BMC Medical Education

RESEARCH Open Access

Exploring the potential of artificial intelligence chatbots in prosthodontics education

Ravza Eraslan[1], Mustafa Ayata[2*], Filiz Yagci[1] and Haydar Albayrak[1]

Abstract

Background The purpose of this study was to evaluate the performance of widely used artificial intelligence (AI) chatbots in answering prosthodontics questions from the Dentistry Specialization Residency Examination (DSRE).

Methods A total of 126 DSRE prosthodontics questions were divided into seven subtopics (dental morphology, materials science, fixed dentures, removable partial dentures, complete dentures, occlusion/temporomandibular joint, and dental implantology). Questions were translated into English by the authors, and this version of the questions were asked to five chatbots (ChatGPT-3.5, Gemini Advanced, Claude Pro, Microsoft Copilot, and Perplexity) within a 7-day period. Statistical analyses, including chi-square and z-tests, were performed to compare accuracy rates across the chatbots and subtopics at a significance level of 0.05.

Results The overall accuracy rates for the chatbots were as follows: Copilot (73%), Gemini (63.5%), ChatGPT-3.5 (61.1%), Claude Pro (57.9%), and Perplexity (54.8%). Copilot significantly outperformed Perplexity ($P = 0.035$). However, no significant differences in accuracy were found across subtopics among chatbots. Questions on dental implantology had the highest accuracy rate (75%), while questions on removable partial dentures had the lowest (50.8%).

Conclusion Copilot showed the highest accuracy rate (73%), significantly outperforming Perplexity (54.8%). AI models demonstrate potential as educational support tools but currently face limitations in serving as reliable educational tools across all areas of prosthodontics. Future advancements in AI may lead to better integration and more effective use in dental education.

Keywords Prosthodontics education, Artificial intelligence applications, Dentistry specialization, AI chatbot evaluation, Clinical decision-support systems

*Correspondence:
Mustafa Ayata
dt.mustafaayata@gmail.com
[1] Department of Prosthodontics, Faculty of Dentistry, Erciyes University, Kayseri, Türkiye
[2] Private Practice, Ortoperio Oral and Dental Health Polyclinic, Kayseri, Türkiye

Background

The rapid advancement of artificial intelligence (AI) chatbots has generated significant interest in their potential applications in medical and dental education [1, 2]. One of the transformative technologies in this field is large language models (LLMs), which offer advanced natural language processing and response generation capabilities, and are increasingly being applied [3, 4]. Trained on extensive datasets, these models can comprehend complex queries and deliver evidence-based responses, which are valuable for education and support both students and professionals in clinical decision-making [5–7]. With these innovative educational functions, LLMs are also gaining attention for exam preparation and quick access to information [8].

AI models are increasingly recognized for their potential to facilitate and complement traditional learning resources [9, 10]. Recent studies have highlighted the role of LLMs as interactive educational tools, facilitating learning by providing access to information and clinical scenarios across various fields of medicine and dentistry [11–13]. Traditionally, medical and dentistry knowledge acquisition has relied on textbooks, academic journals, and search engines. However, for personalized, immediate, and interactive learning experiences, LLMs can be integrated into educational programs and clinical decision-support systems [9]. This integration requires careful consideration [14].

The five major commercial LLMs used in this study are among the most popular models, reaching millions of users worldwide. Chat Generative Pre-Trained Transformer (ChatGPT) 3.5, developed by OpenAI, is a large language model with advanced natural language processing capabilities. Gemini Advanced, produced by Google DeepMind, is a highly trained AI model based on extensive datasets. Claude Pro, created by Anthropic, is a LLM focused on security and interpretability. Microsoft's Copilot model is based on GPT-4 and provides real-time web access to current information. Lastly, Perplexity, developed in 2022 by Aravind Srinivas and his team, functions as an AI-powered search engine that delivers responses in natural language.

Recent studies have highlighted that LLMs, such as ChatGPT and Google Gemini, have demonstrated significant potential in transforming medical and dental diagnostics, especially by improving diagnostic precision and patient outcomes through advanced natural language processing capabilities [15–17]. However, challenges such as hallucinations, outdated information, and accuracy inconsistencies underline the need for ongoing evaluations and refinements in their applications within dental specialties [18, 19].

