Journal of Dental Sciences 20 (2025) 1454–1459

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.e-jds.com

Original Article

Comparison of artificial intelligence systems in answering prosthodontics questions from the dental specialty exam in Turkey

Busra Tosun [a*], Zeynep Sen Yilmaz [b]

[a] Department of Prosthodontics, Faculty of Dentistry, Bolu Abant Izzet Baysal University, Bolu, Turkey
[b] Department of Prosthodontics, Faculty of Dentistry, The University of Atatürk, Erzurum, Turkey

Received 4 January 2025; Final revision received 22 January 2025
Available online 31 January 2025

KEYWORDS
Artificial intelligence;
Large language models;
Multiple-choice question;
Prosthodontics

Abstract *Background/purpose*: Artificial intelligence (AI) is increasingly vital in dentistry, supporting diagnostics, treatment planning, and patient education. However, AI systems face challenges, especially in delivering accurate information within specialized dental fields. This study aimed to evaluate the performance of seven AI-based chatbots (ChatGPT-3.5, ChatGPT-4, Gemini, Gemini Advanced, Claude AI, Microsoft Copilot, and Smodin AI) in correctly answering prosthodontics questions from the Dental Specialty Exam (DUS) in Turkey.
Materials and methods: The dataset for this study consists of 128 multiple-choice prosthodontics questions from the DUS, a national exam administered in Turkey by the Student Selection and Placement Center (ÖSYM) between 2012 and 2021. Chatbot performance was assessed by categorizing the questions into case-based and knowledge-based.
Results: ChatGPT-4 achieved the highest accuracy (75.8 %), while Gemini AI had the lowest (46.1 %). Gemini AI also had more incorrect (69) than correct answers (59). ChatGPT-4 and ChatGPT-3.5 showed significantly higher accuracy in knowledge-based questions compared to case-based ones ($p < 0.05$). For case-based questions, Gemini and Gemini Advanced had the lowest accuracy (36.4 %), while other chatbots averaged 45.5 %. In knowledge-based questions, ChatGPT-4 performed best (78.6 %) and Gemini AI the worst (47 %).
Conclusion: ChatGPT-4 excelled in knowledge-based prosthodontic questions, showing potential to enhance dental education through personalized learning and clinical reasoning support. However, its limitations in case-based scenarios highlight the need for optimization to better address complex clinical situations. These findings suggest that AI models can significantly contribute to dental education and clinical practice.

* Corresponding author. Department of Prosthodontics, Faculty of Dentistry, Bolu Abant Izzet Baysal University, Karaköy Neighborhood, Mehmed Şemseddin Street, Building No:6, Gölköy Campus, Bolu, 14030 Turkey.
E-mail address: dtbusra86@hotmail.com (B. Tosun).

https://doi.org/10.1016/j.jds.2025.01.025

Introduction

AI is revolutionizing healthcare, with large language models serving as essential educational tools for healthcare professionals due to their advanced language capabilities.[1] In dentistry, AI supports the detection of dental diseases with greater accuracy than traditional methods, predicts treatment outcomes using large datasets, and aids in comprehensive treatment planning.[2] The growing availability of electronic dental data has further fueled interest in data-driven AI applications.[3] In recent years, advancements in dentistry have positioned AI as a key player in evolving prosthetic treatment approaches. Prosthetic restorations, which restore function and aesthetics through the design and fabrication of dental prostheses, begin with the impression-taking process.[4] AI has revolutionized this process by enabling next-generation digital systems to analyze impressions rapidly, detect errors instantly, and provide correction suggestions. These innovations reduce errors, accelerate workflows, and significantly enhance productivity.[5]

In dentistry, AI plays a vital role in diagnosing diseases, interpreting radiographs, identifying implants, designing restorations, and detecting caries.[6,7] Despite its growing applications, the accuracy of AI-generated information in prosthodontics remains underexplored. AI is particularly valuable in prosthodontic treatments, reducing human error in tasks such as classification of prostheses, margin line extraction, and implant cementation.[8] CAD/CAM technology, widely used in prosthodontics for fabricating crowns, bridges, and implant-supported restorations, integrates AI and CBCT to enhance implant placement accuracy and optimize prosthetic designs.[9,10] AI-powered CAD/CAM systems leverage machine learning to refine designs, predict material behavior, and add value over traditional systems. In complex aesthetic cases, AI facilitates precise color matching, such as with central incisors or anterior teeth. Additionally, AI enhances implant prosthetics by detecting placement points via intraoral sensors and integrating this data into CAD software for real-time optimization, advancing both design and manufacturing processes.[11]

