Journal of Clinical Periodontology

WILEY

Journal of Clinical Periodontology

ORIGINAL ARTICLE OPEN ACCESS

Development and Comparative Evaluation of a Reinstructed GPT-4o Model Specialized in Periodontology

Francesco Fanelli[1] | Muhammad Saleh[2] | Pasquale Santamaria[3] | Khrystyna Zhurakivska[1] | Luigi Nibali[3] | Giuseppe Troiano[1,4]

[1]Department of Clinical and Experimental Medicine, University of Foggia, Foggia, Italy | [2]Department of Periodontics and Oral Medicine, University of Michigan School of Dentistry, Ann Arbor, Michigan, USA | [3]Centre for Host Microbiome Interactions, Faculty of Dentistry, Oral and Craniofacial Sciences, King's College London, London, UK | [4]Department of Medicine and Surgery, LUM University, Casamassima, Italy

Correspondence: Giuseppe Troiano (giuseppe.troiano@unifg.it)

Received: 29 May 2024 | **Revised:** 15 November 2024 | **Accepted:** 6 December 2024

Funding: The authors received no specific funding for this work.

Keywords: artificial intelligence | ChatGPT | clinical decision support | dental AI | periodontology | retrieval-augmented generation

ABSTRACT

Background: Artificial intelligence (AI) has the potential to enhance healthcare practices, including periodontology, by improving diagnostics, treatment planning and patient care. This study introduces 'PerioGPT', a specialized AI model designed to provide up-to-date periodontal knowledge using GPT-4o and a novel retrieval-augmented generation (RAG) system.

Methods: PerioGPT was evaluated in two phases. First, its performance was compared against those of five other chatbots using 50 periodontal questions from specialists, followed by a validation with 71 questions from the 2023–2024 'In-Service Examination' of the American Academy of Periodontology (AAP). The second phase focused on assessing PerioGPT's generative capacity, specifically its ability to create complex and accurate periodontal questions.

Results: PerioGPT outperformed other chatbots, achieving a higher accuracy rate (81.16%) and generating more complex and precise questions with a mean complexity score of 3.81 ± 0.965 and an accuracy score of 4.35 ± 0.898. These results demonstrate PerioGPT's potential as a leading tool for creating reliable clinical queries in periodontology.

Conclusions: This study underscores the transformative potential of AI in periodontology, illustrating that specialized models can offer significant advantages over general language models for both educational and clinical applications. The findings highlight that tailoring AI technologies to specific medical fields may improve performance and relevance.

1 | Introduction

Rapid developments in artificial intelligence (AI) can provide medical personnel with several advantages, including a better detection, prevention and treatment of diseases (Ali et al. 2023; Liu et al. 2021). Machine learning algorithms, for instance, have the potential to revolutionize radiographic caries detection in dentistry (Schwendicke et al. 2022). Large language models (LLMs) are AI models trained on large text datasets to predict and generate language (Yifan Yao et al. 2024; Vaswani et al. 2023). In addition to producing fluid and cohesive documents, LLMs are also capable of providing language translation services and other language-related duties (Shanahan 2024). Chatbots, such as ChatGPT, were initially trained using a predetermined dataset, which limited their ability to incorporate recent data, such as findings from newer studies. However, with the introduction of GPT-4 Turbo and its specialized variants, such as Scholar GPT, Data Analyst GPT, AO Humanizer, Programming GPT and others, these models now offer more advanced and context-specific responses with up-to-date information synthesis. Additionally, the GPT-4o version, designed for optimized use in business and institutional settings, represents a significant advancement in the GPT model family.

The use of retrieval-augmented generation (RAG) techniques has become increasingly important in enhancing these models (Miao et al. 2024). The RAG approach is a natural language processing (NLP) method that enhances the quality of generated text by combining the advantages of generative and retrieval-based models. Without requiring the model to be retrained, RAG expands the already sophisticated capabilities of LLMs to particular domains or the internal knowledge base of an enterprise. It is a practical method for improving LLM output so that it stays accurate, relevant and helpful in a variety of settings (Lewis et al. 2020).

The RAG system functions in two main phases: information retrieval, where it searches relevant external databases, and answer generation, where it combines retrieved data with existing knowledge to produce coherent, context-specific responses (Miao et al. 2024). The RAG approach is valuable for providing accurate and up-to-date information not included in the model's training, thus reducing inaccuracies and hallucinations by grounding responses in factual data (Kirchenbauer and Barns 2024; Shuster et al. 2021).

ChatGPT has been tested in the dental field for answering questions related to oral medicine, endodontics, implantology and dental prosthetics. Although it shows some potential, its reliability is still insufficient for clinical use (Bagde et al. 2023; Tastan Eroglu et al. 2024; Alhaidry et al. 2023).

Within the evolving landscape of periodontology, the integration of AI-driven technologies presents unparalleled opportunities not only for clinical practice but also for the enhancement of education and training (Rahad et al. 2024). NLP models have been developed to predict periodontal disease progression from electronic health records, improving diagnostic accuracy and risk assessment, but require ongoing refinement for reliable outcomes (Patel et al. 2023, 2022). Similarly, other AI subfields, such as deep learning and machine learning, are advancing diagnostic procedures in periodontology, offering new tools to manage the complexity of periodontal disease, which is characterized by varied causes and treatments (Revilla-León et al. 2023).

The complexity of periodontal disease presents significant challenges in fully understanding its aetiology and treatment options (Sanz et al. 2020). AI has the potential to completely transform dental education by making learning more interactive, personalized and effective. Integrating AI tools into curricula allows students to engage with simulated clinical scenarios, explore diverse cases and receive instant feedback on their decisions (Thurzo et al. 2023). This interactive approach encourages active learning and critical thinking, key components in developing clinical reasoning skills. Healthcare professionals can benefit from chatbots and LLMs, but current versions lack in-depth dental knowledge (Sabri et al. 2024; Giannakopoulos, Kavadella, and Salim 2023). Integrating current clinical data into chatbot models can enhance ChatGPT's accuracy, offering personalized guidance, faster diagnoses and better decision making, which can streamline dental workflows. The aim of this study was to program a RAG system that leverages the GTP-4o language model by OpenAI to create 'PerioGPT', an LLM algorithm specialized in periodontology.

2 | Methods

This study did not involve human or animal subjects and therefore was exempted from ethical committee approval. Application programming interface (API) of Openai was used to access the GPT-4o model through generation of an 'API Key' (Paredes, Machuca, and Claudio 2023). All programming phases were carried out on Visual Studio Code using Python (Python 2021). The LlamaIndex package was used to develop the RAG system. Specifically, the workflow consisted of three fundamental phases: the 'Indexing process', the 'Querying process' and the 'Prompting strategy'.