In dentistry education, the evaluation of broad knowledge using multiple-choice questions (MCQs) holds an important place. Recent studies have compared different LLMs in examinations such as the European Certification in Implant Dentistry Exam [2], Oral and Maxillofacial Surgery Board Exam [20], periodontology in-service exam [21], pediatric dentistry in the Korean National Dental Board Exam [22], Polish Medical-Dental Verification Exam [8] to assess their accuracy and reliability. For the Polish Medical-Dental Verification Exam, ChatGPT and Gemini demonstrated comparable performance to Claude in prosthodontics, with Claude not showing a clear advantage in this specialty despite its overall high accuracy [8]. These results may point out that AI models may lack data in the field of prosthodontics.

In Turkey, the Dentistry Specialization Residency Examination (DSRE) is a test consisting of MCQs with five options each, assessing candidates' knowledge in various fundamental and clinical sciences, including prosthodontics. This exam is mandatory for those seeking to pursue specialty education in dentistry after graduation. This study aimed to assess and compare the performance of ChatGPT-3.5, Gemini Advanced, Claude Pro, Copilot, and Perplexity in the field of prosthodontics, using questions from the DSRE in Turkey. To the best of the authors' knowledge, the present study is the first study comparing performance of five different AI chatbots on questions of DSRE in the literature. The findings of this study aim to offer valuable insights into the integration of AI applications in educational processes and clinical decision-support systems, while exploring their potential as reliable educational tools.

The first null hypothesis of this study posited that there was no statistically significant difference in the accuracy of responses provided by different AI chatbot models for prosthodontics questions. The second null hypothesis assumed that there was no statistically significant difference in the accuracy of responses across subtopics among these models.

Materials and methods

The questions of the DSRE held in Turkey since 2012 can be accessed from the internet link: https://www.osym.gov.tr/TR,15070/dus-cikmis-sorular.html. The system has 13 exams installed and is open to public access. There are 10 questions on prosthodontics in each exam. Although there were 130 questions in total, 4 questions were excluded because they contained images. All questions were translated into English by authors, and this version was used for the study (Additional file 1).

The remaining 126 questions were divided into 7 subtopics according to the content. These categories were dental morphology, materials science, fixed dentures, removable partial dentures, complete dentures, occlusion/temporomandibular joint (TMJ), and dental implantology (Fig. 1). The questions consisting of MCQs with

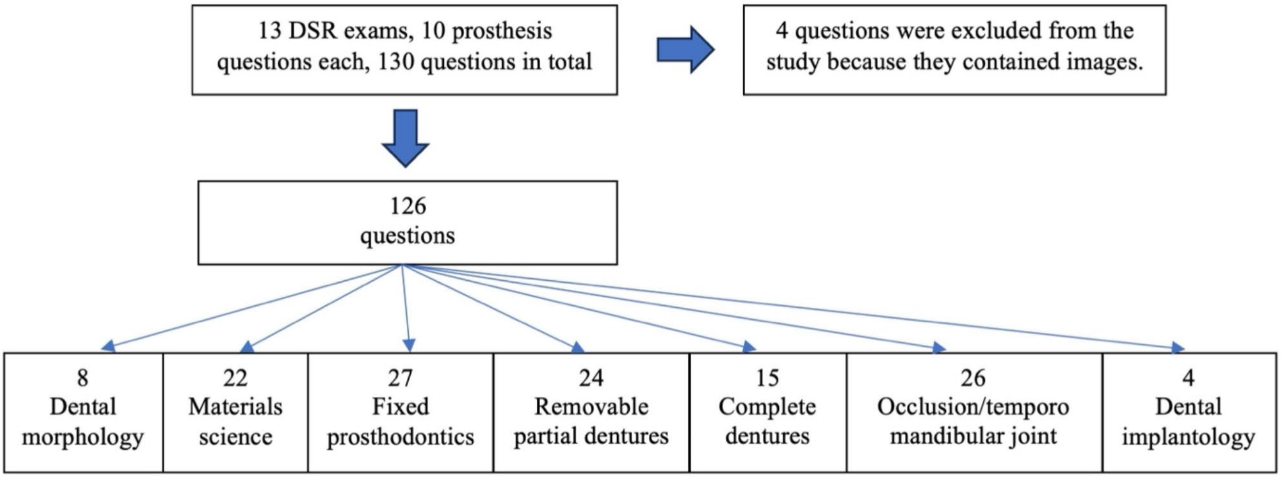
13 DSR exams, 10 prosthesis questions each, 130 questions in total
4 questions were excluded from the study because they contained images.
126 questions
8 Dental morphology
22 Materials science
27 Fixed prosthodontics
24 Removable partial dentures
15 Complete dentures
26 Occlusion/temporo mandibular joint
4 Dental implantology

Fig. 1 Distribution and categorization of prosthodontics questions from the DSRE

five options each and a correct answer were asked to 5 different AI chatbots. The answers were recorded as correct or incorrect. Following the identification of correct and incorrect responses, the performance of the AI models was evaluated using accuracy metrics. Accuracy was calculated as the percentage of correct responses relative to the total number of questions. In this study, the accuracy percentage was determined by dividing the number of correct answers provided by each model by the total number of questions and multiplying the result by 100.