AI models can support students by generating ideas for education and research.[12] While chatbots are valuable educational tools, their occasional inaccuracies pose a concern in precision-focused fields like dentistry. In Turkey, dental graduates take the Dental Specialty Examination (DUS) to qualify for specialist training, a rigorous test organized by the Ministry of Health and ÖSYM. The DUS evaluates knowledge in basic and clinical sciences through 120 multiple-choice questions, covering 8 specialties, and has been held annually since 2015. Medical education relies on multiple-choice question (MCQ) format examinations to assess knowledge in various disciplines.[13] MCQs are accepted and widely used tools in education that can promote learning strategies.[14] Analytical thinking and problem-solving skills are crucial in training competent physicians, and educators globally are designing MCQ formats to evaluate these abilities.[15] Vegi et al. found that the majority of students viewed MCQ-based examinations positively.[16]

As a more advanced model, GPT-4 can handle complex instructions and has a larger knowledge base. While most studies have focused on ChatGPT, research on other large language models such as Gemini (Google), Claude (Anthropic), Copilot (Microsoft) and Smodin AI is limited.[17] Although GPT-4 has shown improved performance in dentistry knowledge compared to ChatGPT-3.5, both models' understanding of dentistry topics is still limited.[18] The lack of research on chatbot performance in prosthodontics makes it necessary to assess the strengths and weaknesses in this field. Therefore, comparing AI models can help to select the most suitable model for specific applications.

Prosthodontics, a complex specialty that integrates theoretical knowledge with clinical decision-making, provides an ideal framework for assessing AI tools. While most studies focus on single AI models and general dentistry questions, limited research has specifically addressed prosthodontics. This highlights the need for comparative studies on multiple AI models within this specialty. This study evaluates the performance of seven artificial intelligence models—ChatGPT-3.5, ChatGPT-4, Gemini, Gemini Advanced, Claude AI, Microsoft Copilot, and Smodin AI—in answering prosthodontics questions from Turkey's Dentistry Specialization Examination (DUS).

Materials and methods

Sample size determination

In this study, a power analysis was conducted to determine the minimum sample size to ensure adequate statistical power. The analysis showed that a minimum of 97 questions were required with a significance level (α) of 0.05, statistical power (1-β) of 0.80, and an assumed effect size (d) of 0.5. However, in order to strengthen the validity and comprehensiveness of the findings, all 130 available multiple-choice questions from the specialty of prosthodontics were included in the study. The inclusion of the entire dataset provided a more robust assessment of the performance of the AI models and eliminated potential biases due to selective sampling.

Inclusion and exclusion criteria for questions and dataset

This study used 130 multiple-choice prosthodontics questions from the DUS exam (2012—2021) published by ÖSYM, with five answer options per question. Questions from 2022 to 2023, as well as two invalidated 2017 questions, were excluded, leaving 128 questions. These were categorized into case-based and knowledge-based types by two prosthodontic specialists for consistency.

Operation of LLMs

New accounts were created for each AI program. The language models (LLMs) were evaluated using their default configurations without any parameter changes or additional prompts. Multiple-choice questions were directly input as they appeared in the test to assess performance under natural conditions. Prosthodontic questions from the DUS

were uploaded to each chatbot and asked only once to prevent learning biases and performance improvements from repetition.[19]

Performance evaluation method

On August 15, 2024, all questions were posed to the AI models simultaneously, and responses were classified as correct or incorrect using ÖSYM's official answer keys. Accuracy rates were calculated based on correct answers, and two prosthodontic experts independently reviewed all responses to ensure consistency. During the calibration process, the consistency of the scores of the two raters was analyzed with the intraclass correlation coefficient (ICC). As a result of the analysis, it was determined that the correlation coefficient obtained in the inter-observer measurements was above 0.700 and these results indicate that both assessors were sufficient to conduct the study. To avoid learning biases, the chatbots were not given follow-up questions or feedback during the evaluation process.

