Heliyon 10 (2024) e28198

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

The Quality of AI-Generated Dental Caries Multiple Choice Questions: A Comparative Analysis of ChatGPT and Google Bard Language Models

Walaa Magdy Ahmed, BDS, MSc, Dip Prosthodontics, PhD, FRCDC, Assistant professor [a], Amr Ahmed Azhari, BDS, MSc, CAGS, MSBI, PhD, Assistant professor [a,*], Amal Alfaraj, BDS, MSD, FRCDC, FACP, Assistant professor [b], Abdulaziz Alhamadani, PhD, Graduate student [c], Min Zhang, MS, Graduate student [c], Chang-Tien Lu, PhD, Professor [c]

[a] *Department of Restorative Dentistry, Faculty of Dentistry, King Abdulaziz University, Jeddah, Saudi Arabia*
[b] *Department of Prosthodontics, School of Dentistry, King Faisal Universality, Al Ahsa, Saudi Arabia*
[c] *Department of Computer Science, Virginia Tech, Northern Virginia Center, USA*

ARTICLE INFO

Keywords:
ChatGPT
Bard
Dental academic assessment
Dental caries
Multiple-choice question
Dental educator

ABSTRACT

Statement of problem: AI technology presents a variety of benefits and challenges for educators.
Purpose: To investigate whether ChatGPT and Google Bard (now is named Gemini) are valuable resources for generating multiple-choice questions for educators of dental caries.
Material and methods: A book on dental caries was used. Sixteen paragraphs were extracted by an expert consultant based on applicability and potential for developing multiple-choice questions. ChatGPT and Bard language models were used to produce multiple-choice questions based on this input, and 64 questions were generated. Three dental specialists assessed the relevance, accuracy, and complexity of the generated questions. The questions were qualitatively evaluated using cognitive learning objectives and item writing flaws. Paired sample t-tests and two-way analysis of variance (ANOVA) were used to compare the generated multiple-choice questions and answers between ChatGPT and Bard.
Results: There were no significant differences between the questions generated by ChatGPT and Bard. Moreover, the analysis of variance found no significant differences in question quality. Bard-generated questions tended to have higher cognitive levels than those of ChatGPT. Format error was predominant in ChatGPT-generated questions. Finally, Bard exhibited more absolute terms than ChatGPT.
Conclusions: ChatGPT and Bard could generate questions related to dental caries, mainly at the cognitive level of knowledge and comprehension.
Clinical significance: Language models are crucial for generating subject-specific questions used in quizzes, tests, and education. By using these models, educators can save time and focus on lesson preparation and student engagement instead of solely focusing on assessment creation. Additionally, language models are adept at generating numerous questions, making them particularly

* Corresponding author. Department of Restorative Dentistry, Faculty of Dentistry, King Abdulaziz University, Jeddah, Saudi Arabia, 23233. Tel.: +966565553332.
E-mail address: aaaazhari@kau.edu.sa (A.A. Azhari).

https://doi.org/10.1016/j.heliyon.2024.e28198
Received 12 October 2023; Received in revised form 5 March 2024; Accepted 13 March 2024
Available online 19 March 2024

valuable for large-scale exams. However, educators must carefully review and adapt the questions to ensure they align with their learning goals.

1. Introduction

Artificial intelligence (AI) has the potential to revolutionize the field of dentistry in general [1] and education in particular in numerous ways, such as personalizing learning experiences, Natural Language Processing (NLP) to aid conversations [2], and the automation of administrative tasks. AI technology presents a variety of benefits and challenges for educators. A significant advantage of AI is its ability to provide students with individualized learning experiences, including feedback generation and suggestions for further study that can adapt to individual learning styles and speed. This helps educators enhance student engagement and intentional learning outcomes [3]. Moreover, AI's ability to automate administrative duties, such as grading assignments and evaluating student performance [4,5] can save educators considerable time, allowing them to focus on instruction and student interactions.[6]AI can assist teachers [6] by boosting objectivity [4] and generate adaptive assessments and study tools [7].

Large language models (LLMs) are artificial intelligence technology that can improve healthcare and education [8]. Reinforcement and supervised learning using large amounts of data train LLMs to generate unique word sequences based on human language patterns [9,10]. LLMs offer a strong and versatile tool for processing natural languages that can accurately interpret, produce, and manipulate human-like languages.

To fully realize the potential of LLMs in education, it is important for educators to stay informed about the latest developments in this field and welcome new approaches to teaching and learning. Furthermore, educators can facilitate the development of LLM-powered systems [11,12]. Moreover, LLM-powered systems must be developed and used in an ethical and responsible manner by advocating transparency, fairness, and accountability in the design and implementation, including avoiding bias [13]. However, artificial intelligence in education has some challenges, such as concerns of AI replacing human educators and resulting in job displacement and the need to ensure that AI-powered systems are equitable and impartial, as these systems can perpetuate biases in data and algorithms [14].

