Joint Diseases and Related Surgery

Jt Dis Relat Surg
2025;36(2):304-310

ORIGINAL ARTICLE

Assessing the performance of ChatGPT-4o on the Turkish Orthopedics and Traumatology Board Examination

Hilal Yağar, MD[1], Ender Gümüşoğlu, MD[2], Zeynel Mert Asfuroğlu, MD[2]

[1]Department of Orthopedics and Traumatology, Ömer Halisdemir University Faculty of Medicine, Niğde, Türkiye
[2]Department of Orthopedics and Traumatology, Division of Hand Surgery, Mersin University Faculty of Medicine, Mersin, Türkiye

ABSTRACT

Objectives: This study aims to assess the overall performance of ChatGPT version 4-omni (GPT-4o) on the Turkish Orthopedics and Traumatology Board Examination (TOTBE) using actual examinees as a reference point to evaluate and compare the performance of GPT-4o with that of human participants.

Materials and methods: In this study, GPT-4o was tested with multiple-choice questions that formed the first step of 14 TOTBEs conducted between 2010 and 2023. The assessment of image-based questions was conducted separately for all exams. The questions were classified based on the subspecialties for the five exams (2010-2014). The performance of GPT-4o was assessed and compared to those of actual examinees of the TOTBE.

Results: The mean total score of GPT-4o was 70.2±5.64 (range, 61 to 84), whereas that of actual examinees was 58±3.28 (range, 53.6 to 64.6). Considering accuracy rates, GPT-4o demonstrated 62% accuracy on image-based questions and 70% accuracy on text-based questions. It also demonstrated superior performance in the field of basic sciences, whereas actual examinees performed better in the specialty of reconstruction. Both GPT-4o and actual examinees exhibited the lowest scores in the subspecialty of lower extremity and foot.

Conclusion: Our study results showed that GPT-4o performed well on the TOTBE, particularly in basic sciences. While it demonstrated accuracy comparable to actual examinees in some areas, these findings highlight its potential as a helpful tool in medical education.

Keywords: Board exam, ChatGPT, multiple-choice questions, orthopedics and traumatology.

Large language models (LLMs) employ computational artificial intelligence (AI) techniques to construct language that mimics human output.[1,2] These models are trained on extensive text data sourced from the internet to respond to questions.[3] The LLMs examine patterns and relationships within their training data to forecast the subsequent words or phrases likely to occur in a given scenario. Nevertheless, concerns have arisen regarding misinformation, privacy, biases in the training data, and risk of misuse.[4,5] Of note, LLMs can be utilized in various areas of medical science, including imaging analysis and diagnosis. These models have been also tested for facilitating communication between patients and physicians, converting medical records into text and enabling remote patient care.[6] In recent years, LLMs have become a crucial component of healthcare education and have been implemented by numerous medical institutions worldwide.[7,8]

Received: August 26, 2024
Accepted: January 07, 2025
Published online: April 05, 2025

Correspondence: Zeynel Mert Asfuroğlu, MD. Mersin Üniversitesi Tıp Fakültesi, Ortopedi ve Travmatoloji Anabilim Dalı, El Cerrahisi Bilim Dalı, 33110, Yenişehir, Mersin, Türkiye.

E-mail: z.mert.asfuroglu@gmail.com

Doi: 10.52312/jdrs.2025.1958

Citation: Yağar H, Gümüşoğlu E, Asfuroğlu ZM. Assessing the performance of ChatGPT-4o on the Turkish Orthopedics and Traumatology Board Examination. Jt Dis Relat Surg 2025;36(2):304-310. doi: 10.52312/jdrs.2025.1958.

©2025 All right reserved by the Turkish Joint Diseases Foundation

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes (http://creativecommons.org/licenses/by-nc/4.0/).