Statistical analysis

Descriptive statistics (numbers and percentages) were provided in this study. To test the relationship between categorical variables, the Pearson chi-square test was used when the sample size assumption (expected value > 5) was met. In cases where the sample size assumption was not met, Fisher's exact test was applied. The analyses were conducted using IBM SPSS version 25, and the significance level was set at $P < 0.05$.

Results

The distribution of AI-generated answers was analyzed, and relationships between programs were evaluated using the Pearson chi-square test. A statistically significant relationship was found between AI applications and answer accuracy ($P < 0.05$). ChatGPT-4 and ChatGPT-3.5 provided the most correct answers, with ChatGPT-4 achieving 97 correct responses and the highest accuracy percentage (75.8 %), while Gemini AI had the lowest performance, with 59 correct answers and a 46.1 % accuracy rate. Notably, Gemini AI gave more incorrect answers than correct ones (Table 1). A histogram showing the distribution of correct and incorrect answers provided by the AI applications is presented in Fig. 1.

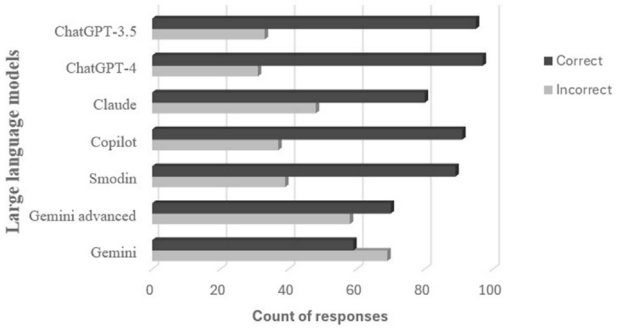

Figure 1 A histogram showing the distribution of correct and incorrect answers provided by AI applications.

A statistically significant difference was found in the accuracy percentages of the seven chatbots. ChatGPT-4 achieved the highest accuracy at 75.8 %, while Gemini AI recorded the lowest at 46.1 %. Significant differences were observed between Gemini and both ChatGPT-3.5 and ChatGPT-4, as well as between Gemini Advanced and these ChatGPT models ($P < 0.05$). However, no significant difference was noted between Gemini AI and Gemini Advanced ($P > 0.05$). Fig. 2 illustrates the accuracy percentages of the AI applications.

The response distribution by question type for each AI application was analyzed using Pearson chi-square and Fisher's exact tests. Statistically significant relationships were found for ChatGPT-4 and ChatGPT-3.5 ($P < 0.05$), with

Table 1 Distribution of responses by chatbots based on question types and total responses.

		Case-based questions			Information-based questions			Total response count			*P*
		n	%	%QT.	n	%	%QT.	n	%	Test statistics	
Gemini	Incorrect	7	10.1	63.6	62	89.9	53.0	69	53.9	0.459[b]	0.498
	Correct	4	6.8	36.4	55	93.2	47.0	59	46.1		
Gemini Advanced	Incorrect	7	12.1	63.6	51	87.9	43.6	58	45.3	—	0.223
	Correct	4	5.7	36.4	66	94.3	56.4	70	54.7		
Smodin	Incorrect	6	15.4	54.5	33	84.6	28.2	39	30.5	—	0.089
	Correct	5	5.6	45.5	84	94.4	71.8	89	69.5		
Copilot	Incorrect	6	16.2	54.5	31	83.8	26.5	37	28.9	—	0.077
	Correct	5	5.5	45.5	86	94.5	73.5	91	71.1		
Claude	Incorrect	6	12.5	54.5	42	87.5	35.9	48	37.5	—	0.329
	Correct	5	6.3	45.5	75	93.8	64.1	80	62.5		
ChatGPT-4	Incorrect	6	19.4	54.5	25	80.6	21.4	31	24.2	—	0.024[a]
	Correct	5	5.2	45.5	92	94.8	78.6	97	75.8		
ChatGPT-3.5	Incorrect	6	18.2	54.5	27	81.8	23.1	33	25.8	—	0.033[a]
	Correct	5	5.3	45.5	90	94.7	76.9	95	74.2		

%: Row percentage and %QT: Column percentage for question type.
[a] $P < 0.05$.
[b] Pearson chi-square test.