The primary application programming interface (API) of each chatbot was used for the simulation of real-world interactions. The following approaches were employed: The APIs were accessed using https://copilot.microsoft.com/; https://chat.openai.com/; https://claude.ai.com/; https://gemini.google.com/; https://www.perplexity.ai/ respectively for Copilot (Microsoft Copilot, Model: GPT-4, Microsoft, Redmond, WA, USA), ChatGPT (Chat Generative Pre-Trained Transformer 3.5, Model: GPT-3.5, OpenAI, San Francisco, CA, USA), Gemini (Gemini Advanced, Google DeepMind, London, United Kingdom), Claude (Claude Pro, Anthropic, San Francisco, CA, USA), and Perplexity (Perplexity AI, Aravind Srinivas et al., San Francisco, CA, USA). To simulate the most used form, free versions were selected except Claude and Gemini. All the questions were asked of each chatbot within seven days (September 28–October 5, 2024). The chat was reset prior to each new question. A new chat was initiated for each question.

The questions were grouped based on the number of AI models that provided incorrect answers. Group 2 included questions for which two AI models gave incorrect answers; Group 3 included questions for which three AI models gave incorrect answers; Group 4 included questions for which four AI models gave incorrect answers; and Group 5 included questions for which all five AI models gave incorrect answers.

Additionally, a level classification was made based on the number of AI models that provided incorrect answers simultaneously. Level 0 referred to questions where all AI models simultaneously provided correct answers; Level 1 referred to questions where one AI model provided an incorrect answer; Level 2 referred to questions where two AI models simultaneously provided incorrect answers; Level 3 referred to questions where three AI models simultaneously provided incorrect answers; Level 4 referred to questions where four AI models simultaneously provided incorrect answers; and Level 5 referred to questions where all five AI models simultaneously provided incorrect answers.

The relationship between the responses given to all questions or questions of each subtopic and the type of AI was analyzed by Pearson Chi-Square tests and Fisher's Exact Test. Comparisons between columns were performed with z-tests. All analyses were performed in a statistical program (SPSS 20) at a significance level of 0.05.

Ethical approval was not applicable to this study as the study does not involve any humans or animals.

Results

When the accuracy of all the answers given to the questions was evaluated using the chi-square analysis, a statistically significant difference was found among AI chatbots (Table 1). The overall accuracy rates for the chatbots were as follows: Copilot achieved the highest accuracy rate of 73%, answering 92 out of 126 questions correctly. ChatGPT followed with an accuracy rate of 61.1% (77/126), while Gemini showed a similar performance with an accuracy rate of 63.5% (80/126). Claude demonstrated an accuracy rate of 57.9% (73/126), and Perplexity had the lowest performance with an accuracy

Table 1 Accuracy rates of AI chatbots in Prosthodontics questions

	Copilot	ChatGPT	Claude	Gemini	Perplexity	*P*
Correct	92 (73%) ^{a}	77 (61.1%) $^{a, b}$	73 (57.9%) $^{a, b}$	80 (63.5%) $^{a, b}$	69 (54.8%) ^{b}	0.035
Incorrect	34 (27%)	49 (38.9%)	53 (42.1%)	46 (36.5%)	57 (45.2%)	

Each subscript letter denotes AI chatbots whose column proportions do not differ significantly from each other

