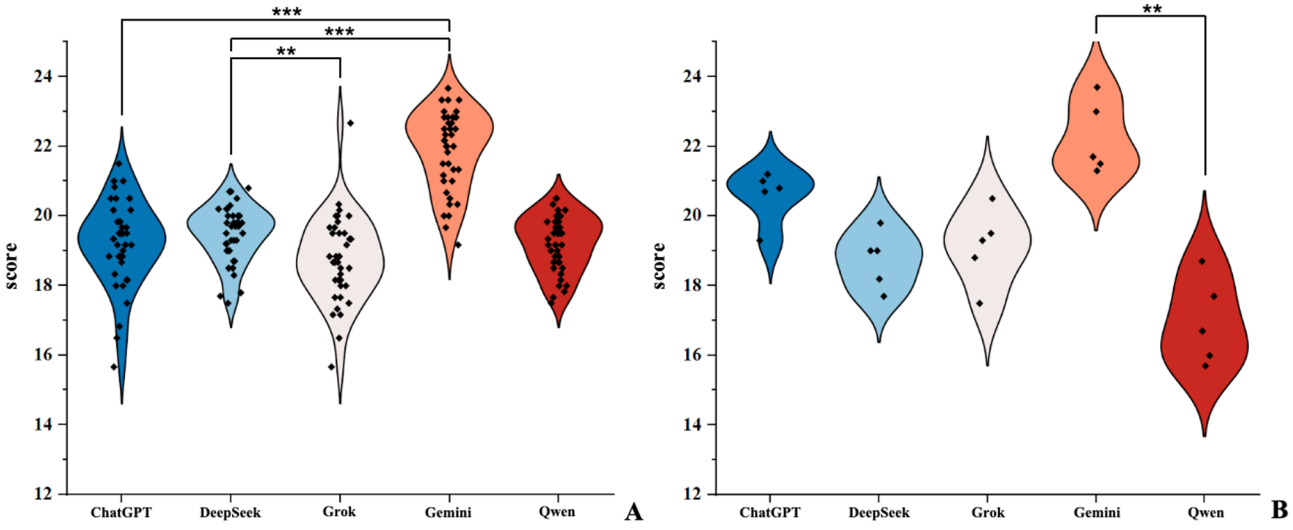

Fig. 4 Matrix plot of correlation analysis of the mean scores of the three raters for the professional questions answering(**A**) and cases analysis(**B**) of each LLMs. Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Significance was adjusted for multiple tests with the use of a Bonferroni correction

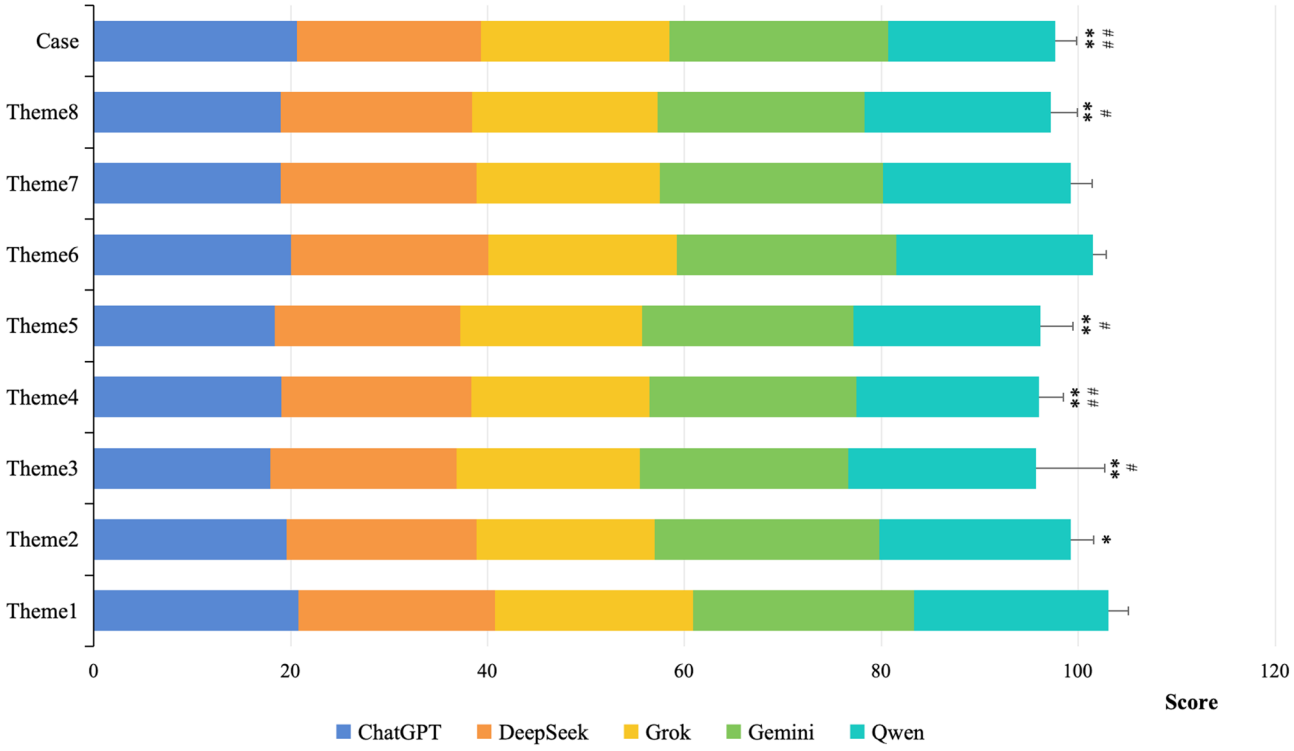

Fig. 5 Comparison of the difficulty of professional questions and cases in different topics and stacking plots of scores of each LLMs for the part of professional questions answering and cases analysis in different topics. Significance codes: Compared with Theme 1, ** $p < 0.01$, * $p < 0.05$; Compared with Theme 6, $^{\#\#}p < 0.01$, $^{\#}p < 0.05$