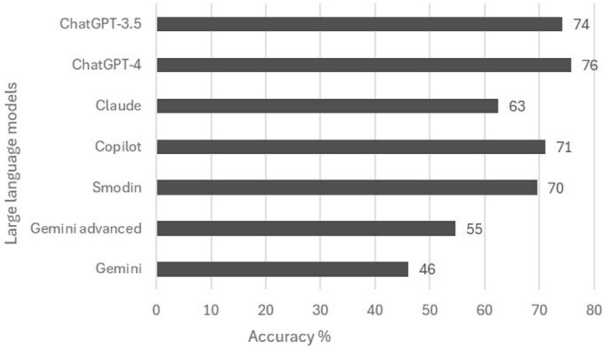

Figure 2 A histogram showing the accuracy percentage of correct answers provided by AI applications.

higher accuracy on knowledge-based questions than case-based ones. No significant relationships were observed for Gemini, Gemini Advanced, Smodin, Copilot, or Claude ($P > 0.05$). Gemini and Gemini Advanced had the lowest accuracy in case-based questions (36.4 %), while other chatbots averaged 45.5 %. For knowledge-based questions, Gemini AI had the lowest accuracy (47 %), and ChatGPT-4 the highest (78.6 %). Detailed results are presented in Table 1.

Discussion

This study evaluates the accuracy of seven artificial intelligence chatbots (ChatGPT-3.5, ChatGPT-4, Gemini, Gemini Advanced, Claude AI, Microsoft Copilot, and Smodin AI) in answering prosthodontics questions from Turkey's Dentistry Specialization Examination (DUS) (2012–2021). The findings reveal the potential applications of these AI models in health education and highlight the risks associated with misinformation. Significant differences in chatbot accuracy rates were observed ($P < 0.05$); therefore, the null hypothesis was rejected.

Large language models (LLMs) are optimized for human-like responses through extensive training.[20] ChatGPT, widely used with high user traffic,[21] showed 75.8 % accuracy in this study, outperforming ChatGPT-3.5 (74.21 %). Similar findings in periodontology,[22] dermatology,[23] and plastic surgery exams[24] confirm its strong potential in health education.

Gemini AI, developed by Google, is a multimodal language model, while Gemini Advanced, based on the Gemini 1.5 Pro model, offers a wider context window for improved recall. In this study, Gemini AI achieved 46.1 % accuracy in prosthodontics questions, with Gemini Advanced slightly higher at 54.68 %, highlighting their limitations in clinical expertise. Despite improving information access, their inaccuracies raise concerns about misinformation and ethical risks.[25,26] Further development is needed to enhance their accuracy and responsible use in education and healthcare.

Claude AI, developed by Anthropic, has limited research on its performance in multiple-choice questions. Agarwal et al.[27] reported that Claude 2 AI outperformed ChatGPT-3.5 in medical physiology questions due to its advanced features, such as a wider context window and ethical safeguards. In this study, Claude AI achieved 62.5 % accuracy, below ChatGPT-3.5's 74.21 %, reflecting the limitations of its earlier version. Similarly, in nephrology exams, Claude AI scored 54.4 %, while ChatGPT-4 achieved 73.3 %.[28] In the Peruvian National Medical Licensure Examination, Claude AI had the lowest performance, whereas ChatGPT-4 excelled.[29] These findings underline Claude AI's struggles with specialized questions, highlighting the potential of improved versions like Claude 2 AI for medical education.