Chat Generative Pre-trained Transformer (ChatGPT; OpenAI Global LLC, San Francisco, CA, USA) is a LLM launched as a prototype on November 30th, 2022, and is currently one of the most popular AI models.[9] It can generate text responses which closely resemble human language. The latest model, ChatGPT 4-omni (GPT-4o), released on May 13th, 2024, exhibits superior capabilities in visual and auditory

comprehension compared to its predecessors.[10] Numerous studies have focused on the ability of various ChatGPT models to successfully pass demanding professional examinations in several medical fields, primarily through the analysis and comprehension of written questions.[11-14]

The purpose of board exams in medicine is to assess the ability to effectively use current knowledge, concepts, and principles to identify skilled physicians. The Turkish Orthopedics and Traumatology Education Council (TOTEC) has been conducting the Turkish Orthopedics and Traumatology Board Examination (TOTBE) annually since 2003 to achieve these goals.[15] The examination consists of two parts. The first is a written assessment, known as the Objective Structured Multiple-Choice Question, and the second is an oral assessment, referred to as the Objective Structured Clinical/Practical Examination. This examination is crucial for assessing the competence and proficiency of orthopedic surgeons in Türkiye.[15,16]

In the present study, we aimed to assess the overall performance of GPT-4o on the TOTBE using actual examinees as a reference point to evaluate and compare the performance of GPT-4o with that of human participants. Additionally, we aimed to assess the performance of GPT-4o across various subspecialties of orthopedics and traumatology.

MATERIALS AND METHODS

Study design

Prior to the study, permission to use TOTBE data were obtained from the Turkish Society of Orthopedics and Traumatology (2024/119). In addition, the study protocol was approved by the Niğde Ömer Halisdemir University Non-Interventional Clinical Research Ethics Committee (date: 02.08.2024, no: 2024/82). The study was conducted in accordance with the principles of the Declaration of Helsinki.

The first steps of 14 board examinations conducted between 2010 and 2023 were included in the study. The oral exam questions which comprised the second step in the examination were excluded from the study. Exams conducted prior to 2010 were also excluded due to inadequate data.

Questions in their original language (Turkish) were inputted into the interface of GPT-4o in a format consisting of a question and multiple-choice options. The interface of GPT-4o was accessed via the webpage, and the zero-shot prompting method was used. Each question had five possible options labeled A-E, with only one correct answer. All responses were recorded, and the total score of GPT-4o was calculated for each exam.

