Application of large language models in oral health education: a review of the literature

Nazlee Sharmin*, PhD, MEd; Ava K Chow§, PhD

PRACTICAL IMPLICATIONS OF THIS RESEARCH

- The recent emergence of large language models (LLMs) has necessitated educators to be aware of and cautious about applying these technologies in dental and dental hygiene education.
- Task-specific training and prompt engineering can tailor LLMs to meet the specific needs of dental and dental hygiene education.

ABSTRACT

Background: The recent emergence of artificial intelligence and large language models (LLMs) has caused educators to be cautious about applying these technologies in education. This narrative review explores the current literature to identify the existing applications of LLMs in dental and dental hygiene education. **Methods:** An extensive literature search in PubMed, CINAHL Plus, and Education Research Complete was conducted. The search string was (((Large language model) OR (Chatbot)) OR (ChatGPT)) AND (Dental education)). Primary research articles published in English and relevant to the research questions were included. Articles were screened by title and then by full-text review. Data were extracted from the eligible studies. **Results:** After 2 rounds of screening, 28 articles were selected for review. The LLMs used in the studies were ChatGPT versions 3, 3.5, 4, 4o, 4V; Bing Chat; Bard; Gemini; Copilot; Llama 2; Claude3-Opus and custom chatbots developed by the authors. Data analysis revealed 2 major themes in the research: 1) the performance of LLMs on standardized exams and 2) LLMs as teaching tools. **Discussion:** Many studies reported that LLMs can pass high-stakes dental exams, raising concerns about current assessment methods. However, findings that LLMs perform poorly in critically appraising literature and interpretation-type questions are insightful for educators when designing new assignments and assessment plans for dental and dental hygiene students. **Conclusion:** LLMs are rapidly developing as artificial intelligence advances. Repeated studies are needed to assess the impact of LLMs on teaching, learning, and assessment experiences.

RÉSUMÉ

Contexte : L'émergence récente de l'intelligence artificielle et des grands modèles linguistiques (GML) a poussé les éducateurs à se montrer prudents quant à l'application de ces technologies dans le domaine de l'enseignement. Cette revue narrative explore la documentation actuelle pour déterminer les applications existantes des GML dans l'enseignement dentaire et la formation en hygiène dentaire. **Méthodes :** Une recherche documentaire approfondie a été menée dans PubMed, CINAHL Plus et Education Research Complete. La chaîne de recherche était « (((Large language model) OR (Chatbot)) OR (ChatGPT)) AND (Dental education)) ». Les principaux articles de recherche publiés en anglais et pertinents aux questions de recherche ont été inclus. Les articles ont été sélectionnés par titre, puis par analyse du texte intégral. Les données ont été extraites des études admissibles. **Résultats :** Après 2 rondes de présélection, 28 articles ont été inclus dans l'analyse. Les GML utilisés dans les études étaient ChatGPT versions 3, 3.5, 4, 4 o, 4V; Bing Chat; Bard; Gemini; Copilot; Llama 2; Claude3-Opus et des agents conversationnels personnalisés développés par les auteurs. L'analyse des données a révélé 2 grands thèmes de la recherche : 1) le rendement des GML aux examens normalisés et 2) les GML comme outils d'enseignement. **Discussion :** De nombreuses études ont révélé que les GML peuvent réussir des examens dentaires à enjeux élevés, ce qui soulève des préoccupations au sujet des méthodes d'évaluation actuelles. Toutefois, les constatations selon lesquelles les GML obtiennent de mauvais résultats dans l'évaluation critique de la documentation et des questions exigeant une interprétation sont pertinentes pour les formateurs lorsqu'ils conçoivent de nouveaux travaux et plans d'évaluation pour les étudiants en médecine dentaire et en hygiène dentaire. **Conclusion :** Les GML se développent rapidement à mesure que l'intelligence artificielle progresse. Des études répétées sont nécessaires pour évaluer l'incidence des GML sur l'enseignement, l'apprentissage et les expériences d'évaluation.

Keywords: artificial intelligence; AI; dental education, teaching; dental students; education; educational activities
CDHA Research Agenda category: capacity building of the profession

INTRODUCTION

The recent surge of artificial intelligence (AI) and large language models (LLMs) has perplexed educators across the world. Although the application of AI is becoming a norm in modern life, it is not easy to define what AI is. The scientific discipline of AI has been around since the 1950s, yet the common understanding of AI has evolved since then.[1] In its current form, the term AI refers to the computational capability of interpreting large amounts of information to make decisions.[1] Many define AI as the study of building or programming computers or machines to do what the human mind can.[2,3] Another broader definition of AI includes "the theory and development of computer systems able to perform tasks that normally require human intelligence, such as

*Associate teaching professor, Mike Petryk School of Dentistry, Faculty of Medicine & Dentistry, College of Health Sciences, University of Alberta, Edmonton, AB, Canada
§Associate professor, Mike Petryk School of Dentistry, Faculty of Medicine & Dentistry, College of Health Sciences, University of Alberta, Edmonton, AB, Canada

Correspondence: Nazlee Sharmin, PhD, MEd; nazlee@ualberta.ca
Manuscript submitted 10 July 2024; revised 16 October 2024; accepted 28 November 2024

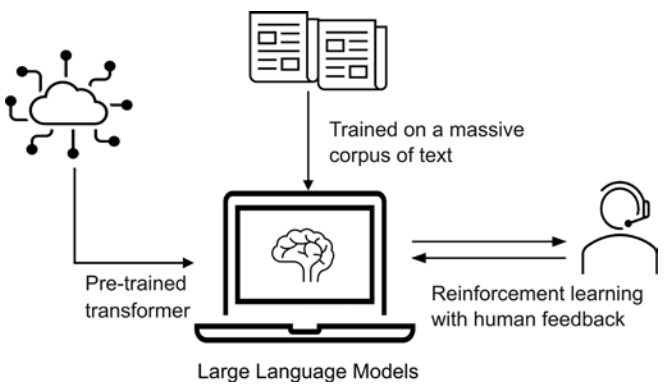
visual perception, speech recognition, decision-making, and translation between languages."[4] The use of AI and LLMs in education is being explored with great interest, although its potential impact on pedagogy is unclear to many educators.[5] LLMs are a recent advancement in AI. Trained with extensive data, LLMs can generate human-like text and conversation, answer questions, translate, and complete language-related tasks with high accuracy.[6] LLMs, such as ChatGPT, have potential as teaching tools that may create opportunities for personalized learning, lesson planning, report writing, language learning, and assessment.[6] However, they can also become a "Pandora's Box" if instructors and students are not cognizant of the limitations of this technology and ethical considerations related to its use are not thoroughly explored.[7]

What are LLMs? How do they work?

