



# Performance of Chat Generative Pretrained Transformer and Bard on the Questions Asked in the Dental Specialty Entrance Examination in Turkey Regarding Bloom's Revised Taxonomy

Türkiye'deki Diş Hekimliğinde Uzmanlık Eğitimi Giriş Sınavı Sorularına İlişkin ChatGPT ve Bard'ın Bloom'un Revize Edilmiş Taksonomisine Dayalı Performansı

Rana TURUNÇ OĞUZMAN<sup>1</sup>   
Zeliha Zuhal  
YURDABAKAN<sup>2</sup> 

<sup>1</sup>Department of Prosthodontics,  
Faculty of Dentistry, Altınbaş  
University, Istanbul, Turkey  
<sup>2</sup>Department of Oral and  
Dentomaxillofacial Radiology,  
Faculty of Dentistry, Altınbaş  
University, Istanbul, Turkey

## ABSTRACT

**Objective:** This study aimed to compare the performance of chat generative pretrained transformer (ChatGPT) (GPT-3.5) and Bard, 2 large language models (LLMs), through multiple-choice dental specialty entrance examination (DUS) questions.

**Methods:** Dental specialty entrance examination questions related to prosthodontics and oral and dentomaxillofacial radiology up to 2021, excluding visually integrated questions, were prompted into LLMs. Then the LLMs were asked to choose the correct response and specify Bloom's taxonomy level. After data collection, the LLMs' ability to recognize Bloom's taxonomy levels and the correct response rate in different subheadings, the agreement between LLMs on correct and incorrect answers, and the effect of Bloom's taxonomy level on correct response rates were evaluated. Data were analyzed using McNemar, Chi-square, and Fisher-Freeman-Halton exact tests, and Yate's continuity correction and Kappa agreement level were calculated ( $P < .05$ ).

**Results:** Notably, the only significant difference was observed between ChatGPT's correct answer rates for oral and dentomaxillofacial radiology subheadings ( $P = .042$ ;  $P < .05$ ). For total prosthodontic questions, ChatGPT and Bard achieved correct answer rates of 35.7% and 38.9%, respectively, while both LLMs achieved a 52.8% correct answer rate for oral and dentomaxillofacial radiology. Moreover, there was a statistically significant agreement between ChatGPT and Bard on correct and incorrect answers. Bloom's taxonomy level did not affect the correct response rates significantly.

**Conclusion:** The performance of ChatGPT and Bard did not demonstrate a reliable result on DUS questions, but considering rapid advancements in these LLMs, this performance gap will probably be closed soon, and these LLMs can be integrated into dental education as an interactive tool.

**Keywords:** ChatGPT, Bard, artificial intelligence, large language models, dental education, multiple choice questioning

## ÖZ

**Amaç:** Bu çalışmanın amacı, iki büyük dil modeli (LLM) olan ChatGPT (GPT-3,5) ve Bard'ın Diş Hekimliğinde Uzmanlık Eğitimi Giriş Sınavındaki (DUS) çoktan seçmeli sorular üzerindeki performansını karşılaştırmaktır.

**Yöntemler:** Görsel içerikli sorular hariç olmak üzere, 2021 yılına kadar olan protetik diş tedavisi ve ağız, diş ve çene radyolojisi ile ilgili DUS soruları LLM'lere sorulmuştur. Daha sonra LLM'lerden doğru yanıtı seçmeleri ve Bloom'un taksonomi düzeyini belirtmeleri istenmiştir. Veriler toplandıktan sonra, LLM'lerin Bloom taksonomi düzeylerini belirleyebilme becerileri ve farklı alt başlıklardaki

Received/Geliş Tarihi: 24.08.2023

Accepted/Kabul Tarihi: 31.10.2023

Publication Date/Yayın Tarihi: 18.01.2024

Corresponding Author/Sorumlu Yazar:  
Rana TURUNÇ OĞUZMAN  
E-mail: ranaturunc@gmail.com

Cite this article as: Turunç-Oğuzman R,  
Yurdabakan ZZ. Performance of  
ChatGPT and Bard on the questions  
asked in the dental specialty entrance  
examination in Turkey regarding  
bloom's revised taxonomy. *Curr Res  
Dent Sci.* 2024;34(1):25-34.



Content of this journal is licensed under  
a Creative Commons Attribution-  
NonCommercial-NoDerivatives 4.0  
International License.

doğru yanıt oranları, LLM'ler arasında doğru ve yanlış yanıtlara ilişkin uyumu ve Bloom taksonomi düzeyinin doğru yanıt oranları üzerindeki etkisi değerlendirilmiştir. Veriler Mc Nemar, Ki-kare ve Fisher Freeman Halton Exact testleri kullanılarak analiz edilmiştir, Yate's Continuity Düzeltmesi ve Kappa uyum düzeyi hesaplanmıştır ( $P < .05$ ).

**Bulgular:** ChatGPT'nin doğru cevap oranları arasında tek anlamlı fark ağız, diş ve çene radyolojisi alt başlıkları arasında gözlenmiştir ( $P: .042$ ;  $P < .05$ ). Toplam protez soruları için ChatGPT ve Bard sırasıyla %35,7 ve %38,9 oranında doğru cevap verirken, her iki LLM de ağız, diş ve çene radyolojisi için %52,8 oranında doğru cevap vermiştir. Ayrıca, ChatGPT ve Bard arasında doğru ve yanlış cevaplar konusunda istatistiksel olarak anlamlı bir uyum saptanmıştır. Bloom'un taksonomi düzeyi doğru yanıt oranlarını anlamlı derecede etkilememiştir.

**Sonuç:** ChatGPT ve Bard, DUS soruları üzerinde güvenilir bir performans göstermemiştir, ancak LLM'lerdeki hızlı gelişmeler göz önünde bulundurulduğunda, performans açıkları muhtemelen yakında kapanacak ve bu LLM'ler interaktif bir araç olarak diş hekimliği eğitimine entegre edilebilecektir.