Students and educators must understand data gathering, use, and protection [15,16]. LLM-powered devices and ethics must be held accountable [17,18]. LLMs may ethically benefit students and dental education by uncovering patterns and trends in dental case reports, clinical notes, and other textual data to enhance dental treatment and patient outcomes [19]. For instance, ChatGPT and Google Bard (now is named Gemini since February, 2024) LLM platforms can offer dental caries prevention, diagnosis, and treatment information to the public [20]. This may help dentists arrange appointments and remind patients about them [21]. ChatGPT and Bard could help dental educators gather ideas for teaching and research.

As demonstrated in Part 1, students studying dental caries exhibited a preference for using ChatGPT and Bard to aid in the educational process. Similarly, educators could utilize these tools for teaching. This study aimed to investigate whether ChatGPT and Bard are valuable resources and user friendly for generating multiple-choice questions (MCQs) for educators on the topic of dental caries incorporating your interest in a performance with lesser format errors and achieving cognitive objectives. The following hypothesis was proposed There are no significant differences in the quality of the MCQs and answers generated by ChatGPT and Bard.

2. Material and methods

A method was created to test the potential of ChatGPT and Bard to generate dental caries multiple-choice exams and quantitively and qualitatively evaluate the method. The methodology emulated the conventional process employed by educators for exam creation, with one notable modification; we replaced the traditional approach with an AI-generated method. In the traditional approach, an educator focuses on an important excerpt from the textbook and formulates a question to test students' comprehension.

MCQs generated by ChatGPT 4.0 and Bard were evaluated using a scientific source on dental caries. Sixteen paragraphs from this source were extracted by an expert consultant based on applicability to the subject of dental caries and potential to be helpful in the development of MCQs. The ChatGPT version utilized in our study was GPT-3.5 LLM and all responses from ChatGPT were generated in May 2023. On the other hand, Bard version utilized in our study was Bard v1.0 and all responses from Bard were generated in May 2023. ChatGPT and Bard automatically created MCQs with four possible answers, of which one was correct. Dental specialists assessed the relevance, accuracy, and complexity of the generated questions. Four dental professionals in this field were included in the panel. Following the validation process, the generated questions were encoded. Assuming a normal distribution of model performance, this decision allowed for 90% power at $\alpha = 0.05$.

Four educators (Teacher A, B, C, and D) were simulated using individual two ChatGPT accounts and two Bard accounts to generate a set of MCQs based on the selected source excerpts. A total of 64 questions were evaluated. A rigorous evaluation method was designed. Evaluation methods were used for each of the four generated question sets, each including 16 questions (Fig. 1). Without knowing the answers or excerpts, Teacher A (Chatgpt) generated questions, and account B (Chatgpt) and account C (Bard) answered them. Also, Teacher B (Chatgpt other account) generated questions and account A (Chatgpt) and account D (Bard other account) answered them. Moreover, Teacher C (Bard) generated questions and account D (Bard other account) and account A (Chatgpt) answered them. Finally, Teacher D (Bard other account) generated questions and account C (Bard) and account B (Chatgpt other account) answered them. A manual qualitative evaluation of the generated questions was performed using cognitive learning objectives (CLO) (Table 1) based on Bloom's taxonomy and item writing flaws (IWF) [22,23] (Table 2).

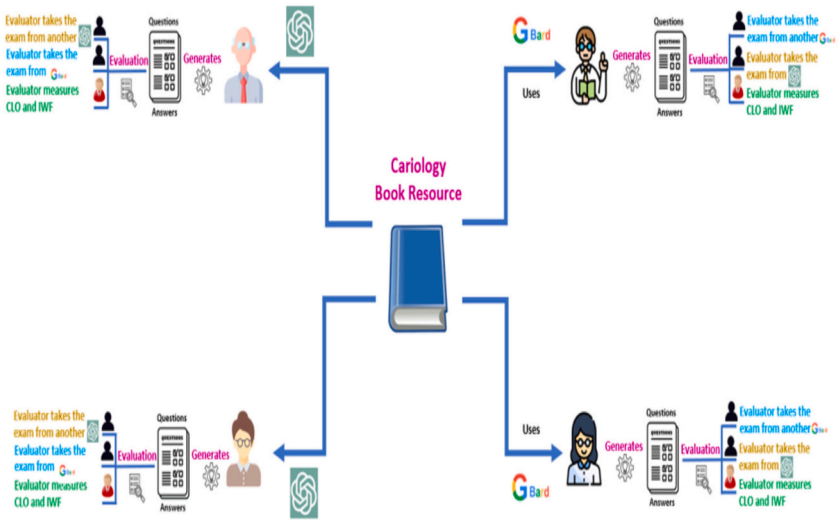

Fig. 1. Human evaluation, cognitive objective learning and item writing flaws.

Step 1 assessed whether the language model retained the questions and answers and supplied the generated answers to the opposite end. We aimed to determine whether students using the same language platform could achieve perfect scores. This outcome would imply that the language model effectively stored and utilized the question-answer pairs, enabling students to attain flawless results on their exams. Step 2 analyzed the capability of a different language model to answer the questions.

