BMC Medical Education

RESEARCH

Open Access

Assessing ChatGPT-4's performance on the US prosthodontic exam: impact of fine-tuning and contextual prompting vs. base knowledge, a cross-sectional study

Mahmood Dashti[1*], Farshad Khosraviani[2], Tara Azimi[3], Delband Hefzi[4], Shohreh Ghasemi[5], Amir Fahimipour[6], Niusha Zare[7], Zohaib Khurshid[8] and Syed Rashid Habib[9*]

Abstract

Background Artificial intelligence (AI), such as ChatGPT-4 from OpenAI, has the potential to transform medical education and assessment. However, its effectiveness in specialized fields like prosthodontics, especially when comparing base to fine-tuned models, remains underexplored. This study evaluates the performance of ChatGPT-4 on the US National Prosthodontic Resident Mock Exam in its base form and after fine-tuning. The aim is to determine whether fine-tuning improves the AI's accuracy in answering specialized questions.

Methods An official sample questions from the 2021 US National Prosthodontic Resident Mock Exam was used, obtained from the American College of Prosthodontists. A total of 150 questions were initially considered, and resources were available for 106 questions. Both the base and fine-tuned models of ChatGPT-4 were tested under simulated exam conditions. Performance was assessed by comparing correct and incorrect responses. The Chi-square test was used to analyze accuracy, with significance set at $p < 0.05$. The Kappa coefficient was calculated to measure agreement between the models' responses.

Results The base model of ChatGPT-4 correctly answered 62.7% of the 150 questions. For the 106 questions with resources, the fine-tuned model answered 73.6% correctly. The Chi-square test showed a significant improvement in performance after fine-tuning ($p < 0.001$). The Kappa coefficient was 0.39, indicating moderate agreement between the models ($p < 0.001$). Performance varied by topic, with lower accuracy in areas such as Implant Prosthodontics, Removable Prosthodontics, and Occlusion, though the fine-tuned model consistently outperformed the base model.

Conclusions Fine-tuning ChatGPT-4 with specific resources significantly enhances its accuracy in answering specialized prosthodontic exam questions. While the base model provides a solid baseline, fine-tuning is essential for improving AI performance in specialized fields. However, certain topics may require more targeted training to achieve higher accuracy.

*Correspondence:
Mahmood Dashti
dashti.mahmood72@gmail.com
Syed Rashid Habib
syhabib@ksu.edu.sa

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

Keywords Artificial intelligence, Prosthodontics, Education, Dental, ChatGPT, Natural Language processing

Background

Artificial intelligence (AI) and large language models (LLMs), such as ChatGPT-4 (OpenAI, San Francisco, CA, USA, 2024), are increasingly used in medical and dental education, including licensing assessments [1, 2]. While their performance has shown promise in general dental exams, their effectiveness in subspecialties like prosthodontics remains underexplored. Assessing AI accuracy in this focused domain may reveal its limitations and potential as a supplementary tool for postgraduate exam preparation [3].

Chau et al. found that ChatGPT 3.5 and 4.0 correctly answered 68.3% and 80.7% of U.S. dental licensing exam questions, respectively [4]. Similarly, Danesh et al. reported 61.3% and 76.9% accuracy for ChatGPT 3.5 and 4 across national U.S. dental exams [5], though AI errors still occur [6]. Other studies have evaluated ChatGPT's performance in Japan's dental hygienist exam [7] and U.S. dental tests like INBDE and ADAT [8], emphasizing the need for targeted training to improve model performance.

This study uses "fine-tuning" to refer to contextual prompting (in-context learning), where domain-specific materials are added during the prompt, not through model retraining [9]. This strategy, supported by recent research, helps LLMs deliver more accurate responses without altering the underlying model [9].

The U.S. National Prosthodontic Resident Exam, administered by the American Board of Prosthodontics (ABP), includes multiple-choice, case-based, and applied questions covering key topics like dental materials and treatment planning [10, 11]. As such, it serves as a robust benchmark for evaluating AI performance in prosthodontics.

This study evaluates ChatGPT-4's accuracy on prosthodontic board-style questions in both its base form and after contextual prompting. It tests two null hypotheses: [1] ChatGPT-4 would not provide accurate responses for the American Prosthodontics Board Examination; and [2] no significant performance difference would exist between the base and contextually prompted models. To our knowledge, no prior study has examined the effects of contextual prompting using prosthodontics-specific resources. This work aims to fill that gap and inform how AI tools can be optimized for specialized medical education.

Materials and methods

Study design

This study was an observational cross-sectional analysis conducted according to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines to ensure transparent and comprehensive reporting [12].