In our study, seven AI programs were evaluated for case-based questions, with Gemini and Gemini Advanced showing the lowest accuracy (36.4 %), while other chatbots averaged 45.5 %. For knowledge questions, Gemini AI had the lowest accuracy (47 %) and ChatGPT-4 the highest (78.6 %). These results indicate that case-based questions, requiring contextual understanding and clinical judgment, pose greater challenges for AI models. Similarly, studies in oral and maxillofacial surgery reported lower accuracy for technical questions requiring similar skills.[30] Buhr et al.[31] reported that ChatGPT underperformed on case-based otolaryngology questions compared to experts in medical competence and conciseness. Similarly, our study highlights that case-based questions are more challenging for AI models than knowledge-based ones due to contextual difficulties, raising concerns about transparency, accountability, and potential biases in clinical decision-making.[32] Pinto et al.[33] evaluated ChatGPT's accuracy on 10 conceptual and 10 case-based questions about urinary incontinence, highlighting ChatGPT-4's significant margin of error and difficulties with contextual integration in case-based questions. They emphasized the need for caution when incorporating this technology into practical applications and stressed the importance of aligning LLM outputs with evidence-based practice. A global survey similarly found that less than 20 % of respondents used ChatGPT in clinical practice, with most citing limitations in academic contexts.[34] Further efforts are needed to enhance LLMs' performance on knowledge-based questions and improve their contextual understanding and clinical reasoning for case-based scenarios.

Our study highlights that large language models excel in structured theoretical knowledge but struggle with tasks requiring higher-order cognitive skills. ChatGPT-4's high accuracy in knowledge-based questions shows its potential as a training tool for dental students, though its limitations in case-based questions call for improvements in clinical reasoning capabilities. This study highlights the benefits and limitations of AI tools in prosthodontic education. ChatGPT-4's high accuracy in knowledge-based questions demonstrates its value as a training tool for dental students. However, its low performance in case-based questions underscores the need for improvements in handling complex clinical scenarios. Critical evaluation of AI-generated answers remains essential.

ChatGPT-4 can aid decision-making in dental clinics by providing quick access to information, analyzing treatment options, and offering recommendations for complex cases. It also enhances patient education through explanations and visualizations. While improving efficiency in treatment planning, these tools should be used cautiously under expert supervision, showcasing their potential as supportive technologies in clinical practice. The growing use of

large language models in healthcare can improve patient access to accurate information, support informed decisions, and enhance treatment adherence.[35,36] However, the low accuracy of models like Gemini AI in this study highlights that not all AI tools are equally effective for education, requiring caution from students and educators. Further development is needed to enhance their consistency and reliability.

LLMs can enhance case-based learning in dental education by generating medical notes, compiling patient information, and assisting in diagnosis and treatment planning.[37] These capabilities enable virtual case simulations and clinical scenarios for students. ChatGPT's success in the US Medical Licensing Examination (USMLE),[38] demonstrates the potential for integrating AI into educational processes in medicine and dentistry, highlighting its contributions across various fields. AI models can analyze incorrect answers, explain correct approaches, and improve students' clinical reasoning skills. Customized training modules offer case-based questions tailored to individual deficiencies. Additionally, LLMs support medical education through interactive Q&A sessions,[39] transferring knowledge and preparing students for real clinical challenges. These capabilities position AI as an indispensable tool in dental education.

Studies on ChatGPT's performance in dental exams provide valuable insights into AI's potential in education. A recent study,[40] comparing ChatGPT-3.5 and 4 on US dental exams (INBDE, DAT, ADAT) found ChatGPT-4 more reliable, with superior accuracy in knowledge-based questions. These findings align with our results, highlighting the need to enhance AI performance for knowledge and case-based questions, demonstrating their potential as innovative tools in dental education globally.

Our study highlights the potential of AI models in dental education, though improvements are needed. Specialized algorithms for case-based questions, training with high-quality dental datasets, and addressing performance weaknesses are essential. Real exam testing and expert feedback will further enhance their effectiveness. This study has some limitations, focusing only on multiple-choice questions and prosthodontics, which restricts generalizability. Future research should evaluate AI tools using diverse question formats, larger datasets, and other dental specialties.