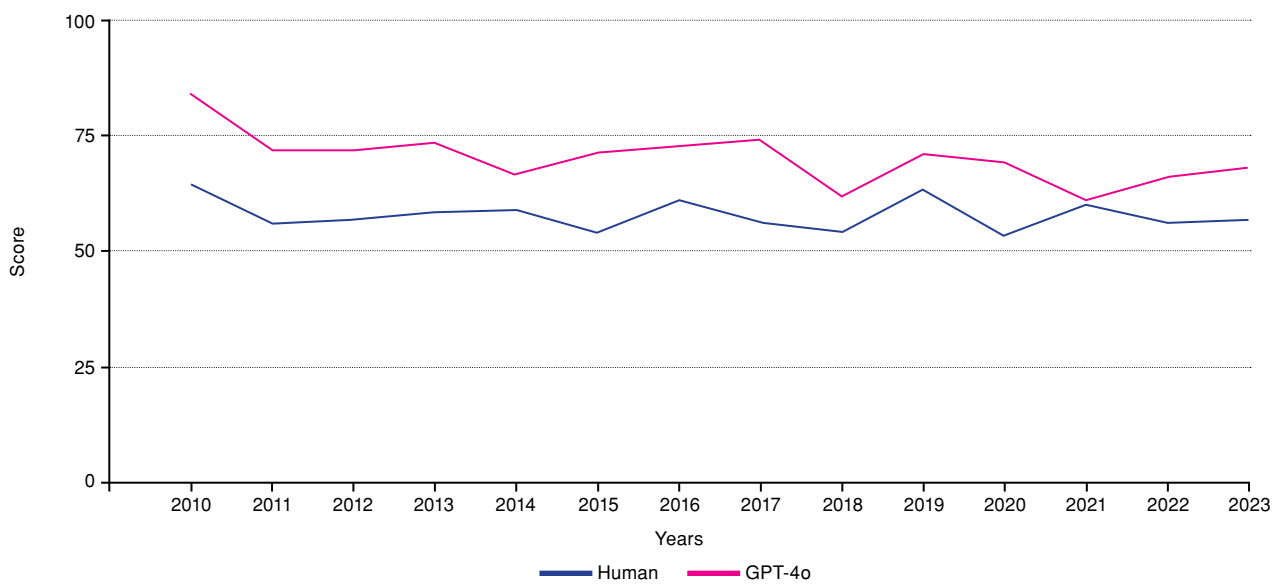

FIGURE 1. Relationship between GPT-4o and actual examinees' exam performance. Line graphs indicate the scores of GPT-4o (red line) and actual examinees (blue line).
GPT-4o: ChatGPT version 4-omni.

TABLE I
Performances of GPT-4o and actual examinees

Year	Total score of GPT-4o	The score of actual examinees	
		Mean	Min-Max
2010	84	64.62	40-82
2011	72	56.41	43-73
2012	72	57.01	37-71
2013	73	58.39	38-81
2014	67	58.79	42-79
2015	71	54.25	35-75
2016	73	60.68	39-86
2017	74	56.61	34-79
2018	62	54.65	35-69
2019	71	63.49	40-79
2020	69	53.58	23-78
2021	61	60.01	34-88
2022	66	56.53	29-80
2023	68	56.94	31-78

GPT-4o: ChatGPT version 4-omni.

General information about TOTBE

The first stage of the TOTBE follows a structured multiple-choice format, comprising 100 questions, each with a single correct answer. It includes questions covering all subspecialties of orthopedics and traumatology and is graded on a 100-point scale. Instead of a predetermined passing score, the Nedelsky method is used to determine the passing score for each examination.[17]

Collection of actual exam data

The actual exam data were obtained from the open-access term books of the TOTEC, which are published every two years.[18] Data unavailable in the term books were acquired by contacting the TOTEC. The questions were grouped into two categories based on their content: text-based (n=1327; 94.8%) and image-based (n=73; 5.2%). In addition, the questions were classified by subject (basic science, trauma, pediatrics, upper extremity/hand surgery, lower extremity, reconstruction, sports medicine, spine, and tumor/infection) for the five exams (2010-2014) where TOTEC categorized the exam questions and shared comprehensive results on their website. The performance of GPT-4o in these exams was assessed based on the percentage of correct responses in each subspecialty. The results of actual examinees were used as a reference point to assess and compare

TABLE II
Performances of GPT-4o and actual examinees according to subspecialty

Sub-specialties	Years										Total	
	2014		2013		2012		2011		2010			
	Percentage of correct response		Percentage of correct response		Percentage of correct response		Percentage of correct response		Percentage of correct response		Percentage of correct response	
	GPT-4o	Human	GPT-4o	Human	GPT-4o	Human	GPT-4o	Human	GPT-4o	Human	GPT-4o	Human
	%	%	%	%	%	%	%	%	%	%	%	%
Basic science	80	51	90	55	100	55	93	55	100	59	93	55
Trauma	47	53	73	62	67	58	74	64	77	63	68	60
Pediatrics	67	53	62	53	54	55	63	49	77	75	65	57
Upper extremity-hand	91	54	64	61	82	59	80	66	100	71	83	62
Lower extremity and foot	55	40	62	54	55	44	50	37	80	70	60	49
Reconstruction	83	67	83	78	50	66	57	51	78	69	70	66
Arthroscopy and sport medicine	83	76	50	43	67	56	57	51	70	65	65	58
Spine	67	56	83	70	100	40	72	67	100	58	84	58
Tumor-infection	71	43	71	47	100	51	67	59	80	65	78	53

GPT-4o: ChatGPT version 4-omni.