Ethical approval

As the study involved secondary data analysis without human subjects or patient interaction, ethical approval was not required [13].

Data source

The dataset consisted of questions and answers from the 2021 National Prosthodontic Resident Mock Exam, obtained from the American College of Prosthodontists (ACP) website. Due to ACP limitations, the full set of questions cannot be included as a supplementary file. The questions were knowledge-based and derived from a variety of established prosthodontic references, which were cited individually for each question.

The 2021 US National Prosthodontic Resident Mock Exam included 150 multiple-choice questions, each with only one correct answer. The official answers were taken directly from the ACP website. Of these 150 questions, supporting resources were available for 106 questions. The base model received all 150 questions using a specific prompt, while the fine-tuned model received only the 106 questions that had accompanying references.

Study phases

The questions were submitted to ChatGPT-4 Plus (April 2024 version) using one Plus account. A new chat session was created for each model, and upon completion of base model testing, that session was deleted before starting the fine-tuned model testing. Each model received each question only once, without any indication of whether the prior answer was correct or incorrect, in order to replicate a real exam environment and minimize response bias.

The study was conducted in two main phases:

Base model testing

- The initial phase involved testing the base model of ChatGPT-4 without any prior fine-tuning. At first the initial prompt was placed, then on each next prompt one question was inserted and asked the model to answer the question. After answering the question, without any response to whether it was correct or wrong we asked the next question. The prompt used was:

"You are a prosthodontist taking the 2021 US National Prosthodontic Resident Exam. Please choose the best answer for the following question based on your current knowledge."

Fine-tuned model testing

- In the subsequent phase, the model was fine-tuned through contextual prompting, also referred to as in-context learning. This involved providing domain-specific reference materials (such as textbook excerpts, clinical guidelines, peer-reviewed journal content) during the same prompt session as the corresponding question. These contextual resources were drawn from the authoritative prosthodontic sources listed in the ACP's answer key. For each question, the prompt included both the question and a relevant excerpt from its cited resource. This method relied strictly on contextual input provided at inference time, without any model retraining or architectural modifications. The following prompt was used:

"You are a prosthodontist taking the 2021 US National Prosthodontic Resident Exam. Please read the provided PDF file and then choose the best answer for the following question."

Statistical analysis

The responses from both the base model and the fine-tuned model were collected and compared to evaluate the AI's ability to answer the exam questions accurately and to assess the impact of fine-tuning on performance.

The Chi-square test was employed to compare the response accuracy of the two models of the software. Additionally, the Kappa agreement coefficient was calculated to assess the level of agreement between the responses. The significance level for all statistical analyses was set at 0.05.

Results

Overall performance

For the base model testing, all 150 exam questions were presented to ChatGPT-4. However, for the fine-tuned model, we were unable to locate reference materials for 44 of the questions, limiting the fine-tuned evaluation to 106 questions. In the base model assessment, ChatGPT-4 correctly answered 94 out of 150 questions (62.7%), while 56 questions (37.3%) were incorrect (Table 1). In comparison, the fine-tuned model (ChatGPT-tuned) achieved 78 correct answers out of 106 questions (73.6%), with 28 questions (26.4%) answered incorrectly. (Table 2; Fig. 1).

In Table 2, the analysis included 106 questions common to both models of the software.

Comparative analysis

The Chi-square test was applied to compare the accuracy of responses between the two models. The results indicated a significant improvement in performance with fine-tuning ($p < 0.001$), demonstrating that the ChatGPT-tuned model performed better overall.

The Kappa agreement coefficient between the base model and the fine-tuned model was 0.39, which is statistically significant ($p < 0.001$) and indicates moderate agreement between the two models in their responses (Table 3).