2.1 | Indexing Phase

In the first phase, LlamaIndex prepared a knowledge base from the data. The OpenAI API key was configured as an environment variable, enabling the notebook to use OpenAI's API for operations such as accessing the advanced linguistic model GPT-4o and embedding models. Initially, the code was provided with hundreds of documents in PDF format related to periodontology, totaling 85,818 pages. This phase was implemented using the Python package 'Simple Directory Reader', which allowed for the uploading of files into the code (data ingestion). These documents consisted of open access articles from some of the leading scientific journals in periodontology, ensuring a solid and authoritative foundation for the chatbot. Specifically, the following journals were included: *Journal of Periodontal Research*, *Periodontology 2000*, *Journal of Periodontology*, *Journal of Clinical Periodontology*, *Clinical Oral Implants Research* and *Clinical Implant Dentistry and Related Research*. The most cited publications between 2000 and the present were also downloaded by two of the authors (F.F. and K.Z.). The selection was guided not only by the number of citations but also by scientific quality criteria, including clinical guidelines and internationally recognized consensus documents. The selected articles were critically reviewed by two authors with expertise in periodontology and research methodology (G.T. and M.S.) to ensure the inclusion of the most robust and relevant evidence for the field of periodontology. Given that 'PerioGPT' was trained exclusively with periodontology documents in PDF format, the PyPDF2 Python library was used to extract the raw text. The extracted text was then segmented into chunks of up to 1024 tokens, with a default overlap of 20 tokens. Each chunk was encoded using OpenAI's text-embedding-ada-002-v2 model, converting it into numerical vectors (embeddings) for semantic querying. These vectors were saved in a 'Vector Store Index' to facilitate semantic search by the LLM model (Figure 1).

2.2 | Querying Phase

In the querying phase, a process unfolds that transforms a simple user question into an informative interaction with the AI system. This phase begins with the user's entry of a query, an act that triggers the conversion of words into embeddings. Specifically, the 'text-embedding-ada-002-v2' model processes the textual query into a vector of numbers that represents the semantic meaning of the user's question. At this point, with the query in a form that can be processed, the Chat Engine comes into play. The Chat Engine is a search engine that has been

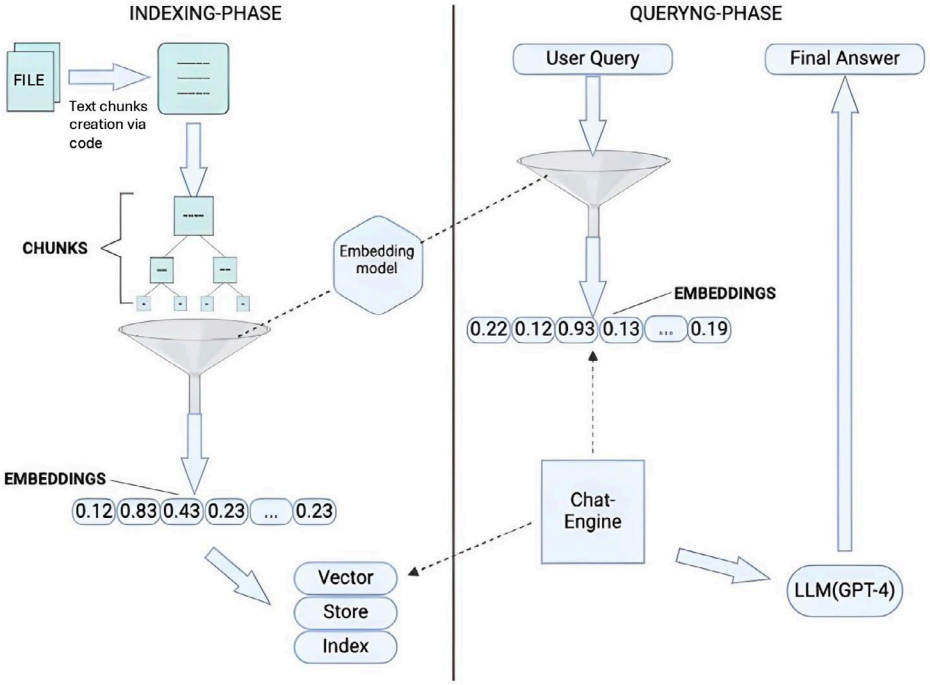

FIGURE 1 | Programming process of RAG using the Llama index. Initially, there is the indexing phase: In this part of the process, textual files are provided to the code and broken down into data fragments, or chunks. These nodes are then transformed into a series of numerical vectors through an embedding model. All generated vectors are stored in a vector store index, ready to be queried. Subsequently, during the querying phase, a query provided by a user undergoes a similar process: It is transformed into a vector by the same embedding model. This vector is then used to search the index for the most similar vector, via the chat engine. The search engine then sends the selected information to GPT-4, which produces a textual response for the user. Thus, the depicted system allows for efficient and intelligent management and retrieval of information, providing relevant and contextualized answers to user queries.

programmed through code to compare the query's embedding with the embeddings in the vector store index by calculating the cosine distance. This process identifies the most relevant matches that correspond to the context of the question. The data selected by the Chat Engine are not yet ready to be presented to the user; this is where the capabilities of the GPT-4o language model come into play (Figure 1). This advanced AI receives the corresponding embeddings and uses them to deeply understand the context of the request. With a 'temperature' set to 0.5, the model generates answers that balance creativity and precision, leading to text that not only responds to the user's query but does so in a coherent and informative way.

2.3 | Prompting Strategy

Through prompting, instructions were provided to structure responses to user queries. Two main prompt templates were implemented: the 'Text_qa_template' for generating an initial response based on the specified context and GPT-4o's knowledge, and the 'Refine_template' for revising responses by incorporating new information or refining the query's interpretation. This enhances the accuracy and relevance of the final answer.