Figure (1) shows the proposed method from an educator's perspective. The methods examined the capability of ChatGPT and Bard to assess dental caries educators and evaluate the outcomes of the generated materials.

Statistical analyses were performed by counting CLO and IWF. Paired sample t-tests were used to compare the accuracy of the multiple-choice answers generated by ChatGPT and Bard. A two-way analysis of variance (ANOVA) was conducted for the CLO and IWF scores for the four teachers to identify differences in the quality of the generated MCQs between ChatGPT and Bard.

Table 1
Cognitive learning objectives description based on Bloom's taxonomy.

Levels of cognitive learning according to Bloom's Taxonomy (Bloom, 1956; Krathwohl)	
Levels	Cognitive learning objectives (CLO)
1	Knowledge: Simple recognition or recall of material
2	Comprehension: Restating or recognizing material to show understanding
3	Application: Problem-solving or applying ideas in new situations
4	Analysis: Separating ideas into component parts, examining relationships
5	Synthesis: Combining ideas into the statement or product new to the learning
6	Evaluation: Judging by using self-produced criteria or established standars

Table 2
Item writing flaws description.

Items writing flaws (IWF) chart 2001 (Haladyna et al., 2002)	
Format/technical errors	Not equal length Answer in parallel format Stem at the end of statement Excessive verbiage Format item horizontally Long answer Numeric data is not in order
Spelling error	
Grammar error	
Logic que	Except
Negative/double negative	
Absolute options	All or none of the above Always, must, never, only
Ambiguity	Sometimes, possible, may, often, commonly, rarely, usually, can, a few,
Repetition	
Uncommon abbreviations	

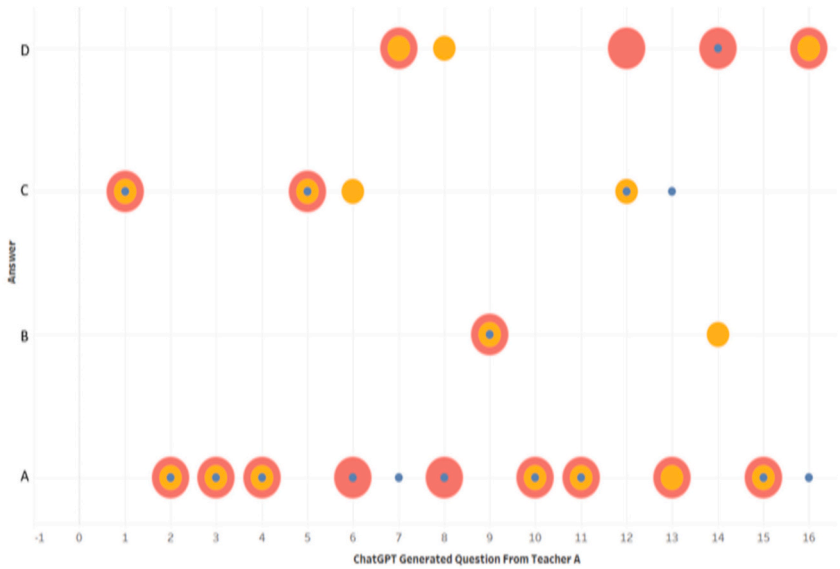

Fig. 2. ChatGPT-generated questions by Teacher A. The blue is the groundtruth, where the question was generated by ChatGPT (Teacher A) and answered using the same prompt by the same ChatGPT (Teacher A). The orange represents the answers that were also selected by the same LLM which is ChatGPT but from another account (teacher B). The Red represents the answers generated by a different LLM which is Google Bard (Teacher C). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

3. Results

3.1. Descriptive statistics: Answers from different account or platform

For the ChatGPT-generated questions from Teacher A, the end user using a different ChatGPT account received ten correct answers (questions 1, 2, 3, 4, 5, 9, 10, 11, 12 and 15) out of the 16 questions (Fig. 2). However, the end user using Bard to answer the same ChatGPT-generated questions from teacher A received 12 correct answers out of the 16 questions (except questions 7, 12, 13 and 16) (Fig. 2). Thus, the answers generated by Bard were more accurate (87.5%) than those generated by ChatGPT (62.5%) Table 3.

For the ChatGPT-generated questions from Teacher B, the end user using a different ChatGPT received 14 correct answers out of the 16 questions (except questions 7 and 10) (Fig. 3). However, the end user using Bard to answer the same ChatGPT-generated questions received 10 correct answers out of the 16 questions (except questions 5, 6, 7, 9, 13, and 15) (Fig. 3). For Teacher B, ChatGPT-generated answers were more accurate (87.5%) than those generated by Bard (68.75%) Table 3.

For the Bard-generated questions from Teacher C, the end user using a different Bard obtained nine correct answers out of the 16 questions (Fig. 2). However, the end user using ChatGPT to answer the same Bard-generated questions from Teacher C received only five correct answers out of the 16 questions (Fig. 4). For the Bard-generated questions from Teachers C, the answers generated from Bard were more accurate (56.25%) than those generated by ChatGPT (at 31.25%) Table 3.