Table 2 Subtopic-wise Accuracy Rates of AI chatbots in Prosthodontics

		Copilot	ChatGPT	Claude	Gemini	Perplexity	Total	*P*
Dental morphology (*N* = 40)	**Correct**	5^{a} (62.5%)	5 ^{a} (62.5%)	4 ^{a} (50%)	7 ^{a} (87.5%)	5 ^{a} (62.5%)	26 (65%)	0.672*
	Incorrect	3 (37.5%)	3 (37.5%)	4 (50%)	1 (12.5%)	3 (37.5%)	14 (35%)	
Materials science (*N* = 110)	**Correct**	17 ^{a} (77.3%)	13 ^{a} (59.1%)	15 ^{a} (68.2%)	15 ^{a} (68.2%)	13 ^{a} (59.1%)	73 (66.4%)	0.684
	Incorrect	5 (22.7%)	9 (40.9%)	7 (31.8%)	7 (31.8%)	9 (40.9%)	37 (33.6%)	
Fixed dentures (*N* = 135)	**Correct**	23 ^{a} (85.2%)	18 ^{a} (66.7%)	14 ^{a} (51.9%)	19 ^{a} (70.4%)	16 ^{a} (59.3%)	90 (56.7%)	0.105
	Incorrect	4 (14.8%)	9 (33.3%)	13 (48.1%)	8 (29.6%)	11 (40.7%)	45 (33.3%)	
Removable partial dentures (*N* = 120)	**Correct**	14 ^{a} (58.3%)	13 ^{a} (54.2%)	11 ^{a} (45.8%)	13 ^{a} (54.2%)	10 ^{a} (41.7%)	61 (50.8%)	0.772
	Incorrect	10 (41.7%)	11 (45.8%)	13 (54.2%)	11 (45.8%)	14 (58.3%)	59 (49.2%)	
Complete dentures (*N* = 75)	**Correct**	9 ^{a} (60%)	7 ^{a} (46.7%)	7 ^{a} (46.7%)	11 ^{a} (73.39%)	7 ^{a} (46.7%)	41 (54.7%)	0.487
	Incorrect	6 (40%)	8 (53.3%)	8 (53.3%)	4 (26.7%)	8 (53.3%)	34 (45.3%)	
Occlusion and TMJ (*N* = 130)	**Correct**	20 ^{a} (76.9%)	18 ^{a} (69.2%)	19 ^{a} (73.1%)	14 ^{a} (53.8%)	15 ^{a} (57.7%)	86 (66.2%)	0.330
	Incorrect	6 (23.1%)	8 (30.8%)	7 (26.9%)	12 (46.2%)	11 (42.3%)	44 (33.8%)	
Dental implantology (*N* = 20)	**Correct**	3 ^{a} (75%)	3 ^{a} (75%)	3 ^{a} (75%)	3 ^{a} (75%)	3 ^{a} (75%)	15 (75%)	1.0*
	Incorrect	1 (25%)	1 (25%)	1 (25%)	1 (25%)	1 (25%)	5 (25%)	

*Fisher's Exact test. Each subscript letter denotes a subset of groups whose column proportions do not differ significantly from each other

Table 3 Accuracy Rates Across Subtopics in Prosthodontics regardless of AI type

	Dental morphology	Materials science	Fixed dentures	Removable partial dentures	Complete dentures	Occlusion TMJ	Dental implantology	*P*
Correct	26 $^{a, b, c, d, e, f, g, h, i}$ (65.0%)	73$^{f, g, h, i}$ (66.4%)	90$^{d, e, h, i}$ (66.7%)	61^{c} (50.8%)	41$^{b, c, e, g, i}$ (54.7%)	86 $^{a, b, d, e, f, g, h, i}$ (66.2%)	16 $^{a, d, f, h}$ (80.0%)	0.013
Incorrect	14 (35.0%)	37 (33.6%)	45 (33.3%)	59 (49.2%)	34 (45.3%)	44 (33.8%)	4 (20.0%)	

Each subscript letter denotes a subset of groups whose column proportions do not differ significantly from each other

rate of 54.8% (69/126). Table 1 summarizes these results. Copilot answered a significantly higher number of questions correctly compared to Perplexity ($P = 0.035$). For all subtopics, the rate of correct answers did not differ statistically among the AI chatbots (Table 2).

A statistically significant relationship was found between the subtopics and the rate of correct answers provided (Table 3). The rate of the correct answers to questions on removable partial dentures was found to be lower than that of fixed dentures, materials science, occlusion/TMJ, and dental implantology ($P = 0.013$). The rate of the correct answers to questions on dental implantology was found to be significantly higher than that of complete dentures and removable partial dentures ($P = 0.013$).