within dental implantology tasks, a finding corroborated by Taymour et al. [24]. and our team's previous investigations [28]. Through a more robust and multifaceted evaluation, incorporating demanding case analysis scenarios, we provide further evidence for the efficacy of Google Gemini series models in this specialized medical field. Quantifiable results demonstrate Gemini-2.0-flash-Thinking's significant outperformance of other LLMs in both professional question answering and case analysis, with notable strength in the latter, suggesting strong domain knowledge coupled with advanced knowledge integration, clinical reasoning, and decision-making in complex clinical contexts. PCA results further validate Gemini-2.0-flash-Thinking's superior overall capabilities, particularly in dimensions crucial for medical LLMs [49], such as relevance, diversity, and interpretability in question answering, and diagnostic accuracy, treatment plan appropriateness, knowledge application, and logical inference in case analysis. This suggests significant potential for Gemini-2.0-Flash-Thinking in areas like high-level clinical decision support, intricate case analysis, and intelligent patient education within dental implantology.

In contrast, the ChatGPT-o3-mini and DeepSeek-R1 models demonstrated robust and reliable performance, exhibiting dependable professional knowledge retrieval and preliminary case analysis capabilities. ChatGPT-o3-mini achieved upper-mid-range scores in both professional questions answering and case analysis. The potential of ChatGPT series models in applying dental expertise has also been corroborated by Revilla-León et al. [14], who found that ChatGPT-4.0 even outperformed human dentists in the EAO implant dentistry certification exam. Prior research [28] also indicated that ChatGPT-4 demonstrated the most stable and consistent performance in dental implantology question answering. However, the results of this study suggest that the ChatGPT-o3-mini model still has room for improvement in knowledge depth and reasoning ability. This aligns with the findings of Taymour et al. [24] regarding the reliability and effectiveness of LLMs in answering dental implant-related questions. These discrepancies further emphasize the performance differentiation of LLMs across various clinical application scenarios and task types. DeepSeek-R1 model's performance in professional question answering was comparable to ChatGPT-o3-mini. It even surpassed Gemini-2.0-flash-Thinking in specific topics, suggesting that its professional knowledge depth and breadth are similar to leading models, potentially with an advantage in procedural knowledge and clinical adaptability. This study represents the first validation of the clinical application efficacy of the DeepSeek series models in the field of dentistry. DeepSeek-R1's enhanced reasoning ability, achieved through reinforcement learning and multi-stage training, explains

Table 9 Mixed-effects model analysis of each LLMs for the difficulty of professional questions answering and cases analysis scores

Effect	Estimate	Std. Error	df	t	*p*
Intercept	20.7333	0.4195	122.2811	49.425	< 0.001***
Model Main Effects					
DeepSeek	-0.7333	0.4814	142.97	-1.523	0.130
Grok	-0.5667	0.4814	142.97	-1.177	0.241
Gemini	1.7	0.4814	142.97	3.531	< 0.001***
Qwen	-1	0.4814	142.97	-2.077	0.040*
Theme Main Effects					
Theme 2	-1.1333	0.5933	122.2811	-1.91	0.058
Theme 3	-2.8	0.5933	122.2811	-4.72	< 0.001***
Theme 4	-1.7	0.5933	122.2811	-2.866	0.005**
Theme 5	-2.3667	0.5933	122.2811	-3.989	< 0.001***
Theme 6	-0.7	0.5933	122.2811	-1.18	0.240
Theme 7	-1.8	0.5933	122.2811	-3.034	0.003**
Theme 8	-1.8	0.5933	122.2811	-3.034	0.003**
Case	-0.1333	0.5933	122.2811	-0.225	0.823
Model × Theme Significant Interaction Effects					
Gemini × Theme 2	1.5667	0.6809	142.97	2.301	0.023*
DeepSeek × Theme 3	1.72	0.6809	142.97	2.526	0.013*
Gemini × Theme 3	1.6	0.6809	142.97	2.35	0.020*
Qwen × Theme 3	2.1333	0.6809	142.97	3.133	0.002**
Gemini × Theme 5	1.4	0.6809	142.97	2.056	0.042*
Qwen × Theme 5	1.6	0.6809	142.97	2.35	0.020*
DeepSeek × Theme 7	1.7	0.6809	142.97	2.497	0.014*
Qwen × Case	-2.6888	0.7053	144.2371	-3.812	<0.001***

The benchmark model is ChatGPT-o3-mini, and the benchmark topic is Theme1. Only significant results were shown for interaction effects. Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

its strong performance in professional question answering. Notably, DeepSeek-R1 is released in an open-source format, offering flexibility and customization for research and application. This open-source advantage may facilitate the continued improvement of its reasoning capabilities. However, DeepSeek-R1 exhibited mediocre performance in case analysis, with overall performance still slightly inferior to Gemini-2.0-flash-Thinking and ChatGPT-o3-mini models. PCA analysis also confirmed the gap in overall capabilities between ChatGPT-o3-mini and DeepSeek-R1 models compared to Gemini-2.0-flash-Thinking while simultaneously revealing their advantages in specific dimensions such as treatment plan rationality and knowledge application ability. ChatGPT-o3-mini and DeepSeek-R1 models are more suitable for the majority of application scenarios in dental implantology, such as professional knowledge Q&A, learning operational protocols, interpreting clinical guidelines, and preliminary case analysis.