For the Bard-generated questions from Teacher D, the end user using a different Bard received 11 correct answers out of the 16 questions (Fig. 5). However, the end user using ChatGPT to answer the same Bard-generated questions from Teacher D received only nine correct answers out of the 16 questions (Fig. 5). For the Bard-generated questions from Teachers D, the answers generated from

Table 3
Descriptive statistics comparing ChatGPT and Bard answers' scores.

Answers			
Questions	Teachers	Scores (%)	
		ChatGPT	Bard
ChatGPT Generated	A	62.5	87.5
	B	87.5	68.75
Bard Generated	C	31.25	56.25
	D	56.25	68.75
Mean		59.375	70.313
SD		23.105	12.885

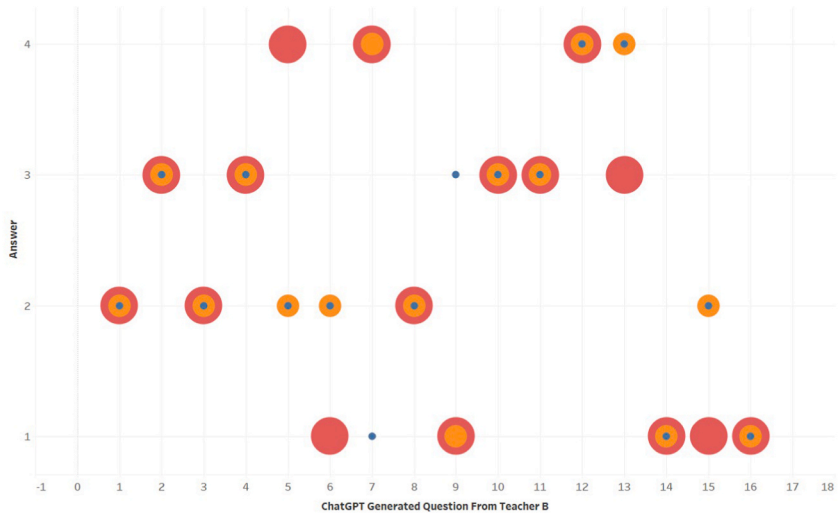

Fig. 3. ChatGPT-generated questions by Teacher B. The blue is the groundtruth, where the question was generated by ChatGPT (Teacher B) and answered using the same prompt by the same ChatGPT (Teacher A). The orange represents the answers that were also selected by the same LLM which is ChatGPT (teacher A). The Red represents the answers generated by a different LLM which is Google Bard (Teacher D). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Bard were more accurate (68.755) than those generated by ChatGPT (56.25%) Table 3. For all 64 questions, the Bard answers had 70.313% (12.885) mean (SD) of correct answers, whereas the ChatGPT answers had (59.375) mean (SD) of correct answers.

3.2. Human evaluation using CLO and IWF

The evaluators' CLO and IWF scores for 16 ChatGPT-generated questions from Teacher A had means (SD) of 2.875 (0.992) and 1.875 (1.025), respectively (Table 3). Out of the CLO scores for all 16 questions from Teacher A generated using ChatGPT, only four qualified as Knowledge level, two were at the Comprehension level, three were at the Application level, six were at the Analysis level, and one was at the Synthesis level. The IWF scores for the ChatGPT-generated questions showed that all questions had 0–3 errors, mainly in ambiguity (68%), followed by format (62%) and repetition (25%). Format errors were related to long questions, unequal length, parallel format, and long answers. Two questions exhibited no errors. Therefore, the ChatGPT-generated questions by Teacher A were of high quality.

The 16 questions from Teacher B had a CLO mean (SD) of 2.75 (1.392) and an IWF mean (SD) of 1.625 (1.025) (Table 3). From the CLO scores for the16 questions generated using ChatGPT by Teacher B, only three qualified as Knowledge level, five were at the Comprehension level, four were at the Application level, two were at the Analysis level, one was at the Synthesis level, and one was at the Evaluation level. The IWF scores for the generated questions showed that the questions had format (62%), vague terms (31 %), and

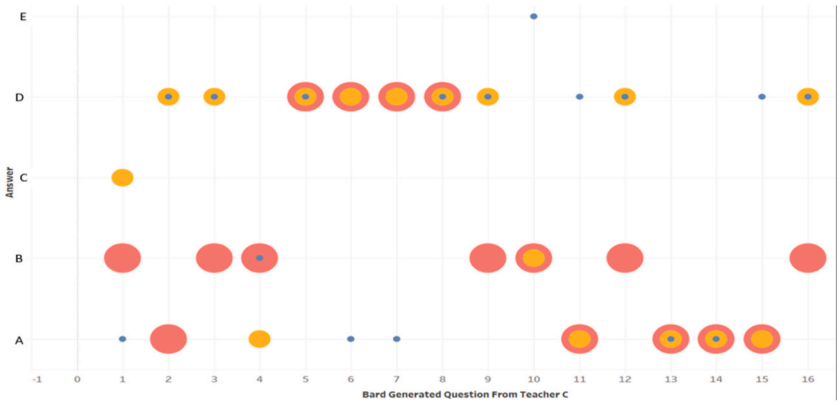

Fig. 4. Bard-generated questions by Teacher C. The blue is the groundtruth, where the question was generated by Bard (Teacher C) and answered using the same prompt by the same Bard (Teacher C). The orange represents the answers that were also selected by the same LLM which is Bard but from another account (teacher D). The Red represents the answers generated by ChatGPT (Teacher A). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