When the similarity ratios of incorrect answers were analyzed, the rate in Group 5, where all AI models

Table 4 Similarity ratios of incorrect answers by groups

	Group 2	Group 3	Group 4	Group 5	Total
Mean ± SD (%)	78.95 ± 25.36	74.20 ± 17.64	75 ± 18.26	81.82 ± 20.89	76.95 ± 20.46
Range	50	67	50	60	67
Min-max	50–100	33–100	50–100	40–100	33–100

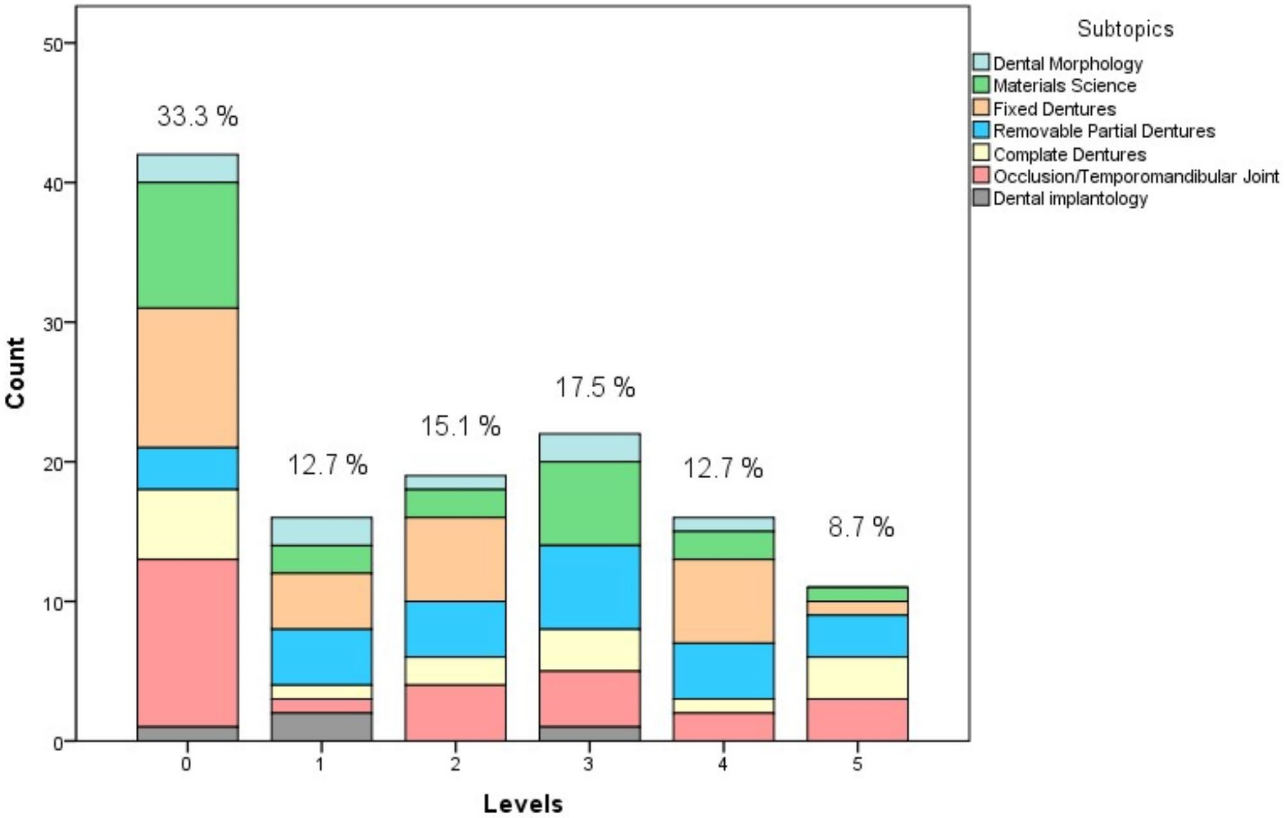

Fig. 2 Grouping of questions with simultaneous incorrect answers by AIs: A summary table of the number of incorrect answers to the questions: Level 0 includes questions with no incorrect answers, where all AIs answered correctly simultaneously; Level 1 consists of questions where 1 AI provided an incorrect answer; Level 2 includes questions where 2 AIs provided incorrect answers simultaneously; Level 3 covers questions where 3 AIs provided incorrect answers simultaneously; Level 4 consists of questions where 4 AIs provided incorrect answers simultaneously; and Level 5 includes questions where all 5 AIs provided incorrect answers

simultaneously provided incorrect answers, was found to be the highest at 81.82% (Table 4).

At least one AI gave an incorrect answer to 67.7% of the questions. All AIs answered incorrectly simultaneously for 8.7% of the questions, while all AIs answered correctly simultaneously for 33.3% of the questions (Fig. 2).