2.4 | Response to Questions

To assess the performance of the PerioGPT mode, a test was conducted using a set of 50 real questions formulated by two of the authors (P.S. and L.N.) at King's College London specialized in periodontology (Data S1). These questions were selected to cover a wide range of topics related to periodontology, thus ensuring a comprehensive evaluation of the chatbot's ability to provide relevant and accurate answers. The author responsible for programming PerioGPT was blind to which were the 50 questions composing the evaluation test. This was to ensure that the author had no prior knowledge of the correct answers, to avoid any potential bias in evaluating the system's responses. To assess the internal reliability and reproducibility of the chatbot, each question was asked five times, resulting in a total of 250 responses for the 50 questions. The generated answers were then collected on an Excel datasheet and re-sent to one of the two authors who assessed the correctness of the answers for each individual question, and a third author who assessed the aggregate performance of each model (G.T.). Concurrently, the same questions were submitted to five of the most widely used AI-based chatbots: ChatGPT-3.5, ChatGPT-4, ChatGPT-4o, ScholarGPT and Gemini. This comparison aimed to benchmark the performance of the PerioGPT against existing solutions widely recognized for their efficiency and reliability in handling text-based queries.

To conduct a further validation of PerioGPT's performance, a multiple-choice test was replicated using 71 quizzes about 'Periodontal Etiology and Pathogenesis' drawn from the 2023 and 2024 "In-Service Examination" of the American Academy of Periodontology (AAP). The same test has been recently used by other studies in the literature to validate chatbot performance

(Sabri et al. 2024). Each question was tested five times, resulting in a total of 355 responses for each chatbot evaluated.

2.5 | Generation of Questions

All the models (Perio-GPT, ChatGPT-3.5, ChatGPT-4, ChatGPT-4o, ScholarGPT and Gemini) were asked to generate a set of 30 multiple-choice questions related to periodontology. The questions were required to cover three main topics: periodontal anatomy, the new classification of periodontitis and the microbiological pathogenesis of periodontitis. All questions generated by the chatbots, totalling 180 questions, were collected by one of the authors responsible for this phase of the study (F.F.). Subsequently, two Google Forms were created containing the 180 generated questions to assess complexity and accuracy, according to a Likert scale (Ankur Joshi, Chandel, and Pal 2015) ranging from 0 to 5 (very low, low, moderate, high, very high). For each question, complexity and accuracy evaluations were then provided by three blinded authors (P.S., K.Z. and M.S.), assigning scores from 0 to 5. The aggregate performance of each model was then assessed by a third author (G.T.) in an unblinded manner.

2.6 | Statistical Analysis

For both the London and the AAP tests, the response performance of the models in answering questions was evaluated using the chi-square test to compare the frequencies of correct/incorrect answers of the chatbots, establishing a statistical significance threshold with a corrected p-value of <0.05. Subsequently, the Benjamini–Hochberg procedure was performed to adjust for multiple comparisons. The reproducibility of each chatbot's responses was assessed by calculating the average intra-class correlation coefficient (ICC) for each chatbot, taking into account the attempts at answering each question. This statistical measure was used to evaluate the consistency and agreement of the chatbots' responses to the same question across five attempts. Additionally, a one-way analysis of variance (ANOVA) was conducted to compare the average ICC values among the different chatbots. This was followed by a Tukey HSD (honestly significant difference) test to compare pairs of chatbots and precisely identify which differences in ICC means were significant. The same statistical tests were also used for assessing differences in complexity and accuracy among the chatbots about the generation of multiple-choice questions related to periodontology. To assess the inter-rater agreement, Fleiss' kappa (κ) was calculated for both complexity and accuracy scores, aiming to quantify the level of agreement among evaluators.

3 | Results

3.1 | Performance Evaluation in Answering Questions

Performance on the quiz was evaluated using both the London and AAP tests. Specifically, on the London Test, PerioGPT emerged as the most accurate chatbot, correctly answering 72.8% of the questions (182/250 responses). This performance was significantly higher than that of the other chatbots tested (Table S1A, Figure S1A), with ScholarGPT ranking second by correctly answering 65.6% (164/250) of questions, while ChatGPT-4o achieved a 60% (150/250) accuracy rate as depicted in Figure S2A.

Subsequently, the chatbots were tested on the AAP In-Service Examination, which served as an external validation of the models' performance. In this test, PerioGPT once again demonstrated its superiority by correctly answering 87% of the questions (309/355) (Table S1B, Figure S1B). ChatGPT-4o ranked second with a 79.2% accuracy rate (281/355), followed by ScholarGPT and ChatGPT-4, with 71.5% (254/355) and 70.1% (249/355) accuracy rates, respectively (Figure 2B).

When considering the cumulative results from both the London Test and the AAP In-Service Examination, PerioGPT maintained its leading position with a total accuracy of 81.2% (491/605), followed by ChatGPT-4o with an accuracy of 71.2% (431/605) and ScholarGPT with 69.1% of correct answers (418/605). The other models showed weaker performance (Figure 2C, Table 1, Table S1C, Figure S1C, Data S2).

3.1.1 | Evaluation of the Responses' Reproducibility

In both tests, the chatbots exhibited very similar mean ICC values, with a notable decline observed only in the performance of ChatGPT-3.5, as shown in Table 2. When aggregating the results from all questions and testing multiple times, the mean ICC values remained consistent. Specifically, PerioGPT achieved an ICC of 0.977 (95% CI: 0.970–0.980) and ChatGPT-4O obtained an ICC of 0.980 (95% CI: 0.975–0.980). ChatGPT-4 maintained a consistent ICC of 0.892 (95% CI: 0.845–0.930). Additionally, ScholarGPT and Gemini reported ICC values of 0.980 (95% CI: 0.975–0.980) and 0.986 (95% CI: 0.980–0.990), respectively. In contrast, ChatGPT-3.5 continued to exhibit lower ICC values, recording an ICC of 0.831 (95% CI: 0.790–0.870). Detailed ICC values and pairwise comparisons for the individual tests (London and AAP) are reported on Table 2 and Table S2A–C.

3.1.2 | Complexity and Accuracy of Question Generation

Focusing on the generation of periodontal-related questions (Table 3), PerioGPT ranked first (Figure 3, Table S3A, Figure S2A) in producing more complex questions with a mean complexity score of 3.81 ± 0.965, which is statistically significantly higher than those of ChatGPT-4o (3.20 ± 0.839), ChatGPT-4 (3.28 ± 0.993), ScholarGPT (2.95 ± 0.964), Gemini (2.75 ± 0.782) and ChatGPT-3.5 (2.86 ± 0.904).

In terms of accuracy, PerioGPT also showed the best score of 4.35 ± 0.898, followed by ChatGPT-4o (4.12 ± 0.715), ChatGPT-4 (3.96 ± 0.822), ScholarGPT (3.78 ± 0.738), Gemini (3.15 ± 0.671) and ChatGPT-3.5 (3.22 ± 0.745) (Figure 4). The accuracy levels between PerioGPT and ChatGPT-4o were comparable, suggesting similar performance, whereas differences with other models were statistically significant (Table S3B, Figure S2B). Additionally, the inter-rater agreement analysis showed a moderate level of agreement, with Fleiss' κ values of 0.394 for complexity and 0.378 for accuracy.