LLMs, including the subcategory of generative pre-trained transformers (GPTs), came about from years of study on natural language processing, machine learning, and neural networks.[8,9] A transformer, which is the building block of large-scale natural language processing, is an essential component of LLMs. The transformer, a simple network architecture based solely on attention mechanisms, was first introduced in 2017 by Vaswani et al.[10] GPTs are trained on immense amounts of data, making them capable of understanding and generating natural language to perform various tasks.[8,9] These models are trained to predict the next word in a sentence based on the context of the preceding words by attributing a probability score to the recurrence of words.[11] They are also trained on a massive corpus of text, allowing them to teach topics including grammar, semantics, and conceptual relationships.[7]

The performance of LLMs can be improved through prompt engineering, prompt tuning, and reinforcement learning with human feedback[12] (Figure 1). Prompt engineering and prompt tuning are 2 closely related, but different techniques. Prompt engineering refers to crafting prompts that guide the LLM in understanding the language and the intent of the query. Prompt tuning, in contrast, leverages optimization techniques to find the best prompt for a given task.[13] Currently the most popular LLMs are versions of ChatGPT, Bing Chat, Copilot, Claude, Gemini, and LLaMA.

Figure 1. Development of large language models (LLMs)



Advantages and disadvantages of using LLMs in education

With the ability to rapidly produce human-like text, LLMs can play roles in developing teaching materials, summarizing documents, and identifying gaps to ensure comprehensive coverage of topics in a curriculum.[12] LLMs can also provide real-time explanations of lecture content, create questions, and facilitate small group discussions.[14] They have the potential to help students by providing personalized learning materials and practice questions.

The capabilities of LLMs come with many caveats. The ability to use LLMs to answer questions and write reports can be exploited, resulting in cheating and academic dishonesty[15] and consequently preventing educators from accurately evaluating student learning. LLMs are also not entirely dependable or sound. They can create text with incorrect references and provide citations that do not exist or are irrelevant,[16,17] rendering their output unreliable. Another limitation of current LLMs is the generation of different responses for the same prompt. This feature can make LLMs more human-like, and consequently makes AI-generated text challenging to identify. This feature can also create different or contradictory responses on the same topic, making this technology less trustworthy as a teaching aid.[14]

Objective of the review

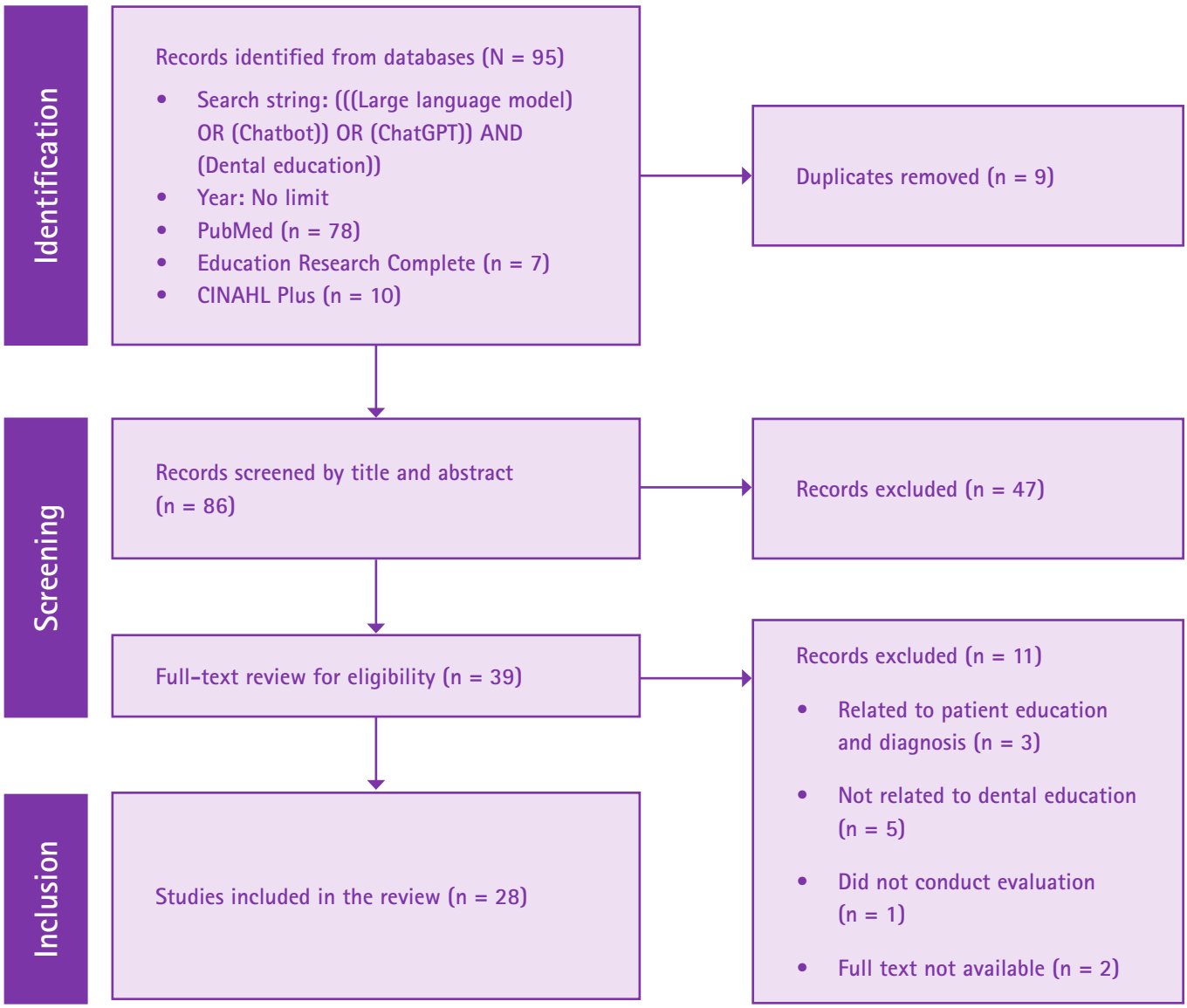
The potential of LLMs in health sciences education, both admirable and nefarious, is vast. It is essential for health professional educators to be well versed in the current trends, evolution, and application of AI technology to manage the rapidly changing educational landscape. This narrative review explores the literature to identify the application of LLMs in dental and dental hygiene education. It aims to answer the following questions:

1. What is the focus of existing research on the application of LLMs in dental education?
2. What are the key findings of the research on the application of LLMs in dental education?

