ACTA ORTHOPAEDICA et TRAUMATOLOGICA TURCICA

www.aott.org.tr

Research Article

Exploring the role of artificial intelligence in Turkish orthopedic progression exams

Gokhan Ayik[1], Ulas Can Kolac[2], Taha Aksoy[2], Abdurrahman Yilmaz[2], Mazlum Veysel Sili[2], Mazhar Tokgozoglu[2], Gazi Huri[2,3]

[1]*Department of Orthopedics and Traumatology, Yuksek Ihtisas University Faculty of Medicine, Ankara, Türkiye*
[2]*Department of Orthopedics and Traumatology, Hacettepe University Faculty Of Medicine, Ankara, Türkiye*
[3]*Aspetar, Orthopedic and Sports Medicine Hospital, FIFA Medical Center of Excellence, Doha, Qatar*

ARTICLE INFO

Article history:
Submitted Jun 12, 2024
Received in revised form July 29, 2024
Last revision received form October 20, 2024
Accepted January 3, 2025
Publication Date March 17, 2025

Keywords:
Artificial Intelligence
Medical Education
Orthopedics
Educational Technology
Clinical Decision-Making.

ORCID iDs of the authors:
G.A. 0000-0003-1454-7157;
U.C.K. 0000-0003-0502-3351;
T.A. 0000-0002-8900-9858;
A.Y. 000-0003-0060-6654;
M.V.S. 0009-0006-9237-2160;
M.T. 0000-0002-1375-8115;
G.H. 0000-0002-7036-8455.

Corresponding author:
Gökhan Ayık
drgayik@gmail.com

ABSTRACT

Objective: The aim of this study was to evaluate and compare the performance of the artificial intelligence (AI) models ChatGPT-3.5, ChatGPT-4, and Gemini on the Turkish Specialization Training and Development Examination (UEGS) to determine their utility in medical education and their potential to improve patient care.

Methods: This retrospective study analyzed responses of ChatGPT-3.5, ChatGPT-4, and Gemini to 1000 true or false questions from UEGS administered over 5 years (2018-2023). Questions, encompassing 9 orthopedic subspecialties, were categorized by 2 independent residents, with discrepancies resolved by a senior author. Artificial intelligence models were restarted for each query to prevent data retention. Performance was evaluated by calculating net scores and comparing them to orthopedic resident scores obtained from the Turkish Orthopedics and Traumatology Education Council (TOTEK) database. Statistical analyses included chi-squared tests, Bonferroni-adjusted Z tests, Cochran's Q test, and receiver operating characteristic (ROC) analysis to determine the optimal question length for AI accuracy. All AI responses were generated independently without retaining prior information.

Results: Significant differences in AI tool accuracy were observed across different years and subspecialties ($P < .001$). ChatGPT-4 consistently outperformed other models, achieving the highest overall accuracy (95% in specific subspecialties). Notably, ChatGPT-4 demonstrated superior performance in Basic and General Orthopedics and Foot and Ankle Surgery, while Gemini and ChatGPT-3.5 showed variability in accuracy across topics and years. Receiver operating characteristic analysis revealed a significant relationship between shorter letter counts and higher accuracy for ChatGPT-4 ($P=.002$). ChatGPT-4 showed significant negative correlations between letter count and accuracy across all years ($r=-0.099$, $P=.002$), outperformed residents in basic and general orthopedics ($P=.015$) and trauma ($P=.012$), unlike other AI models.

Conclusion: The findings underscore the advancing role of AI in the medical field, with ChatGPT-4 demonstrating significant potential as a tool for medical education and clinical decision-making. Continuous evaluation and refinement of AI technologies are essential to enhance their educational and clinical impact.

Introduction

The use of artificial intelligence (AI) has attracted substantial interest in healthcare.[1] Large language models (LLMs), which do not require deep learning techniques as the preceding AI technologies did,[2] have become popular in the field of medicine.[3] Current research suggests that LLMs have the potential to enhance various aspects of patient care. These areas of improvement include diagnosis, analysis of patient information and imaging studies, and the development of individualized treatment plans, as well as patient education.[4,5] Furthermore, LLMs have been increasingly used for teaching purposes in physician education, which is one of the most critical aspects in medicine.[6,7]

The evaluation of healthcare providers' medical proficiency traditionally relies on standardized examinations. The performance of LLMs for those standardized training and licensing examinations has been the subject of several studies.[8-10] It should not be regarded as a paradigm shift to rely solely on LLMs; rather, they should be considered significant tools that can improve education systems.

There are several generative AI chatbots that can be used for the comparison of medical knowledge between medical professionals. Their superiority over one another could be a key determinant for employing these different LLMs.[11-13] Moreover, assessment of chatbots with sample clinical scenarios could provide valuable data to guide decision-making processes. Regardless of the outcomes, these analyses would be beneficial in deciding on whether to include LLMs in clinical care or not.

The main aim of the present study was to evaluate and compare ChatGPT-3.5 (OpenAI, San Francisco, Calif, USA), ChatGPT-4 (OpenAI), and Gemini (Alphabet, Mountain View, Calif, USA) on the Specialization Training and Development Examination (abbreviated as UEGS in Turkish) held by the Turkish Orthopedics and Traumatology Education Council (TOTEK).

Cite this article as: *Ayik G, Can Kolac U, Aksoy T, et al. Exploring the role of artificial intelligence in Turkish orthopedic progression exams. Acta Orthop Traumatol Turc., 2025;59(1):18-26.*

DOI: 10.5152/j.aott.2025.24090

Materials and methods

The UEGS consists of 200 questions covering several orthopedic subspecialties in each examination. All are true or false questions. The subspecialty areas covered in the examination include Spinal Surgery; Pediatric Orthopedics; Foot and Ankle Surgery; Basic and General Orthopedics; Trauma; Adult Reconstructive Surgery, Hand, Wrist, and Upper Extremity Surgery; Orthopedic Oncology; and Sports Traumatology, Arthroscopy, and Knee Surgery. The questions do not contain any figures or tables. The final net score was calculated by subtracting the number of incorrect answers from the number of correct responses.

Two authors (M.S. and T.A.) obtained and scanned the questions and answer keys from the last 5 years' UEGS (2018, 2019, 2021, 2022, and 2023). Each examination included a total of 200 questions, resulting in a cumulative total of 1000 questions throughout the study. In order to identify the specific subspecialty of the questions, the initial evaluator, a postgraduate year (PGY) 4 resident, selected the appropriate subspecialty from the outset. Subsequently, the second observer (PGY5) also selected an appropriate subspecialty for each question. We accepted the decisions when they coincided between the 2 observers. If there was no consensus, the senior author (G.H.) made the final decision. Receiver operating characteristic (ROC) analysis was used to determine the cut-off value for the number of letters for questions. All questions were posed to ChatGPT-3.5, ChatGPT-4, and Gemini in Turkish to determine whether a statement is true or false. We restarted the AI models for every query, preventing their capacity to retain information in order to improve their performance.[14] We subtracted the number of incorrect answers from the number of correct responses to determine the final examination score for the LLMs. The average results of each PGY for residents (PGY1, PGY2, PGY3, PGY4, PGY5, and PGY6) were obtained from the TOTEK database. Due to the absence of animals, humans, and human tissue, ethical approval was deemed unnecessary for our study.