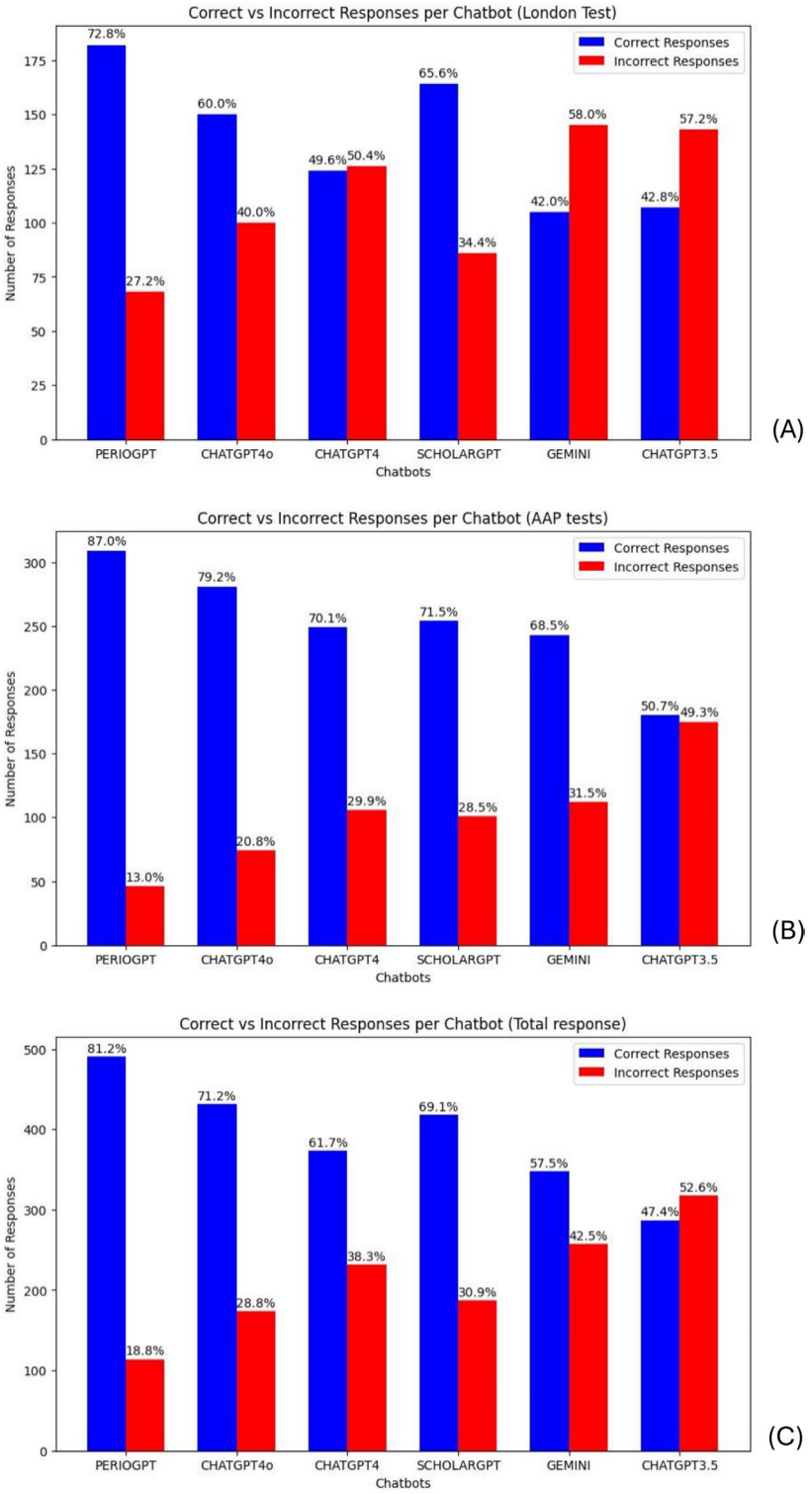

FIGURE 2 | (A–C) Distribution of correct (blue bars) versus incorrect (red bars) responses provided by various chatbots across different tests. (A) The bar chart shows the performance of each chatbot in the London Test, highlighting the differences in accuracy between the models. (B) Results for the AAP In-Service Examination test. (C) Cumulative correct and incorrect responses across both the London Test and the AAP In-Service Examination.

4 | Discussion

LLMs are AI models trained on vast text data to process and generate natural language (Thirunavukarasu, Ting, and Elangovan 2023; Kim et al. 2023). LLMs could enhance dental research, diagnostics and treatment planning, supporting personalized medicine and reducing costs while improving effectiveness in dentistry (Huang et al. 2023). According to Eggmann et al., there are several potential advantages of using LLMs as supplemental instruments in the field of dental medicine, such as clinical decision support, dental

TABLE 1 | Comparative summary of the performance of six different chatbots (PerioGPT, ChatGPT-4o, ChatGPT-4, ScholarGPT, Gemini and ChatGPT-3.5) in terms of the number and percentage of correct and incorrect responses across two assessments: The London Test and the AAP In-Service Examination. The table also includes the cumulative results for both tests, providing a comprehensive overview of each chatbot's accuracy and error rates.

	PerioGPT	ChatGPT-4o	ChatGPT-4	ScholarGPT	Gemini	ChatGPT-3.5
London Test						
Total correct answers	182	150	124	164	105	107
Total wrong answers	68	100	126	86	145	143
Percentage of correct answers	72.80	60	49.60	65.60	42	42.80
Percentage of wrong answers	27.20	40	50.40	34.40	58	57.20
AAP In-Service Examination						
Total correct answers	309	281	249	254	243	180
Total wrong answers	46	74	106	101	112	175
Percentage of correct answers	87.04	79.15	70.14	71.55	68.45	50.70
Percentage of wrong answers	12.96	20.85	29.86	28.45	31.55	49.30
Total questions						
Total correct answers	491	431	373	418	348	287
Total wrong answers	114	174	232	187	257	318
Percentage of correct answers	81.16	71.24	61.65	69.09	57.52	47.44
Percentage of wrong answers	18.84	28.76	38.35	30.91	42.48	52.56

TABLE 2 | Comparison of the intra-class correlation coefficient (ICC) values for six different chatbots across two different tests: the London Test and the AAP In-Service Examination, as well as the cumulative results from all questions. The table shows the consistency of each chatbot's responses, with ICC values and their confidence intervals. It also indicates where statistically significant differences in performance were observed between the chatbots, particularly noting that one chatbot consistently differed from the others across all tests.