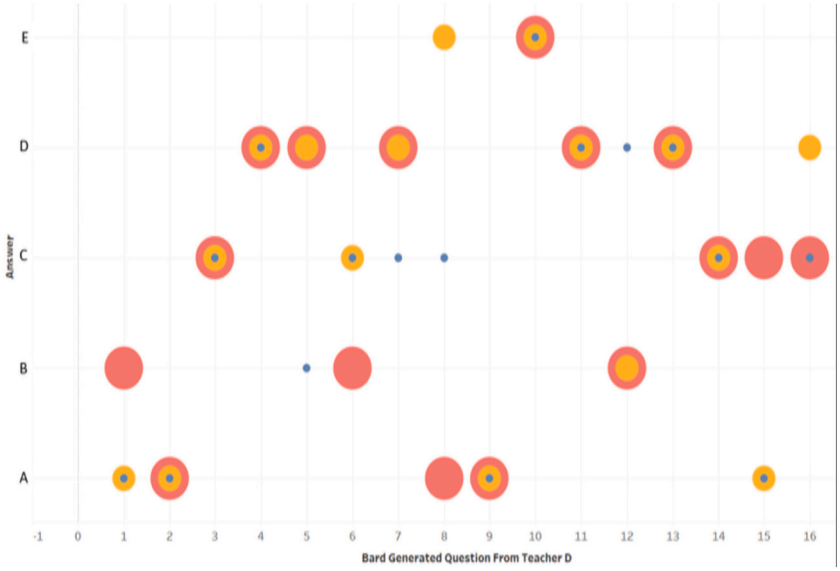

Fig. 5. Bard-generated questions by Teacher D. The blue is the groundtruth, where the question was generated by Bard (Teacher D) and answered using the same prompt by the same Bard (Teacher D). The orange represents the answers that were also selected by the same LLM which is Bard (teacher C). The Red represents the answers generated by ChatGPT (Teacher B). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

ambiguity (18%) errors. ChatGPT-generated questions from Teacher B were of medium to high quality.

The CLO and IWF scores for the 16 questions generated by Bard from Teacher C had means (SD) of 3.563 (1.171) and 1.25 (0.683), respectively (Table 3). From the CLO scores for the16 questions generated using Bard by Teacher C, none qualified as the Knowledge level, five were at the Comprehension level, one was at the Application level, six were at the Analysis level, and four were at the Synthesis level. The IWF scores for questions by Teacher C showed that questions had absolute (75%), logic que (25%), and negative stem (18%) errors. Bard-generated questions by Teacher C were of good quality.

The 16 questions from Teacher D had means (SD) of 2.688 (0.982) and 1.25 (0.856) for CLO and IWF, respectively (Table 3). The CLO scores for questions generated using Bard by Teacher D indicated that only one question qualified as the Knowledge level, eight were at the Comprehension level, five were at the Application level, two were at the Analysis level, and one was at the Synthesis level. The IWF scores for questions by Teacher D showed that most questions had format (37%), absolute (31%), and ambiguity (255) errors. Of the 16 questions, seven had only format errors, five had only spelling errors, and three had no writing flaws. Bard-generated questions by Teacher D were mostly of good quality. For all 64 questions, the overall means (SD) of CLO and IWF scores from the evaluators were 2.969 (0.403) and 1.50 (0.306), respectively. Teachers' descriptive statistics for answer and question quality are shown in Tables 3 and 4.

To test the difference in multiple-choice answers generated by ChatGPT and Bard compared to the input given by the dental specialists, we assigned every correct answer 1 and incorrect answer 0 for every teacher. A paired sample *t*-test was conducted between the two sets of answers for the four teachers. For Teacher A (ChatGPT-generated questions), Teacher B (ChatGPT-generated questions), Teacher C (Bard-generated questions), and Teacher D (Bard-generated questions), there were no statistically significant differences at $\alpha = 0.05$ ($t = -1.732$, $p = 0.104$; $t = 1.86$, $p = 0.0825$; $t = -1.732$, $p < 0.104$; $t = -1.00$, $p = 0.333$, respectively). This demonstrated that there were no statistically significant differences in the accuracy of multiple-choice answers generated by ChatGPT and Bard.

To test the quality of the MCQs generated by ChatGPT and Bard, we used the CLO and IWF scores for the MCQs from the four teachers and conducted a two-way ANOVA. Results for CLOs between ($F(3, 63) = 1.637$, $p = 0.101$) and within groups ($F(15, 63) = 1.884$, $p = 0.146$) were not significant at $\alpha = 0.05$. Likewise, results for IWFs between ($F(3, 63) = 1.714$, $p = 0.0832$) and within groups ($F(15, 63) = 2.143$, $p = 0.108$) were not significant at $\alpha = 0.05$. As such, there were no statistically significant differences in the quality of the MCQs generated by ChatGPT and Bard (Fig. 6).

