RESEARCH ARTICLE

F.U. Med.J.Health.Sci.
2025; 39 (1): 56 - 62
http://www.fusabil.org

Are Artificial Intelligence Chatbots Fast and Safe For Patients and Dentists?

Kübra BİLGE [1, a]
İrem İPEK [2, b]
Enes Mustafa AŞAR [3, c]

[1] Firat University,
Faculty of Dentistry,
Department of Restorative Dentistry,
Elazig – TURKIYE

[2] Firat University,
Faculty of Dentistry,
Department of Pediatric Dentistry,
Elazig – TURKIYE

[3] Selçuk University,
Faculty of Dentistry,
Department of Pediatric Dentistry,
Konya, TURKIYE

[a] ORCİD: 0000-0002-4323-9316
[b] ORCİD: 0000-0002-3542-7122
[c] ORCİD: 0000-0003-3432-8584

Objective: The aim of this study was to evaluate the accuracy and response time of three different artificial intelligence (AI) chatbots' answers to questions frequently asked by clinicians and patients.

Materials and Methods: In this study, dentists specialized in Oral and Maxillofacial Surgery, Pediatric Dentistry, Restorative Dentistry, Endodontics, Orthodontics, Oral and Maxillofacial Radiology, Periodontology and Prosthodontics were asked to prepare a total of 80 questions (n=10). The questions were multiple choice questions requiring text-based answers. 64 questions frequently asked by patients and 16 questions were selected from the WebMD website. Each question was asked once by one of the authors to each large language model (LLM). The answer to each branch's question was evaluated by two experienced experts who specialize in their field. LLM's answers were graded on Likert scale of 1 (minimum) to 5 (maximum) points according to the scoring key.

Results: When the answers given by 3 LLMs in all branches of dentistry were evaluated in terms of time, Bing gave the latest answer, while ChatGPT gave the fastest answer and LLMs are evaluated in terms of points; no statistically significant difference was found ($p<0.001$).

Conclusion: Considering the answers AI chatbots provide to dentistry-related questions and questions frequently asked by patients, they are likely to have a significant impact on various aspects of dentistry in the near future.

Key Words: *Artificial intelligence, chatbots, ChatGPT, gemini, bing, dentistry*

Yapay Zeka Sohbet Robotları Hastalar ve Diş Hekimleri İçin Hızlı ve Güvenli mi?

Amaç: Bu çalışmanın amacı, üç farklı yapay zeka (YZ) sohbet robotunun klinisyenler ve hastalar tarafından sıkça sorulan sorulara verdiği yanıtların doğruluğunu ve yanıt süresini değerlendirmektir.

Gereç ve Yöntem: Bu çalışmada, Ağız ve Çene Cerrahisi, Çocuk Diş Hekimliği, Restoratif Diş Hekimliği, Endodonti, Ortodonti, Ağız ve Çene Radyolojisi, Periodontoloji ve Protez alanında uzmanlaşmış diş hekimlerinden toplam 80 soru (n=10) hazırlamaları istendi. Sorular, metin tabanlı yanıtlar gerektiren çoktan seçmeli sorulardı. Hastalar tarafından sıkça sorulan 64 soru ve WebMD web sitesinden 16 soru seçildi. Her bir branşın sorusuna verilen yanıt, kendi alanında uzmanlaşmış iki deneyimli uzman tarafından değerlendirildi. LLM'lerin cevapları puanlama anahtarına göre 1 (minimum) ile 5 (maksimum) arasında Likert ölçeğinde derecelendirildi.

Bulgular: Diş hekimliğinin tüm dallarındaki 3 LLM'nin verdiği cevaplar zaman açısından değerlendirildiğinde, Bing en son cevabı verirken, ChatGPT en hızlı cevabı verdi ve LLM'ler puan açısından değerlendirildi; istatistiksel olarak anlamlı bir fark bulunamadı ($p<0,001$).