Chatbot	London Test ICC (95% CI)	AAP Test ICC (95% CI)	Total questions ICC (95% CI)	Tukey's significant differences
PerioGPT	0.974 (0.963–0.980)	0.980 (0.970–0.990)	0.977 (0.970–0.980)	Significant difference with GPT-3.5 (in all tests)
ChatGPT-4o	0.991 (0.990–0.995)	0.972 (0.965–0.980)	0.980 (0.975–0.980)	Significant difference with ChatGPT-4 (London) and GPT-3.5 (all tests).
ChatGPT-4	0.892 (0.845–0.930)	0.982 (0.975–0.990)	0.892 (0.845–0.930)	Significant difference with ChatGPT-4o (London) and GPT-3.5 (AAP and total questions)
ScholarGPT	0.985 (0.975–0.990)	0.964 (0.950–0.975)	0.980 (0.975–0.980)	Significant difference with GPT-3.5 (in all tests)
Gemini	0.982 (0.970–0.990)	0.987 (0.980–0.995)	0.986 (0.980–0.990)	Significant difference with GPT-3.5 (in all tests)
ChatGPT-3.5	0.829 (0.763–0.888)	0.827 (0.770–0.875)	0.831 (0.790–0.870)	—

education and administrative work. LLMs can assist professionals in clinical decision making by synthesizing complex information and speeding up the review of medical records, thus improving efficiency and accuracy. They also enhance telemedicine in dentistry by aiding in patient information collection, symptom analysis and real-time translation, making dental care more accessible, especially in underserved areas (Cascella et al. 2023).

NLP, in particular LLMs, can enhance automated data extraction from dental clinical notes, thus improving access to electronic dental records (EDRs) for research purposes. However, its use is limited by issues with data standardization and privacy concerns due to differing vocabularies in EDRs. With further development, these tools can improve clinical decision making and documentation but still face challenges regarding generalizability and real-world reliability (Pethani and Dunn 2023; Hossain et al. 2023; Büttner et al. 2024).

From an administrative perspective, LLMs can streamline tasks such as drafting pre-authorization requests and managing communications, thereby increasing efficiency and reducing costs. This enables professionals to focus more on clinical activities and patient care. In education, LLMs aid by generating complex exam questions and relevant teaching materials, supporting educators and enhancing student preparation for clinical practice (Puladi et al. 2024; Eggmann et al. 2023).

ChatGPT is actually the best known LLM (its acronym GPT means 'generative pre-trained transformer'). The term 'transformer' describes the kind of neural network architecture employed, which is well known for its efficiency in processing data sequences, including text, because it can handle long-range dependencies seen in the text (Vaswani et al. 2023). The performance of various LLMs in the dental field has been recently evaluated, with ChatGPT achieving the best performance across different specialties; however, results varied depending on the area of application (Garg et al. 2023; Sabri et al. 2024). The findings of such studies highlight that ChatGPT is not yet capable of replacing the professional judgement of dentists, and caution is advised regarding its limitations. In endodontics, ChatGPT achieved 57.3% accuracy and 85.4% consistency when answering dichotomous questions, suggesting that while it can support decision making, it cannot replace human clinical expertise (Suárez, García, and Algar 2024). ChatGPT demonstrated a low level of accuracy in prosthodontics, achieving only 25.6%. Its reliability in answering questions related to dental prostheses was deemed insufficient for clinical decision making, highlighting the importance of its cautious use (Freire et al. 2024). In oral medicine, ChatGPT shows a moderate level of understanding, especially regarding potentially malignant oral disorders (OPMDs). However, it often presents outdated or inaccurate information, with 58%

TABLE 3 | Average accuracy and complexity values from the evaluation of question generation by various chatbots, namely PerioGPT, ChatGPT-4, ChatGPT-4o, ChatGPT-3.5, ScholarGPT and Gemini. The table presents the mean accuracy and complexity scores for each chatbot, along with their respective standard deviations, providing an overview of the performance differences in generating accurate and complex questions.

	Average accuracy value	Average complexity value
PerioGPT	4.35 (±0.898)	3.81 (±0.965)
ChatGPT-4	3.96 (±0.822)	3.28 (±0.993)
ChatGPT-4o	4.12 (±0.715)	3.2 (±0.839)
ChatGPT-3.5	3.6 (±0.905)	3.08 (±1.154)
Scholar	3.78 (±0.738)	2.95 (±0.964)
Gemini	3.05 (±0.852)	2.36 (±0.662)

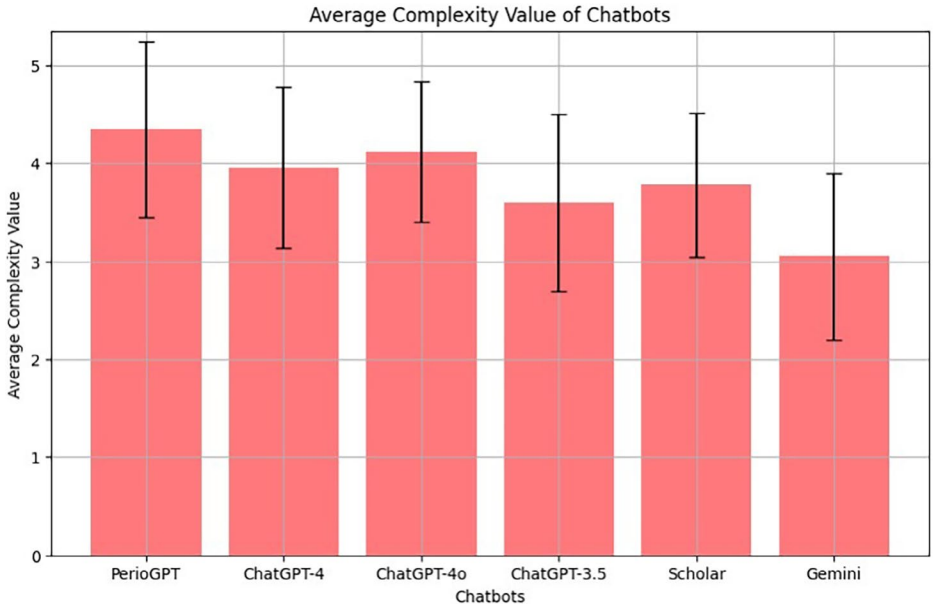

FIGURE 3 | Average complexity value of questions generated by various chatbots, namely PerioGPT, ChatGPT-4, ChatGPT-4o, ChatGPT-3.5, ScholarGPT and Gemini. Each bar represents the average complexity value for the respective chatbot, with error bars indicating the standard deviation from the mean. This chart suggests that PerioGPT tends to generate more complex questions compared to the other chatbots, with Gemini generating the least complex ones.