**Anahtar Kelimeler:** ChatGPT, Bard, yapay zeka, büyük dil modelleri, diş hekimliği eğitimi, çoktan seçmeli soru

## INTRODUCTION

The rapid progress in artificial intelligence (AI) has given rise to optimistic prospects for its utilization within the medical domain. Among the various applications of AI, one noteworthy implementation involves large language models (LLMs). These models possess the ability to produce text resembling human language and respond to prompts by leveraging patterns acquired from extensive training on substantial volumes of textual data.<sup>1,2</sup> This potential of LLMs spans across diverse fields, encompassing medical education and aiding clinical decision-making through multilingual interaction.<sup>1</sup> In the domain of education, AI's evolution has introduced novel prospects for the transformation of established learning methodologies. Conventionally, medical education has depended on resources like textbooks, academic journals, and search engines such as PubMed (National Library of Medicine, Bethesda, Md, USA) for knowledge acquisition. Nonetheless, there has been growing emphasis on integrating multidisciplinary AI-based training to adapt to the evolving landscape of medical practices, and LLMs can also be used as a part of this AI-based training.<sup>3</sup>

Regarding LLMs, chat generative pretrained transformer (ChatGPT) also known as GPT-3.5 (OpenAI, San Francisco, Calif, USA) became the most popular one with more than a hundred million users.<sup>4,5</sup> This conversational AI embodies a lineage of LLMs termed the GPT series, underpinned by deep learning methodologies. ChatGPT delivers direct responses to queries rather than merely directing users to various websites as web search engines do, thereby enhancing the engagement and immediacy of the interaction. It is easy to access and available online, and it exhibits proficiency in addressing queries across various languages, including English, Turkish, and several other languages.<sup>1,2</sup> Following its public release on November 30, 2022, ChatGPT has garnered considerable prominence, particularly within the realm of education.<sup>3</sup> The favorable outcomes exhibited by ChatGPT in these evaluations suggest its potential utility as an educational tool within the medical domain.<sup>3</sup> Several studies have highlighted the coherent and informative nature of ChatGPT's responses, signifying its potential as an interactive tool for medical education, capable of augmenting learning and enhancing comprehension of intricate subjects.<sup>1,5-8</sup> However, it is reported that the outputs it produces are rooted in data acquired prior to September 2021, so some of the information it is revealing may not be relevant today. Therefore, to overcome this problem, companies

are actively developing alternative LLMs and one such solution is Bard (Google Inc., Mountain View, Calif, USA). Released in March 2023, Bard distinguishes itself by its capacity to instantly access and assimilate real-time information from the internet while formulating responses. This unique trait fuels anticipation of Bard's efficacy across diverse domains that demand up-to-date insights.<sup>1,9</sup> Nonetheless, while the proficiency of ChatGPT in specialized medical multiple-choice question (MCQ) assessments across diverse medical domains, encompassing the United States Medical Licensing Exam, and other exams on orthopedics, cardiology, microbiology, gynecology, family medicine have been documented investigations into Bard's performance in medical education remain in their infancy, and direct comparisons between Bard and established LLMs are just commencing.<sup>6,10-14</sup> Importantly, there exists a notable absence of head-to-head evaluations between Bard and ChatGPT, particularly in the specific context of multiple-choice dental examinations.

Multiple-choice examinations represent the prevailing mode of assessment for gauging student learning due to their capacity for objective evaluation. They offer the advantage of efficiently covering a broad spectrum of concepts within a constrained time frame, affording students immediate formative feedback and supplying educators with achievement data, while also informing learning developers about student engagement levels.<sup>15,16</sup> Additionally, responding to MCQs, enables students to swiftly pinpoint gaps in their knowledge, which is valuable for directing future learning.<sup>17</sup> However, it's important to prepare well-crafted MCQs to direct the students' future learning through critical thinking. In this pursuit, the application of Bloom's revised taxonomy, which is a hierarchical classification of cognitive learning objectives, has emerged as a strategy to design MCQs that effectively assess critical thinking competencies, with evidence suggesting that the incorporation of higher-order MCQs supports the cultivation of a profound comprehension of scientific processes.<sup>16,18</sup> This taxonomy not only finds utility in preparing MCQs but also in the realms of other assessments, teaching, and learning, providing an easily comprehensible and practical guideline for curriculum development.<sup>19</sup> According to this guideline, the verbs "remember" (formerly labeled as knowledge or recall in the original version of Bloom's taxonomy) and "understand" (previously called comprehension) constitute the lower-order questions, whereas the verbs "apply," "analyze," "evaluate," and "create" form the higher-order questions. These higher-order questions necessitate the application of advanced cognitive skills,

compelling students to employ their foundational knowledge in intricate ways.<sup>16,19,20</sup> Consequently, it is imperative for an LLM to adeptly discern the distinction between lower and higher-order questions, thus ensuring its capability to interpret these categories effectively for future utilization in the educational field, especially for preparing questions.<sup>19</sup>

Several studies have investigated LLMs' performance on their correct response rates in different medical fields according to lower and higher-order questions, and they reported them to be promising to be used in medical education.<sup>5,9,12,18</sup> However, to the best of the authors' knowledge, there is no study examining the ChatGPT's and Bard's performances on Bloom's taxonomy and their correct response rates in the dental field, so it's unclear whether these LLMs can be used confidently as a tool for dental education. Therefore, in this study, the Dental Specialty Entrance Examination (abbreviated as DUS in Turkish), which is taken in Turkey by candidates who want to receive specialty education in the schools of dentistry, is evaluated. This exam consists of MCQs, with 5 options for each question. It comprises 40 questions of basic sciences and 10 questions of each clinical science, which prosthodontics, oral and maxillofacial radiology, pediatric dentistry, endodontics, orthodontics, periodontology, oral and maxillofacial surgery, and restorative dentistry.<sup>21</sup> In this study, DUS questions of prosthodontics and oral and dentomaxillofacial radiology that the authors specialized on are evaluated to assess: 1- the performance of ChatGPT and Bard on the recognition of Bloom's taxonomy level of the MCQs, 2- the performance of ChatGPT and Bard on the correct response rate on MCQs related to subheadings of prosthodontics and oral and dentomaxillofacial radiology, 3- the agreement of ChatGPT and Bard on correct/incorrect answers, 4- the effect of Bloom's taxonomy level (lower or higher-order) on the correct response rate of LLMs and the LLMs' agreement on correct/incorrect answers to all questions according to Bloom's taxonomy level specified by the authors and thereby identify the LLMs' strengths and weaknesses, create awareness and pave the way for further research and development of LLMs to be used in dental education. Consequently, the null hypotheses were set as follows: (1) Bard outperforms ChatGPT on the recognition of Bloom's taxonomy level of the MCQs, (2) Bard outperforms ChatGPT on the correct response rate on MCQs related to subheadings of prosthodontics and oral and dentomaxillofacial radiology, (3) The LLMs have no agreement on correct and incorrect answers, (4) Bloom's taxonomy level, specified by the authors, affects the correct answer rate of the LLMs and the LLMs have no agreement on correct/incorrect answers to all questions according to Bloom's taxonomy level specified by the authors.