METHODS

An extensive literature search was conducted in PubMed, CINAHL Plus, and Education Research Complete, using the search string (((Large language model) OR (Chatbot)) OR (ChatGPT)) AND (Dental education)) with no limit on year of publication. Primary research articles reporting on experimental and intervention studies, published in English, and relevant to the research questions were included. The initial search from the 3 databases identified 95 records, which were first screened by title and then by full-text review to exclude reports that were not English, unavailable in full-text, did not involve dental faculty or students, were focused on patient education, or did not

Figure 2. Flow diagram explaining the study selection process



apply LLMs or any AI technologies (Figure 2). Review studies, perspectives, and editorials were excluded (Table 1). After the article selection was finalized, data were extracted from the studies, including authors, year and origin of publication, research method, study participants, the LLM used in the study, and key findings related to the research question (Table 2, Table 3).

RESULTS

After 2 rounds of screening, 28 articles were included in the review, all of which were either experimental or intervention studies focused on applying or evaluating LLMs in dental education. All studies applied quantitative research methods and were published between 2023 and 2024. The LLMs used in the studies were ChatGPT versions 3, 3.5, 4.0, 4o, and 4V (n = 24), Bing Chat (n = 3), Bard (n = 1), Gemini (n = 4), Copilot (n = 1), Llama 2 (n = 1), Claude3-Opus (n = 1), and custom Chatbots developed by the authors (n = 2). Twelve studies[18-20,23-25,27,28,30,31,36,44] compared either multiple LLMs or different versions of the same LLM (Table 2, Table 3). Analysis of the data revealed 2 major themes: 1) the performance of LLMs on standardized exams and 2) LLMs as teaching tools.

LLM performance on standardized exams

Fifty-seven percent (57%, n = 16) of the included studies[18-32,44] focused on evaluating the performance of multiple LLMs on dental or dental hygiene student assessments (Table 2). As reported in these studies, ChatGPT 4.0 and 4V achieved passing scores when presented with dental board exam questions from around the world, including a dental licensing examination,[18] periodontic in-service examinations administered by the American Academy of Periodontology (AAP),[19,25] the Swiss Federal

Table 1. Inclusion and exclusion criteria

	Inclusion	Exclusion
Language	English	Non-English
Study focus	Dental education	Non-dental education
Article type	Primary research studies	Conference proceedings, Reviews (including systematic reviews), Opinion, Editorial
	Peer reviewed	Non-peer reviewed
Study design	Any	Nil
Setting	Any	Nil

Licensing Examination in Dental Medicine (SFLEDM),[20] the Integrated National Board Dental Examination (INBDE) of Iran,[28] the Korean Dental Licensing Examination (KDLE),[30] and the Japanese National Dental Examination.[32] When the exam performance was compared between versions of ChatGPT, ChatGPT 4.0, 4o, and 4V outperformed ChatGPT 3 and 3.5. Bing and Bard also achieved acceptable scores on the exams (Table 2).

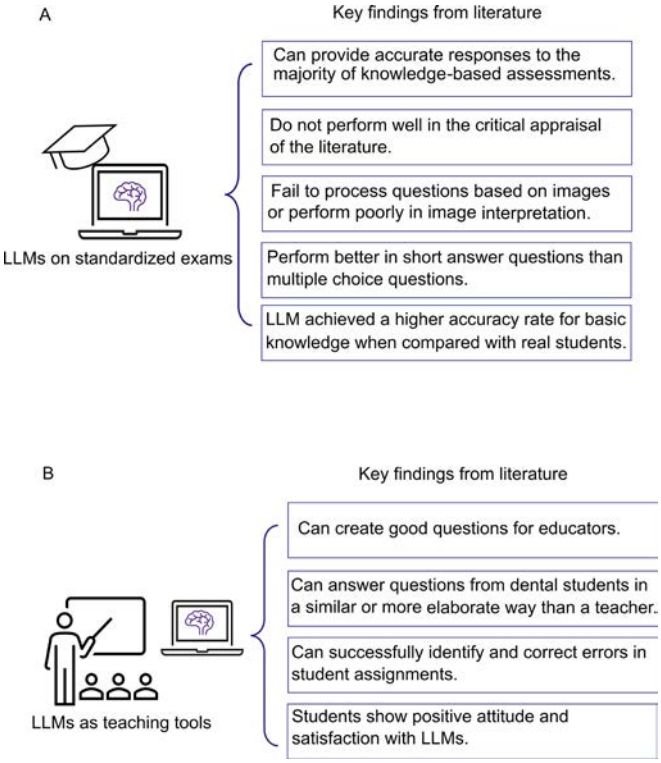
LLMs as teaching tools

Forty-three percent (43%, n = 12) of studies[33-43,45] focused on evaluating the effectiveness of LLMs as a teaching tool to improve students' learning experiences (Table 3). The evaluation of LLMs as self-directed learning tools yielded mixed results. Bhatia et al.[33] reported that the students who used conventional methods for studying performed better on exams compared to the student groups who studied using ChatGPT. However, the opposite findings were reported by Kavadella et al.[34] and Roganović[35].

Several studies explored the ability of LLMs to perform tasks typically done by an instructor in a course.[36-39,45] After receiving textbook or other content as input, LLMs were able to produce questions from the given text input.[36,37] Apart from minimal errors, ChatGPT and Gemini produced good-quality questions from the provided content as evaluated by faculty and content experts.[36,37] Hultgren et al.[38] prompted ChatGPT with questions dental students asked during an online discussion forum. ChatGPT answered those questions equally well or in more depth than the actual instructor. Rahad et al.[39] prompted ChatGPT to check student assignments with intentionally embedded errors. ChatGPT was successful in identifying and correcting all the errors in student assignments.

A number of studies explored dental student and educator interactions with and perceptions of LLMs.[40-43] Students found ChatGPT to be helpful in writing assignments,[42] creating opportunities for practice,[41] improving interactions with patients, enhancing receptiveness, and reducing anxiety.[40] The only study exploring educators' perceptions of LLMs reported that most educators recognize their potential in dental education but are concerned about the reduction of human interaction.[43]

Figure 3. Key findings from the literature. (A) Performance of LLMs on standardized exams; (B) LLMs as teaching tools.