The Grok-3 and Qwen2.5-max models demonstrated a need for performance improvement in dental implantology, indicating relatively limited application potential within this specialized field. Both Grok-3 and Qwen2.5-max exhibited similar overall performance in professional question answering and case analysis, scoring lower than the top three models. While Grok-3-Preview-02-24 ranks highly on the Chatbot Arena LLMs leaderboard, suggesting strong general capabilities [50], its performance in the highly specialized domain of dental implantology was not fully realized. This may stem from its training data and methodologies, which prioritize generalizability and conversational ability over in-depth professional knowledge and rigorous medical reasoning [51]. Instead, these features might divert resources away from specialized knowledge acquisition and clinical skill enhancement, hindering its practical application in the highly specialized field of dental implantology. Qwen2.5-max, as a Chinese language model, also underperformed, potentially indicating limitations in its professional medical knowledge and reasoning, particularly in English-language professional domains. The application potential of the Qwen2.5-max model in dental implantology appears constrained, especially regarding case analysis capabilities. Prior research [28] also noted the insufficient performance of the Qwen 2.0-72B model in professional dental implantology tasks. PCA analysis further corroborated the gap in overall capabilities between Grok-3 and Qwen2.5-max models and the higher-performing tiers.

Notably, the mixed-effects model analysis in this study revealed performance variations among LLMs across different topics and task difficulties. Dental implantology topics can be broadly categorized into two types: easily

Table 10 Descriptive statistics of the scores in five different evaluation dimensions in the professional questions answering of each LLMs

	ChatGPT					DeepSeek					Grok					Gemini					Qwen				
Criterion dimensions	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
Mean	3.6	3.9	4.2	3.8	3.7	3.6	4.0	4.3	3.8	3.8	3.5	3.8	4.1	3.7	3.7	4.0	4.5	4.7	4.3	4.4	3.5	3.9	4.2	3.8	3.8
SE of the mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Median	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	3.0	4.0	4.0	4.0	4.0	4.0	5.0	5.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0
Mode	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	3.0	4.0	4.0	4.0	4.0	4.0	5.0	5.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0
Minimum	3.0	3.0	3.0	2.0	2.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	2.0	3.0	4.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0
Maximum	5.0	5.0	5.0	5.0	5.0	4.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	4.0	5.0	5.0	5.0	5.0
standard deviation	0.5	0.4	0.5	0.5	0.5	0.5	0.4	0.5	0.4	0.5	0.5	0.4	0.4	0.5	0.5	0.3	0.5	0.5	0.5	0.5	0.5	0.3	0.4	0.4	0.4
variance	0.3	0.1	0.2	0.3	0.3	0.2	0.2	0.2	0.2	0.2	0.3	0.2	0.2	0.3	0.3	0.1	0.3	0.2	0.3	0.3	0.3	0.1	0.2	0.2	0.2

Table 11 Descriptive statistics of the scores in five different evaluation dimensions in the cases analysis of each LLMs

	ChatGPT					DeepSeek					Grok					Gemini					Qwen				
Criterion dimensions	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
Mean	4.2	4.1	4.1	4.1	4.1	4.0	3.7	3.6	3.9	3.5	3.9	3.9	4.0	3.6	3.8	4.3	4.8	4.3	4.5	4.3	3.3	3.2	3.6	3.5	3.3
SE of the mean	0.1	0.1	0.0	0.1	0.1	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Median	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	3.0	4.0	4.0	4.0	4.0	4.0	4.0	5.0	4.0	4.0	4.0	3.0	3.0	4.0	4.0	3.0
Mode	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	3.0	4.0	4.0	4.0	4.0	4.0	4.0	5.0	4.0	4.0	4.0	3.0	3.0	4.0	4.0	3.0
Minimum	4.0	3.0	4.0	3.0	3.0	4.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	4.0	4.0	4.0	4.0	4.0	3.0	3.0	3.0	2.0	3.0
Maximum	5.0	5.0	5.0	5.0	5.0	4.0	4.0	4.0	5.0	4.0	5.0	4.0	5.0	4.0	4.0	5.0	5.0	5.0	5.0	5.0	4.0	4.0	4.0	4.0	4.0
standard deviation	0.4	0.5	0.3	0.6	0.5	0.0	0.4	0.5	0.5	0.5	0.6	0.3	0.3	0.5	0.4	0.5	0.4	0.5	0.5	0.5	0.5	0.4	0.5	0.6	0.4
variance	0.2	0.2	0.1	0.4	0.2	0.0	0.2	0.2	0.2	0.3	0.3	0.1	0.1	0.2	0.2	0.2	0.2	0.2	0.3	0.2	0.2	0.2	0.2	0.4	0.2