## MATERIAL AND METHODS

All DUS questions on prosthodontics and oral and dentomaxillofacial radiology, up to the year 2021, were downloaded from the database of ÖSYM, which is a governmental institution established by the Turkish parliament to assess and place proficient applicants who seek admission to higher education programs by means of centralized examinations. The questions after 2021 were excluded since ChatGPT covers data up to 2021. In prosthodontics, 2 figure-containing and 2 canceled questions were excluded, and 126 questions were asked; in radiology, 7 questions containing radiography were excluded, and only 123 questions with text content were asked in Turkish, the same as in the ÖSYM database. Questions with images, charts, or tables were excluded

since ChatGPT is adapted according to text input and the current version of Bard allows images only in English prompts.

Both authors independently categorized the questions as lower-order or higher-order according to Bloom's taxonomy. All categorization was performed blindly, without the knowledge of any LLM's responses to the questions. In addition, prosthodontic questions were classified under 7 main headings: dental morphology, complete dentures, removable partial dentures, fixed partial prostheses, materials science, implant-supported prostheses, temporomandibular joint (TMJ) disorders, and occlusion.<sup>22</sup> Oral and dentomaxillofacial radiology questions were classified under 2 main headings: (1) oral medicine and oral diagnosis, (2) oral radiology.<sup>23</sup>

The DUS questions, with the original text in Turkish, were entered individually into ChatGPT and Bard as prompts, and both were asked to respond from July 24 to 29, 2023. The prompt "Can you give the correct answer to the multiple-choice question above and categorize this multiple-choice question according to Bloom's taxonomy?" was entered also in Turkish, just after the DUS question, as can be seen in Figure 1. The "regenerate response" button was not used, and the first response was taken to be the final answer. The official answers provided by the ÖSYM were used to establish the proper answer to each question. Answers were categorized as correct (1) or incorrect (0) in a binary format. Questions that an LLM prefers not to answer were classified as incorrect.

## Statistical Analysis

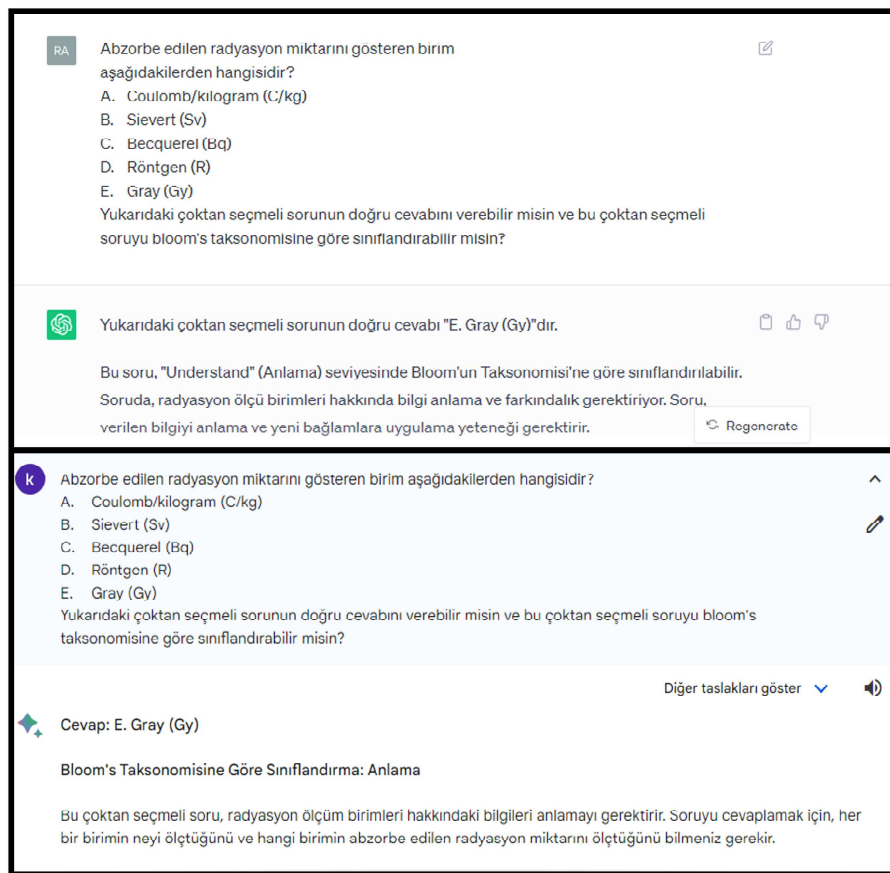
After data collection, IBM Statistical Package for the Social Sciences Statistics, version 22.0 software (IBM Corp.; Armonk, NY, USA) program was used for statistical analyses to evaluate the null hypotheses according to the findings in the study. In addition to descriptive statistical methods (frequency), the McNemar test, Chi-square test, Fisher-Freeman-Halton exact test, and Yate's continuity correction were used to compare qualitative data, and the Kappa agreement level was calculated. Significance was evaluated at  $P < .05$  level.

## RESULTS

According to the McNemar test results, as seen in Table 1, the agreement between the authors', ChatGPT's, and Bard's lower/higher-order classifications is statistically not significant ( $P: .001$ ;  $P < .05$ ). The Kappa agreement levels between the authors' and ChatGPT's and Bard's classifications were 66.1% and 35%, respectively, while the Kappa agreement level between ChatGPT and Bard's classifications was 21.4%. The examples of agreement and disagreement in the classification can be seen in Figure 1 and 2.

When the correct/incorrect answer rates according to the subheadings are evaluated (Table 2), it is found that there is no statistically significant difference between the correct answer rates of LLMs among the subheadings of prosthodontics ( $P > .05$ ). When all prosthodontic questions are taken into consideration, the correct answer rate of ChatGPT is 35.7%, while the correct answer rate of Bard is 38.9%. Between the oral and dentomaxillofacial radiology subheadings, it was found that there is statistically no significant difference in terms of Bard correct answer rates ( $P > .05$ ), but there is a significant difference in terms of ChatGPT correct answer rates ( $P: .042$ ;  $P < .05$ ). The correct answer rate of ChatGPT for oral medicine and oral diagnosis questions (64.7%) is significantly higher than for oral radiology questions (44.4%). In





**Figure 1.** ChatGPT (upper screenshot) and Bard (lower screenshot) choose the correct answer (E.Gray) and define the same classification with the authors to the same question which is translated as, "Which of the following is the unit indicating the amount of absorbed radiation? A. Coulomb/kilogram (C/kg), B. Sievert (Sv), C. Becquerel (Bq), D. x-ray (R), E. Gray (Gy).

addition, for total oral and dentomaxillofacial radiology questions, the correct answer rates of both LLMs are 52.8%. The examples of correct and incorrect responses by both LLMs can be seen in Figure 1 and 2.