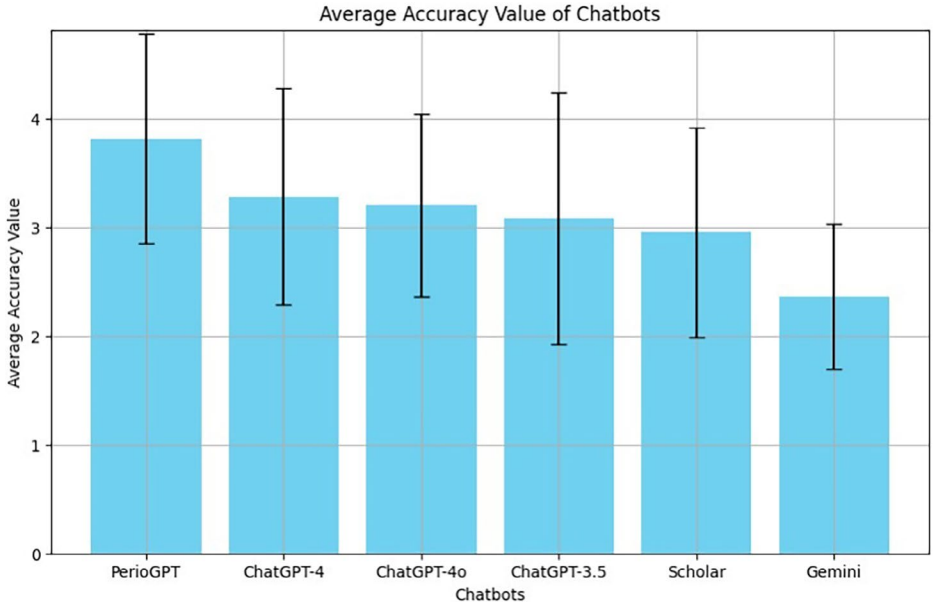

FIGURE 4 | Average accuracy value of questions generated by various chatbots, namely PerioGPT, ChatGPT-4, ChatGPT-4o, ChatGPT-3.5, ScholarGPT and Gemini. Each bar represents the average accuracy value for the respective chatbot, with error bars indicating the standard deviation from the mean. This chart suggests that PerioGPT tends to generate more accurate questions compared to the other chatbots, with Gemini generating the least accurate ones.

of its responses being partially or entirely incorrect. Its frequent omission of key conditions further highlights its limitations in providing comprehensive clinical guidance (Freitas, Mundiña, and Iglesias 2023). In periodontology, ChatGPT was assessed for its ability to accurately diagnose periodontitis based on the 2018 classification. It performed moderately well, correctly identifying the stage in 59.5% of cases, the grade in 50.5% and the extent in 84.0% of cases (Tastan Eroglu et al. 2024). However, LLMs such as ChatGPT-3.5 and ChatGPT-4 risk spreading misinformation, with ChatGPT-4 still requiring supervision because of occasional inaccuracies. In addition, the inability of the model to analyse clinical images further limits its utility in the field (Danesh et al. 2024).

While LLMs are advanced tools for supporting clinical decision making, they cannot replace professional judgement in dentistry. Their limitations are significant in a field where accurate and complete information is crucial for correct diagnoses and treatments. The risk of generating inaccurate or incomplete responses, commonly referred to as 'hallucination', is a critical concern (Sallam 2023).

Hallucinations in LLMs refer to the generation of responses that, while appearing plausible, are actually inaccurate or completely fabricated. This phenomenon poses a significant challenge in using these models in critical fields such as medicine and dentistry, where the accuracy of information is essential (Weston, Dinan, and Miller 2018).

Hallucinations in LLMs can result in inaccurate diagnoses, undermine trust in these tools among healthcare professionals and contribute to the spread of misinformation, creating significant risks in clinical settings (Prathiksha Rumale et al. 2024). To mitigate the problem of hallucinations in LLMs, several strategies can be employed. An effective method is RAG, where the LLM is paired with a retrieval system that searches a database of verified information before generating a response (Kirchenbauer and Barns 2024). This approach ensures that the model's output is based on accurate and up-to-date references, thus reducing hallucinations. Human feedback and rigorous validation further enhance reliability, especially in specialized fields such as dentistry (Prathiksha Rumale et al. 2024).

Several articles report that chatbots developed using RAG systems demonstrate excellent performance, with a significant reduction in hallucinations, making them promising tools in both clinical and educational settings (Rau et al. 2023, 2024; Zhou et al. 2024; Russe, Rau, and Ermer 2024).

This study aimed to develop PerioGPT, a specialized GPT model for periodontology, using GPT-4o and a RAG system to ensure precise and updated knowledge. PerioGPT's performance was compared with five other chatbots on periodontal questions, with statistical analysis to evaluate its effectiveness and accuracy in addressing complex periodontology topics. Specifically, PerioGPT outperformed all other compared chatbots in both the London Test and the AAP external validation, achieving a total accuracy rate of 81.16% for correct answers. This was followed closely by ChatGPT-4o, which answered 71.24% of the questions correctly. These positive results highlight how PerioGPT outperformed the other tested chatbots in terms of response accuracy.

The evaluation of question generation by chatbots revealed that PerioGPT excelled in both accuracy and complexity, significantly outperforming other chatbots except for ChatGPT-4o in accuracy. PerioGPT's ability to provide precise and intricate explanations deepens understanding and elevates the learning experience for both students and professionals. The evaluation of PerioGPT's question generation aimed to assess its potential in university teaching. The goal was to see whether AI could create

high-quality educational materials for periodontology. The generated questions could be used in exams, quizzes and simulations, thereby enhancing student preparation and allowing for customized learning. This study highlights the transformative potential of AI in reshaping traditional education, enhancing teaching effectiveness and better preparing students for clinical practice in periodontology. PerioGPT can support clinical decision making by synthesizing complex information and speeding up medical record review, primarily serving as an educational tool for university faculty and dental students rather than replacing clinical expertise. Integrating AI such as PerioGPT into periodontal education can revolutionize learning with interactive platforms, personalized instruction and improved clinical preparation, thereby enhancing educational quality, patient care and practitioners' understanding of periodontal concepts and decision-making skills. While PerioGPT shows promise, its limitations include reliance on training dataset quality, potential for inaccuracies and a limited study scope. Future efforts should focus on improving dataset diversity, reducing inaccuracies, validating the model in clinical settings and involving more dental professionals in evaluations to enhance the model's reliability and real-world applicability.