Statistical analysis

The data were analyzed with IBM SPSS V23 ((IBM SPSS Corp.; Armonk, NY, USA). The chi-squared test was used to compare categorical variables across groups, and multiple comparisons for ratios were examined with the Bonferroni-adjusted Z test. Cochran's Q test was used to compare the accuracy of responses by the AI tools. Receiver operating curve analysis was used to determine the cut-off value for the number of letters. The point-biserial correlation coefficient was used to examine the relationship between the number of letters and the accuracy of responses. Repeated measures analysis of variance was used to compare the scores of 3 or more dependent groups that were normally distributed, and multiple comparisons were examined with the LSD test and the Bonferroni test. Friedman's test was used to compare the scores of 3 or more dependent groups that were not normally distributed, while multiple comparisons were examined with Dunn's test. The analysis results for categorical variables were presented as frequency (percentage). The level of significance was set at $P < .05$.

HIGHLIGHTS

- This study evaluates the performance of ChatGPT-3.5, ChatGPT-4, and Gemini AI models on the Turkish Orthopedic Progress Testing Examination (UEGS).
- Analysis of 1000 true or false questions from 2018 to 2023 showed significant accuracy differences among AI models and years.
- ChatGPT-4 consistently outperformed other models, achieving higher accuracy, especially with shorter inputs.
- The study underscores the potential of ChatGPT-4 in medical education and clinical decision support.
- Continuous evaluation of AI technologies is crucial for maximizing their educational and clinical utility.

Results

Comparison of interyear and intrayear correct answers by the artificial intelligence tools

Foot and ankle surgery

The accuracy of the AI tools in this section significantly varied across years (ChatGPT-3.5: $P=.013$, Gemini: $P=.002$). Both ChatGPT-3.5 and Gemini achieved their highest performance in 2021 (95% and 100% accuracy, respectively), with significant differences compared to the other years. Overall, when not considering yearly variations, AI tool accuracy did not show significant differences ($P=.334$).

Hand, wrist, and upper extremity surgery

This section showed significant differences in AI tool accuracy within 2021 ($P=.012$). ChatGPT-4 achieved the highest accuracy (95%).

Adult reconstructive surgery

There were no significant differences in the number of correct answers provided by the AI tools across years ($P=.066$, $P=.129$, $P=.995$, respectively). No significant differences were found between the AI tools in any of the individual years.

Spinal surgery

No significant differences in accuracy were found between the AI tools ($P=.2$) regardless of year. ChatGPT-4 achieved the highest accuracy (75%), followed by ChatGPT-3.5 and Gemini (both at 45.8%).

Orthopedic oncology

No significant yearly or overall differences were found in the accuracy of the AI tools.

Pediatric orthopedics

The accuracy of the AI tools significantly varied across years (ChatGPT-3.5: $P=.003$, ChatGPT-4: $P < ..001$, Gemini: $P=.013$). Notably, 2021 showed the highest performance for ChatGPT-4 (95%) and Gemini (95%) compared to the other years. ChatGPT-3.5 achieved its peak accuracy of 57.1% in 2023.

Sports traumatology, arthroscopy, and knee surgery

Significant yearly variations in ChatGPT-3.5 tool accuracy were found ($P < .001$). While all tools achieved their peak performance in 2021 (ChatGPT-3.5: 100%, ChatGPT-4: 84%), Gemini's accuracy remained consistently lower across all years ($P=.007$).

Basic and general orthopedics

AI tool performance in basic and general orthopedics significantly varied across years ($P < .001$). ChatGPT-4 achieved the highest overall accuracy (80.5%), with its accuracy peaking in 2021 (100%) and declining afterward ($P=.003$). Gemini's accuracy also showed significant yearly variations ($P=.024$), with the best performance observed in 2021. Notably, in 2019, all the AI tools exhibited significantly different accuracies ($P=.01$).

Trauma

Artificial intelligence tool accuracy showed significant variations overall ($P=.006$). ChatGPT-4 achieved the highest accuracy (66.4%)

with a statistically significant difference compared to ChatGPT-3.5 (P=.006) across all years.

Across all topics, the accuracy of all the AI tools significantly varied across years (P < .008). Notably, 2019, 2021, and 2023 saw significant variations in tool performance. ChatGPT-4 achieved the highest overall accuracy. Detailed results are presented in Table 1.

Comparison by subspecialties

Across the various orthopedic subspecialties in 2021, ChatGPT-3.5's accuracy significantly varied (P < .001) with the highest performance in Foot and Ankle Surgery (95%) and the lowest in Trauma (32%). ChatGPT-4 also exhibited significant variations in accuracy across subspecialties in 2021 (P < .001), achieving the highest performance in Basic and General Orthopedics (100%) but showing a significant difference compared to Spinal Surgery and Orthopedic Oncology.

Significant differences in Gemini's correct answer distribution across topics were found (P < .001), with accuracies varying notably from 10% in Hand, Wrist, and Upper Extremity Surgery to 100% in Foot and Ankle Surgery. ChatGPT-3.5's distribution of correct answers significantly varied across different topics without year distinction (P=.049), with differences noted between Trauma (46.7%) and Sports Traumatology, Arthroscopy, and Knee Surgery (67.2%), whereas ChatGPT-4 showed no significant difference across topics (P=.123). Gemini's correct answer distribution significantly differed across topics (P=.02), with variances noted after multiple comparisons, and correct answer percentages ranging from 47.6% in Adult Reconstructive Surgery to 67.3% in Pediatric Orthopedics. A summary of the results is given in Table 2.

Receiver operating characteristic analysis of letter count for predicting the probability of correct answers

In 2018 and 2019, ChatGPT-4's ability to predict correct answers based on letter count was significant, with AUC values of 0.588 (P=.042) at a cutoff of 72 letters and 0.617 (P=.009) at a cutoff of 73 letters, respectively; lower letter counts indicated higher probabilities of correct answers. Across all years, a significant AUC of 0.561 (P=.002) at a cutoff of 74 letters was noted, unlike the other AI models (P > .050), as detailed in Table 3.

The correlation between letter count and the accuracy of the artificial intelligence models in generating correct responses

Statistically significant negative correlations were found between letter count and ChatGPT-4's accuracy in 2018 (r=−0.145; P=.044), 2019 (r=−0.201; P=.004), and across all years (r=−0.099; P=.002), with no similar correlation for the other AI models (P > .050), as summarized in Table 4.