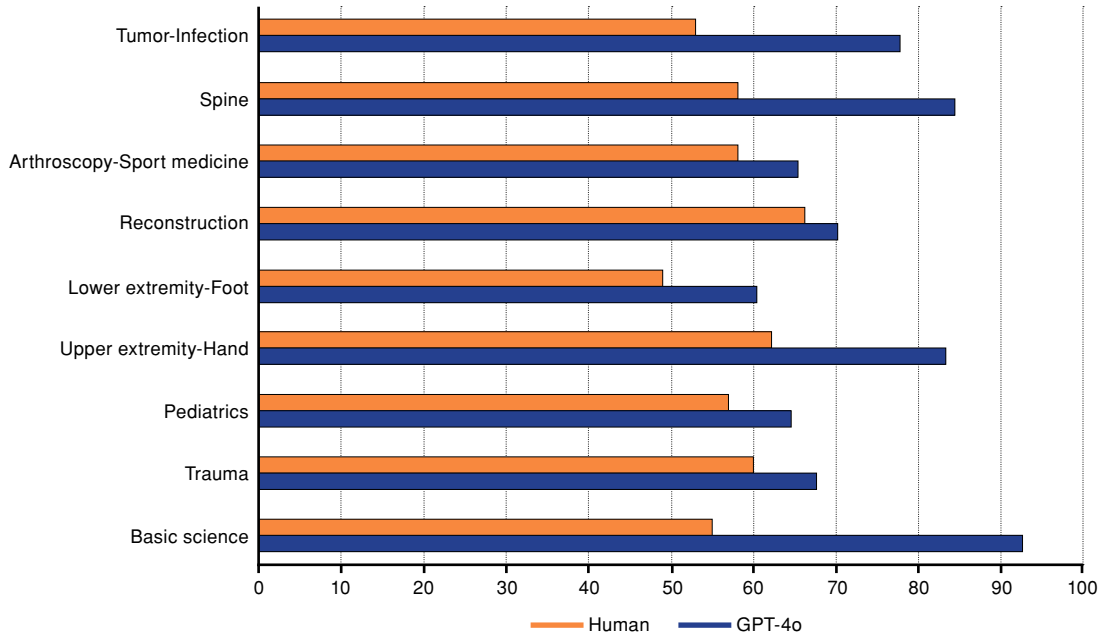

FIGURE 2. Correct response rates of GPT-4o and actual examinees according to subspecialty.
GPT-4o: ChatGPT version 4-omni.

the performance of GPT-4o with that of human participants.

Statistical analysis

Statistical analysis was performed using the IBM SPSS version 26.0 software (IBM Corp., Armonk, NY, USA). Descriptive data were expressed in mean ± standard deviation (SD), median (min-max) or number and frequency, where applicable. The Pearson chi-square test was used to compare the number of correct image-based and text-based questions. A p value of <0.05 was considered statistically significant.

RESULTS

The mean total score of GPT-4o was 70.2±5.64 (range, 61 to 84), whereas that of actual examinees was 58±3.28 (range, 53.6 to 64.6). The minimum and maximum scores of actual examinees across all examinations were 29 and 88, respectively (Figure 1 and Table I).

In addition, GPT-4o demonstrated an accuracy rate of 62% on image-based questions and 70% on text-based questions (p=0.433).

Table II presents the percentages of correct responses for each of the five examinations (2010-2014), categorized by subspecialty. Based on the total correct response calculated as the average across five years, GPT-4o demonstrated the highest performance in basic science, spine, and upper extremity-hand, respectively. The actual examinees performed superior results in the fields of reconstruction, upper extremity-hand, and trauma, respectively. Both GPT-4o and actual examinees demonstrated the lowest performance in the subspecialty of lower extremity and foot (Figure 2 and Table II).

DISCUSSION

The LLM chatbot technology has rapidly advanced, leading to its widespread utilization across multiple platforms in society.[19] The AI involves replicating human intelligence by teaching specialized programs and computers to mimic human cognitive capabilities. Machine learning, a subset of AI, uses algorithms implemented on computers to learn and improve performance by analyzing and processing new data.[20] Several recent studies have focused on the utilization of AI in orthopedic surgery. However, most have attempted to improve patient outcomes and reduce the workload on healthcare professionals.[21] In the future, LLM chatbots have the potential to be utilized in orthopedics and traumatology education. To achieve this, it is necessary to assess their theoretical knowledge. In the present study, we evaluated the suitability of ChatGPT for orthopedics and traumatology

education by assessing its knowledge competency through the TOTBE. Our study results showed that GPT-4o performed well on the TOTBE.