DISCUSSION

This review explored the literature to identify ongoing research trends, applications, and impacts of LLMs in dental and dental hygiene education. The technology of LLMs is still in its infancy, with most studies published within the last year. The studies were undertaken in locations across the world, indicating a global interest in LLMs.

For educators, the most crucial concern with LLMs is their use by students to exploit academic integrity. Many studies thus have explored how well LLMs can answer questions that are usually used in high-stakes exams. The findings that LLMs can pass and, in many cases, perform even better than real students raise concerns about current assessment methods. However, findings also show that LLMs perform poorly in critical appraisal of literature and interpretation-type questions, shedding light on how to make assessments "AI-proof". Figure 3 outlines the key findings from the literature.

One feature of LLMs is that their responses to prompts are inconsistent. Although this feature makes these LLMs more human-like, it raises questions about the application of LLMs in education and health care. LLMs were found to provide detailed and satisfactory answers to most students,[38] yet they failed to consider patient demographics or clinical context when making recommendations to patients.[44] It has also been reported that LLMs fabricate references to support scientific facts in their text-based responses.[16,17] Educators must consider these underlying limitations of LLMs when considering them as teaching assistants (Figure 3B).

Table 2. Performance of large language models (LLMs) on dental examinations and assignments

	Author, year, country	LLM used	Exam/ question type/ assignments	Aim of the study	Research method	Key findings
1	Chau et al. (2024)[18] China	ChatGPT 3.5 ChatGPT 4.0	Dental licensing examination	Assess the performance of ChatGPT on the dental licensing examination.	146 multiple-choice questions (MCQ) from question books of the US and the UK dental licensing examinations were input into ChatGPT 3.5 and 4.0.	The passing grades of the US and UK dental examinations were 75% and 50%, respectively. ChatGPT 3.5 correctly answered 68.3% and 43.3% of questions from the US and UK dental licensing examinations, respectively. The scores for ChatGPT 4.0 were 80.7% and 62.7%, respectively. ChatGPT 4.0 passed both written dental licensing examinations; ChatGPT 3.5 failed.
2	Brozovic et al. (2024)[21] Croatia	Bing Chat artificial intelligence	(a) Exam questions for dental students (b) Guidelines for dental practitioners (c) Frequently asked questions by patients	Assess the performance of Bing Chat in: (a) exam questions for dental students (b) providing guidelines for dental practitioners (c) answering patients' frequently asked questions.	Bing Chat was presented with: (a) 532 MCQs (b) 15 questions, each with 2 follow-up questions on clinical protocols (c) 15 patients' frequently asked questions. The answers were assessed by 4 reviewers	Bing Chat achieved 71.99% on the dental exam. For outlining clinical protocols for practitioners, Bing Chat achieved 81.05%. For patients' frequently asked questions, Bing Chat scored 83.8%.
3	Ali et al. (2024)[22] Qatar	ChatGPT	Multiple recognized assessments in health care education curricula.	Investigate the accuracy of ChatGPT in questions in a variety of formats.	A total of 50 questions with 50 different learning outcomes were developed by the research team. Question formats including multiple-choice; short-answers; short essay; true/false; and fill in the blanks. Questions were presented to ChatGPT.	ChatGPT provided accurate responses to majority of knowledge-based assessments. ChatGPT could not process questions based on images. Responses generated by ChatGPT to written assignments were satisfactory. ChatGPT received borderline scores for critical appraisal of literature.
4	Cung et al. (2024)[44] USA	ChatGPT 4.0 BingAI Bard	(a) Basic and translational skeletal biology (b) Clinical practitioner management of skeletal disorders (c) Patient queries	To assess the performance of ChatGPT 4.0, BingAI, and Bard in addressing questions in 3 categories: basic and translational skeletal biology, clinical practitioner management of skeletal disorders, and patient queries.	Thirty questions from each category were posed to the chatbots, and responses were independently graded for their degree of accuracy by 4 reviewers.	ChatGPT 4.0 had the highest overall median score in each category. Each chatbot displayed distinct limitations that included inconsistent, incomplete or irrelevant responses, inappropriate utilization of lay sources in a professional context, a failure to take patient demographics or clinical context into account when providing recommendations, and an inability to consistently identify areas of uncertainty in the relevant literature.

Continued...

	Author, year, country	LLM used	Exam/ question type/ assignments	Aim of the study	Research method	Key findings
5	Danesh et al. (2024)[19] USA	ChatGPT 3.5 ChatGPT 4.0	Periodontic in-service examination administered by the American Academy of Periodontology (AAP).	To explore ChatGPT's foundation of knowledge in the field of periodontology.	ChatGPT 3.5 and ChatGPT 4.0 were evaluated on 311 multiple-choice questions obtained from the 2023 in-service examination administered by the AAP.	ChatGPT 3.5 and ChatGPT 4.0 answered 57.9% and 73.6% of in-service questions correctly on the 2023 Periodontics In-Service Written Examination, respectively.
6	Danesh et al. (2023)[23] USA	ChatGPT 3.5 ChatGPT 4.0	Board-style dental knowledge assessment	To evaluate the performance of ChatGPT on a board-style multiple-choice dental knowledge assessment	ChatGPT 3.5 and ChatGPT4.0 were asked questions from: INBDE Bootcamp, ITDOnline, and a list of board-style questions. Image-based questions were excluded.	ChatGPT 3.5 and ChatGPT 4.0 answered 61.3% and 76.9% of the questions correctly on average, respectively.
7	Fuchs et al. (2024)[20] Switzerland	ChatGPT 3 ChatGPT 4.0	Swiss Federal Licensing Examination in Dental Medicine (SFLEDM)	To evaluate the performance of ChatGPT 3 and ChatGPT 4.0 on self-assessment questions for dentistry from the SFLEDM. To assess the impact of priming on ChatGPT's performance.	The SFLEDM multiple-choice questions from the University of Bern's Institute for Medical Education platform were administered to both ChatGPT versions, with and without priming.	The average accuracy rate in the SFLEDM was 63.3%, with ChatGPT 4.0 outperforming ChatGPT 3. ChatGPT 3's performance exhibited a significant improvement with priming.
8	Jeong et al. (2024)[24] Korea	ChatGPT, ChatGPT Plus, Bard, Bing Chat	Oral and maxillofacial radiology examination	To evaluate the performance of 4 LLM-based chatbots by comparing their test results with those of dental students.	Chatbots were tested on 52 questions from regular dental college examinations. Questions were categorized into basic knowledge; imaging and equipment; and image interpretation. The accuracy rates of the chatbots were compared with the performance of students.	The students' overall accuracy rate was 81.2%, while that of the chatbots varied: 50.0% for ChatGPT, 65.4% for ChatGPT Plus, 50.0% for Bard, and 63.5% for Bing Chat. ChatGPT Plus achieved a higher accuracy rate for basic knowledge than the students (93.8% vs. 78.7%). All chatbots performed poorly in image interpretation, with accuracy rates below 35.0%. All chatbots scored less than 60.0% on MCQs, but performed better on SAQs.
9	Sabri et al. (2024)[25] USA	ChatGPT 4.0 ChatGPT 3.5 Gemini	Annual in-service examination by the APP.	To evaluate the performance of LLMs in professional exams and compare with the human control group.	1312 questions from the annual AAP examination were presented to the LLMs. Their responses were analyzed using chi-square tests and compared with the scores of periodontal residents as the human control group.	ChatGPT 4.0 outperformed all human control groups, ChatGPT 3.5, and Gemini in all exam years ($p < 0.001$).