In conclusion, this study represents one of the first efforts to compare multiple artificial intelligence systems within prosthodontics, addressing a critical gap in this field. The results showed that ChatGPT-4 outperformed other models, particularly in knowledge-based questions, highlighting its potential to enhance prosthodontics education through personalized learning modules and virtual case simulations. Additionally, ChatGPT-4 may serve as a valuable tool in fostering clinical reasoning skills.

However, the challenges encountered with case-based questions emphasize the need for further optimization to handle complex clinical scenarios. These findings suggest that AI models like ChatGPT-4 could play a transformative role in dental education by providing innovative tools and supporting clinical decision-making. To achieve widespread adoption in education and healthcare, AI tools must be rigorously evaluated and refined for complex applications. Future research should focus on developing AI systems specifically designed for case-based learning and clinical use.

Declaration of competing interest

The authors have no conflicts of interest relevant to this article.

References

1. Joda T, Bornstein MM, Jung RE, Ferrari M, Waltimo T, Zitzmann NU. Recent trends and future direction of dental research in the digital era. *Int J Environ Res Publ Health* 2020; 17:1987.
2. Alqutaibi AY. Artificial intelligence models show potential in recognizing the dental implant type, predicting implant success, and optimizing implant design. *J Evid Base Dent Pract* 2023;23:101836.
3. Shan T, Tay F, Gu L. Application of artificial intelligence in dentistry. *J Dent Res* 2021;100:232—44.
4. Fogel AL, Kvedar JC. Artificial intelligence powers digital medicine. *NPJ Digit Med* 2018;1:5.
5. Cicciù M, Fiorillo L, D'Amico C, et al. 3D digital impression systems compared with traditional techniques in dentistry: a recent data systematic review. *Mater* 2020;13:1982.
6. Orhan K, Bayrakdar I, Ezhov M, Kravtsov A, Özyürek T. Evaluation of artificial intelligence for detecting periapical pathosis on cone-beam computed tomography scans. *Int Endod J* 2020; 53:680—9.
7. Ender A, Mörmann WH, Mehl A. Efficiency of a mathematical model in generating CAD/CAM-partial crowns with natural tooth morphology. *Clin Oral Invest* 2011;15:283—9.
8. Singi SR, Sathe S, Reche AR, Sibal A, Mantri N. Extended arm of precision in prosthodontics: artificial intelligence. *Cureus* 2022;14:e30962.
9. Ghaffari M, Zhu Y, Shrestha A. A Review of advancements of artificial intelligence in dentistry. *Dent Rev* 2024;4:100081.
10. Dobrzański LA, Dobrzański LB. Dentistry 4.0 concept in the design and manufacturing of prosthetic dental restorations. *Process* 2020;8:525.
11. Bernauer SA, Zitzmann NU, Joda T. The use and performance of artificial intelligence in prosthodontics: a systematic review. *Sens* 2021;21:6628.
12. Ahmed WM, Azhari AA, Alfaraj A, Alhamadani A, Zhang M, Lu CT. The quality of AI-generated dental caries multiple choice questions: a comparative analysis of chatGPT and google bard language models. *Heliyon* 2024;10:e28198.
13. Ali R, Sultan AS, Zahid N. Evaluating the effectiveness of'MCQ development workshop using cognitive model framework: a pre-post study. *J Pakistan Med Assoc* 2021;71:119.
14. Grainger R, Dai W, Osborne E, Kenwright D. Medical students create multiple-choice questions for learning in pathology education: a pilot study. *BMC Med Educ* 2018;18:1—8.
15. Meo SA, Al-Masri AA, Alotaibi M, Meo MZS, Meo MOS. ChatGPT knowledge evaluation in basic and clinical medical sciences: multiple choice question examination-based performance. *Healthc* 2023;11:2046.
16. Vegi VAK, Sudhakar P, Bhimarasetty DM, et al. Multiple-choice questions in assessment: perceptions of medical students from low-resource setting. *J Educ Health Promot* 2022;11:103.
17. Ittarat M, Cheungpasitporn W, Chansangpetch S. Personalized care in eye health: exploring opportunities, challenges, and the road ahead for chatbots. *J Personalized Med* 2023;13:1679.
18. Danesh A, Pazouki H, Danesh K, Danesh F, Danesh A. The performance of artificial intelligence language models in