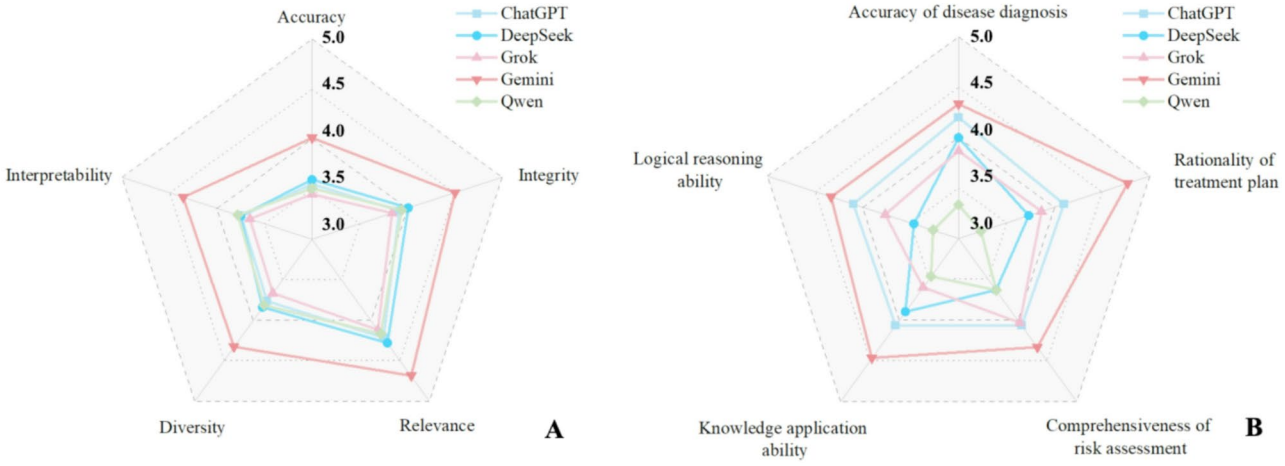

Fig. 6 Radar plots of scores in five different Criterion dimensions for the professional questions answering (A) and cases analysis (B) of each LLMs

addressed topics (such as implant structural design, long-term care, and maintenance), where all LLMs achieved relatively high scores; and challenging topics (such as pre-implantation assessment, implant surgical procedures, bone augmentation techniques, post-implantation healing and restoration, pathophysiology, and case analysis). Across these challenging topics, LLMs scores generally decreased, indicating that LLMs still face challenges when handling complex clinical scenarios, procedural knowledge, and high-difficulty tasks like case analysis. The Gemini-2.0-flash-Thinking model and the DeepSeek-R1 model exhibited significant positive interaction effects in specific high-difficulty topics (such as pre-implantation assessment, implant surgical procedures, post-implantation healing and restoration, and management of unique situations and complications). This suggests that Gemini and DeepSeek models are more suitable for handling complex clinical scenarios. Conversely, the Qwen2.5-max model demonstrated a significant negative interaction effect in case analysis tasks, indicating its limitations in complex case analysis. These interaction effects corroborate the performance differentiation among different LLMs based on professional domain and task difficulty.

Furthermore, the PCA results not only showcased the performance disparities among LLMs in dental implantology tasks but also provided more profound insights into the underlying structure behind the dimensional scores, revealing subtle nuances in the models' strengths and focuses. In professional question answering, PC1 reflected the LLMs' "patient-centred" answering capability, particularly the integrated performance across relevance, diversity, and interpretability dimensions. The Gemini-2.0-flash-Thinking model's prominence on PC1 suggests its superior ability to quickly grasp the core of clinical queries and provide more comprehensive, multi-faceted, and easily understandable answers when responding to clinical inquiries, which is crucial for enhancing doctor-patient communication efficiency. PC2 pertained to the LLMs' capabilities in "knowledge depth and rigour" and revealed a potential trade-off between pursuing answer completeness and maintaining accuracy and interpretability. The PCA results for professional question answering emphasized that an exceptional LLMs should not only possess solid professional knowledge but also be patient-oriented, providing relevant, informative, and easily comprehensible answers. In case analysis, PC1 centrally reflected the LLMs' comprehensive level of "core clinical decision-making ability," as embodied by key dimensions such as diagnostic accuracy, treatment plan rationality, knowledge application ability, and logical reasoning ability. The leading position of the Gemini-2.0-flash-Thinking model on PC1 foreshadows its significant potential in complex case analysis and intelligent treatment planning. PC2 was associated with "risk management and knowledge depth," suggesting that the Grok-3 model may possess a comparative advantage in risk assessment. The PCA results for case analysis emphasized that the core value of LLMs in clinical decision-making lies in the accuracy of disease diagnosis, the scientific validity of treatment plans, the effectiveness of knowledge application, and the rigour of logical reasoning. The inherent connections between dimensions highlighted the comprehensiveness and complexity of clinical decision-making, underscoring the need for high-quality clinical decisions to rely on the synergistic efforts of LLMs across multiple dimensions.