Discussion

In this study, the accuracy of responses provided by five different AI models to prosthodontics questions in DSRE was compared. Statistical analysis revealed a significant difference among the chatbots, and the first null hypothesis was rejected. Copilot, which had the highest correct answer rate (73%), demonstrated a notably better performance compared to Perplexity (54.8%) ($P = 0.035$). The lowest correct answer rate that was observed in Perplexity suggests that this model has a limited database in the context of prosthodontics. However Hancı et al. [4] highlighted that Perplexity stood out in reliability and quality evaluations (e.g., JAMA and Modified DISCERN scores) compared to other artificial intelligence models. These findings highlight Perplexity's potential to deliver detailed and comprehensive information within a limited range of topics, rather than covering a wide breadth of knowledge.

Ahmad et al. investigated the performance of ChatGPT-4.0, Claude 3 Opus, and Gemini Advanced, using the performance of second-year periodontics residents in the periodontology in-service examination. ChatGPT-4 demonstrated the highest accuracy rate at 92.7%, followed by Gemini Advanced with 81.6% and Claude 3 Opus with 78.5%, whereas second-year periodontics residents achieved a rate of only 61.9% ($P<0.001$) [21]. Similarly, recent studies have shown that LLMs like ChatGPT and Google Gemini can significantly enhance diagnostic accuracy and patient outcomes in various dental specialties, including prosthodontics and endodontics. For example, a study evaluating Google Gemini's performance in endodontics highlighted its potential for accurate diagnostics, despite occasional errors such as incomplete responses or hallucinations in complex scenarios [15, 17]. These findings align with the current study's results, where Gemini Advanced demonstrated moderate accuracy in prosthodontics.

Revilla-León et al. evaluated the performance of ChatGPT-3.5, ChatGPT-4.0, and licensed dentists in responding to the European Certification Exam in implant dentistry [2]. They found a statistical difference between ChatGPT-4.0 (mean score of 84%) and ChatGPT-3.5 (mean score of 72%) ($P<0.001$); and between ChatGPT-4.0 and dentists (mean score of 74%) ($P=0.010$). However, no statistically significant difference was found in the mean score between the ChatGPT-3.5 and dentist groups ($P=0.828$). ChatGPT-4.0 was revealed the highest score among the groups tested [2]. In the present study ChatGPT-3.5 revealed the third highest score after Copilot and Gemini. However, the results might differ if ChatGPT-4, the subscription-based version, were used. To simulate mostly used form by the public, the free version was included in this study. However, the advanced versions of the Gemini and Claude chatbots were utilized for expanded functionality due to the limited number of questions the free versions could answer.

In addition, a systematic review by Umer et al. emphasized that LLMs are increasingly being utilized to assist with patient queries and clinical decision-making, particularly in areas like prosthodontics and oral surgery [10]. However, challenges such as the reliance on outdated data and accuracy inconsistencies underscore the need for further refinements in these technologies. This aligns with the current study's findings, where varying performance levels were observed among the models, highlighting the importance of database quality and model training.

Chen et al. evaluated ChatGPT-4.0 by posing 509 neurology questions from an online question bank. The cumulative performance of ChatGPT compared to users' total correct answers showed a success rate of 65.82% in a single attempt and 75.25% across three attempts. The total correct answers provided by users were reported to be 72.63%. A statistically significant difference was not found between Chat-GPT 4.0 and the users [13].

Wójcik et al. [8] highlighted that Claude provided higher accuracy rates in integrated medical fields and emphasized the importance of performance alignment across different languages. However, in the present study, Claude demonstrated a moderate accuracy rate, indicating that it may not be sufficiently optimized for specific areas of dentistry.

When the number of correct answers was investigated according to the subtopic no statistically difference was found among AIs. Therefore, the second null hypothesis has been accepted. Interestingly Gemini revealed an 87% correct answer rate on dental morphology and Copilot revealed an 85.2% correct answer rate on fixed dentures. Guerra et al. were evaluated of the performance of ChatGPT-4 on 643 Congress of Neurological Surgeons Self-Assessment Neurosurgery Exam (SANS) board-style questions by comparing SANS users, medical students, and neurosurgery residents. ChatGPT-4 (79%) outperformed medical students (26.3%), neurosurgery residents (61.5%), and the national average of SANS users (69.3%) across all categories [7].

In the study conducted by Warwas and Heim on oral and maxillofacial surgery examinations, ChatGPT-4 was reported to perform better in general medical topics such as pharmacology (92.8%) and anatomy (73.3%), but showed significantly weaker performance in specialized topics like implants (37.5%) and reconstructive surgery (42.9%) [20]. Farhadi Nia et al. highlighted that while AI models like ChatGPT have improved accessibility and efficiency in dentistry, they often show limitations in specialized areas such as removable partial dentures [16]. Similarly, in this study, lower performance was observed, particularly in specific topics such as removable partial dentures (50.8%) and complete dentures (54.7%) regardless of AI type.