Continued...

Table 2. *Continued*

	Author, year, country	LLM used	Exam/ question type/ assignments	Aim of the study	Research method	Key findings
10	Brondani et al. (2024)[26] Canada	ChatGPT	Reflection assignments	To evaluate if university instructors can differentiate reflection assignments created by ChatGPT and students.	Hardcopies of 20 reflections (10 generated by undergraduate dental students and 10 generated by ChatGPT) were distributed to 3 instructors with at least 5 years of teaching experience. Instructors were asked to assign either "ChatGPT" or "student" to each reflection.	ChatGPT can write reflection assignments equally to dental students. However, instructors can differentiate between reflections generated by ChatGPT or by students most of the time.
11	Quah et al. (2024)[27] Singapore	ChatGPT 3.5 ChatGPT 4.0 Llama 2 Gemini Copilot	Oral and maxillofacial surgery examinations	To evaluate the accuracy of LLMs in answering MCQ from the oral and maxillofacial surgery examination.	259 questions from the university's question bank were answered by the 5 LLMs.	ChatGPT 4.0 performed the best (76.8%), followed by Copilot (72.6%), ChatGPT 3.5 (62.2%), Gemini (58.7%, 95%), and Llama 2 (42.5%).
12	Dashti et al. (2024)[28] Iran	ChatGPT 3.5 ChatGPT 4.0	Integrated National Board Dental Examination (INBDE), Dental Admission Test (DAT), Advanced Dental Admission Test (ADAT)	To investigate the effectiveness of ChatGPT in answering dentistry exam questions.	ChatGPT 3.5 and 4.0 were tested with 253 questions from the INBDE, ADAT, and DAT exams.	For the INBDE, both versions achieved 80% accuracy in knowledge-based questions and 66% to 69% in case history questions. ChatGPT 4.0 excelled on the DAT, with 94% accuracy in knowledge-based questions, 57% in mathematical analysis items, and 100% in comprehension questions.
13	Jaworsk et al. (2024)[29] Poland	ChatGPT 4.0	Polish Final Dentistry Examination (LDEK)	To evaluate the performance of ChatGPT 4.0 on the LDEK exam and compare it with human.	200 multiple choice type questions from the spring 2023 LDEK exam were used to test ChatGPT 4.0	ChatGPT 4.0 correctly answered 70.85% questions. The GPT performed better in endodontics (71.74%) and prosthetic dentistry (80%) but showed lower accuracy in pediatric dentistry (62.07%) and orthodontics (52.63%). A statistically significant difference was observed between ChatGPT's performance on clinical case-based questions (36.36% accuracy) and other factual questions (72.87% accuracy), with a p value of 0.025.
14	Kim et al. (2024)[30] USA	ChatGPT 3.5 ChatGPT 4.0 Claude3-Opus	Korean Dental Licensing Examination (KDLE)	To evaluate the performance of LLMs in KDLE	KDLE questionnaires from 2019 to 2023 were used as inputs to the LLMs.	Claude3-Opus performed best among the LLMs used in the study. Claude3-Opus and ChatGPT 4.0 surpassed the cut-off scores in all the years considered; indicating that Claude3-Opus and ChatGPT 4.0 passed the KDLE, whereas ChatGPT 3.5 did not.
15	Künzle et al. (2024)[31] Germany	ChatGPT 3.5 ChatGPT 4.0 ChatGPT 4.o Gemini 1.0	Restorative Dentistry and Endodontics (RDE) student assessment	To evaluate the performance of LLMs on solving RDE student assessment questions.	151 questions from a RDE question pool were prepared for prompting, entered into LLMs, and answers recorded for analysis.	The total answer accuracy of ChatGPT 4.0o was the highest, followed by ChatGPT 4.0, Gemini 1.0 and ChatGPT 3.5 (72%, 62%, 44%, and 25%, respectively) with significant differences between all LLMAs except ChatGPT 4.0 models.

Continued...

	Author, year, country	LLM used	Exam/ question type/ assignments	Aim of the study	Research method	Key findings
16	Morishita et al. (2024)[32] Japan	ChatGPT 4V	Japanese National Dental Examination (JNDE)	To assess the capabilities of ChatGPT 4V with image recognition in answering image-based questions from the JNDE.	The input dataset for the ChatGPT 4V used questions from the JNDE, with a focus on image-related queries.	The overall correct response rate of ChatGPT 4V for image-based JNDE questions was 35.0%. The correct response rates were 57.1% for compulsory questions, 43.6% for general questions, and 28.6% for clinical practical questions.