5 | Conclusions

This study highlights the potential of LLMs to transform periodontology and dental medicine, showcasing PerioGPT's strengths while acknowledging challenges like data quality and AI inaccuracies. Future improvements in LLMs could make AI central to advancing patient care and education in dentistry, setting new standards for accuracy and efficiency.

Author Contributions

F.F. contributed to conception and design of the study, data acquisition, interpretation and formal analysis, and drafted the manuscript. M.S. contributed to the conception of the study, acquisition, analysis, funding, interpretation of data and critical revision of the manuscript. P.S. contributed to data acquisition, interpretation and analysis, and drafted the manuscript. K.Z. contributed to the design of the study, interpretation of data and drafting and critically revising the manuscript. L.N. contributed to the design of the study, acquisition and interpretation of data and drafting and critically revising the manuscript. G.T. contributed to the conception and design of the study, acquisition, analysis and interpretation of data, and drafting and critically revising the manuscript.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Alhaidry, H. M., B. Fatani, J. O. Alrayes, and A. M. Almana. 2023. "ChatGPT in Dentistry: A Comprehensive Review." *Cureus* 15.

Ali, O., W. Abdelbaki, A. Shrestha, E. Elbasi, M. A. A. Alryalat, and Y. K. Dwivedi. 2023. "A Systematic Literature Review of Artificial Intelligence in the Healthcare Sector: Benefits, Challenges, Methodologies, and Functionalities." *Journal of Innovation & Knowledge* 8, no. 1: 100333. https://doi.org/10.1016/j.jik.2023.100333.

Ankur Joshi, S. K., S. Chandel, and D. K. Pal. 2015. "Likert Scale: Explored and Explained." *British Journal of Applied Science & Technology* 7, no. 4: 396–403.

Bagde, H., A. Dhopte, M. K. Alam, and R. Basri. 2023. "A Systematic Review and Meta-Analysis on ChatGPT and Its Utilization in Medical and Dental Research." *Heliyon* 9: e23050.

Büttner, M., U. Leser, L. Schneider, and F. Schwendicke. 2024. "Natural Language Processing: Chances and Challenges in Dentistry." *Journal of Dentistry* 141: 104796. https://doi.org/10.1016/j.jdent.2023.104796.

Cascella, M., J. Montomoli, V. Bellini, and E. Bignami. 2023. "Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios." *Journal of Medical Systems* 47: 33.

Danesh, A., H. Pazouki, F. Danesh, A. Danesh, and S. Vardar-Sengul. 2024. "Artificial Intelligence in Dental Education: ChatGPT's Performance on the Periodontic In-Service Examination." *Journal of Periodontology* 95, no. 7: 682–687. https://doi.org/10.1002/JPER.23-0514.

Eggmann, F., R. Weiger, N. U. Zitzmann, and M. B. Blatz. 2023. "Implications of Large Language Models Such as ChatGPT for Dental Medicine." *Journal of Esthetic and Restorative Dentistry* 35: 1098–1102. https://doi.org/10.1111/jerd.13046.

Freire, Y., A. S. Laorden, J. O. Pérez, and M. G. Sánchez. 2024. "ChatGPT Performance in Prosthodontics: Assessment of Accuracy and Repeatability in Answer Generation." *Journal of Prosthetic* 131: 659.e1–e6.

Freitas, M. D., B. R. Mundiña, and J. R. G. Iglesias. 2023. "How ChatGPT Performs in Oral Medicine: The Case of Oral Potentially Malignant Disorders." minerva.usc.es.

Garg, R. K., V. L. Urs, A. A. Agarwal, S. K. Chaudhary, V. Paliwal, and S. K. Kar. 2023. "Exploring the Role of ChatGPT in Patient Care (Diagnosis and Treatment) and Medical Research: A Systematic Review." *Health Promotion Perspective* 13, no. 3: 183–191. https://doi.org/10.34172/hpp.2023.22.

Giannakopoulos, K., A. Kavadella, and A. A. Salim. 2023. "Evaluation of the Performance of Generative AI Large Language Models ChatGPT, Google Bard, and Microsoft Bing Chat in Supporting Evidence-Based." *Journal of Medical Internet Research* 25: e51580.

Hossain, E., R. Rana, N. Higgins, et al. 2023. "Natural Language Processing in Electronic Health Records in Relation to Healthcare Decision-Making: A Systematic Review." *Computers in Biology and Medicine* 155: 106649. https://doi.org/10.1016/j.compbiomed.2023.106649.

Huang, H., O. Zheng, D. Wang, J. Yin, and Z. Wang. 2023. "ChatGPT for Shaping the Future of Dentistry: The Potential of Multi-Modal Large Language Model." *International Journal of Oral Science* 15, no. 1: 29.

Kim, J. K., M. Chua, M. Rickard, and A. Lorenzo. 2023. "ChatGPT and Large Language Model (LLM) Chatbots: The Current State of Acceptability and a Proposal for Guidelines on Utilization in Academic Medicine." *Journal of Pediatric Urology* 19: 598–604.

Kirchenbauer, J., and C. Barns. 2024. "Hallucination Reduction in Large Language Models With Retrieval-Augmented Generation Using Wikipedia Knowledge." files.osf.io. https://files.osf.io/v1/resources/pv7r5/providers/osfstorage/6657166cd835c421594ce333?format=pdf&action=download&direct&version=1.

Lewis, P., E. Perez, A. Piktus, et al. 2020. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." *Advances in Neural Information Processing Systems* 33: 9459–9474.

Liu, P. R., L. Lu, J. Y. Zhang, T. T. Huo, S. X. Liu, and Z. W. Ye. 2021. "Application of Artificial Intelligence in Medicine: An Overview." *Current Medical Science* 41, no. 6: 1105–1115. https://doi.org/10.1007/s11596-021-2474-3.

Miao, J., C. Thongprayoon, S. Suppadungsuk, O. A. Garcia Valencia, and W. Cheungpasitporn. 2024. "Integrating Retrieval-Augmented Generation With Large Language Models in Nephrology: Advancing Practical Applications." *Medicina (Kaunas, Lithuania)* 60, no. 3: 445. https://doi.org/10.3390/medicina60030445.

Paredes, C. M., C. Machuca, and Y. M. Claudio. 2023. "ChatGPT API: Brief Overview and Integration in Software Development." *International Journal of Engineering Insights* 1: 25–29.