Jung et al. investigated the performance of ChatGPT-3.5 and Gemini in answering pediatric dentistry questions from the Korean National Dental Board Exam. It was stated that both AI-based chatbots (ChatGPT-3.5 35.3 ± 5.6% and Gemini 33.0 ± 4%) could not sufficiently answer the questions [22]. Mahajan et al. investigated the potential use and reliability of ChatGPT for answering questions in the otolaryngology-head and neck surgery exam. The study demonstrated a correct answer rate of 53% and a correct explanation rate of 54%. They found that with increasing difficulty of questions there was a decreasing rate of answer and explanation accuracy [6]. Table 4; Fig. 2 highlight the shared challenges AI models face in specific areas and reveal their limitations in these domains. The group data presented in Table 4, particularly in Group 5 (where all models provided incorrect answers), shows a high similarity rate (81.82%), indicating that some question types are especially challenging for the models. This can be linked to knowledge deficiencies in areas such as removable partial dentures, which demonstrated a low accuracy rate (50.8%). The proportion of questions where all models provided correct answers simultaneously (Level 0, 33.3%) is significantly higher than the proportion of questions where all models provided incorrect answers (Level 5, 8.7%) (Fig. 2). These results indicate that while AI models are effective in general topics relying on broad knowledge bases, their performance in specific and in-depth topics remains limited. In the context of education, it is evident that AI needs to be optimized to enhance knowledge transfer in complex and specialized areas.

This study has several limitations that warrant consideration. Firstly, the limited dataset of 126 questions from the DSRE may not comprehensively represent the full scope of prosthodontics knowledge. Secondly, only two of the five AI models evaluated, (Claude and Gemini) were tested in their advanced versions, while the other three were assessed in their standard versions. This discrepancy in model versions may have contributed to the observed performance differences. Additionally, the study focused solely on a specific field of expertise, which limits the generalizability of the findings. Moreover, repeatability was not assessed in this study, which constitutes another limitation. Future research should consider larger and more diverse datasets, evaluate all models using their advanced versions, and explore a broader range of specialized topics for a more comprehensive assessment.

Conclusions

Copilot achieved the highest accuracy rate, with 73% (92/126) correct answers, while Perplexity had the lowest, at 54.8% (69/126). Among the subtopics, dental implantology demonstrated the highest accuracy rate (75%), whereas removable partial dentures had the lowest (50.8%).

AI models demonstrate potential as educational support tools but currently face limitations in serving as reliable educational or clinical decision-support tools across all areas of prosthodontics. However, as models like Copilot demonstrate higher accuracy rates, selecting the appropriate AI model is crucial for the effective integration of AI into educational processes. As AI applications continue to evolve, it is likely that more successful outcomes will be achieved in the future.

Abbreviations

AI	Artificial intelligence
DSRE	Dentistry Specialization Residency Examination
LLMs	Large language models
ChatGPT	Chat Generative Pre-Trained Transformer
MCQs	Multiple-choice questions
TMJ	Temporomandibular joint
API	Application programming interface

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12909-025-06849-w.

Supplementary Material 1

Acknowledgements
Not applicable.

Author contributions

Conceptualization: R.E. and M.A. Formal analysis: R.E. and M.A. Investigation: R.E., M.A., F.Y., and H.A. Methodology: R.E., M.A., F.Y., and H.A. Supervision: R.E. Writing-Original Draft: R.E. and M.A. Writing-Review & Editing: R.E., M.A., F.Y., and H.A. Visualization: M.A. Validation: F.Y. and H.A. All authors have read and agreed to the published version of the manuscript.

Funding
This study received no financial support.

Data availability

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

Declarations

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare no competing interests.

Informed consent

Informed consent were not required.

Institutional review board

This study does not involve human or animal subjects and therefore did not require approval from the Institutional Ethical Board.