Table 3. Performance of large language models (LLMs) as teaching tools

	Author, year, country	LLM used	Aim of the study	Research method	Study participants	Key findings
1	Bhatia et al. (2024)[33] USA	ChatGPT	To determine whether ChatGPT is more effective than conventional methods in teaching undergraduate dental students.	Students were randomly divided into 2 groups. Group A was given textbooks to read and Group B used ChatGPT. The pre- and post-test scores were compared.	100 dental students	The mean test scores for students from the conventional method group were higher than the mean scores for the ChatGPT group on the post-test. Traditional teaching methods are more effective for learning and understanding than ChatGPT.
2	Ahmed et al. (2023)[36] Saudi Arabia	ChatGPT Google Bard (Gemini)	To investigate the effectiveness of ChatGPT and Google Bard in generating MCQs for educators of dental caries.	Sixteen paragraphs from a textbook were used as input in ChatGPT and Bard to produce MCQs. Three dental specialists assessed the relevance, accuracy, and complexity of the generated questions.	NA	No significant differences were found between the questions generated by ChatGPT and Bard. Bard-generated questions tended to have higher cognitive levels than those of ChatGPT. Format error was predominant in ChatGPT-generated questions. Bard exhibited more absolute terms than ChatGPT.
3	Fang et al. (2024)[40] USA	Custom-developed chatbot (CB)	To investigate the awareness and perceptions of AI, interaction experiences, and concerns about a CB compared with the traditional Blackboard (BB) online platform.	Students were randomly divided into a custom-developed chatbot (CB) group (n = 43) and a traditional blackboard (BB) group (n = 43). The groups were asked to engage with their designated platforms for 10 to 15 minutes by focusing on clinical inquiries in a predoctoral implant clinic. After the interaction, participants responded on a 5-point Likert scale to a 19-item survey.	86 dental students	The CB group demonstrated improved timeliness ($p < 0.001$), more interaction ($p < 0.001$), enhanced receptiveness ($p = 0.002$), and less anxiety ($p < 0.001$) and was more satisfied ($p < 0.001$) when compared with the BB group.
4	Hultgren et al. (2023)[38] Sweden	ChatGPT 3.5	Compared the ability of ChatGPT 3.5 and teachers to answer questions from dental students.	The questions from the students and replies from the teachers were obtained from an online discussion forum during a course in microbial pathogenesis for dental students. The same questions were administered to ChatGPT 3.5.	22 dental students who took the course on microbial pathogenesis	ChatGPT 3.5 answered the questions from dental students in a similar or even more elaborate way compared to the answers that had previously been provided by a teacher.

Continued...

Table 3. *Continued*

	Author, year, country	LLM used	Aim of the study	Research method	Study participants	Key findings
5	Kavadella et al. (2024)[34] Cyprus	ChatGPT	To evaluate the implementation of ChatGPT in the educational process.	Students were devided into 2 groups and were asked to perform an assignment. One group searched the internet for scientific resources and the other group used ChatGPT for this purpose. Both groups developed a PowerPoint presentation based on their research and presented it in class. Seventy students undertook a knowledge examination	77 dental students	In the knowledge test, students in the ChatGPT group performed significantly better than students in the literature research group.
6	Or et al. (2024)[41] Australia	Custom-developed history-taking chatbot	To assess student perception and acceptance of a history-taking chatbot.	A history-taking chatbot was developed for students to act as "clinician" and the chatbot as "patient." A survey was conducted.	13 Doctor of Dental Medicine students	Most students agreed that they participated more with the chatbot. Most students also agreed that the chatbot would provide more opportunities for them to practise.
7	Ozbay (2024)[37] Turkey	ChatGPT 4.0	To evaluate the ability of ChatGPT 4.0 to generate clinical case-based MCQs.	International Association of Dental Traumatology guidelines for the management of traumatic dental injuries were introduced to ChatGPT 4.0 as an information source and prompted to generate 20 questions on fractures and luxations, avulsion of permanent teeth, injuries in the primary dentition. Questions were evaluated by 2 endodontists.	NA	52% of the questions were usable without modification or with minor changes. 28% questions were incorrect.
8	Rahad et al. (2023)[39] USA	ChatGPT 3.5	To assess ChatGPT's utilities for enhancing pedagogical aspect of dental education.	Student essays were collected and errors were embedded regarding dental terminologies. The essays were presented to ChatGPT to check if it can identify and correct the dental-specific terms.	NA	ChatGPT successfully identified and corrected all the errors in student assignments.
9	Roganović (2024)[35] Serbia	ChatGPT	To investigate how reading ChatGPT features/descriptions influences the willingness and expectations of using this technology.	Students were asked to learn about side effects of drugs used in dental practice via reading recommended literature or ChatGPT. Expectations for ChatGPT were measured by survey, before and after reading of a system features description. Learning outcomes were evaluated via pharmacology quiz.	104 dental students	Students who used ChatGPT (YG group) showed better results on the pharmacology quiz than students who neither read the description nor used ChatGPT for learning (NN condition). Students who read the description of ChatGPT features yet did not use it (NG) showed better results on the pharmacology quiz compared with the NN condition. The NG students compared to the YG students had less trust in AI system assistance in learning, and after the AI system description reading, their expectations changed significantly.

Continued...

	Author, year, country	LLM used	Aim of the study	Research method	Study participants	Key findings
10	Saravia-Rojas et al. (2024)[42] Peru	ChatGPT	To assess the influence of ChatGPT on the academic tasks performed by dental students.	Participants were asked to complete scientific writing assignments using ChatGPT and conventional search methods. The assignments were reviewed by professors. An anonymous questionnaire was administered to the students regarding the usefulness of ChatGPT.	55 dental students	64.29% of the students found ChatGPT useful, 33.33% found it very useful. Regarding its application in further academic activities, 54.76% considered it useful, 40.48% found it very useful. All students provided positive feedback.
11	Uribe et al. (2024)[43] Latvia	Any artificial intelligence (AI) chatbots	To explore dental educators' perceptions of AI chatbots and LLMs	A global cross-sectional survey was conducted to evaluate dental educators' perceptions of AI chatbots and their influence on dental education.	428 dental educators	31% of the participants already use AI tools. 64% recognize their potential in dental education. Educators stated that AI chatbots could enhance knowledge acquisition (74.3%), research (68.5%), and clinical decision-making (63.6%) but expressed concern about the potential reduction of human interaction (53.9%).
12	Quah et al. (2024)[45] Singapore	ChatGPT 4	To explore how reliable ChatGPT is in automated essay scoring (AES) for oral and maxillofacial surgery (OMS) examinations compared to human assessors	Sixty-nine undergraduate dental students participated in a closed-book examination comprising 2 essays. Using pre-created assessment rubrics, 3 assessors independently performed manual essay scoring, while 1 separate assessor performed AES using ChatGPT 4. Intraclass correlation coefficient and Cronbach's α were calculated to evaluate the reliability and inter-rater agreement of the test scores among all assessors.	69 dental students	A strong correlation between all manual scorers was observed for one question ($r = 0.752–0.848$, $p < 0.001$) whereas a moderate correlation was observed for the other question ($r = 0.527–0.571$, $p < 0.001$).The results indicated a potential of ChatGPT for essay marking.