In summary, the performance variations observed among LLMs in dental implantology tasks likely originate from fundamental differences in model architecture, training data, and methodologies. Newer generation LLMs, such as Gemini, DeepSeek, and ChatGPT, leveraging technological advantages, including multimodal information processing and complex logical reasoning,

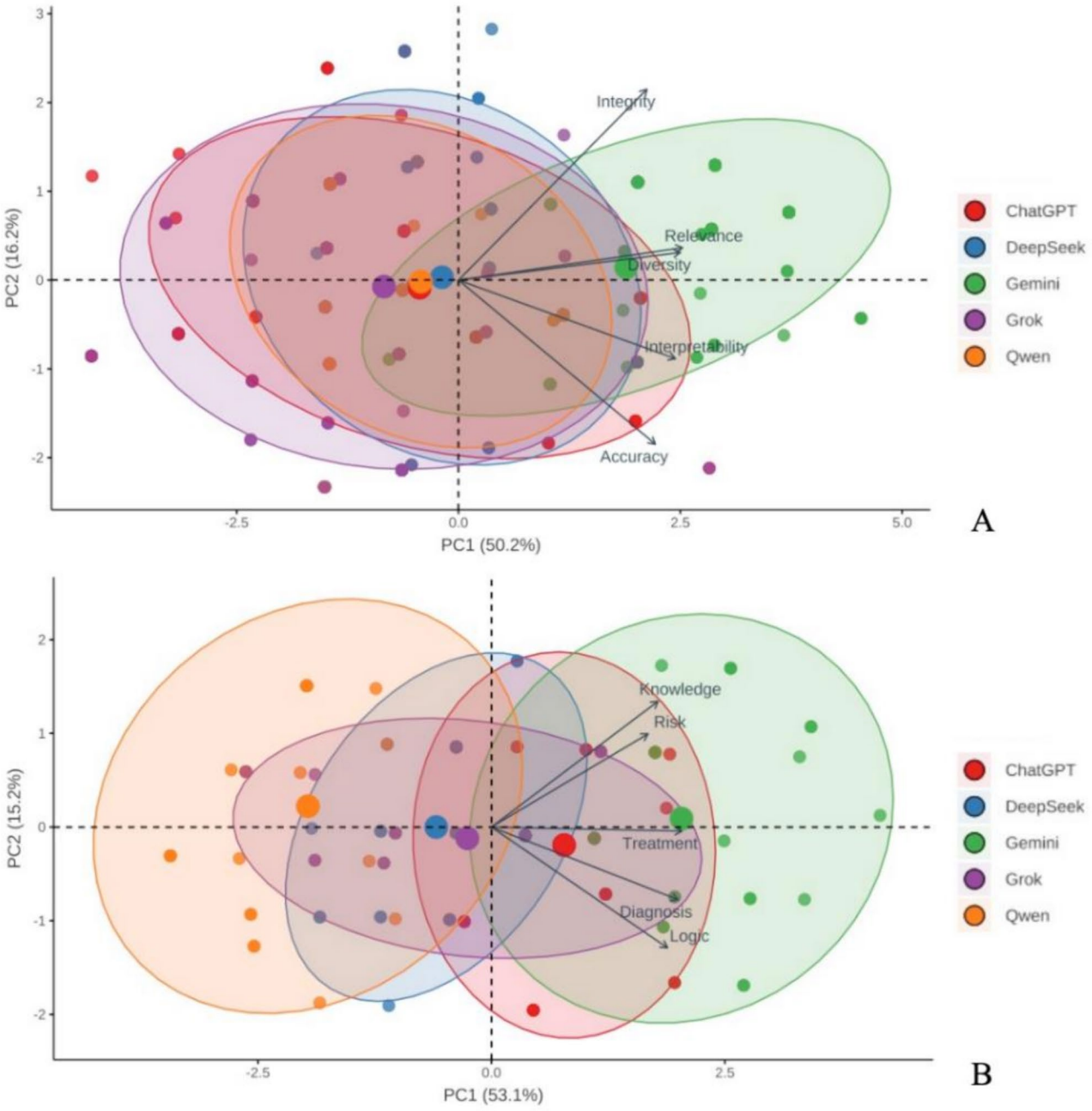

Fig. 7 PCA of scores in five different Criterion dimensions for the professional questions answering (**A**) and cases analysis (**B**) of each LLMs

demonstrate superior professional capabilities and represent the evolving trajectory of LLMs in the medical domain. Models like Grok and Qwen, however, still exhibit limitations in addressing complex clinical decision-making tasks. From a clinical application standpoint, this study offers valuable insights for dental implantologists in selecting appropriate LLMs tools. Clinicians should selectively utilize the advantageous functionalities of different LLMs based on specific application scenarios and requirements. Nevertheless, in high-stakes clinical decision-making contexts involving complex cases, it remains imperative to prioritize personal professional knowledge and experience, treating LLMs outputs as supplementary references. While LLMs hold considerable promise in dentistry, risks such as data bias, insufficient explainability, and "hallucinations" persist. Healthcare professionals must maintain vigilance and ethical awareness, positioning LLMs as adjunctive tools to enhance diagnostic and treatment efficiency and optimize patient management rather than as replacements for independent clinical judgment. This approach is crucial to ensure the healthy and sustainable development of

LLMs technology within dental implantology and to realize its genuine benefits for patients [52].