4. Discussion

To the best of our knowledge, this was the first study to investigate the effectiveness of ChatGPT and Bard language models for generating MCQs for to assist educators in the field of dental caries with exam development. The results indicated that if a teacher at one end of ChatGPT generates questions with answers, students at the other end can achieve high grades.

Conversely, the results for teachers using Bard showed a random behavior. This indicated that Bard could not answers ChatGPT-generated questions unless Bard and ChatGPT were trained using a similar dataset. By employing two distinct educators on the same platform, we analyzed the level of similarity or dissimilarity in the generated questions. This comparative approach provided insights

Table 4
Descriptive statistics comparing ChatGPT and Bard quality of generated questions.

Question Quality			
Questions	Teachers	Quality Scores	
		CLO	IWF
		Mean	Mean
ChatGPT-Generated	A	2.875 (0.992)	1.875 (1.025)
	B	2.75 (1.392)	1.625 (1.025)
Bard-Generated	C	3.563 (1.171)	1.25 (0.683)
	D	2.688 (0.982)	1.25 (0.856)
Mean		2.969 (0.403)	1.5 (0.306)

Note. CLO, cognitive learning objectives; IWF, item writing flaws.

into the unique characteristics and potential divergences between educators and their outputs.

4.1. Quality assessment: IWF

ChatGPT errors were mainly in the format category, followed by ambiguity and repetition. Format error was related to long answers, unequal length, and parallel format. Bard errors was mainly related to absolute options (all or none of the above), negative questions, logic que, and format errors (unequal length). Therefore, if teachers prefer to formulate questions devoted to absolute terms, clear instructions must be provided to avoid Bard not including absolute terms or logical ques. Alternatively, the number of choices can be limited to four, as this will provide absolute terms. As longer paragraphs were provided for both platforms, there was less ambiguity. Selecting an appropriate paragraph length is essential to reduce vagueness and ambiguity.

There were no statistically significant differences in the ability to answer MCQs generated by ChatGPT and Bard from different accounts. This is logical because questions were created with AI assistance. This may be due to self-training during the first formulation of the question.

There were no statistically significant differences in the quality of the MCQs generated by ChatGPT and Bard; however, differences were observed in the types of questions. ChatGPT exhibited longer questions (format), repetition in wording, zero-logic que, and vague terms. Conversely, Bard exhibited more negative and logic que and absolute terms compared to ChatGPT. Both ChatGPT and Bard had zero spelling and grammatical errors, which could be attributed to efficient training during development.

A rigorous methodology were implemented by repeating each question three times to assess the consistency of responses from both ChatGPT and Bard. Additionally, a robust evaluation was ensured by repeating the prompts using two different language model accounts. This approach allowed to account for any inherent randomness in the answers generated by these generative AI models and provides a comprehensive analysis of stability. We believe that these adjustments enhance the reliability of our results and contribute to a more thorough understanding of the performance of ChatGPT and Bard in answering multiple-choice questions. However, in our study, small sample size was acknowledged as a limitation.

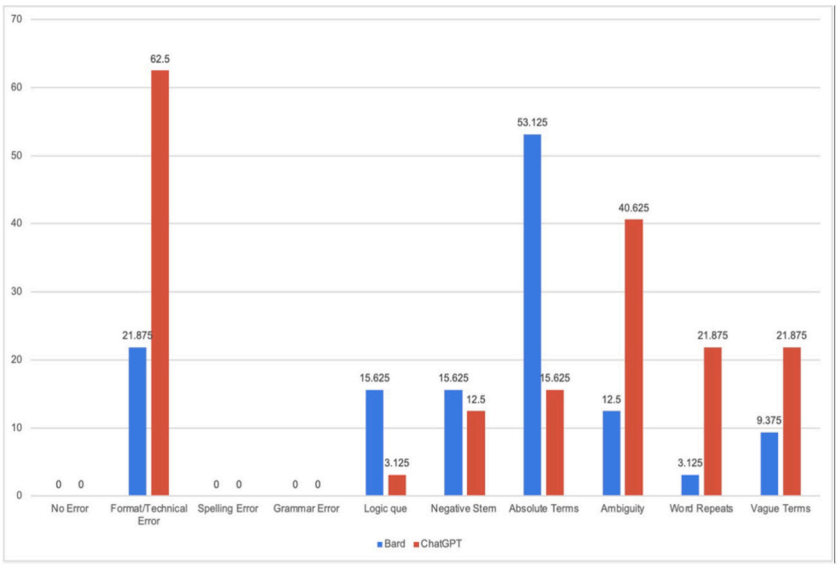

Fig. 6. Overall quality assessment item writing flaws for ChatGPT and Bard (percentages).