Improving LLMs and fine-tuning their responses by further training is possible and can be very helpful to meet specific needs. Two studies in this review reported developing and evaluating custom-built chatbots.[40,41] LLMs, trained to perform specific tasks, may improve their performance and reduce inconsistency, an area that is still open for further research. Educators must carefully consider LLMs as teaching tools.

Though some studies showed that LLMs respond to student queries with more sophisticated, detailed answers than instructors,[38] such responses do not necessarily consider the learner level and may result in confusion if the answers are more complex or nuanced than the students can appreciate. Student characteristics such as learner level, amount of clinical experience, and pre-requisite knowledge are not incorporated into LLM responses and may result in the generation of an answer that is either overly complex or too simple.

Dental hygiene curricula across the world consists of foundational and clinical learnings. Rahad et al.[39] reported the ability of ChatGPT to aid dental and dental hygiene students with their writing by effectively recognizing and accurately rectifying dental-specific terminologies. ChatGPT-generated practice questions also helped dental and dental hygiene students prepare for exams.[39] However, some studies indicate a possible weakness of ChatGPT 4.0 in its knowledge of dental hygiene practices and theories. In a very recent exploration of how 4 different LLMs performed on the Japanese National Dental Hygienist Examination, Yamaguchi et al.[46] showed that LLMs achieved an average score of 68.15%, with ChatGPT 4.0 performing the best at 75.3%, though this result was not statistically significant. Notably, ChatGPT 4.0 performed the worst of all 4 LLMs on questions that addressed the theory of preventive dental procedures and introduction of dental hygiene, both of which are critical aspects of the dental hygiene profession.

This finding suggests that, although ChatGPT 4.0 may have more robust dental knowledge,[6] the training data lack material related specifically to the dental hygiene body of knowledge and may consequently disadvantage dental hygiene students who elect to use the most current GPT as part of their learning. It is essential to ensure that dental hygiene students and practitioners are equipped with critical analysis skills to carefully evaluate the information they may receive outside the classroom and as part of their continuing education.

The latest version of ChatGPT can now recognize image and audio inputs and can talk back.[47] Given that dental hygiene students extensively study macro- and micro-anatomy of tooth and facial structures, the image recognition ability of LLMs may potentially help them to identify and study anatomical structures. In addition, the ability to receive and return audio may make the newest versions of ChatGPT valuable tools for dental and dental hygiene students when practising patient interactions and preparing for the Observed Structured Clinical Examination (OSCE).

In dental hygiene education, LLMs could also be leveraged for back-end administrative and management tasks. For example, LLMs are particularly adept at big data learning analytics. The information can track performance, improve pedagogical tools, and/or identify at-risk students. LLMs with the latest upgrades can create case-based scenarios for students and reduce standardized patient costs. The key findings from this literature review will be a helpful guide for many dental and dental hygiene educators (Figure 3).

CONCLUSION

Like many new technologies, LLMs are still evolving. The underlying limitations have obscured their potential thus far. Educators must be vigilant when applying this technology to reduce their workload and improve student learning experiences. Further studies are needed in this area to enhance and fine-tune the performance of LLMs and explore their impact on students' learning experiences.

CONFLICTS OF INTEREST

The authors of this study have declared no conflicts of interest.

REFERENCES

1. Dignum V. What is artificial intelligence? In: *Responsible artificial intelligence: How to develop and use AI in a responsible way*. Cham (CH): Springer Nature; 2019.
2. Boden MA, editor. *Artificial intelligence:Handbook of perception and cognition*, 1st edition. Elsevier; 1996.
3. Mondal B. Artificial intelligence: state of the art. In: Balas VE, Kumar R, Srivastava R, editors. *Recent trends and advances in artificial intelligence and internet of things*. Cham, Switzerland: Springer; 2020. pp. 389–425.
4. Oxford English Dictionary. Artificial intelligence [Internet]. ©2024 [cited 2024 June 19]. Available from: oed.com/
5. Zawacki-Richter O, Marín VI, Bond M, Gouverneur F. Systematic review of research on artificial intelligence applications in higher education–where are the educators? *International Journal of Educational Technology in Higher Education*. 2019;16(1):1–27.
6. Kasneci E, Seßler K, Küchemann S, Bannert M, Dementieva D, Fischer F, et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*. 2023;103:102274.
7. Májovský M, Černý M, Kasal M, Komarc M, Netuka D. Artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: Pandora's box has been opened. *J Med Internet Res*. 2023;25:e46924.
8. Floridi L, Chiriatti M. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*. 2020;30:681–94.
9. IBM. What are large language models (LLMs)? [Internet]. ©2024 [cited 2024 June 19]. Available from: ibm.com/think/topics/large-language-models
10. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. *Advances in neural information processing systems 30* (NIPS 2017). Proceedings of the 31st Conference on Neural Information Processing Systems, 2017 Dec 4–9, Long Beach, CA, USA.
11. Christiano PF, Leike J, Brown T, Martic M, Legg S, Amodei D. Deep reinforcement learning from human preferences. In: Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. *Advances in neural information processing systems 30* (NIPS 2017). Proceedings of the 31st Conference on Neural Information Processing Systems, 2017 Dec 4–9, Long Beach, CA, USA.
12. Lee H. The rise of ChatGPT: Exploring its potential in medical education. *Anat Sci Educ*. 2024;17(5):926–31.
13. Abhishek A. Prompt Engineering vs Prompt Tuning: A Detailed Explanation [Internet]. ©2024 [cited 2024 Oct 12]. Available from: medium.com/@aabhi02/prompt-engineering-vs-prompt-tuning-a-detailed-explanation-19ea8ce62ac4
14. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large language models in medical education: opportunities, challenges, and future directions. *JMIR Med Educ*. 2023;9(1):e48291.
15. Choi EP, Lee JJ, Ho MH, Kwok JY, Lok KY. Chatting or cheating? The impacts of ChatGPT and other artificial intelligence language models on nurse education. *Nurse Educ Today*. 2023;125:105796.
16. Manohar N, Prasad SS. Use of ChatGPT in academic publishing: A rare case of seronegative systemic lupus erythematosus in a patient with HIV infection. *Cureus*. 2023;15(2):e34616.
17. Akhter HM, Cooper JS. Acute pulmonary edema after hyperbaric oxygen treatment: a case report written with ChatGPT assistance. *Cureus*. 2023 Feb;15(2):e34752.