Sonuç: Yapay zeka sohbet robotlarının diş hekimliğiyle ilgili sorulara ve hastalar tarafından sıkça sorulan sorulara verdiği cevaplar göz önüne alındığında, yakın gelecekte diş hekimliğinin çeşitli yönleri üzerinde önemli bir etkiye sahip olmaları muhtemeldir.

Anahtar Kelimeler: *Yapay zeka sohbet robotları, ChatGPT, gemini, bing, diş hekimliği*

Received : 01.11.2024
Accepted : 19.12.2024

Correspondence

Kübra BİLGE
Firat University,
Faculty of Dentistry,
Department of Restorative Dentistry,
Elazig - TURKIYE

kubratnyl@gmail.com

Introduction

Artificial intelligence (AI) refers to the ability of a computer system to perform certain activities, typically those that can easily mimic human intelligence. Its fundamental principle relies on the development of algorithms enabling machines to process vast datasets, learn from them, tackle problems, adapt, and enhance their performance progressively over time (1). Advancing artificial intelligence technologies have begun to have an important role in dentistry, as in other fields such as technology, industry and medicine, and a number of simple tasks in dentistry can be performed by artificial intelligence with more precision, fewer errors and less personnel employment (2). Dentistry applications and tools of AI are experiencing rapid growth, aiming to help professionals deliver consistently improved oral health care. Today, such tools can support image analysis, interpretation of radiographs and diagnoses, data synthesis, and information on clinical techniques, patient record management, forensic dentistry, orthodontics, periodontics, endodontics, caries diagnosis, treatment planning and patient communication and interaction (3).

Through data synthesis and identification of risk factors, AI can help systematically evaluate clinically relevant scientific evidence. Thus, when integrated with the dentist's clinical expertise in addition to the patient's treatment needs and preferences, it can help clinicians overcome the challenges associated with implementing an evidence-based dentistry approach (4, 5). Thus, AI can promote individualized patient-centered care and support a more efficient, reliable and standardized clinical practice (6).

AI-based chatbots are software applications that can interact with users using natural language and provide various services, usually through text-based communication interfaces. With increasing access to technological devices (smartphones and computers) and the internet, AI chatbots have the potential to provide accessible health-related information and services (7). AI-powered chatbots are predicted to be widely used outside the clinical setting to address patients' personal health concerns (8).

Introduced by OpenAI in late 2022, Chatbot Generative Pre-trained Transformer (ChatGPT), alongside Google's 'Gemini' and Microsoft's Bing Chat, which debuted in early 2023, represent a class of Large Language Models (LLMs). These models stem from natural language processing (NLP), a branch of AI dedicated to facilitating interaction between computers and human language. NLP enables machines to comprehend, interpret, and generate text resembling human speech, leveraging training data to produce novel content (9-11).

LLMs hold significant promise for various applications in dentistry, ranging from streamlining dental records to assisting in clinical decision-making (12). However, the potential for AI to provide completely incorrect answers, generate nonsensical content, and propagate misinformation and disinformation as factual poses serious concerns, particularly in critical fields like healthcare (13).

The aim of this study was to evaluate the accuracy and response time of three different AI chatbots' answers to questions frequently asked by clinicians and patients.

Materials and Methods

Research and Publication Ethics: Ethical approval was not required as no human participants were included in the study.

80 multiple choice questions about clinical dentistry were asked to three different LLMs. These LLMs were: ChatGPT- 3.5, Google Gemini and Microsoft Bing–chat functionality. The flowchart of the study is shown in Figure 1.

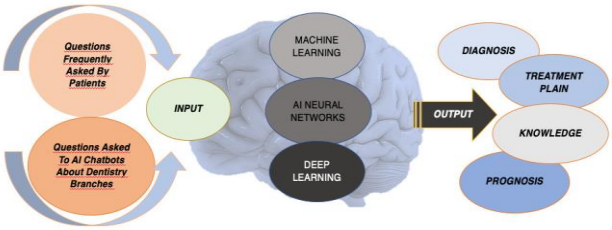

Figure 1. The flowchart of the study

Design Of Questions Asked To AI Chatbots About Dentistry Branches: In this study, two experienced experts from each branch of dentistry were asked to ask multiple-choice questions that were clinically meaningful and had strong evidence to support the answers. A pool of 80 questions was created from the disciplines of Oral and Maxillofacial Surgery, Pediatric Dentistry, Restorative Dentistry, Endodontics, Orthodontics, Oral and Maxillofacial Radiology, Periodontology and Prosthodontics (n=10).