In recent years, multiple studies have assessed the performance of ChatGPT on various examinations related to orthopedics and traumatology. Massey et al.[22] found that orthopedic residents outperformed ChatGPT-3.5 and GPT-4 in accurately answering questions on orthopedic assessment exams. Kung et al.[23] assessed the performance of ChatGPT on an in-training evaluation and found that ChatGPT-3.5 and ChatGPT-4 achieved accuracies of 54.3% and 73.6%, respectively. Isleem et al.[12] demonstrated that ChatGPT provided a correct answer rate of 60.8% for the American Board of Orthopedics exam-style questions. It is evident that different versions of ChatGPT (GPT-3.5, GPT-4, and GPT-4o) exhibit varying levels of performance; thus, only the most recent version, GPT-4o, was utilized in our study. The mean total score of GPT-4o for TOTBE was 70.2±5.64, whereas the mean score of actual examinees was 58±3.28.

Previous versions of ChatGPT had limited capability to answer image-based questions due to the lack of support for visual content. Currently, GPT-4o, introduced by OpenAI, is the new flagship model capable of applying real-time logic over voice, image, and text.[10] As GPT-4o was used in our study, we were able to assess image-based questions as well. In their study, Ghanem et al.[24] found no significant differences in the accuracy of the responses of ChatGPT-4 to image-based and non-image-based questions on the American Hand Surgery Board Examination. Massey et al.[22] reported that ChatGPT demonstrated superior results for text-based questions about orthopedics and traumatology. In our study, GPT-4o demonstrated a 62% accuracy rate for image-based questions and a 70% accuracy rate for text-based questions.

In the current study, GPT-4o performed well across all subspecialties. It achieved a correct response rate exceeding 80% in three subspecialties: basic science (93%), spine (84%), and upper extremity/hand (83%). The lowest correct response rate for GPT-4o was in the lower extremity and foot subspecialty (60%). Among actual examinees, questions related to reconstruction had the highest correct response rate (66%), whereas those related to the lower extremity and foot (49%) had the lowest correct response rate. In Isleem et al.'s[12] study, the correct response rate of ChatGPT was high in basic sciences, the spine, shoulder/elbow, sports medicine, and oncology. In Lum et al.'s[25] study, ChatGPT performed well in basic sciences and sports medicine. Another study found that ChatGPT had a higher rate of correct responses to questions about hand surgery, sports medicine, and pediatrics.[22] Studies analyzing subspecialties of orthopedic in-training exams showed that questions related to basic sciences had the highest percentage of knowledge recall, while questions related to adult reconstruction were more complex and multistep.[26,27]

Although there is no conclusive method to identify the factors that result in superior performance in certain subspecialties, we believe that GPT-4o provides more accurate responses to questions which test theoretical knowledge objectively and do not require clinical interpretation. In their study, Atik et al.[28] emphasized that LLM-powered chatbots are not designed to replace the nuanced expertise and clinical judgment of experienced orthopedic surgeons, particularly in complex decision-making scenarios involving treatment indications. While our study was not designed to evaluate clinical scenarios, we believe that the high performance by GPT-4o on board exams does not imply proficiency in clinical decision-making.

Nevertheless, this study has several limitations. First, it exclusively assessed the performance of ChatGPT without undertaking any comparative assessments with other AI models. Second, TOTBE questions were inputted into GPT-4o in their original language (Turkish), and the responses were also received in Turkish. While the Turkish language support of GPT-4o functioned efficiently, it is uncertain whether this had any impact on the study's results. Third, we were unable to find any data regarding the responses provided by humans to the image-based questions. Consequently, a comparison of the success rates between GPT-4o and humans for image-based questions could not be conducted. In addition, it is of utmost importance to note that ChatGPT undergoes periodic updates, and the version used in our study may not necessarily represent the most up-to-date iteration at the time of publication. Similarly, literature information on orthopedics and traumatology is also constantly updated. In our study, it was not checked whether the questioned information and their response are still relevant according to the current literature. Despite these limitations, our study is the first to test the TOTBE with GPT-4o and 14 different exams including image-based questions were assessed individually. The large number of questions

assessed, the use of the results of actual examinees as reference point, and the categorization of questions by subspecialties distinguish our study from other studies in the literature.