Patel, J. S., R. Brandon, M. Tellez, et al. 2022. "Developing Automated Computer Algorithms to Phenotype Periodontal Disease Diagnoses in Electronic Dental Records." *Methods of Information in Medicine* 61, no. S 02: e125–e133. https://doi.org/10.1055/s-0042-1757880.

Patel, J. S., D. Shin, L. Willis, A. Zai, K. Kumar, and T. P. Thyvalikakath. 2023. "Comparing Gingivitis Diagnoses by Bleeding on Probing (BOP) Exclusively Versus BOP Combined With Visual Signs Using Large Electronic Dental Records." *Scientific Reports* 13, no. 1: 3. https://doi.org/10.1038/s41598-023-44307-z.

Pethani, F., and A. G. Dunn. 2023. "Natural Language Processing for Clinical Notes in Dentistry: A Systematic Review." *Journal of Biomedical Informatics* 138: 104282. https://doi.org/10.1016/j.jbi.2023.104282.

Prathiksha Rumale, S. T., T. G. Naik, S. Gupta, et al. 2024. "Faithfulness Hallucination Detection in Healthcare AI."

Puladi, B., C. Gsaxner, J. Kleesiek, F. Hölzle, R. Röhrig, and J. Egger. 2024. "The Impact and Opportunities of Large Language Models Like ChatGPT in Oral and Maxillofacial Surgery: A Narrative Review." *International Journal of Maxillofacial Surgery* 53, no. 1: 78–88.

Python, W. 2021. "Python." Python releases for windows.

Rahad, K., K. Martin, I. Amugo, and S. Ferguson. 2024. "ChatGPT to Enhance Learning in Dental Education at a Historically Black Medical College." *Dental Research and Oral Health* 7: 8.

Rau, A., S. Rau, D. Zoeller, A. Fink, H. Tran, and C. Wilpert. 2023. "A Context-Based Chatbot Surpasses Radiologists and Generic ChatGPT in Following the ACR Appropriateness Guidelines." *Radiology* 308: e230970. https://doi.org/10.1148/radiol.230970.

Rau, S., A. Rau, J. Nattenmüller, et al. 2024. "A Retrieval-Augmented Chatbot Based on GPT-4 Provides Appropriate Differential Diagnosis in Gastrointestinal Radiology: A Proof of Concept Study." *European Radiology Experimental* 8, no. 1: 60. https://doi.org/10.1186/s41747-024-00457-x.

Revilla-León, M., M. Gómez-Polo, A. B. Barmak, et al. 2023. "Artificial Intelligence Models for Diagnosing Gingivitis and Periodontal Disease: A Systematic Review." *Journal of Prosthetic Dentistry* 130, no. 6: 816–824. https://doi.org/10.1016/j.prosdent.2022.01.026.

Russe, M. F., A. Rau, and M. A. Ermer. 2024. "A Content-Aware Chatbot Based on GPT 4 Provides Trustworthy Recommendations for Cone-Beam CT Guidelines in Dental Imaging." *Dento Maxillo Facial Radiology* 53: 109–114.

Sabri, H., M. H. A. Saleh, P. Hazrati, et al. 2024. "Performance of Three Artificial Intelligence (AI)-based Large Language Models in Standardized Testing; Implications for AI-Assisted Dental Education." *Journal of Periodontal Research*. https://doi.org/10.1111/jre.13323.

Sallam, M. 2023. "ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns." *Healthcare (Basel)* 11, no. 6: 887. https://doi.org/10.3390/healthcare11060887.

Sanz, M., D. Herrera, M. Kebschull, et al. 2020. "Treatment of Stage I-III Periodontitis-The EFP S3 Level Clinical Practice Guideline." *Journal of Clinical Periodontology* 47: 4–60. https://doi.org/10.1111/jcpe.13290.

Schwendicke, F., C. Grano, J. de Oro, et al. 2022. "Artificial Intelligence for Caries Detection: Value of Data and Information." *Journal of Dental Research* 101, no. 11: 1350–1356. https://doi.org/10.1177/00220345221113756.

Shanahan, M. 2024. "Talking About Large Language Models." *Communications of the ACM* 67: 68–79. https://doi.org/10.1145/3624724.

Shuster, K., S. Poff, M. Chen, D. Kiela, and J. Weston. 2021. "Retrieval Augmentation Reduces Hallucination in Conversation." Findings of the Association for Computational Linguistics: EMNLP 2021, pages 3784–3803 November 7–11, 2021. ©2021 Association for Computational Linguistics.

Suárez, A., V. D. García, and J. Algar. 2024. "Unveiling the ChatGPT Phenomenon: Evaluating the Consistency and Accuracy of Endodontic Question Answers." *International Endodontic Journal* 57: 108–113. https://doi.org/10.1111/iej.13985.

Tastan Eroglu, Z., O. Babayigit, D. Ozkan Sen, and F. Ucan Yarkac. 2024. "Performance of ChatGPT in Classifying Periodontitis According to the 2018 Classification of Periodontal Diseases." *Clinical Oral Investigations* 28, no. 7: 407. https://doi.org/10.1007/s00784-024-05799-9.

Thirunavukarasu, A. J., D. S. J. Ting, and K. Elangovan. 2023. "Large Language Models in Medicine." *Nature Medicine* 29: 1930–1940.

Thurzo, A., M. Strunga, R. Urban, and J. Surovková. 2023. "Impact of Artificial Intelligence on Dental Education: A Review and Guide for Curriculum Update." *Education Sciences* 13, no. 2: 150.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, and L. Jones. 2023. "Attention Is all You Need Study Notes." awesome.qubitpi.org.

Weston, J., E. Dinan, and A. Miller. 2018. "Retrieve and Refine: Improved Sequence Generation Models For Dialogue." Proceedings from Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI, Stroudsburg, PA, USA.

Yifan Yao, J. D., X. Kaidi, Y. Cai, Z. Sun, and Y. Zhang. 2024. "A Survey on Large Language Model (LLM) Security and Privacy: The Good, the Bad, and the Ugly. High-Confidence Computing, A Survey on Large Language Model (LLM) Security and Privacy: The Good, the Bad, and the Ugly." *High-Confidence Computing* 4: 100211.

Zhou, Q., C. Liu, Y. Duan, et al. 2024. "GastroBot: A Chinese Gastrointestinal Disease Chatbot Based on the Retrieval-Augmented Generation." *Frontiers in Medicine* 11: 1392555. https://doi.org/10.3389/fmed.2024.1392555.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.