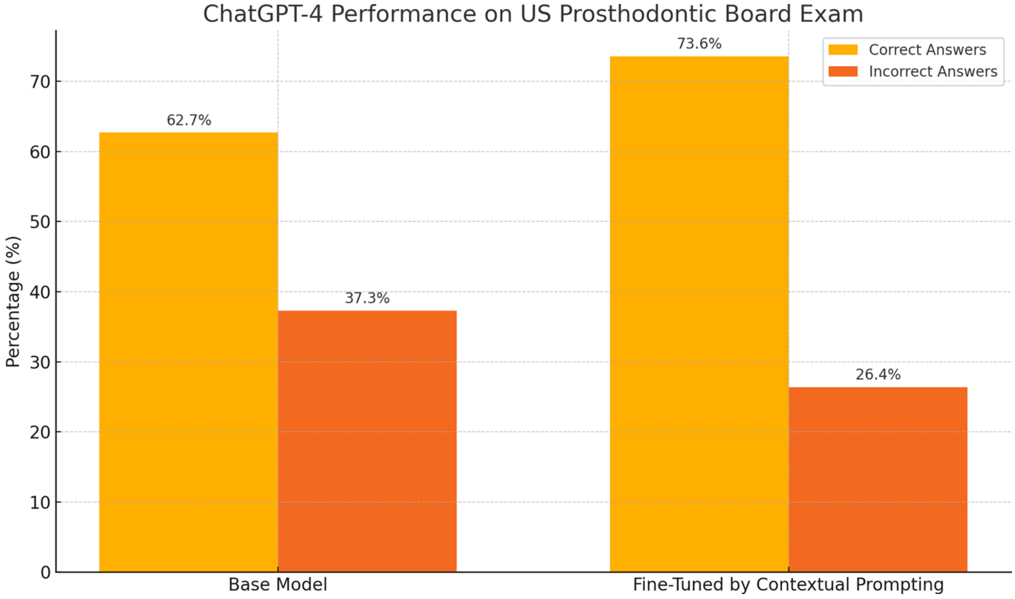

Fig. 1 Performance comparison of ChatGPT-4 in base and contextually prompted conditions. Bars show both correct and incorrect answer percentages

Topic-specific performance

Further analysis by question type, as presented in Table 2, showed no significant difference in response accuracy between the two models across specific topics. This indicates that while there is a general improvement in accuracy with fine-tuning, it is not related to the specific type of questions. For the base model (ChatGPT), the lowest correct response rates were observed in Implant Prosthodontics (60%), Removable Prosthodontics (57.9%), and Occlusion (57.1%), corresponding to incorrect response rates of 40%, 42.1%, and 42.9%, respectively. In the contextually prompted model, the lowest correct response rates were also seen in Implant Prosthodontics (64.3%), Removable Prosthodontics (66.7%), and Occlusion (66.7%), corresponding to incorrect response rates of 35.7%, 33.3%, and 33.3%, respectively.

Discussion

Main findings

The evaluation of ChatGPT-4's performance on the 2021 US National Prosthodontic Resident Mock Exam provides important insights into the AI's capabilities and limitations. This study aimed to assess how well the base model of ChatGPT-4 performs in a specialized field like prosthodontics and to evaluate the improvements gained through fine-tuning with targeted resources. The first null hypothesis stated that ChatGPT-4 would not be able to provide accurate responses for the American Prosthodontic Board Examination. This hypothesis was rejected because the base model of ChatGPT-4 achieved a correct answer rate of 62.7% on 150 questions, demonstrating a moderate level of accuracy even without additional support. This result suggests that the model, even in its base form, possesses substantial domain-relevant knowledge.

The second null hypothesis proposed that there would be no significant performance difference between the base model and the contextually prompted model. This hypothesis was also rejected. The contextually prompted model correctly answered 73.6% of 106 questions, significantly outperforming the base model's 64.2% accuracy on the same set. The difference was statistically significant according to the Chi-square test ($p < 0.001$), indicating that contextual prompting substantially improved the model's performance.

The results show that the non-fine-tuned model of the ChatGPT-4 achieved a correct response rate of 62.7% on the 150 questions from the prosthodontic exam (Table 1). This baseline performance is notable given the complexity and specificity of the exam content. The fact that nearly two-thirds of the answers were correct suggests that ChatGPT-4, even without fine-tuning, has a substantial amount of general knowledge relevant to the field of prosthodontics. However, the 37.3% incorrect response rate highlights the limitations of a general-purpose AI when applied to specialized domains without additional training.

Fine-tuning with specific resources for each question led to a significant improvement in ChatGPT-4's performance. As shown in Table 2, the fine-tuned model (ChatGPT-tuned) correctly answered 73.6% of the 106 questions, compared to 64.2% by the non-fine-tuned model. This enhancement underscores the importance of providing targeted information to AI models, which improves their ability to deliver accurate and relevant responses in specialized fields. Statistical analysis using the Chi-square test confirmed that this improvement in performance was significant ($p < 0.001$).

A detailed analysis of performance across different topics reveals additional insights. Both models of the ChatGPT-4 showed varying levels of accuracy across subjects. For instance, the non-fine-tuned model struggled particularly with topics like Esthetics, Implant Surgery, and Occlusion, where the correct response rates were notably lower (Table 1). Although the fine-tuned model exhibited improvements in these areas, challenges persisted.

For example, in the category of Implant Prosthodontics, the non-fine-tuned model had a correct response rate of 60%, which increased to 64.3% after fine-tuning (Table 2).