Comparison of artificial intelligence tools with orthopedic residents

In the Basic and General Orthopedics section, a significant difference was determined in mean scores between assistants and AI tools (P=.015), with ChatGPT-4 outperforming Gemini significantly. In Trauma, a notable difference was observed (P=.012), especially between residents and ChatGPT-3.5 and Gemini. Overall, total scores reveal a significant disparity between residents and AI tools (P < .001), with residents and ChatGPT-4 showing significant differences, while no significant differences were found in other categories (P > .050), as detailed in Table 5. The mean scores of the residents and AI tools are shown in the Appendix.

Comparison of scores between the artificial intelligence tools and years of residency within each year

Throughout 2018 to 2023, significant differences were found in median net scores between the AI tools and PGY (P < .001), with ChatGPT-4 consistently achieving the highest scores among the AI tools, and senior residents scoring higher than junior residents, indicating a notable variance in performance linked to PGY.

Discussion

In the present study, we investigated the competency of 3 LLMs, namely ChatGPT-3.5, ChatGPT-4, and Gemini, on the UEGS. The analysis of correct answers across different orthopedic subspecialties within the UEGS (2018-2023) reveals interesting patterns. The performance of the AI models varied significantly. In some subspecialties, like Foot and Ankle Surgery and Pediatric Orthopedics, the models performed considerably better in specific years (2021) compared to others. This might suggest fluctuations in the difficulty of the examination within those subspecialties over time. Conversely, subspecialties like Adult Reconstructive Surgery and Spinal Surgery showed consistent performance across models and years, potentially indicating a more stable level of difficulty in those areas. One interesting observation is that all the AI tools exhibited a decrease in performance in 2022 compared to 2021 across most subspecialties (Table 6). This could have been due to a genuine increase in examination difficulty in 2022.

Evaluation of these AI models across different orthopedic subspecialties in 2021 reveals both strengths and weaknesses. ChatGPT-3.5 excels in specific areas like Foot and Ankle Surgery (95% accuracy) but struggles with subspecialties that may require broader medical knowledge, like Trauma (32%). ChatGPT-4, while gaining the highest overall score in Basic and General Orthopedics (100%), shows inconsistency across subspecialties, needing improvement in Spinal Surgery and Orthopedic Oncology. Gemini exhibits the most dramatic discrepancies, achieving a perfect score in Foot and Ankle Surgery but performing poorly in Hand, Wrist, and Upper Extremity Surgery (10%). This highlights significant knowledge gaps in Gemini. Interestingly, analysis of ChatGPT-3.5's performance across all years shows it performs better in general trauma concepts compared to the more specialized Sports Traumatology. Islem et al[9] evaluated ChatGPT's performance on Orthopaedic In-Training Examinations, finding that it correctly answered 60.8% of the questions, with particularly successful results in the basic science, oncology, shoulder and elbow, and sports subspecialties.[9] The analysis indicates that while ChatGPT can provide accurate clinical conclusions for a majority of board-style questions, its application in clinical education requires cautious validation of its reasoning and accuracy.[9] Overall, these findings emphasize the importance of considering subspecialties when evaluating AI models in orthopedics. While ChatGPT-4 demonstrates promise with consistent performance across various areas, the others require focused training to address knowledge gaps and ensure well-rounded competency for real-world application. Saad et al[3] assessed ChatGPT-4's ability to pass the FRCS Orth Part A examination and found it achieved a 67.5% score, falling short of the pass mark due to limitations in critical thinking, clinical expertise, and meeting rigorous examination requirements. Thibaut et al[11] investigated whether Google's Bard Chatbot could pass ChatGPT on the European Board of Hand Surgery (EBHS) examination, finding that Bard did not achieve significantly higher scores than ChatGPT in answering the EBHS's multiple-choice questions. Consequently, in their current iterations neither Bard nor ChatGPT was capable of passing the first part of the

Table 1. Comparison of correct answers between and within years

		Years						
		2018	**2019**	**2021**	**2022**	**2023**	**Total**	**P***
Foot and Ankle Surgery	ChatGPT-3.5							
	False	10 (50)	6 (30)	1 (5)	9 (45)	10 (50)	36 (36)	**.013**
	True	10 (50)a	14 (70)ab	19 (95)b	11 (55)a	10 (50)a	64 (64)	
	ChatGPT-4							
	False	5 (25)	7 (35)	2 (10.5)	6 (30)	9 (45)	29 (29.3)	.203
	True	15 (75)	13 (65)	17 (89.5)	14 (70)	11 (55)	70 (70.7)	
	Gemini							
	False	10 (50)	9 (45)	0 (0)	11 (55)	7 (35)	37 (37)	**.002**
	True	10 (50)a	11 (55)a	20 (100)b	9 (45)a	13 (65)a	63 (63)	
	P**	.103	.584	.368	.205	.584	.334	
Hand, Wrist, and Upper Extremity Surgery	ChatGPT-3.5							
	False	6 (30)	6 (30)	8 (40)	7 (35)	9 (45)	36 (36)	.831
	True	14 (70)	14 (70)	12 (60)AB	13 (65)	11 (55)	64 (64)	
	ChatGPT-4							
	False	8 (40)	7 (35)	1 (5)	7 (35)	9 (45)	32 (32)	.061
	True	12 (60)	13 (65)	19 (95)A	13 (65)	11 (55)	68 (68)	
	Gemini							
	False	8 (40)	10 (50)	9 (45)	10 (50)	8 (40)	45 (45)	.937
	True	12 (60)	10 (50)	11 (55)B	10 (50)	12 (60)	55 (55)	
	P**	.695	.338	**.012**	.526	.926	.121	
Adult Reconstructive Surgery	ChatGPT-3.5							
	False	8 (40)	5 (25)	9 (45)	14 (70)	13 (52)	49 (46.7)	.066
	True	12 (60)	15 (75)	11 (55)	6 (30)	12 (48)	56 (53.3)	
	ChatGPT-4							
	False	9 (45)	5 (25)	4 (20)	11 (55)	10 (40)	39 (37.1)	.129
	True	11 (55)	15 (75)	16 (80)	9 (45)	15 (60)	66 (62.9)	
	Gemini							
	False	11 (55)	10 (50)	10 (50)	11 (55)	13 (52)	55 (52.4)	.995
	True	9 (45)	10 (50)	10 (50)	9 (45)	12 (48)	50 (47.6)	
	P**	.607	.210	.127	.589	.589	.084	
Spinal Surgery	ChatGPT-3.5							
	False	10 (50)	4 (20)	10 (50)	9 (45)	13 (54.2)	46 (44.2)	.180
	True	10 (50)	16 (80)	10 (50)	11 (55)	11 (45.8)	58 (55.8)	
	ChatGPT-4							
	False	9 (45)	7 (35)	8 (40)	6 (30)	6 (25)	36 (34.6)	.666
	True	11 (55)	13 (65)	12 (60)	14 (70)	18 (75)	68 (65.4)	
	Gemini							
	False	9 (45)	9 (45)	10 (50)	7 (35)	13 (54.2)	48 (46.2)	.779
	True	11 (55)	11 (55)	10 (50)	13 (65)	11 (45.8)	56 (53.8)	
	P**	.936	.305	.766	.627	**.047**	.200	
Orthopedic Oncology	ChatGPT-3.5							
	False	10 (50)	7 (35)	8 (40)	8 (40)	6 (40)	39 (41.1)	.911
	True	10 (50)	13 (65)	12 (60)	12 (60)	9 (60)	56 (58.9)	
	ChatGPT-4							
	False	8 (40)	7 (35)	9 (45)	4 (20)	3 (20)	31 (32.6)	.345
	True	12 (60)	13 (65)	11 (55)	16 (80)	12 (80)	64 (67.4)	
	Gemini							
	False	8 (40)	5 (25)	8 (40)	6 (30)	5 (33.3)	32 (33.7)	.825
	True	12 (60)	15 (75)	12 (60)	14 (70)	10 (66.7)	63 (66.3)	
	P**	.751	.766	.939	.301	.368	.399	
Pediatric orthopedics	ChatGPT-3.5							
	False	7 (35)	11 (55)	1 (5)	12 (60)	9 (42.9)	40 (39.6)	**.003**
	True	13 (65)abc	9 (45)c	19 (95)b	8 (40)ac	12 (57.1)ac	61 (60.4)	
	ChatGPT-4							
	False	5 (26.3)	4 (20)	1 (5)	13 (65)	10 (47.6)	33 (33)	**<.001**
	True	14 (73.7)abc	16 (80)bc	19 (95)c	7 (35)a	11 (52.4)ab	67 (67)	
	Gemini							
	False	10 (50)	6 (30)	1 (5)	10 (50)	6 (28.6)	33 (32.7)	**.013**
	True	10 (50)a	14 (70)ab	19 (95)b	10 (50)a	15 (71.4)ab	68 (67.3)	
	P**	.150	.101	1.000	.558	.307	.390	