Design Of Questions Frequently Asked By Patients: Two experts from each branch of dentistry were asked to record 8 questions frequently asked by patients and forward them to the researcher. In addition, the WebMD (www.webmd.com) website was examined for the questions frequently asked by patients, and 2 questions were selected among these questions, each suitable for the relevant branch. A question pool of 80 questions in total was created.

Generating Answers In LLMs: The questions asked about the branches of dentistry are multiple choice questions written in Turkish scientific language using appropriate terminology, requiring text-based answers, and do not contain photographs or illustrations. The format of the questions consisted of a question or a case scenario followed by 5 potential answers. The task was to determine the single most appropriate answer. Frequently asked questions by patients were asked directly in Turkish without any changes.

Tests involving ChatGPT, Gemini and Bing took place on February 19-24.2024, respectively. Each question was asked once to each LLM by one of the authors. The entire chat history was cleared before starting each trial. A new chat window was then opened to eliminate possible content transfer. When LLMs did not answer; there were no follow-up questions, restatements, or additional clarifications. It was also not asked a second time by another author. After each question was asked, the stopwatch was started simultaneously and the answer time was recorded in seconds. For an objective evaluation of seconds, an internet connection with the same speed (100 megabytes per second) was used when calculating seconds in the study.

The answers to each question prepared by experts in the branches of dentistry were evaluated by two experienced experts in the field. The answers given to the frequently asked questions of patients regarding the branches of dentistry were evaluated by two experienced experts in the field. When inconsistencies emerged in some answers, the inconsistencies were resolved with the help of a third expert who consulted reference articles to reach consensus. LLM's answers were graded on a scale of 1 (minimum) to 5 (maximum) points according to the scoring key. Answers were blindly scored by assigning a letter to each LLM; so they were not aware of which LLM they were scoring at the time.

A modified version of the Likert scale (7) was used to assign scores based on the content and context of the answers:

According to Likert scale;

[5] The application gave correct and adequate answers;

[4] The application responded correctly but not sufficiently;

[3] The application did not directly answer the question, but suggested references where the correct answer to the question could be found;

[2] The application could not provide an adequate answer to the question and did not suggest a source regarding the question.

[1] The application answered the question incorrectly.

Statistical Analysis: The data obtained as a result of the research were evaluated with the statistical package program (SPSS 22.0). Normality analysis of numerical data was calculated through Kolmogorov Smirnov test. It was determined that the data did not show a normal distribution. The Kruskal-Wallis test was used to evaluate the data, and the Mann-Whitney U test was used to find the group or groups that differed as a result of the analysis.

Results

Descriptive statistics for the answers provided by the 3 LLMs are shown in Table 1.

When the answers given by 3 LLMs to questions in all branches of dentistry and to questions frequently asked by patients are evaluated in scores; There was no statistically significant and shown in Figure 2 (p>0.05).

Table 1. Descriptive statistics for the answers provided by the three LLMs

AI	LLMs answers to questions				LLMs answers to questions frequently asked by patients			
	Score		Seconds		Score		Seconds	
	Mean ±SD	*p*	Mean ±SD	*p*	Mean ±SD	*p*	Mean ±SD	*p*
ChatGPT	2.75±1.21		6.05±2.95		3.68±1.90		5.98±2.09	
Gemini	2.73±1.24	0.814	7.99±2.65	<0.001*	3.93±1.71	0.736	8.23±1.63	<0.001*
Bing	2.62±1.23		13.28±4.56		3.12±1.32		12.54±1.87	

*p< 0.05 was accepted as significance level.