Upon evaluation of LLM's agreement on correct/incorrect answers to prosthodontics (Table 3) and oral and dentomaxillofacial radiology (Table 4) questions, it has been discovered that there is a statistically significant agreement on correct answer rates between the LLMs in terms of subheadings of both prosthodontics and oral and dentomaxillofacial radiology and in terms of total questions of both clinical sciences ( $P > .05$ ). The inter-rater

agreement, as indicated by the Kappa coefficient, between the 2 LLMs varies across different clinical sciences and their subheadings. For prosthodontics, the agreement level is 32.2% for all questions with the same answer rate of 86/126. In the realm of oral and dentomaxillofacial radiology, it stands at 28.2% for all questions with 79/123 having the same answer rate. Delving into specific subheadings, the agreement percentages are as follows: 52.6% for dental morphology, 45.1% for complete dentures, 14.3% for removable partial dentures, 15.4% for fixed partial prostheses, 20.8% for materials science, and 53.2% for TMJ disorders and occlusion. Meanwhile, for oral medicine and oral diagnosis, the agreement level is 29.7%, and for oral radiology, it is 24.8%.

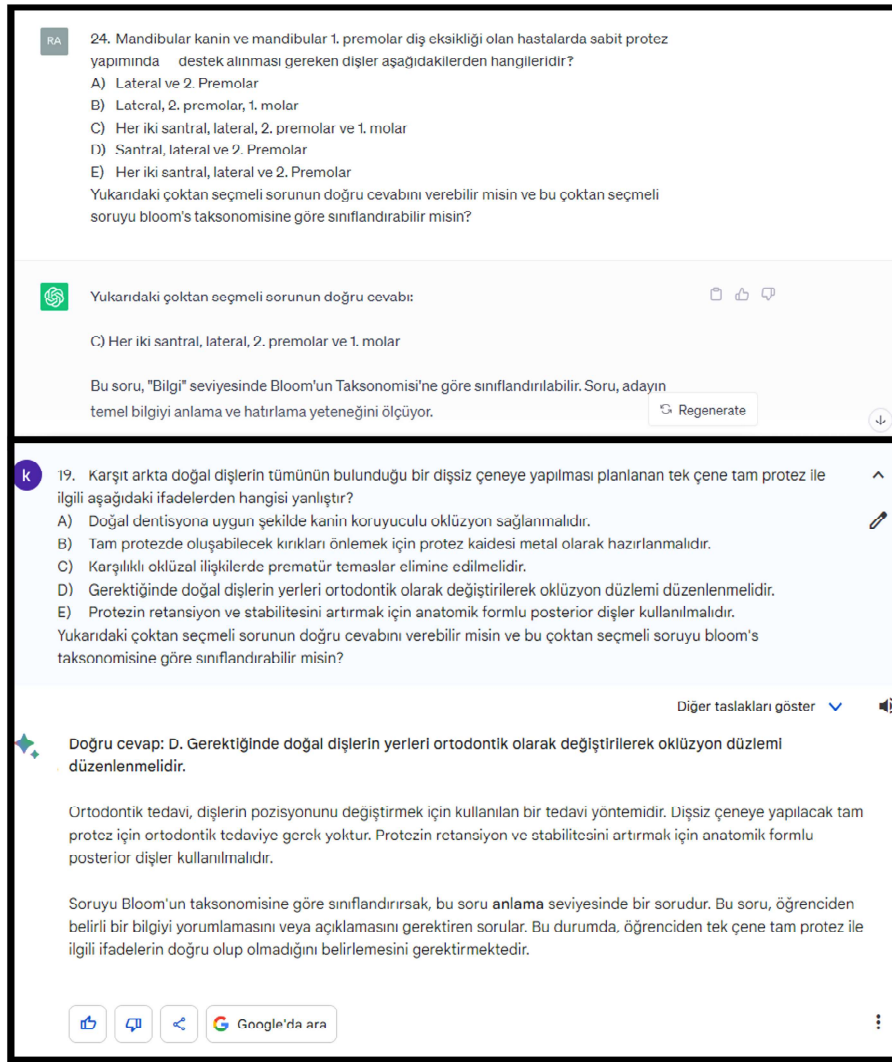
**Table 1.** Agreement between the authors', ChatGPT's, and Bard's lower- and higher-order classifications (N = 249)

		Authors' Classification			P
		Lower n (%)	Higher n (%)	Total n (%)	
Chat GPT's classification	Lower	200 (80.3)	19 (7.6)	219 (88)	<.001*
	Higher	3 (1.2)	27 (10.8)	30 (12)	
	Total	203 (81.5)	46 (18.5)	249 (100)	
Bard's classification	Lower	142 (57.0)	9 (3.6)	151 (60.6)	<.001*
	Higher	61 (24.5)	37 (14.9)	98 (39.4)	
	Total	203 (81.5)	46 (18.5)	249 (100)	
Bard's classification		ChatGPT's Classification			P
		Lower	Higher	Total	
		Lower	Higher	Total	
		Higher	Higher	Total	
Total	Total	Total	Total	Total	

ChatGPT, Chat generative pretrained transformer.

\* $P < .05$ —McNemar test.





**Figure 2.** The upper screenshot presents ChatGPT choosing the incorrect answer (C. Both central, lateral, second premolar, and first molar) for the question "Which of the following are the teeth that should be prepared to support a fixed prosthesis in patients with mandibular canine and mandibular first premolar tooth loss?" (The correct answer is option D. Central, lateral, and second premolar) and defining the different classifications of Bloom's taxonomy (knowledge/remembers classified as lower-order) compared to the authors' (applications classified as higher-order). The lower screenshot presents a question translated as "Which of the following statements about a single jaw complete denture planned for an edentulous jaw with all-natural teeth in the opposite arch is incorrect? A. Canine-protected occlusion should be provided in accordance with the natural dentition. B. The denture base should be prepared as metal to prevent fractures that may occur in the complete denture. C. Premature contacts in interocclusal relationships should be eliminated. D. When necessary, the plane of occlusion should be adjusted by orthodontically replacing the natural teeth. E. Anatomically shaped posterior teeth should be used to increase the retention and stability of the prosthesis. This lower screenshot presents Bard choosing the incorrect answer (option D) whereas the correct answer is option A Bard further explains incorrectly that, "Orthodontic treatment is a treatment method used to change the position of the teeth. Orthodontic treatment is not required for a complete denture for an edentulous jaw. Anatomically shaped posterior teeth should be used to increase the retention and stability of the prosthesis."and defines the classification of Bloom's taxonomy differently (comprehension classified as lower-order) compared to the authors' (application classified as higher-order).