In conclusion, GPT-4o performed well on the TOTBE and provided more accurate responses in subspecialties which do not require clinical interpretation, such as basic sciences. The rate of correct responses in all subspecialties was satisfactory. Based on these findings, we can speculate that GPT-4o possesses a high level of knowledge in the field of orthopedics and traumatology and GPT-4o can play a role in educating orthopedics and traumatology residents. By providing instant access to a vast amount of medical information, GPT-4o can help residents understand complex concepts, review surgical procedures, and stay updated with the latest research and clinical guidelines. Fortunately, orthopedic residents currently have the opportunity to utilize LLM-chatbots to help them answer test questions.

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

Author Contributions: Idea/concept, design, data collection and/or processing, analysis and/or interpretation: H.Y.; Idea/concept, design, data collection and/or processing, literature review: E.G.; Idea/concept, design, control/supervision, writing the article: Z.M.A.

Conflict of Interest: The authors declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding: The authors received no financial support for the research and/or authorship of this article.

REFERENCES

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29:1930-40. doi: 10.1038/s41591-023-02448-8.
2. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, et al. The future landscape of large language models in medicine. Commun Med (Lond) 2023;3:141. doi: 10.1038/s43856-023-00370-1.
3. Helm JM, Swiergosz AM, Haeberle HS, Karnuta JM, Schaffer JL, Krebs VE, et al. Machine learning and artificial intelligence: Definitions, applications, and future directions. Curr Rev Musculoskelet Med 2020;13:69-76. doi: 10.1007/s12178-020-09600-8.
4. Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. NPJ Digit Med 2021;4:93. doi: 10.1038/s41746-021-00464-x.
5. Civaner MM, Uncu Y, Bulut F, Chalil EG, Tatli A. Artificial intelligence in medical education: A cross-sectional needs assessment. BMC Med Educ 2022;22:772. doi: 10.1186/s12909-022-03852-3.
6. Atalar E, Üreten K, Kanatlı U, Çiçeklidağ M, Kaya İ, Vural A, et al. The diagnosis of femoroacetabular impingement can be made on pelvis radiographs using deep learning methods. Jt Dis Relat Surg 2023;34:298-304. doi: 10.52312/jdrs.2023.996.
7. Sapci AH, Sapci HA. Artificial intelligence education and tools for medical and health informatics students: Systematic review. JMIR Med Educ 2020;6:e19285. doi: 10.2196/19285.
8. Yapar D, Demir Avcı Y, Tokur Sonuvar E, Eğerci ÖF, Yapar A. ChatGPT's potential to support home care for patients in the early period after orthopedic interventions and enhance public health. Jt Dis Relat Surg 2024;35:169-76. doi: 10.52312/jdrs.2023.1402.
9. Open AI. Introducing ChatGPT. Available at: https://openai.com/blog/chatgpt/. [Accessed: 16.07.2024]
10. Open AI. Hello GPT-4o. Available at: https://openai.com/index/hello-gpt-4o/. [Accessed: 16.07.2024]
11. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2:e0000198. doi: 10.1371/journal.pdig.0000198.
12. Isleem UN, Zaidat B, Ren R, Geng EA, Burapachaisri A, Tang JE, et al. Can generative artificial intelligence pass the orthopaedic board examination? J Orthop 2023;53:27-33. doi: 10.1016/j.jor.2023.10.026.
13. Stengel FC, Stienen MN, Ivanov M, Gandía-González ML, Raffa G, Ganau M, et al. Can AI pass the written European board examination in neurological surgery? - Ethical and practical issues. Brain Spine 2024;4:102765. doi: 10.1016/j.bas.2024.102765.
14. Arango SD, Flynn JC, Zeitlin J, Lorenzana DJ, Miller AJ, Wilson MS, et al. The performance of ChatGPT on the American Society for Surgery of the Hand Self-Assessment Examination. Cureus 2024;16:e58950. doi: 10.7759/cureus.58950.
15. Gönen E, Berk H. TOTEK (Türk Ortopedi ve Travmatoloji Eğitim Konseyi) ve EBOT (European Board of Orthopaedics and Traumatology Fellowship) yeterlik sınavlarının karşılaştırılması. TOTBİD Dergisi 2014;13:488-99.
16. Benli İ, Acaroğlu E. Türk Ortopedi ve Travmatoloji Birliği Derneği (TOTBİD) Türk ortopedi ve travmatoloji eğitim konseyi yeterlik sınavları. Acta Orthopaedica et Traumatologica Turcica 2011;45.
17. Violato C, Marini A, Lee C. A validity study of expert judgment procedures for setting cutoff scores on high-stakes credentialing examinations using cluster analysis. Eval Health Prof 2003;26:59-72. doi: 10.1177/0163278702250082.
18. TOTEK. Dönem Kitapları. Available at: https://totek.totbid.org.tr/tr/donem-kitaplari/. [Accessed: 17.07.2024]
19. Chakraborty C, Pal S, Bhattacharya M, Dash S, Lee SS. Overview of Chatbots with special emphasis on artificial intelligence-enabled ChatGPT in medical science. Front Artif Intell 2023;6:1237704. doi: 10.3389/frai.2023.1237704.
20. Mavrogenis AF, Scarlat MM. Artificial intelligence publications: Synthetic data, patients, and papers. Int Orthop 2023;47:1395-6. doi: 10.1007/s00264-023-05830-w.
21. Wu L, Yang X, Wu J, Zhao X, Lu Z, Li P. Short-term outcomeof artificial intelligence-assisted preoperative three-dimensional planning of total hip arthroplasty for