(*Continued*)

Table 1. Comparison of correct answers between and within years (*Continued*)

		Years						
		2018	2019	2021	2022	2023	Total	P^*
Sports Traumatology, Arthroscopy, and Knee Surgery	ChatGPT-3.5							
	False	14 (56)	7 (28)	0 (0)	11 (44)	9 (36)	41 (32.8)	**<.001**
	True	11 (44)a	18 (72)a	25 (100)bB	14 (56)a	16 (64)a	84 (67.2)B	
	ChatGPT-4							
	False	13 (52)	10 (40)	4 (16)	7 (28)	6 (24)	40 (32)	.056
	True	12 (48)	15 (60)	21 (84)B	18 (72)	19 (76)	85 (68)B	
	Gemini							
	False	13 (52)	12 (48)	12 (48)	13 (52)	10 (40)	60 (48)	.916
	True	12 (48)	13 (52)	13 (52)A	12 (48)	15 (60)	65 (52)A	
	*P***	.946	.327	**.001**	.174	.273	**.007**	
Basic and General Orthopedics	ChatGPT-3.5							
	False	10 (33.3)	16 (53.3)	8 (26.7)	12 (40)	11 (36.7)	57 (38)	.289
	True	20 (66.7)	14 (46.7)B	22 (73.3)A	18 (60)	19 (63.3)	93 (62)B	
	ChatGPT-4							
	False	5 (17.2)	4 (13.3)	0 (0)	11 (36.7)	9 (30)	29 (19.5)	**.003**
	True	24 (82.8)ab	26 (86.7)abA	30 (100)bB	19 (63.3)a	21 (70)a	120 (80.5)A	
	Gemini							
	False	10 (33.3)	15 (50)	5 (16.7)	16 (53.3)	10 (33.3)	56 (37.3)	**.024**
	True	20 (66.7)ab	15 (50)abB	25 (83.3)aAB	14 (46.7)b	20 (66.7)ab	94 (62.7)B	
	*P***	.210	**.010**	**.012**	.311	.846	**<.001**	
Trauma	ChatGPT-3.5							
	False	13 (52)	13 (52)	17 (68)	10 (40)	11 (55)	64 (53.3)	.405
	True	12 (48)	12 (48)	8 (32)A	15 (60)	9 (45)	56 (46.7)B	
	ChatGPT-4							
	False	10 (41.7)	8 (32)	4 (16)	11 (44)	7 (35)	40 (33.6)	.246
	True	14 (58.3)	17 (68)	21 (84)B	14 (56)	13 (65)	79 (66.4)A	
	Gemini							
	False	13 (52)	15 (60)	12 (48)	9 (36)	9 (45)	58 (48.3)	.540
	True	12 (48)	10 (40)	13 (52)AB	16 (64)	11 (55)	62 (51.7)AB	
	P^{**}	.766	.170	**.001**	.819	.397	**.006**	
Total	ChatGPT-3.5							
	False	88 (44)	75 (37.5)	62 (31)	92 (46)	91 (45.5)	408 (40.8)	**.008**
	True	112 (56)ab	125 (62.5)abAB	138 (69)bB	108 (54)a	109 (54.5)aA	592 (59.2)A	
	ChatGPT-4							
	False	72 (36.5)	59 (29.5)	33 (16.6)	76 (38)	69 (34.5)	309 (31)	**<.001**
	True	125 (63.5)a	141 (70.5)aB	166 (83.4)bA	124 (62)a	131 (65.5)aB	687 (69)B	
	Gemini							
	False	92 (46)	91 (45.5)	67 (33.5)	93 (46.5)	81 (40.5)	424 (42.4)	**.040**
	True	108 (54)	109 (54.5)A	133 (66.5)B	107 (53.5)	119 (59.5)AB	576 (57.6)A	
	P^{**}	.088	**.008**	**<.001**	.120	**.046**	**<.001**	

a-c, there is no difference between years with the same letter in each line; A-B, there is no difference between the artificial intelligence tools within each year, frequency (percentage). *Chi-squared test, **Cochran's Q test.

EBHS diploma examination. This outcome underlined the need for targeted improvements.

Analysis of ChatGPT-4's response duration revealed an intriguing trend: shorter answers were associated with higher accuracy, particularly in 2018 and 2019. This suggests that conciseness might have indicated a better grasp of the question in those years. Interestingly, this correlation between letter count and accuracy persisted across all years for ChatGPT-4 but not for the other AI models. Oztemerli et al's observational study assessed ChatGPT's performance on the last 5 medical specialty examinations (abbreviated as TUS in Turkish), revealing it could achieve rankings between 1787th and 4428th out of thousands of candidates, with a correct answer percentage ranging from 54.3% to 70.9%.[10] ChatGPT showed no significant preference in answering clinical vs. basic science questions but performed better on short questions and single select multiple-choice questions than on long or multiselect ones.[10] Despite its success in these challenging examinations, ChatGPT still falls short of expert human performance, prompting curiosity about future improvements and applications in healthcare. Kaarre et al[15] demonstrated moderate levels of accuracy and adaptability in ChatGPT-4 in providing information on anterior cruciate ligament surgery to both patients and nonorthopedic medical doctors, with correct responses generated in approximately 65% of inquiries. This suggests that while ChatGPT can serve as a supplementary tool for disseminating orthopedic knowledge, it cannot replace the expertise of professional orthopedic surgeons due to limitations in understanding and potential for inaccuracies.[15] This unique characteristic suggests that focusing on generating concise answers improved ChatGPT-4's performance. Further investigation is needed to understand the reasons behind this. Analyzing the content of shorter vs. longer responses could reveal if conciseness reflects focused knowledge or simply less information. Additionally, examination of the training data might expose potential biases influencing ChatGPT-4's response generation. By understanding this unique behavior, researchers can potentially improve ChatGPT-4's training and optimize its response generation for better accuracy in medical applications.