This study provided a prospective evaluation of the application potential of LLMs in dental implantology, but it is not without limitations. Firstly, this research only assessed five specific LLMs, and due to the rapid pace of technological iteration, the current conclusions may not fully represent the capabilities of future models. Secondly, the research was mainly focused on dental implantology. Additionally, the clinical cases used for evaluation in this study were sourced from the database of a single center, which might restrict the general applicability of the research results. Furthermore, the inherent subjectivity of expert ratings and the evaluation dimensions employed may not have captured the clinical performance of LLMs entirely, and the number of assessment questions and cases was limited. To more accurately evaluate the true capabilities of LLMs, including Gemini, future research should expand the sample size, use prompts that are more suitable for specific clinical tasks, explore more objective and comprehensive assessment methodologies, and strengthen validation in real-world clinical settings, thereby better guiding practical applications.

Conclusion

Within the limitation of the present study, the Gemini-2.0-flash-Thinking model exhibited comparatively better overall performance and further demonstrated the clinical acceptability of LLMs responses to inquiries pertinent to dental implantology. Across specific tasks and evaluation metrics, individual LLMs displayed their respective strengths and unique attributes, clinicians should apply their expertise to evaluate the outcomes provided by LLMs critically.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12903-025-06619-6.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Supplementary Material 5

Supplementary Material 6

Acknowledgements

The authors gratefully acknowledge all experts and scholars who participated in this study for their valuable support and assistance in the successful completion of this research, and we would like to express our sincere gratitude to the Fund Committee of the Pioneer and Leading Goose Technology Project of Zhejiang Province.

Author contributions

Xing Wu: Conceptualization, Methodology, Software, Data curation, Investigation, Formal analysis, Visualization, and Writing - original draft. Guofei Cai: Conceptualization, Data curation, Project administration. Bin Guo: Methodology, Software, Resources, and Writing - review & editing. Leizi Ma: Conceptualization, Methodology, Data curation, and Investigation. Siqi Shao: Conceptualization, Data curation, Formal analysis, Supervision. Jun Yu: Formal analysis, Supervision, Funding acquisition, Resources. Yuchen Zheng: Conceptualization, Software, Data curation, Validation, Visualization, Resources, Writing - original draft, and Writing - review & editing. Linhong Wang: Formal analysis, Supervision, Funding acquisition, Resources, and Writing - review & editing. Fan Yang: Conceptualization, Validation, Supervision, Funding acquisition, Project administration, Resources, and Writing - review & editing. All authors reviewed the manuscript.

Funding

This work was supported by the Pioneer and Leading Goose Technology Project of Zhejiang Province (2024C03094).

Data availability

The original data of this paper are available in the appendix and can be used.

Declarations

Ethics approval and consent to participate

The design and implementation of the present study were conducted in accordance with the Declaration of Helsinki. The study protocol and the use of clinical case data have been approved by the Medical Ethics Committee of Zhejiang Provincial People's Hospital (Approval No. QT2025050) on March 4th, 2025. The Medical Ethics Committee of Zhejiang Provincial People's Hospital has waived informed consent for participants in this study.

Consent for publication

All participants were fully informed about the research procedures and provided written consent.

Competing interests

The authors declare no competing interests.

Clinical trial number

Not applicable.

Author details

[1]School of Stomatology, Zhejiang Chinese Medical University, Hangzhou, Zhejiang, China
[2]Center for Plastic and Reconstructive Surgery, Department of Stomatology, Affiliated People's Hospital, Zhejiang Provincial People's Hospital, Hangzhou Medical College, Hangzhou, Zhejiang, China
[3]Stomatology department, Zhuji second people's Hospital, Zhuji, Zhejiang, China
[4]School of Stomatology, Hangzhou Normal University, Hangzhou, Zhejiang, China
[5]School of Mathematics and Statistics, Jiangxi Normal University, Nanchang, Jiangxi, China
[6]School of Intelligence Science and Engineering, Harbin Institute of Technology (Shenzhen), Shenzhen, Guangdong, China

Received: 4 May 2025 / Accepted: 14 July 2025
Published online: 28 July 2025

References

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. Nat Med. 2023;9(8):1930–40. https://doi.org/10.1038/s41591-023-02448-8
2. Meng X, Yan X, Zhang K, et al. The application of large language models in medicine: a scoping review. iScience. 2024;27(5):109713. https://doi.org/10.1016/j.isci.2024.109713