board-style dental knowledge assessment: a preliminary study on ChatGPT. *J Am Dent Assoc* 2023;154:970—4.
19. Schwendicke F, Singh T, Lee JH, et al. Artificial intelligence in dental research: checklist for authors, reviewers, readers. *J Dent* 2021;107:103610.
20. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. *Commun Med* 2023;3:141.
21. Sarkar S. AI industry analysis: 50 most visited AI tools and their 24B+ traffic behavior. *Writer* 2023.
22. Danesh A, Pazouki H, Danesh F, Danesh A, Vardar-Sengul S. Artificial intelligence in dental education: ChatGPT's performance on the periodontic in-service examination. *J Periodontol* 2024;95:682—7.
23. Lewandowski M, Łukowicz P, Świetlik D, Barańska-Rybak W. ChatGPT-3.5 and ChatGPT-4 dermatological knowledge level based on the specialty certificate examination in Dermatology. *Clin Exp Dermatol* 2024;49:686—91.
24. Hsieh CH, Hsieh HY, Lin HP. Evaluating the performance of ChatGPT-3.5 and ChatGPT-4 on the Taiwan plastic surgery board examination. *Heliyon* 2024;10:e34851.
25. Stokel-Walker C. AI bot ChatGPT writes smart essays-should academics worry? *Nature* 2022 (in press).
26. Chatterjee J, Dethlefs N. This new conversational AI model can be your friend, philosopher, and guide... and even your worst enemy. *Patterns* 2023;4:1—3.
27. Agarwal M, Goswami A, Sharma P. Evaluating ChatGPT-3.5 and Claude-2 in answering and explaining conceptual medical physiology multiple-choice questions. *Cureus* 2023;15:e46222.
28. Wu S, Koo M, Blum L, et al. A comparative study of open-source large language models, gpt-4 and claude 2: multiple-choice test taking in nephrology. *arXiv* 2023:1—7.
29. Torres-Zegarra BC, Rios-Garcia W, Ñaña-Cordova AM, et al. Performance of ChatGPT, Bard, Claude, and Bing on the Peruvian national licensing medical examination: a cross-sectional study. *J Educ Eval Health Prof* 2023;20:30.
30. Balel Y. Can ChatGPT be used in oral and maxillofacial surgery? *J Stomatol Oral Maxillofac Surg* 2023;124:101471.
31. Buhr CR, Smith H, Huppertz T, et al. ChatGPT versus consultants: blinded evaluation on answering otorhinolaryngology case—based questions. *JMIR Med Educ* 2023;9:e49183.
32. Shen Y, Heacock L, Elias J, et al. ChatGPT and other large language models are double-edged swords. *Radiol* 2023;307:e230163.
33. Pinto VB, de Azevedo MF, Wroclawski ML, et al. Conformity of ChatGPT recommendations with the AUA/SUFU guideline on postprostatectomy urinary incontinence. *Neurourol Urodyn* 2024;43:935—41.
34. Eppler M, Ganjavi C, Ramacciotti LS, et al. Awareness and use of ChatGPT and large language models: a prospective cross-sectional global survey in urology. *Eur Urol* 2024;85:146—53.
35. Clarke MA, Moore JL, Steege LM, et al. Health information needs, sources, and barriers of primary care patients to achieve patient-centered care: a literature review. *Health Inf J* 2016;22:992—1016.
36. Ball MJ, Carla Smith N, Bakalar RS. Personal health records: empowering consumers. *J Healthc Inf Manag* 2007;21:77.
37. Alser M, Waisberg E. Concerns with the usage of Chatgpt in academia and medicine: a viewpoint. *Am J Med Open* 2023;9:100036.
38. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ* 2023;9:e45312.
39. Alhur A. Redefining healthcare with artificial intelligence (AI): the contributions of ChatGPT, Gemini, and Co-pilot. *Cureus* 2024;16:e57795.
40. Dashti M, Ghasemi S, Ghadimi N, et al. Performance of ChatGPT 3.5 and 4 on US dental examinations: the INBDE, ADAT, and DAT. *Imaging Sci Dent* 2024;54:271.