Table 2. Comparison of results by subspecialties

		ChatGPT-3.5		ChatGPT-4		Gemini	
Year	**Subspecialties**	**False**	**True**	**False**	**True**	**False**	**True**
2018	Foot and Ankle Surgery	10 (50)	10 (50)	5 (25)	15 (75)	10 (50)	10 (50)
	Hand, Wrist, and Upper Extremity Surgery	6 (30)	14 (70)	8 (40)	12 (60)	8 (40)	12 (60)
	Adult Reconstructive Surgery	8 (40)	12 (60)	9 (45)	11 (55)	11 (55)	9 (45)
	Spinal Surgery	10 (50)	10 (50)	9 (45)	11 (55)	9 (45)	11 (55)
	Orthopedic Oncology	10 (50)	10 (50)	8 (40)	12 (60)	8 (40)	12 (60)
	Pediatric Orthopedics	7 (35)	13 (65)	5 (26.3)	14 (73.7)	10 (50)	10 (50)
	Sports Traumatology, Arthroscopy, and Knee Surgery	14 (56)	11 (44)	13 (52)	12 (48)	13 (52)	12 (48)
	Basic and General Orthopedics	10 (33.3)	20 (66.7)	5 (17.2)	24 (82.8)	10 (33,3)	20 (66,7)
	Trauma	13 (52)	12 (48)	10 (41.7)	14 (58.3)	13 (52)	12 (48)
	P	.564		.204		.842	
2019	Foot and Ankle Surgery	6 (30)	14 (70)	7 (35)	13 (65)	9 (45)	11 (55)
	Hand, Wrist, and Upper Extremity Surgery	6 (30)	14 (70)	7 (35)	13 (65)	10 (50)	10 (50)
	Adult Reconstructive Surgery	5 (25)	15 (75)	5 (25)	15 (75)	10 (50)	10 (50)
	Spinal Surgery	4 (20)	16 (80)	7 (35)	13 (65)	9 (45)	11 (55)
	Orthopedic Oncology	7 (35)	13 (65)	7 (35)	13 (65)	5 (25)	15 (75)
	Pediatric Orthopedics	11 (55)	9 (45)	4 (20)	16 (80)	6 (30)	14 (70)
	Sports Traumatology, Arthroscopy, and Knee Surgery	7 (28)	18 (72)	10 (40)	15 (60)	12 (48)	13 (52)
	Basic and General Orthopedics	16 (53.3)	14 (46.7)	4 (13.3)	26 (86.7)	15 (50)	15 (50)
	Trauma	13 (52)	12 (48)	8 (32)	17 (68)	15 (60)	10 (40)
	P	.082		.494		.425	
2021	Foot and Ankle Surgery	1 (5)	19 (95)bc	2 (10.5)	17 (89.5)abcd	0 (0)	20 (100)a
	Hand, Wrist, and Upper Extremity Surgery	8 (40)	12 (60)ac	1 (5)	19 (95)abcd	9 (45)	11 (55)b
	Adult Reconstructive Surgery	9 (45)	11 (55)ac	4 (20)	16 (80)abcd	10 (50)	10 (50)b
	Spinal Surgery	10 (50)	10 (50)ac	8 (40)	12 (60)cd	10 (50)	10 (50)b
	Orthopedic Oncology	8 (40)	12 (60)ac	9 (45)	11 (55)bd	8 (40)	12 (60)ab
	Pediatric Orthopedics	1 (5)	19 (95)bc	1 (5)	19 (95)abcd	1 (5)	19 (95)ab
	Sports Traumatology, Arthroscopy, and Knee Surgery	0 (0)	25 (100)b	4 (16)	21 (84)abcd	12 (48)	13 (52)b
	Basic and General Orthopedics	8 (26.7)	22 (73.3)ab	0 (0)	30 (100)a	5 (16,7)	25 (83.3)ab
	Trauma	17 (68)	8 (32)a	4 (16)	21 (84)abcd	12 (48)	13 (52)b
	P	**<.001**		**<.001**		**<.001**	
2022	Foot and Ankle Surgery	9 (45)	11 (55)	6 (30)	14 (70)	11 (55)	9 (45)
	Hand, Wrist, and Upper Extremity Surgery	7 (35)	13 (65)	7 (35)	13 (65)	10 (50)	10 (50)
	Adult Reconstructive Surgery	14 (70)	6 (30)	11 (55)	9 (45)	11 (55)	9 (45)
	Spinal Surgery	9 (45)	11 (55)	6 (30)	14 (70)	7 (35)	13 (65)
	Orthopedic Oncology	8 (40)	12 (60)	4 (20)	16 (80)	6 (30)	14 (70)
	Pediatric Orthopedics	12 (60)	8 (40)	13 (65)	7 (35)	10 (50)	10 (50)
	Sports Traumatology, Arthroscopy, and Knee Surgery	11 (44)	14 (56)	7 (28)	18 (72)	13 (52)	12 (48)
	Basic and General Orthopedics	12 (40)	18 (60)	11 (36.7)	19 (63.3)	16 (53.3)	14 (46.7)
	Trauma	10 (40)	15 (60)	11 (44)	14 (56)	9 (36)	16 (64)
	P	.402		.081		.582	
2023	Foot and Ankle Surgery	10 (50)	10 (50)	9 (45)	11 (55)	7 (35)	13 (65)
	Hand, Wrist, and Upper Extremity Surgery	9 (45)	11 (55)	9 (45)	11 (55)	8 (40)	12 (60)
	Adult Reconstructive Surgery	13 (52)	12 (48)	10 (40)	15 (60)	13 (52)	12 (48)
	Spinal Surgery	13 (54.2)	11 (45.8)	6 (25)	18 (75)	13 (54.2)	11 (45.8)
	Orthopedic Oncology	6 (40)	9 (60)	3 (20)	12 (80)	5 (33.3)	10 (66.7)
	Pediatric Orthopedics	9 (42.9)	12 (57.1)	10 (47.6)	11 (52.4)	6 (28.)	15 (71.4)
	Sports Traumatology, Arthroscopy, and Knee Surgery	9 (36)	16 (64)	6 (24)	19 (76)	10 (40)	15 (60)
	Basic and General Orthopedics	11 (36.7)	19 (63.3)	9 (30)	21 (70)	10 (33.3)	20 (66.7)
	Trauma	11 (55)	9 (45)	7 (35)	13 (65)	9 (45)	11 (55)
	P	.844		.460		.663	
Total	Foot and Ankle Surgery	36 (36)	64 (64)ab	29 (29.3)	70 (70.7)	37 (37)	63 (63)
	Hand, Wrist, and Upper Extremity Surgery	36 (36)	64 (64)ab	32 (32)	68 (68)	45 (45)	55 (55)
	Adult Reconstructive Surgery	49 (46.7)	56 (53.3)ab	39 (37.1)	66 (62.9)	55 (52.4)	50 (47.6)
	Spinal Surgery	46 (44.2)	58 (55.8)ab	36 (34.6)	68 (65.4)	48 (46.2)	56 (53.8)
	Orthopedic Oncology	39 (41.1)	56 (58.9)ab	31 (32.6)	64 (67.4)	32 (33.7)	63 (66.3)
	Pediatric Orthopedics	40 (39.6)	61 (60.4)ab	33 (33)	67 (67)	33 (32.7)	68 (67.3)
	Sports Traumatology, Arthroscopy, and Knee Surgery	41 (32.8)	84 (67.2)b	40 (32)	85 (68)	60 (48)	65 (52)
	Basic and General Orthopedics	57 (38)	93 (62)ab	29 (19.5)	120 (80.5)	56 (37.3)	94 (62.7)
	Trauma	64 (53.3)	56 (46.7)a	40 (33.6)	79 (66.4)	58 (48.3)	62 (51.7)
	P	**.049**		.123		**.020**	