3. Perera Molligoda Arachchige AS. Large language models (LLM) and chatgpt: a medical student perspective. Eur J Nucl Med Mol Imaging. 2023;50(8):2248–9. https://doi.org/10.1007/s00259-023-06227-y
4. Benítez TM, Xu Y, Boudreau JD, et al. Harnessing the potential of large language models in medical education: promise and pitfalls. J Am Med Inf Assoc. 2024;31(3):776–83. https://doi.org/10.1093/jamia/ocad252
5. Wang D, Liang J, Ye J, et al. Enhancement of the performance of large language models in diabetes education through retrieval-augmented generation: comparative study. J Med Internet Res. 2024;26:e58041. https://doi.org/10.2196/58041
6. Shusterman R, Waters AC, O'Neill S, et al. An active inference strategy for prompting reliable responses from large language models in medical practice. NPJ Digit Med. 2025;8(1):119. https://doi.org/10.1038/s41746-025-01516-2
7. Liu X, Liu H, Yang G, et al. A generalist medical language model for disease diagnosis assistance. Nat Med. 2025;31(3):932–42. https://doi.org/10.1038/s41591-024-03416-6
8. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open. 2024;7(10):e2440969. https://doi.org/10.1001/jamanetworkopen.2024.40969
9. Liu J, Shen H, Chen K, Li X. Large language model produces high accurate diagnosis of cancer from end-motif profiles of cell-free DNA. Brief Bioinform. 2024;25(5):bbae430. https://doi.org/10.1093/bib/bbae430
10. Oh Y, Park S, Byun HK, et al. LLM-driven multimodal target volume contouring in radiation oncology. Nat Commun. 2024;15(1):9186. https://doi.org/10.1038/s41467-024-53387-y
11. Benary M, Wang XD, Schmidt M, et al. Leveraging large language models for decision support in personalized oncology. JAMA Netw Open. 2023;6(11):e2343689. https://doi.org/10.1001/jamanetworkopen.2023.43689
12. Li J, Guan Z, Wang J, et al. Integrated image-based deep learning and language models for primary diabetes care. Nat Med. 2024;30(10):2886–96. https://doi.org/10.1038/s41591-024-03139-8
13. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA. 2025;333(4):319–28. https://doi.org/10.1001/jama.2024.21700
14. Revilla-León M, Barmak BA, Sailer I, Kois JC, Att W. Performance of an artificial intelligence-based chatbot (ChatGPT) answering the European certification in implant dentistry exam. Int J Prosthodont. 2024;37(2):221–4. https://doi.org/10.11607/ijp.8852
15. Kurt Demirsoy K, Buyuk SK, Bicer T. How reliable is the artificial intelligence product large language model ChatGPT in orthodontics? Angle Orthod. 2024;94(6):602–7. https://doi.org/10.2319/031224-207.1
16. Huang H, Zheng O, Wang D, 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. https://doi.org/10.1038/s41368-023-00239-y
17. Chatzopoulos GS, Koidou VP, Tsalikis L, Kaklamanos EG. Large language models in periodontology: assessing their performance in clinically relevant questions. J Prosthet Dent. 2024. https://doi.org/10.1016/j.prosdent.2024.10.020. Published online November 18, 2024.
18. Pradhan P. Accuracy of ChatGPT 3.5, 4.0, 4o and gemini in diagnosing oral potentially malignant lesions based on clinical case reports and image recognition. Med Oral Patol Oral Cir Bucal. 2025;30(2):e224–31. https://doi.org/10.4317/medoral.26824
19. Tastan Eroglu Z, Babayigit O, Ozkan Sen D, Ucan Yarkac F. Performance of ChatGPT in classifying periodontitis according to the 2018 classification of periodontal diseases. Clin Oral Investig. 2024;28(7):407. https://doi.org/10.1007/s00784-024-05799-9
20. Jorba-García A, Bara-Casaus JJ, Camps-Font O, et al. Accuracy of dental implant placement with or without the use of a dynamic navigation assisted system: a randomized clinical trial. Clin Oral Implants Res. 2023;34(5):438–49. https://doi.org/10.1111/clr.14050
21. Ding Y, Zhou H, Zhang W, et al. Evaluation of a platform-switched Morse taper connection for all-on-four or six treatment in edentulous or terminal dentition treatment: a retrospective study with 1–8 years of follow-up. Clin Implant Dent Relat Res. 2023;25(5):815–28. https://doi.org/10.1111/cid.13228
22. Yang F, Ruan Y, Liu Y, et al. Abutment mechanical complications of a Morse taper connection implant system: a 1- to 9-year retrospective study. Clin Implant Dent Relat Res. 2022;24(5):683–95. https://doi.org/10.1111/cid.13115
23. Tonetti MS, Sanz M, Avila-Ortiz G, et al. Relevant domains, core outcome sets and measurements for implant dentistry clinical trials: the implant dentistry core outcome set and measurement (ID-COSM) international consensus report. Clin Oral Implants Res. 2023;34(Suppl 25):4–21. https://doi.org/10.1111/clr.14074
24. Taymour N, Fouda SM, Abdelrahaman HH, Hassan MG. Performance of the ChatGPT-3.5, ChatGPT-4, and Google gemini large language models in responding to dental implantology inquiries. J Prosthet Dent 2025 Published Online January. 2025;4. https://doi.org/10.1016/j.prosdent.2024.12.016
25. Kaygisiz ÖF, Teke MT. Can deepseek and ChatGPT be used in the diagnosis of oral pathologies? BMC Oral Health. 2025;25(1):638. https://doi.org/10.1186/s12903-025-06034-x
26. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. 2024;30(9):2613–22. https://doi.org/10.1038/s41591-024-03097-1
27. Wang L, Shen Y. Evaluating causal reasoning capabilities of large language models: a systematic analysis across three scenarios. Electronics. 2024;13(23):4584. https://doi.org/10.3390/electronics13234584
28. Wu Y, Zhang Y, Xu M, et al. Effectiveness of various general large language models in clinical consensus and case analysis in dental implantology: a comparative study. BMC Med Inf Decis Mak. 2025;25(1):147. https://doi.org/10.1186/s12911-025-02972-2
29. DeepSeek-AI, Guo D, Yang D, et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. arXiv.2025. Accessed: March 18th, 2025. https://doi.org/10.48550/arXiv.2501.12948
30. OpenAI ChatGPT. OpenAI o3-mini: Pushing the frontier of cost-effective reasoning. 2025. At: https://openai.com/index/openai-o3-mini/. Accessed: March 18th, 2025.
31. Gibney E. Scientists flock to deepseek: how they're using the blockbuster AI model. Nat 2025 Published Online January. 2025;29. https://doi.org/10.1038/d41586-025-00275-0
32. Lee D, Kader G. The emergence of strategic reasoning of large language models. 2024. At: https://doi.org/10.48550/arXiv.2412.13013. Accessed: March 18th, 2025.
33. Tordjman M, Liu Z, Yuce M, et al. Comparative benchmarking of the deepseek large language model on medical tasks and clinical reasoning. Nat Med Published Online April. 2025;23. https://doi.org/10.1038/s41591-025-03726-3
34. Samara W, Moztarzadeh O, Hauer L, Babuska V. Dental implant placement in medically compromised patients: a literature review. Cureus. 2024;16(2):e54199. https://doi.org/10.7759/cureus.54199
35. Hussein A, Shah M, Atieh MA, et al. Influence of implant surfaces on peri-implant diseases - a systematic review and meta-analysis. Int Dent J. 2025;75(1):75–85. https://doi.org/10.1016/j.identj.2024.10.007
36. Lemos CAA, de Oliveira AS, Faé DS, et al. Do dental implants placed in patients with osteoporosis have higher risks of failure and marginal bone loss compared to those in healthy patients? A systematic review with meta-analysis. Clin Oral Investig. 2023;27(6):2483–93. https://doi.org/10.1007/s00784-023-05005-2
37. Naseri R, Yaghini J, Feizi A. Levels of smoking and dental implants failure: a systematic review and meta-analysis. J Clin Periodontol. 2020;47(4):518–28. https://doi.org/10.1111/jcpe.13257
38. Wu XY, Shi JY, Qiao SC, et al. Accuracy of robotic surgery for dental implant placement: a systematic review and meta-analysis. Clin Oral Implants Res. 2024;35(6):598–608. https://doi.org/10.1111/clr.14255
39. Qian X, Vánkos B, Kelemen K, et al. Comparison of implant placement and loading protocols for single anterior maxillary implants: a systematic review and network meta-analysis. J Prosthet Dent. 2025;133(3):677–88. https://doi.org/10.1016/j.prosdent.2024.05.033
40. Farina R, Franzini C, Trombelli L, et al. Minimal invasiveness in the transcrestal elevation of the maxillary sinus floor: a systematic review. Periodontol 2000. 2023;91(1):145–66. https://doi.org/10.1111/prd.12464
41. Elboraey MO, Alqutaibi AY, Aboalrejal AN, et al. Regenerative approaches in alveolar bone augmentation for dental implant placement: techniques, biomaterials, and clinical decision-making: a comprehensive review. J Dent. 2025;154:105612. https://doi.org/10.1016/j.jdent.2025.105612
42. Soares LFF, Malzoni CMA, da Silveira ML, et al. Evaluation of different approaches for sinus membrane perforation repair during sinus elevation: a systematic review and meta-analysis. Int J Oral Maxillofac Implants. 2024;39(1):107–18. https://doi.org/10.11607/jomi.10180
43. Kensara A, Hefni E, Williams MA, et al. Microbiological profile and human immune response associated with peri-implantitis: a systematic review. J Prosthodont. 2021;30(3):210–34. https://doi.org/10.1111/jopr.13270
44. Carra MC, Blanc-Sylvestre N, Courtet A, et al. Primordial and primary prevention of peri-implant diseases: a systematic review and meta-analysis. J Clin Periodontol. 2023;50(Suppl):77–112. https://doi.org/10.1111/jcpe.13790