developmental dysplasia of the hip compared to traditional surgery. Jt Dis Relat Surg 2023;34:571-82. doi: 10.52312/jdrs.2023.1076.

22. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT-3.5, ChatGPT-4, and orthopaedic resident performance on orthopaedic assessment examinations. J Am Acad Orthop Surg 2023;31:1173-9. doi: 10.5435/JAAOS-D-23-00396.
23. Kung JE, Marshall C, Gauthier C, Gonzalez TA, Jackson JB 3rd. Evaluating ChatGPT performance on the orthopaedic in-training examination. JB JS Open Access 2023;8:e23.00056. doi: 10.2106/JBJS.OA.23.00056.
24. Ghanem D, Nassar JE, El Bachour J, Hanna T. ChatGPT earns American Board Certification in hand surgery. Hand Surg Rehabil 2024;43:101688. doi: 10.1016/j.hansur.2024.101688.
25. Lum ZC, Collins DP, Dennison S, Guntupalli L, Choudhary S, Saiz AM, et al. Generative artificial intelligence performs at a second-year orthopedic resident level. Cureus 2024;16:e56104. doi: 10.7759/cureus.56104.
26. Shen TS, Driscoll DA, Ellsworth BK, Premkumar A, Lebrun DG, Bostrom MPG, et al. Analysis of the basic science questions on the orthopaedic ın-training examination from 2014 to 2019. J Am Acad Orthop Surg 2021;29:e1225-31. doi: 10.5435/JAAOS-D-20-00862.
27. Premkumar A, Lebrun DG, Shen TS, Ellsworth BK, Bostrom MPG, Cross MB. Analysis of hip and knee reconstruction questions on the orthopedic in-training examination. J Arthroplasty 2021;36:1156-9. doi: 10.1016/j.arth.2020.09.018.
28. Atik OŞ, Sezgin EA, Bahadır B. The main goal of ChatGPT is not to replace healthcare professionals Jt Dis Relat Surg 2024;35:471-2. doi: 10.52312/jdrs.2024.57923.