Regarding the effect of Bloom's taxonomy level, as specified by the authors, on the correct response rate of LLMs, as Bloom's taxonomy level increased, the LLM's performance decreased, but the difference was not statistically significant (Table 5). For lower-order questions, the correct response rates of ChatGPT for prosthodontics and oral and dentomaxillofacial radiology questions are 37.5% and 57.6%, while Bard's are 43.3% and 53.5%, respectively. For higher-order questions, ChatGPT's correct response rates for prosthodontics and oral and dentomaxillofacial radiology questions are 27.3% and 33.3%, whereas Bard's are 18.2% and 50%, respectively. In addition, regarding the LLMs' agreement on correct/incorrect answers to all questions according to the authors'

lower or higher-order classification (Table 6), it is found that, in both the categorization of prosthodontics and oral and dentomaxillofacial radiology questions as lower or higher-order, a statistically significant agreement in terms of correct response rates is observed between ChatGPT and Bard ( $P > .05$ ).

## DISCUSSION

In this study, the performance of ChatGPT and Bard on DUS questions, specifically prosthodontics and oral and dentomaxillofacial radiology questions, was evaluated. According to the results, (1) ChatGPT outperformed Bard on the recognition of Bloom's taxonomy level of MCQs, but the agreement between both LLMs

Table 2. Evaluation of Correct/Incorrect Answer Rates of Chat Generative Pretrained Transformer and Bard According to the Subheadings

		Dental Morphology		Complete Dentures		Removable Partial Dentures		Fixed Partial Prosthesis		Materials Science		Implant-Supported Prosthesis		TMJ Disorders and Occlusion		Total			
		n (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)		
ChatGPT	Incorrect	7 (77.8)	11 (57.9)	21 (77.8)	15 (55.6)	15 (68.2)	1 (25)	11 (61.1)	81 (64.3)									.333	
	Correct	2 (22.2)	8 (42.1)	6 (22.2)	12 (44.4)	7 (31.8)	3 (75)	7 (38.9)	45 (35.7)										
	Incorrect	5 (55.6)	12 (63.2)	21 (77.8)	18 (66.7)	10 (45.5)	0 (0)	11 (61.1)	77 (61.1)										.059
	Correct	4 (44.4)	7 (36.8)	6 (22.2)	9 (33.3)	12 (54.5)	4 (100)	7 (38.9)	49 (38.9)										
ChatGPT	Incorrect		Oral Medicine and Oral Diagnosis															.042*	
	Correct		Oral Radiology																
Bard	Incorrect		18 (35.3)															.264	
	Correct		33 (64.7)																
Bard	Incorrect		21 (41.2)															.264	
	Correct		30 (58.8)																

ChatGPT: chat generative pretrained transformer; TMJ, temporomandibular joint.

\*Fisher-Freeman-Halton exact test.

\*rate's continuity correction.

\*Chi-square test.

\*P < .05.

and authors was still not significant; (2) there was no statistically significant difference between the LLMs related to the correct response rate of MCQs on subheadings of prosthodontics and oral and dentomaxillofacial radiology; (3) there is a statistically significant agreement on correct answer rates between the LLMs in terms of subheadings of both prosthodontics and oral and dentomaxillofacial radiology and in terms of total questions of both clinical sciences; (4) Bloom's taxonomy level, specified by the authors, did not have a significant effect on performance of both LLMs, and the LLMs had a significant agreement on correct/incorrect answers to all questions according to the Bloom's taxonomy level specified by the authors. Therefore, the null hypotheses are all rejected.

Large language models can be used in education not only in answering questions but also in preparing questions. However, in order to be able to prepare questions that force students to use their critical thinking skills for an effective evaluation, it is important to prepare questions according to Bloom's taxonomy. Therefore, it is crucial to determine how effective LLMs are at recognizing Bloom's taxonomy in order to benefit from these LLMs when preparing questions.<sup>8</sup> However, according to the results of this study, both LLMs do not have a statistically significant agreement with the authors on specifying Bloom's taxonomy level as lower or higher order. An example of this disagreement is that when Bard was asked, "Which of the following impression materials should not be disinfected with chlorine solutions?" it classified this question in application level and thereby categorized it as higher-order, but the authors classified it in remember level as it was a simple knowledge question not requiring critical thinking and thereby categorized it as lower-order. As far as the authors searched, there is no similar study on LLMs' recognition of Bloom's taxonomy in the literature, but it was reported that there is no gold standard for making a lower or higher-order classification according to Bloom's taxonomy.<sup>20</sup> Also in another study, educators, faculty, and students were asked to classify questions as lower or higher-order according to Bloom's taxonomy, but the agreement between them was not significant, just like the results of the present study.<sup>19</sup>

Regarding the results on the correct response rate on questions related to subheadings of prosthodontics and oral and dentomaxillofacial radiology, the only significant difference was observed between ChatGPT's correct answer rates for oral and dentomaxillofacial radiology subheadings, which is significantly higher for oral diagnosis and oral medicine questions. This might be because ChatGPT was trained more, or users sought more medical advice and provided feedback on that subheading, leading to its continuous improvement.<sup>3,5,24</sup> For total prosthodontic questions, ChatGPT and Bard achieved correct answer rates of 35.7% and 38.9%, respectively, while both LLMs achieved a 52.8% correct answer rate for oral and dentomaxillofacial radiology. These rates are quite low, according to a recent study mentioning that for the LLM to become a reliable and widely acceptable educational tool, it should consistently provide 95% accuracy. In addition, other studies on the MCQ performance of ChatGPT in medical examinations also revealed that it was not successful enough to be used as an educational tool in gastroenterology, neurosurgery, urology, parasitology, and ophthalmology.<sup>3,8,9,11,25,26</sup> Furthermore, a recent report on the performance of ChatGPT and Bard on nephrology concluded that both LLMs had similar scores and were both insufficient.<sup>1</sup> Several factors could account for the relatively insufficient performance of these LLMs in certain

Table 3. Chat Generative Pretrained Transformer's and Bard's Agreement on Correct/Incorrect Answers to Prosthodontics Questions