Received: 3 January 2025 / Accepted: 10 February 2025
Published online: 27 February 2025

References

1. Noda R, Izaki Y, Kitano F, Komatsu J, Ichikawa D, Shibagaki Y. Performance of ChatGPT and Bard in self-assessment questions for nephrology board renewal. Clin Exp Nephrol. 2024;28(5):465–9.
2. 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).
3. Schwendicke Fa, Samek W, Krois J. Artificial intelligence in dentistry: chances and challenges. J Dent Res. 2020;99(7):769–74.
4. Hancı V, Ergün B, Gül Ş, Uzun Ö, Erdemir İ, Hancı FB. Assessment of readability, reliability, and quality of ChatGPT®, BARD®, Gemini®, Copilot®, Perplexity® responses on palliative care. Medicine. 2024;103(33):e39305.
5. 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(11):970–4.
6. Mahajan AP, Shabet CL, Smith J, Rudy SF, Kupfer RA, Bohm LA. Assessment of artificial intelligence performance on the Otolaryngology Residency In-Service exam. OTO open. 2023;7(4):e98.
7. Guerra GA, Hofmann H, Sobhani S, Hofmann G, Gomez D, Soroudi D, et al. GPT-4 artificial intelligence model outperforms ChatGPT, medical students, and neurosurgery residents on neurosurgery written board-like questions. World Neurosurg. 2023;179:e160–5.
8. Wójcik D, Adamiak O, Czerepak G, Tokarczuk O, Szalewski L. A comparative analysis of the performance of chatGPT4, Gemini and Claude for the Polish Medical final diploma exam and Medical-Dental Verification exam. medRxiv (preprint) 2024:2024.07.
9. Suchman K, Garg S, Trindade AJ. Chat generative pretrained transformer fails the multiple-choice American College of Gastroenterology self-assessment test. Am J Gastroenterol. 2023;118(12):2280–2.

10. Umer F, Batool I, Naved N. Innovation and application of large Language models (LLMs) in dentistry–a scoping review. BDJ open. 2024;10(1):90.
11. Ali MJ. ChatGPT and lacrimal drainage disorders: performance and scope of improvement. Ophthalmic Plast Reconstr Surg. 2023;39(3):221–5.
12. Huang H, Zheng O, Wang D, Yin J, Wang Z, Ding S, et al. ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model. Int J Oral Sci. 2023;15(1):29.
13. Chen TC, Multala E, Kearns P, Delashaw J, Dumont A, Maraganore D et al. Assessment of ChatGPT's performance on neurology written board examination questions. BMJ Neurol Open 2023;5(2).
14. Lum ZC. Can artificial intelligence pass the American Board of Orthopaedic Surgery examination? Orthopaedic residents versus ChatGPT. Clin Orthop Relat Res. 2023;481(8):1623–30.
15. Díaz-Flores García V, Freire Y, Tortosa M, Tejedor B, Estevez R, Suárez A. Google Gemini's performance in endodontics: a study on answer precision and reliability. Appl Sci. 2024;14(15):6390.
16. Nia MF, Ahmadi M, Irankhah E. Transforming dental diagnostics with artificial intelligence: advanced integration of ChatGPT and large language models for patient care. 2024;arXiv preprint arXiv:2406.06616.
17. Suárez A, Díaz-Flores García V, Algar J, Gómez Sánchez M, Llorente de Pedro M, Freire Y. Unveiling the ChatGPT phenomenon: evaluating the consistency and accuracy of endodontic question answers. Int Endod J. 2024;57(1):108–13.
18. Freire Y, Laorden AS, Pérez JO, Sánchez MG, García VDF, Suárez A. ChatGPT performance in prosthodontics: Assessment of accuracy and repeatability in answer generation. J Prosthet Dent. 2024;131(4):659–e1.
19. Suárez A, Jiménez J, de Pedro ML, Andreu-Vázquez C, García VDF, Sánchez MG, Freire Y. Beyond the scalpel: assessing ChatGPT's potential as an auxiliary intelligent virtual assistant in oral surgery. Comput Struct Biotechnol J. 2024;24:46–52.
20. Warwas FB, Heim N. Performance of GPT-4 in Oral and Maxillofacial Surgery Board Exams: Challenges in Specialized Questions. Research Square (preprint). 2024.
21. Ahmad B, Saleh K, Alharbi S, Alqaderi H, Jeong YN. Artificial Intelligence in Periodontology: Performance Evaluation of ChatGPT, Claude, and Gemini on the In-service Examination. medRxiv (preprint) 2024:2024.05.
22. Jung YS, Chae YK, Kim MS, Lee H-S, Choi SC, Nam OH. Evaluating the accuracy of artificial intelligence-based chatbots on pediatric dentistry questions in the Korean national dental board exam. J Korean Acad Pediatr Dent. 2024;51(3):299–309.

Publisher's note