45. Zhang S, Zhang X, Li Y, et al. Clinical reference strategy for the selection of treatment materials for maxillofacial bone transplantation: a systematic review and network meta-analysis. Tissue Eng Regen Med. 2022;19(3):437–50. https://doi.org/10.1007/s13770-022-00445-5
46. Stiesch M, Grischke J, Schaefer P, et al. Supportive care for the prevention of disease recurrence/progression following peri-implantitis treatment: a systematic review. J Clin Periodontol. 2023;50(Suppl 26):113–34. https://doi.org/10.1111/jcpe.13822
47. Camps-Font O, Rubianes-Porta L, Valmaseda-Castellón E, et al. Comparison of external, internal flat-to-flat, and conical implant abutment connections for implant-supported prostheses: a systematic review and network meta-analysis of randomized clinical trials. J Prosthet Dent. 2023;130(3):327–40. https://doi.org/10.1016/j.prosdent.2021.09.029
48. Bienz SP, Pirc M, Papageorgiou SN, et al. The influence of thin as compared to thick peri-implant soft tissues on aesthetic outcomes: a systematic review and meta-analysis. Clin Oral Implants Res. 2022;33(Suppl 23):56–71. https://doi.org/10.1111/clr.13789
49. Tian S, Jin Q, Yeganova L, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Brief Bioinform. 2024;25(1):bbad493. https://doi.org/10.1093/bib/bbad493
50. HKUST Library. Emerging AI tools for literature review: comparison of LLMs. 2025. At: https://libguides.hkust.edu.hk/AI-tools-literature-review/compare-llm/. Accessed: March 18th, 2025.
51. xAI AG. 2023. At: https://x.ai/news/grok/. Accessed: March 18th, 2025.
52. Vimalraj S, Sekaran S. ChatGPT: empowering dentistry with future possibilities. Oral Oncol. 2023;144:106496. https://doi.org/10.1016/j.oraloncology.2023.106496

Publisher's note