Comparison with the existing evidence

Despite these gains, certain areas, such as Implant Surgery, continued to show relatively low accuracy even after fine-tuning, suggesting that more intensive or specialized training may be needed to further enhance performance. A study by Freire et al. [14]. highlighted ChatGPT's limitations in generating answers related to Removable Dental Prostheses (RDPs) and tooth-supported Fixed Dental Prostheses (FDPs). Our study found that only 11 out of 19 questions related to RPDs (57.9%) were answered correctly, indicating potential gaps in the AI's knowledge. Dashti et al. [6]. also demonstrated that the ChatGPT software can assist dentists and dental students in preparing for U.S. dental examinations. Chau et al. study on U.S., and U.K. dental licensing examinations, scored 80.7% and 62.7% with ChatGPT 4.0 [4]. Danesh et al., showed that ChatGPT 4 scored 76.9% respectively on questions that obtained from different national U.S. dental examinations sources [5]. They both show that ChatGPT 4 can be used for helping students on taking dental licensing examinations.

A recent study by Eraslan et al. evaluated five widely used AI chatbots, including ChatGPT-3.5, across 126 prosthodontics questions from the Dentistry Specialization Residency Examination (DSRE). Their results showed variable performance, with Microsoft Copilot achieving the highest accuracy at 73%, while ChatGPT-3.5 scored 61.1%. Notably, the lowest accuracy

Table 1 ChatGPT answers to the examination questions

	ChatGPT				
	Total	**Correct**		**Incorrect**	
Overall	150	94	62.7%	56	37.3%
Topic					
Anatomy	7	5	71.4%	2	28.6%
Caries	3	2	66.7%	1	33.3%
Dx & Tx Plan	4	3	75%	1	25.0%
Dx & Tx Planning	1	1	100%	0	0.0%
EBD	3	2	66.7%	1	33.3%
Emerging Tech	2	1	50%	1	50.0%
Emerging Technology	1	1	100%	0	0.0%
Endo	1	0	0.0%	1	100%
Esthetics	5	2	40%	3	60%
Fixed Pros	11	6	54.5%	5	45.5%
Geriatrics	1	1	100%	0	0.0%
Implant Pros	15	9	60%	6	40%
Implant Pros/Esthetics	1	0	0.0%	1	100%
Implant Surg	1	0	0.0%	1	100%
Implant Surgery	8	3	37.5%	5	62.5%
Infection Control	1	0	0.0%	1	100%
Lab Procedures	4	3	75%	1	25.0%
MF	4	1	25%	3	75%
Materials	15	11	73.3%	4	26.7%
Medical Emergencies	2	2	100%	0	0.0%
Occlusion	14	8	57.1%	6	42.9%
Oral Path	4	4	100%	0	0.0%
Pain Control	1	1	100%	0	0.0%
Perio	5	4	80%	1	20%
Pharm	3	2	66.7%	1	33.3%
Pre Pros Surgery	2	1	50%	1	50%
Radiology	2	2	100%	0	0.0%
Removable Pros	19	11	57.9%	8	42.1%
Risk Assessment	1	1	100%	0	0.0%
Sleep Apnea	3	2	66.7%	1	33.3%
TMD	4	3	75%	1	25%
Wound Healing	2	2	100%	0	0.0%

Note: Abbreviations used in the table: EBD = Evidence-Based Dentistry; Dx = Diagnosis; Tx = Treatment; Pros = Prosthodontics; Endo = Endodontics; Fixed Pros = Fixed Prosthodontics; Implant Pros = Implant Prosthodontics; Implant Surg = Implant Surgery; MF = Maxillofacial; Oral Path = Oral Pathology; Perio = Periodontics; Pharm = Pharmacology; Pre Pros Surgery = Pre-Prosthetic Surgery; Removable Pros = Removable Prosthodontics; TMD = Temporomandibular Disorder

was observed in the subtopic of removable partial dentures (50.8%), a trend consistent with our study's findings where both base and fine-tuned ChatGPT-4 models struggled with Removable Prosthodontics [15]. This cross-study similarity suggests that certain prosthodontic subdomains present consistent challenges across AI models and highlights the need for targeted training in these areas. Unlike Eraslan's study, which used static chatbot models without fine-tuning, our work demonstrates how contextual prompting can significantly improve performance, suggesting a potential path forward for optimizing AI-driven educational support tools.