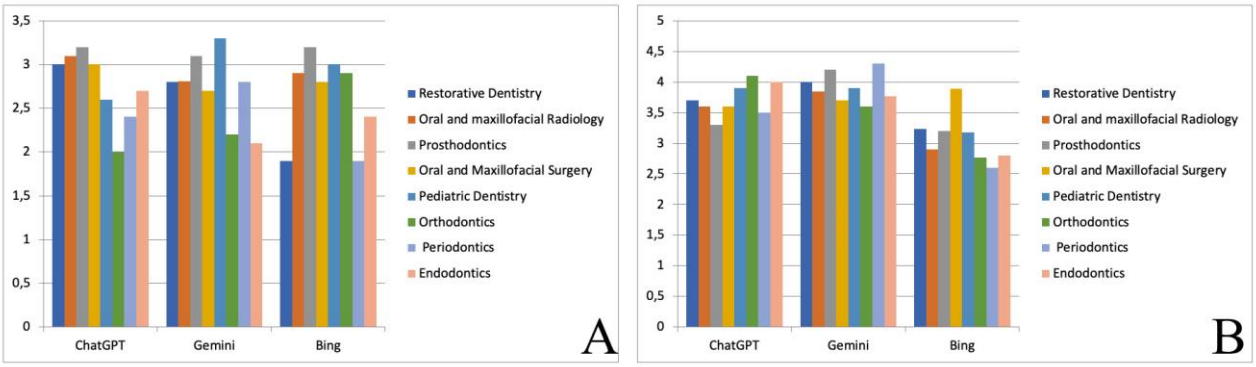

Figure 2. Scoring of the answers given by three LLMs all branches of dentistry (A) and to questions frequently asked by patients (B)

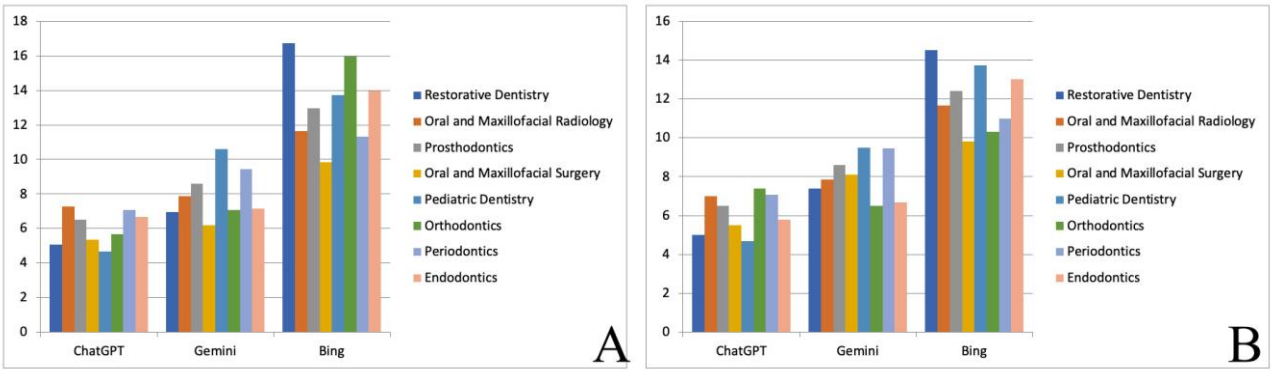

Figure 3. Seconds of the answers given by three LLMs all branches of dentistry (A) and to questions frequently asked by patients (B)

When the questions asked by clinicians are evaluated on a score basis;

- While the accuracy of ChatGPT's answers to questions about prosthodontics was the highest, the accuracy of its answers to orthodontics questions had the lowest score.
- While Gemini's answers to questions about pediatric dentistry had the highest score, the accuracy of answers to endodontics questions had the lowest score.
- Bing's answers to questions about prosthodontics had the highest accuracy score, while answers to questions about restorative dentistry and periodontics had the lowest accuracy scores.