AI has demonstrated potential for enhancing medical educational procedures. Previous studies have shown that ChatGPT has limitations in medical education [20,24,25]. A recent study evaluated the performance of ChatGPT on the United States Medical Licensing Examination (USMLE) and revealed that ChatGPT passed all three exams (Step 1, Step 2 CK, and Step 3) near the threshold without any previous training; however, performance decreased significantly with increased question difficulty [20,23,24]. Another study tested ChatGPT using two multiple-choice question banks for the Ophthalmic Knowledge Assessment Program (OKAP) exam, and ChatGPT achieved 55.8% and 42.7% accuracies in the examinations. ChatGPT's ability to understand the human anatomy remains a limitation. Therefore, while ChatGPT has the potential to change medical education [26,27], further improvements in its application are required to allow regular use in medicine and dentistry.

LLMs present opportunities and challenges for higher education, particularly in disciplines that rely heavily on written assignments [26,27]. One review suggested that LLMs can improve academic writing fluency; however, it is necessary to establish permissible usage boundaries in science [26,27]. This indicates that LLMs are susceptible to malicious use and have severe limitations, including the possibility of disinformation. Moreover, while LLMs have a variety of beneficial applications in dental medicine, the limitations and potential dangers inherent to such artificial intelligence technologies must be carefully considered.

ChatGPT and Bard are LLM that can produce responses that may seem reasonable but may not accurate in terms of facts. These models are constrained by their training data, which may lack the most recent evidence or advancements in research. There is a potential threat where students may employ LLM to complete their projects, compromising the educational experience and giving rise to ethical considerations over ownership and intellectual property. LLMs have the ability to acquire and continue biases in their training data, resulting in biased educational material.

Furthermore, excessive dependence on AI for answers can hinder students' development of critical thinking, problem-solving abilities, and social and emotional skills. Teachers may encounter difficulties when incorporating AI into current curriculums and ensuring it supplements conventional teaching approaches. AI systems are susceptible to periods of inactivity or technical malfunctions, which can interrupt the process of acquiring knowledge. Moreover, the utilization of LLM in education raises concerns regarding the gathering, retention, and utilization of student data, including privacy and permission issues. The potential for data breaches and illegal access to sensitive information poses a significant threat to both students and institutions. LLM may lack accuracy in evaluating student achievement, particularly when it comes to subjective or creative tasks.

To optimize the advantages of LLM in education while addressing these difficulties, it is crucial to consistently update and refine AI models, establish explicit protocols and regulations to effectively tackle ethical issues, promote a well-rounded method to the utilization of AI, allocate resources towards enhancing teacher training, and cultivate a setting that places equal importance on the development of critical thinking and problem-solving skills, as well as the retention of information.

Despite the potential advantages of LLM in dentistry, several limitations need to be considered. The output of LLM may be biased depending on the type and quantity of training data employed. LLM may not be able to respond to inquiries with accuracy or dependability if the training data are not representative or diversified. Due to the lack of information on certain populations or situations, this can be particularly troublesome in the dental field. The ChatGPT both 3.5 and 4 are updated up to only September 2021. However, the chatGPT4.0, used in this study, got a new plugin which uses direct search engine 'bing' results at its disposal. Bard is disposal, which mean that the learning model can readily use the data without any delay or gaps and is real time and accessible.

LLM may not be able to consider individual patient's preferences or circumstances. LLM can offer generic information regarding dental procedures but may not be able to provide patients with individualized guidance based on their medical background, oral health, and lifestyle choices. Most importantly, LLM can help by providing information; however, it cannot replace the discretion and knowledge of dental professionals. However, the use of LLM in dentistry is fraught with ethical issues, particularly regarding informed consent. Consequently, although LLM has the potential to be a helpful tool in dentistry, it is crucial to be aware of its limitations and utilize it in conjunction with the knowledge of dental experts.

5. Conclusions

Within the limitations of this study, it was concluded that:

1. ChatGPT and Bard could generate question related to dental caries mainly at the cognitive level of knowledge and comprehension.
2. Longer paragraphs used for training the AI platform and clearer instruction were associated with higher CLO and clearer questions.
3. Bard-generated questions had higher cognitive levels than those of ChatGPT.
4. Format errors dominated the IWF criteria of ChatGPT, whereas Bard exhibited more absolute terms than ChatGPT.

Data availability statement

Data included in article/supp. material/referenced in article.

CRediT authorship contribution statement

Walaa Magdy Ahmed: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing. **Amr Ahmed Azhari:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing. **Amal Alfaraj:** Investigation, Methodology, Resources, Validation,

Visualization, Writing - original draft. **Abdulaziz Alhamadani:** Investigation, Methodology, Formal analysis, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing. **Min Zhang, MS:** Investigation, Methodology, Formal analysis, Resources; Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Chang-Tien Lu:** Investigation, Methodology, Formal analysis, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] H. Huang, O. Zheng, D. Wang, J. Yin, Z. Wang, S. Ding, H. Yin, C. Xu, R. Yang, Q. Zheng, B. Shi, ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model, arXiv Preprint ArXiv:2304.03086 (2023).