In comparison with the study by Künzle and Paris [16], which assessed ChatGPT-3.5, 4.0, and 4.0o on restorative and endodontics questions, our findings similarly underscore the superior performance of ChatGPT-4.0 on specialized dental assessments. However, while their focus remained on baseline model accuracy across different model versions, our study contributes novel insights by simulating real-world exam settings and applying contextual prompting for performance enhancement within a single model iteration. Similarly, Revilla-León et al. [17] reported that ChatGPT-4.0 outperformed not only ChatGPT-3.5 but also licensed dentists in the European Certification in Implant Dentistry exam. Unlike these prior studies that focused on model comparison or outperforming humans, our study emphasizes fine-tuning strategies and their implications for educational support tools. These distinctions underscore the originality of our

Table 2 Comparing the answers of the ChatGPT and ChatGPT-tuned to the examination questions

	Total	**ChatGPT**				**ChatGPT-tuned**				***P*-value***
		Correct		**Incorrect**		**Correct**		**Incorrect**		
Overall	106	68	64.2%	38	35.8%	78	73.6%	28	26.4%	< 0.001
Topic										
Anatomy	3	3	4.4%	0	0.0%	2	2.6%	1	3.6%	0.708
Caries	3	2	2.9%	1	2.6%	1	1.3%	2	7.1%	0.500
Dx & Tx Plan	3	2	2.9%	1	2.6%	3	3.8%	0	0.0%	0.501
Dx & Tx Planning	1	1	1.5%	0	0.0%	1	1.3%	0	0.0%	-
EBD	2	2	2.9%	0	0.0%	2	2.6%	0	0.0%	-
Emerging Tech	2	1	1.5%	1	2.6%	2	2.6%	0	0.0%	0.500
Emerging Technology	1	1	1.5%	0	0.0%	1	1.3%	0	0.0%	-
Endo	1	0	0.0%	1	2.6%	1	1.3%	0	0.0%	-
Esthetics	3	1	1.5%	2	5.3%	3	3.8%	0	0.0%	0.200
Fixed Pros	8	4	5.9%	4	10.5%	5	6.4%	3	10.7%	0.500
Geriatrics	1	1	1.5%	0	0.0%	1	1.3%	0	0.0%	-
Implant Pros	14	9	13.2%	5	13.2%	12	15.4%	2	7.1%	0.192
Implant Pros/Esthetics	1	0	0.0%	1	2.6%	1	1.3%	0	0.0%	0.500
Implant Surg	1	0	0.0%	1	2.6%	0	0.0%	1	3.6%	-
Implant Surgery	5	2	2.9%	3	7.9%	5	6.4%	0	0.0%	0.083
Infection Control	1	0	0.0%	1	2.6%	0	0.0%	1	3.6%	-
Lab Procedures	2	2	2.9%	0	0.0%	2	2.6%	0	0.0%	-
MF	4	1	1.5%	3	7.9%	1	1.3%	3	10.7%	0.786
Materials	7	5	7.4%	2	5.3%	5	6.4%	2	7.1%	0.720
Medical Emergencies	0	0	0.0%	0	0.0%	0	0.0%	0	0.0%	-
Occlusion	9	6	8.8%	3	7.9%	6	7.7%	3	10.7%	0.690
Oral Path	2	2	2.9%	0	0.0%	2	2.6%	0	0.0%	-
Pain Control	1	1	1.5%	0	0.0%	1	1.3%	0	0.0%	-
Perio	5	4	5.9%	1	2.6%	2	2.6%	3	10.7%	0.262
Pharm	0	0	0.0%	0	0.0%	0	0.0%	0	0.0%	-
Pre Pros Surgery	1	1	1.5%	0	0.0%	1	1.3%	0	0.0%	-
Radiology	2	2	2.9%	0	0.0%	1	1.3%	1	3.6%	0.500
Removable Pros	15	8	11.8%	7	18.4%	10	12.8%	5	17.9%	0.355
Risk Assessment	1	1	1.5%	0	0.0%	1	1.3%	0	0.0%	-
Sleep Apnea	3	2	2.9%	1	2.6%	2	2.6%	1	3.6%	0.800
TMD	2	2	2.9%	0	0.0%	2	2.6%	0	0.0%	-
Wound Healing	2	2	2.9%	0	0.0%	2	2.6%	0	0.0%	-
confidence level: 0.05										

Note: Abbreviations used in the table: EBD = Evidence-Based Dentistry; Dx = Diagnosis; Tx = Treatment; Pros = Prosthodontics; Endo = Endodontics; Fixed Pros = Fixed Prosthodontics; Implant Pros = Implant Prosthodontics; Implant Surg = Implant Surgery; MF = Maxillofacial; Perio = Periodontics; Pharm = Pharmacology; Pre Pros Surgery = Pre-Prosthetic Surgery; Removable Pros = Removable Prosthodontics; TMD = Temporomandibular Disorder