When the questions asked by the patients are evaluated on a score basis;

- While the accuracy of ChatGPT's answers to questions about orthodontics was the highest, the accuracy of answers to questions about prosthodontics had the lowest score.
- While Gemini's answers to periodontics related questions had the highest accuracy score, the accuracy of her answers to orthodontics questions had the lowest score.
- While Bing's answers to questions about oral and maxillofacial surgery had the highest accuracy, the accuracy of answers to periodontics questions had the lowest score.

When the answers given by 3 LLMs to questions in all branches of dentistry and to questions frequently asked by patients are evaluated in seconds; It was statistically significant and shown in Figure 3 ($p<0.001$). While Bing gave the last answer to all branches, ChatGPT gave the fastest answer.

When the questions asked by clinicians are evaluated on a time basis;

- ChatGPT responded fastest to questions about pediatric dentistry and slowest to questions about oral and maxillofacial radiology.
- Gemini responded fastest to questions about oral and maxillofacial surgery and slowest to questions about pediatric dentistry.
- Bing responded fastest to questions about oral and maxillofacial surgery and slowest to questions about restorative dentistry.

When the questions asked by the patients are evaluated on a time basis;

- ChatGPT responded fastest to questions about pediatric dentistry and slowest to questions about orthodontics.
- Gemini responded fastest to questions about orthodontics and slowest to questions about pediatric dentistry.

Bing responded fastest to questions about oral and maxillofacial surgery and slowest to questions about restorative dentistry.

Discussion

In today's dental clinical practice, although the dissemination and development of EBD continues, successful implementation becomes difficult due to difficulties such as rapid scientific and technological developments, outdated guidelines, lack of evidence and implementation (14). AI chatbots, which can theoretically generate immediate evidence-based answers to scientific questions and thus act as the dentist's personal scientific advisor at the chairside, seem to have the potential to be an ideal tool for the successful implementation and development of EBD (15). At the same time, it is an important part of health services that patients can always access their health information accurately and easily (16). With technological advances, patients can access a many of resources to obtain information about healthcare services (such as diagnosis, health promotion, consultancy services) using artificial intelligence systems. Additionally, this easy

access to reliable and accurate health information can also help patients manage their health more effectively (17).

Although LLMs have the ability to identify patterns and organize data, they are known to have limitations in being able to fully understand and grasp the underlying meaning and context of information (18). Therefore, in our study, multiple-choice questions were used instead of open-ended questions to ensure that the LLMs clearly indicate the type of answer requested, thereby preventing the generation of additional or fictitious information.

When the findings of our study were evaluated, although there was no statistical difference in terms of score for all 3 LLMs, ChatGPT and Gemini had higher accuracy values in all branches of dentistry. Taşkın et al. (19) evaluated the success of LLMs in answering common orthodontic questions and reported that ChatGpt gave the most desirable answers. Giannakopoulos et al. (15) evaluated the performance of LLMs (ChatGPT, Gemini and Bing) in supporting evidence-based dentistry and found that although there was no statistical difference between the 3 LLMs, Bing had the lowest score. Suarez et al. (20) evaluated the consistency and accuracy of ChatGPT's endodontic question answers and reported that overall consistency of answers produced by ChatGPT was high, but that it underperformed in correctly answering questions of lower difficulty. In Acar (21)'s study evaluating how LLMs answered to questions about oral surgery complications, ChatGPT showed a higher accuracy score than Bing and Gemini. Differences in the data used to train and teach AI models may be why chatbots vary in their human-like responses (22, 23). In Balel (24)'s study, ChatGPT was asked questions frequently asked by patients about oral and maxillofacial surgery procedures. As a result of this study, he reported that ChatGPT has significant potential as a patient information tool, but may not be completely safe for the time. Because ChatGPT is designed to produce human-like text by predicting the likelihood of a word based on previous words in a sentence, it may have responded more accurately and quickly based on context and patterns learned from previously extensive datasets (25). Howard et al. (26) asked ChatGPT about the statements in the 2020 clinical consensus statement on ankyloglossia and reported that although ChatGPT reflects medical perspectives on ankyloglossia, caution should be exercised in aligning with non-consensus statements and relying on it for medical advice. While AI chatbots have access to information due to their nature, they produce several possible responses and may be selecting fewer possible responses with repeated input, given the randomness in the code. This can be especially dangerous in the medical setting, and patients and families should be warned about this potential limitation of AI chatbots as a resource.