a-d, there is no difference between subjects with the same letter in each column. *Chi-squared test.
Bold P values represents statistical significance.

Table 3. Receiver operating characteristic analysis of the letter count for predicting the probability of a correct answer

Year		AUC (95% CI)	*P*	Cutoff	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
2018	ChatGPT-3.5	0.51 (0.428-0.592)	.809	–	–	–	–	–
	ChatGPT-4	0.588 (0.505-0.671)	**.042**	72	56.6%	60.0%	71.1%	44.2%
	Gemini	0.457 (0.376-0.538)	.300	–	–	–	–	–
2019	ChatGPT-3.5	0.568 (0.487-0.65)	.105	–	–	–	–	–
	ChatGPT-4	0.617 (0.531-0.703)	**.009**	73	70.2%	49.2%	76.7%	40.9%
	Gemini	0.477 (0.396-0.558)	.578	–	–	–	–	–
2021	ChatGPT-3.5	0.459 (0.368-0.55)	.353	–	–	–	–	–
	ChatGPT-4	0.451 (0.332-0.569)	.372	–	–	–	–	–
	Gemini	0.517 (0.431-0.602)	.700	–	–	–	–	–
2022	ChatGPT-3.5	0.561 (0.482-0.641)	.134	–	–	–	–	–
	ChatGPT-4	0.565 (0.483-0.646)	.125	–	–	–	–	
	Gemini	0.507 (0.426-0.588)	.863	–	–	–	–	–
2023	ChatGPT-3.5	0.49 (0.409-0.57)	.803	–	–	–	–	–
	ChatGPT-4	0.515 (0.431-0.6)	.725	–	–	–	–	–
	Gemini	0.449 (0.367-0.532)	.224	–	–	–	–	–
Total	ChatGPT-3.5	0.528 (0.491-0.564)	.136	–	–	–	–	–
	ChatGPT-4	0.561 (0.522-0.599)	**.002**	74	53.7%	56.7%	73.4%	35.4%
	Gemini	0.478 (0.441-0.514)	.231	–	–	–	–	–

AUC (95% CI), area under the curve (95% confidence interval); NPV, negative predictive value; PPV, positive predictive value.

Analysis of AI performance alongside that of residents reveals a nuanced picture. Overall, a significant difference exists, with ChatGPT-4's strong performance in Basic and General Orthopedics showcasing its potential in foundational knowledge areas. Residents, however, outperform the AI tools in total scores, suggesting a well-rounded knowledge base encompassing both foundational and potentially more specialized areas like Trauma. This highlights potential AI weaknesses in handling these specialized subspecialties. Interestingly, the observed improvement in resident scores with residency year suggests a positive correlation between knowledge and training experience. While AI models like ChatGPT-4 are promising, they cannot yet replace the comprehensive knowledge gained through residency. Residents excel in areas that might require broader medical understanding, while AI models show potential in specific knowledge domains. Rizzo et al[8] compared the performance of ChatGPT-3.5 and ChatGPT-4 models on Orthopaedic In-Service Training Exams (OITEs) from 2020 to 2022, finding that ChatGPT-4 consistently outperforms ChatGPT-3.5 across all categories, suggesting potential applications and limitations of AI in orthopedic education. ChatGPT-4's results are likened to the competency level of a second-to-third-year resident, while ChatGPT-3.5 aligns with a first-year resident's performance, indicating that neither model can pass the OITE or replace traditional orthopedic training.[8] Another study by Hoffman et al[16] found that ChatGPT-4 significantly outperformed its predecessor ChatGPT-3.5 on orthopedic surgery board questions, achieving an accuracy rate that aligns it with an average third-year orthopedic surgery resident. Lum et al[17] compared ChatGPT's ability to answer OITE questions; it correctly answered 47% of the questions, performing comparably to a first-year resident but unlikely to pass the orthopedic surgery board examination due to lower performance on higher complexity questions. This suggests that while AI can assist in orthopedic learning and education, its efficacy decreases with the complexity of the questions.[17] Our findings align closely, suggesting that while AI currently does not match the proficiency of a senior resident, it shows promising potential. Future research could explore how AI can be further developed to address these gaps and potentially complement resident education by providing targeted support in areas like Basic and General Orthopedics.

Table 4. The relationship between the number of letters and correct answers of the artificial intelligence tools

		r	*P**
2018	ChatGPT-3.5	−0.014	.849
	ChatGPT-4	−0.145	**.044**
	Gemini	0.064	.376
2019	ChatGPT-3.5	−0.089	.212
	ChatGPT-4	−0.201	**.004**
	Gemini	0.024	.741
2021	ChatGPT-3.5	0.031	.662
	ChatGPT-4	0.044	.539
	Gemini	−0.014	.847
2022	ChatGPT-3.5	−0.101	.154
	ChatGPT-4	−0.095	.179
	Gemini	−0.031	.665
2023	ChatGPT-3.5	0.018	.805
	ChatGPT-4	−0.031	.666
	Gemini	0.099	.162
Total	ChatGPT-3.5	−0.046	.145
	ChatGPT-4	−0.099	**.002**
	Gemini	0.031	.326

*Point-biserial correlation coefficient.