Subheadings	Bard**	ChatGPT*			P
		Incorrect n (%)	Correct n (%)	Total n (%)	
Dental morphology	Incorrect	5 (55.6)	0 (0)	5 (55.6)	.500
	Correct	2 (22.2)	2 (22.2)	4 (44.4)	
	Total	7 (77.8)	2 (22.2)	9 (100)	
Complete dentures	Incorrect	9 (47.4)	3 (15.8)	12 (63.2)	1.000
	Correct	2 (10.5)	5 (26.3)	7 (36.8)	
	Total	11 (57.9)	8 (42.1)	19 (100)	
Removable partial dentures	Incorrect	17 (63)	4 (14.8)	21 (77.8)	1.000
	Correct	4 (14.8)	2 (7.4)	6 (22.2)	
	Total	21 (77.8)	6 (22.2)	27 (100)	
Fixed partial prostheses	Incorrect	11 (40.7)	7 (25.9)	18 (66.7)	.549
	Correct	4 (14.8)	5 (18.5)	9 (33.3)	
	Total	15 (55.6)	12 (44.4)	27 (100)	
Materials science	Incorrect	8 (36.4)	2 (9.1)	10 (45.5)	.180
	Correct	7 (31.8)	5 (22.7)	12 (54.5)	
	Total	15 (68.2)	7 (31.8)	22 (100)	
Implant-supported prostheses	Incorrect	0 (0)	0 (0)	0 (0)	.250
	Correct	1 (25)	3 (75)	4 (100)	
	Total	1 (25)	3 (75)	4 (100)	
TMJ disorders and occlusion	Incorrect	9 (50)	2 (11.1)	11 (61.1)	1.000
	Correct	2 (11.1)	5 (27.8)	7 (38.9)	
	Total	11 (61.1)	7 (38.9)	18 (100)	
All questions	Incorrect	59 (46.8)	18 (14.3)	77 (61.1)	.636
	Correct	22 (17.5)	27 (21.4)	49 (38.9)	
	Total	81 (64.3)	45 (35.7)	126 (100)	

McNemar test.

ChatGPT, chat generative pretrained transformer; TMJ, temporomandibular joint.

\*ChatGPT displays the results of ChatGPT in columns.

\*\*Bard displays the results of Bard in rows.

medical and dental specialties. Primarily, these LLMs were initially developed as general-purpose interactive platforms and weren't specifically tailored to grasp medical literature nuances. Consequently, they lack the medical expertise and contextual comprehension necessary to navigate the intricate interplay between various medical conditions and treatments. Another significant consideration pertains to the training data. Most of the information integrated into these LLMs was derived from publicly accessible sources, potentially limiting access to information requiring paid journal subscriptions. This could be a limitation when addressing specific types of queries. Moreover, the LLMs may draw information from diverse sources, including non-medical ones, and may even retrieve data from outdated references, which may lead to erroneous responses. In addition, the core function of these LLMs revolves around predicting the subsequent words in a text sequence and constructing responses based on available data without assessing their accuracy. However, they lack inherent comprehension of the subjects, merely generating responses based on patterns, which might yield

plausible yet factually incorrect or nonsensical answers, and this phenomenon is called "hallucination."<sup>27</sup> An example of hallucination can be seen in Figure 2 with both LLMs answering incorrectly and Bard further defending its reason of choice with confidence as if it's a fact. Such hallucinations can also be encountered in ChatGPT since there is a significant agreement between ChatGPT and Bard on correct and incorrect answers, which might be because they were equipped with a similar database. Consequently, cross-checking a question's answer between these LLMs does not increase the chance of getting the correct response. Therefore, both LLMs need to be developed to be used as reliable educational tools, as indicated by the studies on MCQ exams that reported ChatGPT achieving a passing score.<sup>3,10,12,13,28</sup>

According to the results of this study, the correct response rate of LLMs decreases as Bloom's taxonomy level increases, though not significantly. Previous studies have also reported that ChatGPT exhibits diminished precision when addressing higher-order inquiries, indicating that even if it possesses knowledge, it cannot apply it critically.<sup>1,6,9</sup> In addition, there was a significant agreement between the LLMs on correct/incorrect answers to all questions according to Bloom's taxonomy level specified by the authors. This might be because these LLMs possess constraints in their capacity to accommodate specific question types or structures, as well as tasks related to constructing arguments and reasoning.<sup>7</sup>

Despite the constraints of LLMs, they are here to stay, and their potential influence on the medical and dental fields is enormous and cannot be ignored.<sup>3,12,26</sup> Therefore, they should be continuously evaluated in terms of advantages and challenges. Consequently, in this study, the performance of ChatGPT and Bard on the recognition of Bloom's taxonomy level and the correct response rate on questions related to subheadings of prosthodontics and oral and dentomaxillofacial radiology, the agreement of LLMs' on correct and incorrect answers, the effect of Bloom's taxonomy level on the

Table 4. Chat Generative Pretrained Transformer's and Bard's Agreement on Correct/Incorrect Answers to Oral and Dentomaxillofacial Questions

Questions	Bard**	ChatGPT*			P
		Incorrect n (%)	Correct n (%)	Total n (%)	
Oral medicine, oral diagnosis	Incorrect	11 (21.6)	10 (19.6)	21 (41.2)	.629
	Correct	7 (13.7)	23 (45.1)	30 (58.8)	
	Total	18 (35.3)	33 (64.7)	51 (100)	
Oral radiology	Incorrect	25 (34.7)	12 (16.7)	37 (51.4)	.701
	Correct	15 (20.8)	20 (27.8)	35 (48.6)	
	Total	40 (55.6)	32 (44.4)	72 (100)	
All questions	Incorrect	36 (29.3)	22 (17.9)	58 (47.2)	1.000
	Correct	22 (17.9)	43 (35.0)	65 (52.8)	
	Total	58 (47.2)	65 (52.8)	123 (100)	

McNemar test.

ChatGPT, chat generative pretrained transformer.

\*ChatGPT displays the results of ChatGPT in columns.

\*\*Bard displays the results of Bard in rows.