When the findings of our study were evaluated in terms of seconds, statistically significant differences were found in all branches of dentistry for all 3 LLMs. In our study, ChatGPT, Gemini and Bing gave the fastest answers to questions in all branches of dentistry, respectively. ChatGPT's speed and greater accuracy can be attributed to its large database, more reliable availability, and extensive training (27). It is also known that Gemini (28) and Bing (29) are based on ChatGPT technologies, but since the exact architectures and technical details of the models are not known, they may perform differently from each other. These differences can also be attributed to the varying design philosophies of AI companies, different algorithms used, the datasets used for training, and the goals that AI is designed to achieve (22, 30).

Despite the generally high validity scores of LLMs, these chatbots are known to make critical errors in some answers. In particular, the length, complexity and difficulty of the questions and the multifaceted nature of dental knowledge may be the cause of these critical errors. LLMs' answers were often more superficial and lacked deep thinking, which revealed their limitations in handling complex queries (31).

Additionally, as with any new technology, there are limitations that must be addressed to ensure the benefits of using this innovation outweigh its risks. These limitations may have the potential to produce biased, outdated or inaccurate content (32). AI-based training tools can compromise the ability of healthcare professionals to develop the skills necessary for human interaction and communication, critical skills, and can also mislead patients into misinformation (33).

In conclusion, LLMs are poised to make a substantial impact on numerous facets of dentistry in the foreseeable future. Nevertheless, they currently lack the capability to replacing dentists in clinical decision-making. Given that AI technologies are still evolving, additional research and development are necessary to unlock their full potential benefits for dental healthcare. With advancements in deep learning, the performance of LLMs is anticipated to enhance, rendering them increasingly valuable and efficient in dentistry. Continuous studies over time are essential to evaluate the learning curve and the capacity of AI to evolve and improve.