It is important to consider the differences in question formats when evaluating AI performance. Studies on AI models like ChatGPT-4 have shown that their performance can vary depending on whether the questions are multiple-choice (MCQ) or true/false. For example, research by Isleem et al[9] found that ChatGPT achieved a 60.8% accuracy rate on MCQs derived from the Orthopedic Board Examination, with performance variations depending on the type of question (e.g., management vs. diagnosis). In contrast, the UEGS exam format used in this study relies exclusively on true/false questions. This format may simplify decision-making for the AI due to the binary nature of the questions, potentially resulting in higher accuracy compared to more complex MCQs. However, it also limits the depth of reasoning required from the AI, as there are no nuanced answer choices to evaluate. Additionally, the UEGS scoring system deducts one point for each incorrect answer, adding a strategic element that can affect AI performance. This penalty for incorrect responses means that precision in answering is crucial, as incorrect answers can directly reduce the overall score.

Lastly, ChatGPT-4's dependence on data memorized before September 2021 may hinder its capability to apply novel problem-solving strategies to situations or datasets it has not been trained on.[18]

Table 5. Comparison of artificial intelligence tools and residents

	Resident	ChatGPT 3.5	ChatGPT 4	Gemini	*P*
Foot and Ankle Surgery	2.6 ± 0.9	2.8 ± 3.8	4.1 ± 2.4	2.6 ± 4.4	.768
Hand, Wrist, and Upper Extremity Surgery	3.4 ± 1.0	2.8 ± 1.3	3.6 ± 3.1	1.0 ± 1.0	.157
Adult Reconstructive Surgery	3.0 ± 0.9	0.7 ± 3.3	2.7 ± 2.9	−0.5 ± 0.5	.05
Spinal Surgery	3.1 ± 1.1	1.2 ± 2.8	3.2 ± 1.9	0.8 ± 1.5	.201
Orthopedic Oncology	3.0 ± 0.8	1.7 ± 1.1	3.3 ± 2.0	3.1 ± 1.3	.178
Pediatric Orthopedics	3.2 ± 1.4	2.1 ± 4.3	3.4 ± 4.7	3.5 ± 3.7	.79
Sports Traumatology, Arthroscopy, and Knee Surgery	3.0 ± 1.1	4.3 ± 5.3	4.5 ± 3.5	0.5 ± 1.2	.109
Basic and General Orthopedics	6.0 ± 1.2ab	3.6 ± 3.0ab	9.1 ± 4.3a	3.8 ± 4.4b	**.015**
Trauma	4.3 ± 0.4a	−0.8 ± 2.5b	3.9 ± 2.8ab	0.4 ± 2.2b	**.012**
Total	3.5 ± 1.4a	2.0 ± 3.3b	4.2 ± 3.4a	1.7 ± 2.8b	**<.001**

a-b, there is no difference between groups with the same letter. *Repeated variance analysis.

Table 6. Specialization training and development examination scores of residents, ChatGPT-3.5, ChatGPT-4, and Gemini

	2018	2019	2021	2022	2023	Total
ChatGPT-3.5	0.0 (−1.5-5.0)a	4.0 (−1.0-6.0)abc	2.0 (−4.5-12.5)ab	1.5 (−4.0-3.0)a	1.0 ± 1.8a	1.5 (−4.5-12.5)a
ChatGPT-4	2.0 (−0.5-9.5)a	3.0 (2.5-11.0)ab	8.5 (1.0-15.0)ab	4.0 (−3.0-6.0)ab	3.4 ± 2.4ab	4.0 (−3.0-15.0)ab
Gemini	0.0 (−1.0-5.0)a	0.5 (−2.5-5.0)a	1.0 (0.0-10.0)a	0.0 (−1.0-4.0)a	2.1 ± 2.0ab	1.0 (−2.5-10.0)a
PGY-1	7.0 (0.8-21.7)ab	6.0 (2.0-16.5)abcd	7.5 (2.5-17.3)ab	5.0 (−2.5-10.0)ab	7.1 ± 3.0ab	6.0 (−2.5-21.7)bc
PGY-2	9.0 (3.6-20.7)ab	7.5 (3.0-17.7)abcd	7.5 (3.5-16.7)abc	6.8 (1.0-12.0)abc	7.2 ± 3.0b	7.5 (1.0-20.7)bc
PGY-3	14.0 (6.0-19.0)ab	10.5 (5.0-20.0)bcde	13.0 (6.0-19.0)abcd	8.0 (0.0-14.0)abcd	10.7 ± 2.9c	10.5 (0.0-20.0)cd
PGY-4	15.5 (8.0-21.5)b	15.0 (6.0-21.0)cde	19.5 (9.5-23.6)bcd	9.0 (3.0-16.0)bcd	13.1 ± 3.5c	14.5 (3.0-23.6)de
PGY-5	16.5 (10.4-26.5)b	16.0 (7.0-22.0)de	22.3 (12.5-26.8)cd	12.8 (6.5-20.0)cd	15.0 ± 3.5d	15.5 (6.5-26.8)ef
PGY-6	19.0 (11.6-30.0)b	19.0 (9.0-24.0)e	23.7 (16.5-28.0)d	12.8 (8.0-22.5)d	17.3 ± 4.2e	18.5 (8.0-30.0)f
Test statistics	63.597	64.387	57.803	61.100	55.027	309.887
P	**<.001***	**<.001***	**<.001***	**<.001***	**<.001****	**<.001***

a-f, there is no difference between groups with the same letter, median (minimum-maximum), mean ± SD. *Friedman's test; **repeated variance analysis.

This inability to integrate post-cutoff developments or understand new trends limits its effectiveness in dynamic fields that evolve rapidly, such as orthopedic literature, potentially reducing its practical utility in real-time decision-making scenarios.[18]

Building on these findings, future investigations could explore LLMs' performance in clinical applications to evaluate their ability to reason, analyze data, and solve real-world problems encountered in medical practice. Integrating LLMs with clinical decision support systems could enhance their effectiveness in real-time patient care by providing on-demand access to vast medical knowledge. As highlighted by previous research, extensive research into decreasing bias within LLM algorithms and establishing clear guidelines for human oversight in LLM-assisted clinical decision-making is paramount. Addressing these concerns is crucial for responsible and ethical integration of LLMs into the future of healthcare.

This study is limited by its exclusive focus on orthopedics and reliance on straightforward true or false questions, potentially not reflecting the depth of medical knowledge or the intricacies of clinical judgment. Concentrating on a singular medical specialty and this question format limits the broader applicability of our findings. Future investigations should evaluate diverse medical fields and introduce more nuanced question types, such as case-based scenarios, to more thoroughly evaluate AI's proficiency in simulating complex clinical reasoning. Additionally, this research did not consider AI's use of real-time patient data, crucial for practical clinical decision-making, nor did it explore ethical concerns related to AI in healthcare, like algorithmic biases and the need for human oversight in AI implementations.

Our study indicates the capability of AI, specifically ChatGPT-4, in achieving a noteworthy level of success in the UEGS. We observed notable differences in AI tool accuracy across various subspecialties and years, with ChatGPT-4 exhibiting consistently superior performance. These findings show the promise of AI in improving clinical decision-making and suggest that optimizing input lengths could further increase AI efficacy. While ChatGPT-4 demonstrates certain advantages in providing detailed explanations and handling standardized exam questions, its role in medical education and practice is best seen as a complementary tool to support, rather than replace, the expertise of orthopedic surgeons. Moving forward, it is essential to explore more complex question formats, integrate AI with decision support systems, and address ethical considerations to fully leverage AI's capabilities in healthcare.