Table 5. The Effect of Lower/Higher-Order Classification, Specified by the Authors, on the Correct Response Rate of Chat Generative Pretrained Transformer and Bard

			Authors' Classification			P
			Lower n (%)	Higher n (%)	Total n (%)	
ChatGPT	Prosthodontics	Incorrect	65 (62.5)	16 (72.7)	81 (64.3)	.506
		Correct	39 (37.5)	6 (27.3)	45 (35.7)	
	Oral and dentomaxillofacial radiology	Incorrect	42 (42.4)	16 (66.7)	58 (47.2)	.057
		Correct	57 (57.6)	8 (33.3)	65 (52.8)	
	Total	Incorrect	107 (52.7)	32 (69.6)	139 (55.8)	.056
		Correct	96 (47.3)	14 (30.4)	110 (44.2)	
Bard	Prosthodontics	Incorrect	59 (56.7)	18 (81.8)	77 (61.1)	.051
		Correct	45 (43.3)	4 (18.2)	49 (38.9)	
	Oral and dentomaxillofacial radiology	Incorrect	46 (46.5)	12 (50)	58 (47.2)	.934
		Correct	53 (53.5)	12 (50)	65 (52.8)	
	Total	Incorrect	105 (51.7)	30 (65.2)	135 (54.2)	.135
		Correct	98 (48.3)	16 (34.8)	114 (45.8)	

Yate's Continuity Correction

ChatGPT, chat generative pretrained transformer.

Table 6. Chat Generative Pretrained Transformer and Bard's Agreement on Correct/Incorrect Answers to All Questions According to the Authors' Lower/Higher-Order Classification

		Bard**	ChatGPT*			P
			Incorrect n (%)	Correct n (%)	Total n (%)	
Prosthodontics	Lower	Incorrect	44 (42.3)	15 (14.4)	59 (56.7)	.405
		Correct	21 (20.2)	24 (23.1)	45 (43.3)	
		Total	65 (62.5)	39 (37.5)	104 (100)	
	Higher	Incorrect	15 (68.2)	3 (13.6)	18 (81.8)	.625
		Correct	1 (4.5)	3 (13.6)	4 (18.2)	
		Total	16 (72.7)	6 (27.3)	22 (100)	
Oral and dentomaxillofacial radiology	Lower	Incorrect	28 (28.3)	18 (18.2)	46 (46.5)	.597
		Correct	14 (14.1)	39 (39.4)	53 (53.5)	
		Total	42 (42.4)	57 (57.6)	99 (100)	
	Higher	Incorrect	8 (33.3)	4 (16.7)	12 (50)	.388
		Correct	8 (33.3)	4 (16.7)	12 (50)	
		Total	16 (66.7)	8 (33.3)	24 (100)	

McNemar test.

ChatGPT, chat generative pretrained transformer.

\*ChatGPT displays the results of ChatGPT in columns.

\*\*Bard displays the results of Bard in rows.

correct response rate of LLMs, and the LLMs' agreement on correct/incorrect answers to all questions according to Bloom's taxonomy level specified by the authors were investigated to assess their reliability in dental education. There is no similar study in the literature comparing ChatGPT's and Bard's performance on such parameters and revealing their strengths and weaknesses on MCQs about dental specialties, which is important, especially since the coronavirus 2019 pandemic because online exams became popular and MCQs are favored types of assessments due to their advantages. Regarding the results of this study, ChatGPT and Bard do not currently provide a sufficient correct response to allow substantial unfair advantage to students taking tests, and they are not reliable enough to be used as an educational tool. This inference, nevertheless, will inevitably change as LLMs undergo ongoing evolution. The advancement of these LLMs, propelled by refined training data and progressively intricate algorithms, foreshadows the emergence of more precise LLMs adept at producing contextually fitting answers. This progression, consequently, introduces new ethical predicaments about their implementation within educational contexts. Despite these possible disadvantages, the main point is to integrate LLMs as they advance into a broader learning strategy and supplement conventional educational resources such as textbooks and lectures.<sup>12</sup>

There are some limitations to this study. First, the result of this study belongs to the data from July 24 to 29, 2023, and as LLMs will likely continue to evolve rapidly, a future trial with the same items may yield different results. However, to pave the way for this rapid development by manufacturers, the shortcomings of

LLMs for each subheading need to be clarified. Second, since the authors specialized in prosthodontics and oral dentomaxillofacial radiology, only questions of these specialties are evaluated, and these results cannot be generalized directly to other dental specialties but may set an example and lead the way to further studies. Third, the questions were asked only in Turkish as they were on the OSYM database because Turkish students would probably prefer to use Turkish, and if translated, there could be deviations from the original text. However, asking them in English could have increased the correct answer rate.<sup>29</sup> Fourth, since ÖSYM did not share statistical data on the correct answer rate of the test takers for each subheading, the results of LLMs could not be compared to human performance. Fifth, the questions with figures and tables were excluded since ChatGPT could not be integrated with multimodal input. In further studies, as these LLMs evolve rapidly, different kinds of questions with such input, different kinds of assessments like open-ended questions, other dental specialties, and various LLMs should be tested to be able to confidently integrate LLMs into dental education.

The agreement between both LLMs and authors on the recognition of Bloom's taxonomy level of MCQs was not significant. There was no statistically significant difference between ChatGPT and Bard, related to the correct response rate of MCQs on prosthodontics (35.7% and 38.9%, respectively) and oral and dentomaxillofacial radiology (52.8% for both LLMs), and there was a significant agreement between ChatGPT and Bard on correct and incorrect answers. As Bloom's taxonomy level increased, the correct response rate of LLMs decreased, though not significantly.

As a result, these LLMs are not yet reliable educational tools but can be used as a supplement to traditional educational methods as they evolve. They can even be integrated into the dental education curriculum after sufficient development, but dental academicians should be aware of their strengths and weaknesses to prepare assessments accordingly and to guide their students in self-learning strategies.

**Ethics Committee Approval:** This study does not require ethics committee approval since it does not involve any living subjects or personal data.

**Informed Consent:** This study does not require informed consent as it does not involve human participation.

**Peer-review:** Externally peer-reviewed.

**Author Contributions:** Conception – R.T.O., Z.Z.Y.; Design – R.T.O.; Supervision – R.T.O., Z.Z.Y.; Data Collection and/or Processing – R.T.O., Z.Z.Y.; Analysis and/or Interpretation – R.T.O., Z.Z.Y.; Literature Search – R.T.O.; Writing Manuscript – R.T.O., Z.Z.Y.; Critical Review – R.T.O., Z.Z.Y.

**Acknowledgements:** The authors used ChatGPT for language editing while writing the manuscript. In addition, the authors wish to express their sincere thanks to statistician Ebru Osmanoğlu for helping with the statistical analyses.

**Declaration of Interests:** The authors declare no potential conflict of interest.

**Funding:** The authors declared that this study has received no financial support.

**Etik Komite Onayı:** Bu çalışma herhangi bir canlı veya kişisel veri içermediği için etik komite onayı gerektirmemektedir.