References

1. Rodrigues JA, Krois J, Schwendicke F. Demystifying artificial intelligence and deep learning in dentistry. Braz Oral Res 2021;35
2. Chen Y-w, Stanley K, Att W. Artificial intelligence in dentistry: current applications and future perspectives. Quintessence Int 2020; 51(3): 248-257.
3. Schwendicke F BM, Uribe S, Cheung W, Verma M, Linton J, Kim YJ. White paper,. Artificial Intelligence for dentistry, FDI 2023.
4. McGlone P, Watt R, Sheiham A. Evidence-based dentistry: An overview of the challenges in changing professional practice. Br Dent J 2001; 190(12): 636-639.
5. Giannakopoulos K, Kavadella A, Stamatopoulos V, Kaklamanos E. Evaluation of generative artificial intelligence large language models chatGPT, google bard, and microsoft bing chat in supporting evidence-based dentistry: A comparative mixed-methods study. J Med Internet Res 2023;
6. Mertens S, Krois J, Cantu AG, Arsiwala LT, Schwendicke F. Artificial intelligence for caries detection: Randomized trial. J Dent 2021; 115: 103849.
7. Aggarwal A, Tam CC, Wu D, Li X, Qiao S. Artificial intelligence–based chatbots for promoting health behavioral changes: Systematic review. J Med Internet Res 2023; 25: e40789.
8. Lyons RJ, Arepalli SR, Fromal O, Choi JD, Jain N. Artificial intelligence chatbot performance in triage of ophthalmic conditions. Can J Ophthalmol 2023; 59(4): e301-e308.
9. Cadamuro J, Cabitza F, Debeljak Z, et al. Potentials and pitfalls of ChatGPT and natural-language artificial intelligence models for the understanding of laboratory medicine test results. An assessment by the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Working Group on Artificial Intelligence (WG-AI). Clin Chem Lab Med 2023; 61(7): 1158-1166.
10. Krishnan C, Gupta A, Gupta A, Singh G. Impact of artificial intelligence-based chatbots on customer engagement and business growth. Deep Learning for Social Media Data Analytics. Springer; 2022: 195-210.
11. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions. JMIR Med Educ 2023; 9(1): e48291.
12. Shen Y, Heacock L, Elias J, et al. ChatGPT and other large language models are double-edged swords. Radiological Society of North America 2023; 307 (2): e230163.
13. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. J Esthet Restor Dent 2023; 35(7):1098-1102.
14. Frantsve-Hawley J, Abt E, Carrasco-Labra A, et al. Strategies for developing evidence-based clinical practice guidelines to foster implementation into dental practice. The Journal of the American Dental Association 2022; 153(11): 1041-1052
15. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the performance of generative ai large language models chatgpt, google bard, and microsoft bing chat in supporting evidence-based dentistry: Comparative mixed methods study. J Med Internet Res 2023; 25: e51580.
16. Snyder CF, Wu AW, Miller RS, Jensen RE, Bantug ET, Wolff AC. The role of informatics in promoting patient-centered care. The Cancer Journal 2011; 17(4): 211-218.
17. de Cock C, Milne-Ives M, van Velthoven MH, Alturkistani A, Lam C, Meinert E. Effectiveness of conversational agents (virtual assistants) in health care: Protocol for a systematic review. JMIR Res Protoc 2020; 9(3): e16934.
18. Malhotra K, Wong BNX, Lee S, et al. Role of artificial intelligence in global surgery: A review of opportunities and challenges. Cureus 2023; 15(8): e43192.
19. Taşkın S, Cesur MG, Mustafa U. Yapay zekâ destekli sohbet robotlarının yaygın ortodontik soruları cevaplama başarısının değerlendirilmesi. SDÜ Tıp Fakültesi Dergisi 2023; 30(4): 680-686.
20. Suárez A, Díaz-Flores García V, Algar J, Gómez Sánchez M, Llorente de Pedro M, Freire Y. Unveiling the ChatGPT phenomenon: evaluating the consistency and accuracy of endodontic question answers. Int Endod J 2024; 57(1): 108-113.
21. Acar AH. Can natural language processing serve as a consultant in oral surgery? J Stomatol Oral Maxillofac Surg 2024; 125(3): 101724.
22. Bhardwaz S, Kumar J. An extensive comparative analysis of chatbot technologies-chatGPT, google BARD and microsoft bing. IEEE 2023: 673-679.
23. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: An analysis of multiple clinical and research scenarios. J Med Syst 2023; 47(1): 33.
24. Balel Y. Can ChatGPT be used in oral and maxillofacial surgery? J Stomatol Oral Maxillofac Surg 2023; 124(5): 101471.
25. Sanmarchi F, Bucci A, Nuzzolese AG, et al. A step-by-step researcher's guide to the use of an AI-based transformer in epidemiology: an exploratory analysis of ChatGPT using the STROBE checklist for observational studies. J Public Health 2023: 1-36.
26. Howard EC, Chong NY, Carnino JM, Levi JR. Comparison of ChatGPT Knowledge Against 2020 Consensus Statement on Ankyloglossia in Children. Int J Pediatr Otorhinolaryngol 2024: 111957.
27. Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. NPJ Digit Med 2021; 4(1): 93.
28. An important next step on our AI journey "https://blog.google/intl/en-africa/products/explore-get-answers/an-important-next-step-on-our-ai-journey/. "
29. Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web "https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/. "
30. Gao J, Galley M, Li L. Neural approaches to conversational AI: Question answering, task-oriented dialogues and social chatbots. Now Foundations and Trends; 2019, doi: 10.1561/1500000074..

31. Mohammad-Rahimi H, Ourang SA, Pourhoseingholi MA, et al. Validity and reliability of artificial intelligence chatbots as public sources of information on endodontics. Int Endod J 2023; 57(3): 305-314.

32. Sallam M. ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. MDPI 2023: 887.

33. Sallam M, Salim NA, Barakat M, Ala'a B. ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations. Narra J 2023; 3(1): e103.