[2] Z. Kastrati, F. Dalipi, A.S. Imran, K. Pireva Nuci, M.A. Wani, Sentiment analysis of students' feedback with NLP and deep learning: a systematic mapping study, Appl. Sci. 11 (2021) 3986.

[3] A. Alam, Should robots replace teachers? Mobilisation of AI and learning analytics in education International Conference on, Advances in Computing, Communication, and Control. (ICAC3) (2021) 112. Mumbai, India.

[4] F. Pedro, M. Subosa, A. Rivas, P. Valverde, Artificial Intelligence in Education: Challenges and Opportunities for Sustainable Development, Available from:, UNESCO, 2019 https://unesdoc.unesco.org/ark:/48223/pf0000366994.

[5] S. Atlas, ChatGPT for Higher Education and Professional Development: a Guide to Conversational AI, 2023. Available from: https://digitalcommons.uri.edu/cba_facpubs/548.

[6] M. Chassignol, A. Khoroshavin, A. Klimova, A. Bilyatdinova, Artificial Intelligence trends in education: a narrative overview, Procedia Comput. Sci. 1 (2018) 1624.

[7] B. Boukenze, H. Mousannif, A. Haqiq, Predictive analytics in healthcare system using data mining techniques, Comput. Sci. Inf. Technol 23 (2016) 1–9.

[8] Open, A. I.; & CHATGPT Optimizing Language Models for Dialogue. Accessed February 15, 2023..

[9] D.A. Mikolov T, D. Povey, L. Burget, J. Černocký, Strategies for training large scale neural network language models, In2011 IEEE Workshop on Automatic Speech Recognition & Understanding 11 (2011) 196–201.

[10] T.H. Davenport, P. Barth, R. Bean, How Big Data Is Different, 2012.

[11] D.K. Kaye, Privacy and artificial intelligence: challenges for protecting health information in a new era, BMC Med. Ethics 22 (2021) 1.

[12] R.R. Althar, D. Samanta, M. Kaur, A.A. Alnuaim, N. Aljaffan, M. Aman Ullah, Software systems security vulnerabilities management by exploring the capabilities of language models using NLP, Comput. Intell. Neurosci. (2021) 8522839.

[13] R.S. Baker, A. Hawn, Algorithmic bias in education, Int. J. Artif. Intell. Educ. 32 (2022) 1052–1092.

[14] J.S.a.J. Manyika, Notes from the AI Frontier: Tackling Bias in AI (And in Humans), 2019.

[15] D. Boiko, A. M.R, G. Gomes, Emergent Autonomous Scientific Research Capabilities of Large Language Models, 2023 arXiv Preprint ArXiv:2304.05332.

[16] M. Hind, S. Houde, J. Martino, A. Mojsilovic, D. Piorkowski, J. Richards, K.R. Varshney, Experiences with improving the transparency of AI models and services, Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (2020) 1–8.

[17] C.M. Gevaert, M. Carman, B. Rosman, Y. Georgiadou, R. Soden, Fairness and accountability of AI in disaster risk management: opportunities and challenges, Patterns (N Y) 2 (2021) 100363.

[18] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S.J. Gershman, D. O'Brien, S. Shieber, J. Waldo, D. Weinberger, A. Wood, Accountability of AI under the law: the role of explanation, SSRN Journal (2017).

[19] T. Shan, F.R. Tay, L. Gu, Application of artificial intelligence in dentistry, J. Dent. Res. 100 (2021) 232–244.

[20] S. Biswas, Role of ChatGPT in Dental Science, March 28, 2023, https://doi.org/10.2139/ssrn.4403581. Available at: SSRN: https://ssrn.com/abstract=4403581.

[21] A. Shafeeg, I. Shazhaev, D. Mihaylov, A. Tularov, I. Shazhaev, Voice assistant integrated with Chat GPT, At. Indones. J. 28 (2023) 1.

[22] K. Shigli, S.S. Nayak, S. Gali, B. Sankeshwari, D. Fulari, K. Shyam Kishore, N. Upadhya P, V. Jirge, Are multiple choice questions for post graduate dental entrance examinations spot on?-Item Analysis of MCQs in Prosthodontics in India, J. Natl. Med. Assoc. 110 (2018) 455–458.

[23] H. Abouelkheir, The criteria and analysis of multiple-choice questions in undergraduate dental examinations, J. Dent. Res. Rev. 5 (2018) 59.

[24] S. Biswas, ChatGPT and the future of medical writing, Radiology 307 (2023) e223312.

[25] S. Sedaghat, Early applications of ChatGPT in medical practice, education and research, Clin. Med. 23 (2023) 278–279.

[26] F. Eggmann, R. Weiger, N.U. Zitzmann, M.B. Blatz, Implications of large language models such as ChatGPT for dental medicine, J. Esthetic Restor. Dent. (2023), https://doi.org/10.1111/jerd.13046 [E-pub ahead of print].

[27] S. Sedaghat, Future potential challenges of using large Language Models like ChatGPT in daily medical practice, J. Am. Coll. Radiol. 2 (2023) 344–345.