Table 3 Kappa agreement results between the ChatGPT and the ChatGPT-tuned

		ChatGPT-tuned		**Total**	**Kappa ***	***P*-value**
		Correct	**Incorrect**			
ChatGPT	**Correct**	59 (55.7%)	9 (8.5%)	68 (64.2%)	0.39	< 0.001
	Incorrect	19 (17.9%)	19 (17.9%)	38 (35.8%)		
Total		78 (73.6%)	28 (26.4%)			

* Measure of agreement

approach in evaluating AI augmentation via in-context learning within a prosthodontics-specific domain.

The Kappa agreement coefficient between the base and fine-tuned models was 0.39 ($p < 0.001$), indicating moderate agreement (Table 3). This suggests that while the fine-tuned model performed better overall, there was still considerable variability in responses between the two models. The moderate Kappa coefficient also reflects the complexity of the subject matter and the inherent challenges in training AI to consistently apply specialized knowledge.

Implications for medical education

These findings have significant implications for the use of AI in medical education. The enhanced performance of the fine-tuned model demonstrates the potential of tailored AI training to improve educational tools. By providing AI models with specific resources and training materials, educators can significantly enhance the accuracy and reliability of AI-based educational aids. While this study did not assess direct educational outcomes such as student performance or learning retention, the improved response accuracy of ChatGPT through fine tuning and contextual prompting suggests a potential role as a supplementary tool for board exam preparation. Future studies should explore its utility in real-world learning. Moreover, the variability in performance across different topics highlights the need for continuous refinement of AI training processes. A one-size-fits-all approach may not be effective for all subject areas within a complex field like prosthodontics. Instead, a dynamic and iterative training process, involving multiple rounds of fine-tuning and the incorporation of diverse and comprehensive training materials, may be necessary to optimize AI performance.

Ethical and pedagogical implications

The ability of large language models such as ChatGPT-4 to answer licensing exam questions with high accuracy raises important questions about the future of assessment in dental and medical education. If AI tools can reliably pass standardized tests, this challenges the extent to which such exams assess deep understanding, clinical reasoning, or critical thinking. There is also an ethical dimension regarding the use of AI during preparation and potentially during assessment, particularly concerning over-reliance by students or inequitable access to such tools. These developments underscore the need to rethink exam design, focusing more on skills that AI cannot replicate and ensuring that AI supports (rather than replaces) human learning.

The increasing use of AI in exam preparation raises ethical considerations, particularly regarding equitable access, over-reliance, and the need for responsible integration into curricula. As Daungsupawong and Wiwanitkit [18] highlight, tools like ChatGPT may be used unethically if clear guidelines are not established. Students may overdepend on AI-generated answers, potentially weakening critical thinking and long-term knowledge retention. Chau et al. [19] similarly advocate for AI to be used as a complement to (not a replacement for) human learning. Developing international guidelines for ethical AI use in dental education, as suggested by organizations such as the WHO and FDI, will be essential for ensuring fair and effective implementation of these tools.

Limitations and future research

The scope of this study was limited to the prosthodontic exam, and the findings may not be generalized to other medical or dental specialties. Future research should examine the performance of AI models across various fields and types of exams to validate these results. Additionally, this study only compared a baseline model with a single session of contextual prompting. No model retraining or technical fine-tuning was performed. Furthermore, this study focused solely on model accuracy in answering exam questions and did not evaluate educational outcomes such as student learning, knowledge retention, or instructional effectiveness.

Another limitation is the relatively modest sample size, particularly the subset of 106 questions used for contextual prompting. While the dataset was based on real, board-style exam questions, a larger and more diverse question pool would provide more robust statistical power and allow for deeper topic-specific analysis. Future studies should include a broader range of exam content to improve generalizability across the full prosthodontic curriculum.

Although contextual prompting offers a practical alternative to full model fine-tuning, its implementation is not without barriers. Access to advanced models like GPT-4 may be restricted due to licensing, API usage limits, and cost, which can limit reproducibility and broader adoption in educational institutions. Additionally, constructing high-quality prompts and curating domain-specific content require expert input and technical familiarity. These factors should be considered when evaluating the scalability and real-world feasibility of implementing such AI tools in academic settings.