**Hasta Onamı:** Bu çalışma insan katılımı içermediği için bilgilendirilmiş onam gerektirmemektedir.

**Hakem Değerlendirmesi:** Dış bağımsız.

**Yazar Katkıları:** Fikir – R.T.O., Z.Z.Y.; Tasarım – R.T.O.; Denetleme – R.T.O., Z.Z.Y.; Veri Toplanması ve/veya İşlemesi – R.T.O., Z.Z.Y.; Analiz ve/veya Yorumlama – R.T.O., Z.Z.Y.; Literatür Taraması – R.T.O.; Makale Yazımı – R.T.O., Z.Z.Y.; Eleştirel İnceleme – R.T.O., Z.Z.Y.

**Teşekkür:** Yazarlar, metnin yazımında lisan düzenlemeleri için ChatGPT'den faydalanmışlardır. Ayrıca istatistiksel analizlere katkılarından dolayı istatistik uzmanı Ebru Osmanoğlu'na içten teşekkürlerini sunmaktadırlar.

**Çıkar Çatışması:** Yazarlar çıkar çatışması bildirmemişlerdir.

**Finansal Destek:** Yazarlar bu çalışma için herhangi bir finansal destek almadığını beyan etmiştir.

## REFERENCES

- Noda R, Izaki Y, Kitano F, Komatsu J, Ichikawa D, Shigabaki Y. Performance of ChatGPT and Bard in Self-Assessment Questions for Nephrology Board Renewal. *medRxiv* [preprint]. 2023. [CrossRef]
- 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. [CrossRef]
- Suchman K, Garg S, Trindade AJ. Chat Generative Pretrained Transformer fails the multiple-choice American College of Gastroenterology self-assessment test. *Am J Gastroenterol*. 2023;118(12):2280-2282. [CrossRef]
- Fatani B. ChatGPT for future medical and dental research. *Cureus*. 2023;15(4):e37285. [CrossRef]
- Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. *Neurosurgery*. 2023;93(6):1353-1365. [CrossRef]
- Lum ZC. Can artificial intelligence pass the American Board of Orthopaedic Surgery examination? Orthopaedic residents versus ChatGPT. *Clin Orthop Relat Res*. 2023;481(8):1623-1630. [CrossRef]
- Giannos P, Delardas O. Performance of ChatGPT on UK standardized admission tests: insights from the BMAT, TMUA, LNAT, and TSA examinations. *JMIR Med Educ*. 2023;9(9):e47737. [CrossRef]
- Deebel NA, Terlecki R. ChatGPT performance on the American Urological Association self-assessment study program and the potential influence of artificial intelligence in urologic training. *Urology*. 2023;177:29-33. [CrossRef]
- Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. *Neurosurgery*. 2023. [CrossRef]
- Skalidis I, Cagnina A, Luangphiphat W, et al. ChatGPT takes on the European Exam in core cardiology: an artificial intelligence success story? *Eur Heart J Digit Health*. 2023;4(3):279-281. [CrossRef]
- Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination? a descriptive study. *J Educ Eval Health Prof*. 2023;20:1. [CrossRef]
- Das D, Kumar N, Longjam LA, et al. Assessing the capability of ChatGPT in answering first- and second-order knowledge questions on microbiology as per competency-based medical education curriculum. *Cureus*. 2023;15(3):e36034. [CrossRef]
- Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*. 2023;9:e45312. [CrossRef]
- Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit Health*. 2023;2(2):e0000198. [CrossRef]
- Gonsalves C. On ChatGPT: what promise remains for multiple choice assessment? *JLDHE*. [CrossRef]
- Zaidi NLB, Grob KL, Monrad SM, et al. Pushing critical thinking skills with multiple-choice questions: does Bloom's taxonomy work? *Acad Med*. 2018;93(6):856-859. [CrossRef]
- Grainger R, Dai W, Osborne E, Kenwright D. Medical students create multiple-choice questions for learning in pathology education: A pilot study. *BMC Med Educ*. 2018;18(1):201. [CrossRef]
- Elsayed S. Towards Mitigating ChatGPT's Negative Impact on Education: Optimizing Question Design through Bloom's Taxonomy. *arXiv* [Preprint]. 2023. [CrossRef]
- Monrad SU, Bibler Zaidi NL, Grob KL, et al. What faculty write versus what students see? Perspectives on multiple-choice questions using Bloom's taxonomy. *Med Teach*. 2021;43(5):575-582. [CrossRef]
- Stringer JK, Santen SA, Lee E, et al. Examining Bloom's taxonomy in multiple choice questions: students' approach to questions. *Med Sci Educ*. 2021;31(4):1311-1317. [CrossRef]
- Çulhaoğlu AK, Kiliçarslan MA, Deniz KZ. Dış hekimliğinde uzmanlık sinavının farklı eğitim seviyelerdeki algı ve tercih durumlarının değerlendirilmesi. *Atatürk Üniversitesi Dış Hekimliği Fakültesi Dergisi*. 2021;31(3):1-1. [CrossRef]
- Ertan AA. *DUS Protetik Dış Tedavisi*. 4th ed. Ankara: TUSEM Tıbbi Yayıncılık; 2017.
- Ongole R, Praveen BN, eds. *Textbook of Oral Medicine, Oral Diagnosis, and Oral Radiology*. 2nd ed. New Delhi: Elsevier Health Sciences; 2013.
- Hoch CC, Wollenberg B, Lüers JC, et al. ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification preparation questions. *Eur Arch Otorhinolaryngol*. 2023;280(9):4271-4278. [CrossRef]
- Huynh LM, Bonebrake BT, Schultis K, Quach A, Deibert CM. New artificial intelligence ChatGPT performs poorly on the 2022 self-assessment study program for urology. *Urol Pract*. 2023;10(4):409-415. [CrossRef]

26. Ali MJ. ChatGPT and lacrimal drainage disorders: performance and scope of improvement. *Ophthal Plast Reconstr Surg*. 2023;39(3):221-225. [\[CrossRef\]](#)
27. 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. [\[CrossRef\]](#)
28. Liévin V, Hother CE, Winther O. Can Large Language Models Reason about Medical Questions? *arXiv [preprint]*. 2023. [\[CrossRef\]](#)
29. Wang YM, Shen HW, Chen TJ. Performance of ChatGPT on the pharmacist licensing examination in Taiwan. *J Chin Med Assoc*. 2023;86(7):653-658. [\[CrossRef\]](#)