18. Chau RC, Thu KM, Yu OY, Hsung RT, Lo EC, Lam WY. Performance of generative artificial intelligence in dental licensing examinations. *Int Dent J.* 2024;74(3):616–21.
19. Danesh A, Pazouki H, Danesh F, Danesh A, Vardar-Sengul S. Artificial intelligence in dental education: ChatGPT's performance on the periodontic in-service examination. *J Periodontol.* 2024;95(7):682–87.
20. Fuchs A, Trachsel T, Weiger R, Eggmann F. ChatGPT's performance in dentistry and allergyimmunology assessments: a comparative study. *Swiss Dent J.* 2024;134(2):1–17.
21. Brozović J, Mikulić B, Tomas M, Juzbašić M, Blašković M. Assessing the performance of Bing Chat artificial intelligence: Dental exams, clinical guidelines, and patients' frequent questions. *J Dent.* 2024;144:104927.
22. Ali K, Barhom N, Tamimi F, Duggal M. ChatGPT—A double-edged sword for healthcare education? Implications for assessments of dental students. *Eur J Dent Educ.* 2024;28(1):206–211.
23. 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):970–74.
24. Jeong H, Han SS, Yu Y, Kim S, Jeon KJ. How well do large language model-based chatbots perform in oral and maxillofacial radiology? *Dentomaxillofac Radiol.* 2024;53(6):390–95.
25. Sabri H, Saleh MH, Hazrati P, Merchant K, Misch J, Kumar PS, et al. Performance of three artificial intelligence (AI)-based large language models in standardized testing: implications for AI-assisted dental education. *J Periodontal Res.* 2024;60(2):121–33.
26. Brondani M, Alves C, Ribeiro C, Braga MM, Garcia RC, Ardenghi T, Pattanaporn K. Artificial intelligence, ChatGPT, and dental education: Implications for reflective assignments and qualitative research. *J Dent Educ.* 2024;88(12):1671–1680.
27. Quah B, Yong CW, Lai CW, Islam I. Performance of large language models in oral and maxillofacial surgery examinations. *Int J Oral Maxillofac Surg.* 2024;53(10):881–86.
28. Dashti M, Ghasemi S, Ghadimi N, Hefzi D, Karimian A, Zare N, et al. Performance of ChatGPT 3.5 and 4 on US dental examinations: the INBDE, ADAT, and DAT. *Imaging Sci Dent.* 2024;54(3):271–75.
29. Jaworski A, Jasiński D, Sławińska B, Błecha Z, Jaworski W, Kruplewicz M, et al. GPT-4o vs. human candidates: Performance analysis in the Polish Final Dentistry Examination. *Cureus.* 2024;16(9):e68813.
30. Kim W, Kim BC, Yeom HG. Performance of large language models on the Korean Dental Licensing Examination: a comparative study. *Int Dent J.* 2025;75(1):176–84.
31. 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.
32. Morishita M, Fukuda H, Muraoka K, Nakamura T, Hayashi M, Yoshioka I, et al. Evaluating GPT-4V's performance in the Japanese national dental examination: a challenge explored. *J Dent Sci.* 2024;19(3):1595–1600.
33. Bhatia AP, Lambat A, Jain T. A comparative analysis of conventional and chat-generative pre-trained transformer-assisted teaching methods in undergraduate dental education. *Cureus.* 2024;16(5):e60006.
34. Kavadella A, Da Silva MA, Kaklamanos EG, Stamatopoulos V, Giannakopoulos K. Evaluation of ChatGPT's real-life implementation in undergraduate dental education: mixed methods study. *JMIR Med Educ.* 2024;10(1):e51344.
35. Roganović J. Familiarity with ChatGPT features modifies expectations and learning outcomes of dental students. *Int Dent J.* 2024;74(6):1456–1462.
36. Ahmed WM, Azhari AA, Alfaraj A, Alhamadani A, Zhang M, Lu CT. The quality of AI-generated dental caries multiple choice questions: A comparative analysis of ChatGPT and Google Bard language models. *Heliyon.* 2024;10(7):e28198.
37. Özbay Y. Evaluation of ChatGPT as a multiple-choice question generator in dental traumatology. *Med Records.* 2024;6(2):235–38.
38. Hultgren C, Lindkvist A, Özenci V, Curbo S. ChatGPT (GPT-3.5) as an assistant tool in microbial pathogenesis studies in Sweden: a cross-sectional comparative study. *J Educ Eval Health Prof.* 2023;20:32.
39. Rahad K, Martin K, Amugo I, Ferguson S, Curtis A, Davis A, et al. ChatGPT to enhance learning in dental education at a historically Black medical college. *Dent Res Oral Health.* 2024;7(1):8–14.
40. Fang Q, Reynaldi R, Araminta AS, Kamal I, Saini P, Afshari FS, et al. Artificial intelligence (AI)-driven dental education: Exploring the role of chatbots in a clinical learning environment. *J Prosthet Dent.* 2024:S0022-3913(24)00231-2.
41. Or AJ, Sukumar S, Ritchie HE, Sarrafpour B. Using artificial intelligence chatbots to improve patient history taking in dental education (pilot study). *J Dent Educ.* 2024;88(Suppl 3):1988–1990.
42. Saravia-Rojas MÁ, Camarena-Fonseca AR, León-Manco R, Geng-Vivanco R. Artificial intelligence: ChatGPT as a disruptive didactic strategy in dental education. *J Dent Educ.* 2024;88(6):872–76.
43. Uribe SE, Maldupa I, Kavadella A, El Tantawi M, Chaurasia A, Fontana M, et al. Artificial intelligence chatbots and large language models in dental education: worldwide survey of educators. *Eur J Dent Educ.* 2024;28(4):865–76.
44. Cung M, Sosa B, Yang HS, McDonald MM, Matthews BG, Vlug AG, et al. The performance of artificial intelligence chatbot large language models to address skeletal biology and bone health queries. *J Bone Miner Res.* 2024;39(2):106–115.
45. Quah B, Zheng L, Sng TJ, Yong CW, Islam I. Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations. *BMC Med Educ.* 2024;24(1):962.
46. Yamaguchi S, Morishita M, Fukuda H, Muraoka K, Nakamura T, Yoshioka I, et al. 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;19(4):2262–2267.
47. OpenAI. ChatGPT Can Now See, Hear, and Speak [Internet]. 2023 Sept 25 [cited 2024 Oct 13]. Available from: openai.com/index/chatgpt-can-now-see-hear-and-speak/