Another limitation relates to the fine-tuning method employed. While contextual prompting provides a flexible and resource-efficient way to enhance model performance, it also presents challenges. As noted by Büttner et al. [20], transformer-based models such as ChatGPT are inherently data-hungry, and static snippets may not adequately reflect the full spectrum of clinical reasoning required in prosthodontics. This may limit the depth and generalizability of AI responses. Additionally, the effectiveness of contextual prompting depends on how prompts are structured, which introduces variability and challenges in standardization. Future work could explore integration of structured data, federated learning, or more dynamic prompt engineering to mitigate these issues and ensure broader applicability in dental education.

Topic-specific challenges and future directions

Despite improvements from fine-tuning and contextual prompting, certain topics such as Implant Prosthodontics, Occlusion, and Removable Prosthodontics continued to yield lower accuracy across both the base and contextually prompted models. This persistent underperformance suggests that these subcategories may involve more complex reasoning or less standardized knowledge, making them more difficult for large language models to handle reliably. These areas often require integration of nuanced, patient-specific variables, which are not easily conveyed through static reference texts. Additionally, the contextual materials used may have lacked sufficient depth or clinical specificity. Future research should explore the impact of incorporating clinically oriented, case-based resources and multimodal inputs (such as radiographs or diagrams) to improve AI comprehension and accuracy in these challenging domains. Tailoring prompts to better reflect diagnostic reasoning pathways may further enhance performance in specialized areas of prosthodontics.

Conclusion

This study found that contextual prompting significantly improved ChatGPT-4's accuracy on the US Prosthodontic Board Exam mock questions, increasing the correct response rate from 62.7 to 73.6%, a statistically significant improvement ($p < 0.001$). These results highlight the benefit of providing domain-specific reference materials during inference to enhance performance in specialized knowledge assessments.

However, persistent challenges were observed in topics such as Implant Prosthodontics, Removable Prosthodontics, and Occlusion, suggesting that even with contextual support, some subject areas may require more complex or multimodal training approaches.

While these findings suggest potential for AI models to support exam preparation in prosthodontics, it is important to note that this study did not assess student learning, engagement, or long-term retention. Future research should explore the educational implications of using AI tools in real-world learning environments and evaluate their practical utility within dental curricula.

Abbreviations

ACP	American College of Prosthodontists
AI	Artificial Intelligence
LLM	Large Language Model
EBD	Evidence-Based Dentistry
Dx	Diagnosis
Tx	Treatment
Pros	Prosthodontics
Endo	Endodontics
Fixed Pros	Fixed Prosthodontics
Implant Pros	Implant Prosthodontics
Implant Surg	Implant Surgery
MF	Maxillofacial
Perio	Periodontics
Pharm	Pharmacology
Pre Pros Surgery	Pre-Prosthetic Surgery
Removable Pros	Removable Prosthodontics
TMD	Temporomandibular Disorder

Acknowledgements

The authors appreciate the support received from the Researchers Supporting Project (number RSPD2025R950), King Saud University, Riyadh, Saudi Arabia.

Author contributions

M.D: Conception and design of study, Acquisition of data, Drafting of article and/or critical revision, Final approval of manuscript.F.Kh.: Conception and designed of study, Final approval of manuscript, Analysis of data.T.A.: Acquisition of data, Analysis of data, Drafting of article and/or critical revision, Final approval of manuscript.D.H.: Conception and design of study, Analysis of data, Drafting of article and/or critical revision, Final approval of manuscript. Sh.Gh.: Conception and design of study, Acquisition of data, Drafting of article and/or critical revision, Final approval of manuscript.A.F.: Conception and design of study, Drafting of article and/or critical revision, Final approval of manuscript.N.Z.: Final approval of manuscript, Analysis of data, Drafting of article and/or critical revision.Z.Kh.: Conception and design of study, Final approval of manuscript, Drafting of article and/or critical revision.S.R.H: Final approval of manuscript, Drafting of article and/or critical revision.

Funding

Researchers Supporting Project number (RSPD2025R950), King Saud University, Riyadh, Saudi Arabia.

Data availability

The data that support the findings of this study are available from the American board of prosthodontics, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the American board of prosthodontics.

Declarations

Ethics approval and consent to participate

This study did not involve human participants, human data, or human tissue, and consisted solely of secondary analysis of publicly available examination questions and answer keys. According to the guidelines of the Research Ethics Committee of Shahid Beheshti University of Medical Sciences, ethical approval is not required for studies involving publicly available data without any human subject interaction. Therefore, the need for ethical approval and informed consent was waived by the committee. The study was conducted in accordance with the Declaration of Helsinki.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Dentofacial Deformities Research Center, Research Institute of Dental Sciences, Shahid Beheshti University of Medical Sciences, Tajrish, District 1, Daneshjou Blvd, Tehran, Tehran Province, Iran
[2]UCLA School of Dentistry, Los Angeles, CA, USA
[3]Orofacial Pain and Disfunction, UCLA School of Dentistry, Los Angeles, CA, USA
[4]School of Dentistry, Tehran University of Medical Science, Tehran, Iran
[5]Trauma and Craniofacial Reconstruction, Queen Mary College, London, England
[6]Discipline of Oral Surgery, Medicine and Diagnostics, Faculty of Medicine and Health, Westmead Hospital, The University of Sydney, Sydney, NSW 2145, Australia
[7]Department of Operative Dentistry, University of Southern California, Los Angeles, USA

[8]Center of Excellence for Regenerative Dentistry, Department of Anatomy, Faculty of Dentistry, Chulalongkorn University, Bangkok 10330, Thailand
[9]Department of Prosthetic Dental Sciences, College of Dentistry, King Saud University, P. O. Box 60169, King Abdullah Road, 11545 Riyadh, Saudi Arabia

Received: 27 March 2025 / Accepted: 19 May 2025
Published online: 23 May 2025

References

1. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large Language models such as ChatGPT for dental medicine. J Esthet Restor Dent. 2023;35(7):1098–102.
2. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large Language models. PLOS Digit Health. 2023;2(2):e0000198.
3. Chiarello F, Giordano V, Spada I, Barandoni S, Fantoni G. Future applications of generative large Language models: A data-driven case study on ChatGPT. Technovation. Volume 133. Elsevier BV; 2024. p. 103002.
4. Chau RCW, Thu KM, Yu OY, Hsung RT-C, Lo ECM, Lam WYH. Performance of generative artificial intelligence in dental licensing examinations. International dental journal. Volume 74. Elsevier BV; 2024. pp. 616–21. 3.
5. Danesh A, Pazouki H, Danesh K, Danesh F, Danesh A. (2023). The performance of artificial intelligence language models in board-style dental knowledge assessment. In The Journal of the American Dental Association (Vol. 154, Issue 11, pp. 970–974). Elsevier BV.
6. Dashti M, Londono J, Ghasemi S, Moghaddasi N. How much can we rely on artificial intelligence chatbots such as the ChatGPT software program to assist with scientific writing? J Prosthet Dent. 2023 Jul;10(23):00371–2.
7. Yamaguchi S, Morishita M, Fukuda H, Muraoka K, Nakamura T, Yoshioka I, Awano S. Evaluating the efficacy of leading large Language models in the Japanese National dental hygienist examination: A comparative analysis of chatGPT, bard, and Bing chat. J Dent Sci; 2024.
8. Dashti M, Ghasemi S, Ghadimi N, Hefzi D, Karimian A, Zare N, Fahimipour A, Khurshid Z, Chafjiri MM, Ghaedsharaf S. Performance of ChatGPT 3.5 and 4 on U.S. Dental examinations: the INBDE, ADAT, and DAT. Imaging Sci Dent. 2024;54.
9. Brown T, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877–901.
10. Sharry JJ. The future of the American board of prosthodontics. J Prosthet Dent. 1976;35(1):79–81.
11. Woody RD, Grisius RJ. (1976). Evaluation of the examination given by the American Board of Prosthodontics. In The Journal of Prosthetic Dentistry (Vol. 35, Issue 1, pp. 74–78).
12. Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med. 2007;4(10):e296.
13. Nowell J. Guide to ethical approval. BMJ. 2009;338:b450.
14. Freire Y, Santamaría Laorden A, Orejas Pérez J, Gómez Sánchez M, Díaz-Flores García V, Suárez A. ChatGPT performance in prosthodontics: assessment of accuracy and repeatability in answer generation. J Prosthet Dent. 2024;131:e6591–6.
15. Eraslan R, Ayata M, Yagci F, et al. Exploring the potential of artificial intelligence chatbots in prosthodontics education. BMC Med Educ. 2025;25:321.
16. Künzle P, Paris S. Performance of large Language artificial intelligence models on solving restorative dentistry and endodontics student assessments. Clin Oral Investig. 2024;28(11):575.
17. 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.
18. Daungsupawong H, Wiwanitkit V. Generative artificial intelligence in dental licensing examinations: comment. Int Dent J. 2024;74(2):361.
19. Chau RCW, Thu KM, Yu OY, Lo ECM, Hsung RT, Lam WYH. Response to generative AI in dental licensing examinations: comment. Int Dent J. 2024;74(4):897–8.
20. Büttner M, Leser U, Schneider L, Schwendicke F. Natural Language processing: chances and challenges in dentistry. J Dent. 2024;141:104796.

Publisher's note