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author.

Ethics Committee Approval: N/A.

Informed Consent: N/A.

Peer-review: Externally peer-reviewed.

Author Contributions: Concept – G.A., M.T., G.H., U.K.; Design – G.A., G.H., T.A.; Supervision – G.H., M.T.; Resources – M.S., T.A.; Materials – A.Y., U.K.; Data Collection and/or Processing – M.S., U.K., A.Y.; Analysis and/or Interpretation – M.S., U.K.; Literature Search – A.Y., M.S.; Writing – G.A., U.K., T.A., A.Y.; Critical Review – M.T., G.H..

Declaration of Interests: Gazi Huri is an Associate Editor at Acta Orthopaedica et Traumatologica, however, his involvement in the peer-review process was solely as an author. The other authors have no conflict of interest to declare.

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

References

1. St Mart JP, Goh EL, Liew I, Shah Z, Sinha J. Artificial intelligence in orthopaedics surgery: transforming technological innovation in patient care and surgical training. *Postgrad Med J.* 2023;99(1173):687-694. [CrossRef]
2. Brown TB, Mann B, Ryder N, et al.Language Models are Few-Shot Learners Neural Information Processing Systems. 2020;33:1877-1901. [CrossRef]

3. Saad A, Iyengar KP, Kurisunkal V, Botchu R. Assessing ChatGPT's ability to pass the FRCS orthopaedic part A exam: a critical analysis. *Surgeon*. 2023;21(5):263-266. [CrossRef]
4. Straw I, Callison-Burch C. Artificial Intelligence in mental health and the biases of language based models. *PLoS One*. 2020;15(12):e0240376. [CrossRef]
5. Arora A, Arora A. Generative adversarial networks and synthetic patient data: current challenges and future perspectives. *Future Healthc J*. 2022;9(2):190-193. [CrossRef]
6. Lee H. The rise of ChatGPT: exploring its potential in medical education. *Anat Sci Educ*. 2024;17(5):926-931. [CrossRef]
7. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. *Commun Med (Lond)*. 2023;3(1):141. [CrossRef]
8. Rizzo MG, Cai N, Constantinescu D. The performance of ChatGPT on orthopaedic in-service training exams: a comparative study of the GPT-3.5 turbo and GPT-4 models in orthopaedic education. *J Orthop*. 2024;50:70-75. [CrossRef]
9. Isleem UN, Zaidat B, Ren R, et al. Can generative artificial intelligence pass the orthopaedic board examination? *J Orthop*. 2024;53:27-33. [CrossRef]
10. Oztermeli AD, Oztermeli A. ChatGPT performance in the medical specialty exam: an observational study. *Med (Baltim)*. 2023;102(32):e34673. [CrossRef]
11. Thibaut G, Dabbagh A, Liverneaux P. Does Google's Bard Chatbot perform better than ChatGPT on the European hand surgery exam? *Int Orthop*. 2024;48(1):151-158. [CrossRef]
12. Cheong RCT, Pang KP, Unadkat S, et al. Performance of artificial intelligence chatbots in sleep medicine certification board exams: ChatGPT versus Google Bard. *Eur Arch Otorhinolaryngol*. 2024;281(4):2137-2143. [CrossRef]
13. Farhat F, Chaudhry BM, Nadeem M, Sohail SS, Madsen DØ. Evaluating large language models for the national premedical exam in India: comparative analysis of GPT-3.5, GPT-4, and Bard. *JMIR Med Educ*. 2024;10:e51523. [CrossRef]
14. 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]
15. Kaarre J, Feldt R, Keeling LE, et al. Exploring the potential of ChatGPT as a supplementary tool for providing orthopaedic information. *Knee Surg Sports Traumatol Arthrosc*. 2023;31(11):5190-5198. [CrossRef]
16. Hofmann HL, Guerra GA, Le JL, et al. The rapid development of artificial intelligence: GPT-4's performance on orthopedic surgery board questions. *Orthopedics*. 2024;47(2):e85-e89. [CrossRef]
17. 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]
18. Narayanan A, Kappor S. GPT-4 and professional benchmarks: the wrong answer to the wrong question, GPT-4 and professional benchmarks: the wrong answer to the wrong question. 2024. Available at: https://www.aisnakeoil.com/p/gpt-4-and-professional-benchmarks. .Accessed April 15, 2024.

Appendix: Yearly performance comparison between AI tools and orthopedic residents

		ChatGPT-3.5	ChatGPT-4	Gemini	Residents
2018	Foot and Ankle Surgery	0.0	5.0	0.0	2.9
	Hand, Wrist, and Upper Extremity Surgery	4.0	2.0	2.0	4.4
	Adult Reconstructive Surgery	2.0	1.0	−1.0	3.3
	Spinal Surgery	0.0	1.0	1.0	4.6
	Orthopedic Oncology	0.0	2.0	2.0	3.2
	Pediatric Orthopedics	3.0	4.5	0.0	2.2
	Sports Traumatology, Arthroscopy, and Knee Surgery	−1.5	-0.5	−0.5	1.6
	Basic and General Orthopedics	5.0	9.5	5.0	6.4
	Trauma	−0.5	2.0	−0.5	3.8
	Total	12.0	26.5	8.0	32.4
2019	Foot and Ankle Surgery	4.0	3.0	1.0	1.9
	Hand, Wrist, and Upper Extremity Surgery	4.0	3.0	0.0	2.1
	Adult Reconstructive Surgery	5.0	5.0	0.0	2.1
	Spinal Surgery	6.0	3.0	1.0	2.2
	Orthopedic Oncology	3.0	3.0	5.0	3.8
	Pediatric Orthopedics	−1.0	6.0	4.0	4.4
	Sports Traumatology, Arthroscopy, and Knee Surgery	5.5	2.5	0.5	2.2
	Basic and General Orthopedics	−1.0	11.0	0.0	6.6
	Trauma	-0.5	4.5	−2.5	4.9
	Total	25.0	41.0	9.0	30.2
2021	Foot and Ankle Surgery	9.0	7.5	10.0	1.5
	Hand, Wrist, and Upper Extremity Surgery	2.0	9.0	1.0	4.3
	Adult Reconstructive Surgery	1.0	6.0	0.0	3.7
	Spinal Surgery	0.0	2.0	0.0	4.0
	Orthopedic Oncology	2.0	1.0	2.0	2.2
	Pediatric Orthopedics	9.0	9.0	9.0	4.2
	Sports Traumatology, Arthroscopy, and Knee Surgery	12.5	8.5	0.5	3.3
	Basic and General Orthopedics	7.0	15.0	10.0	7.6
	Trauma	−4.5	8.5	0.5	4.4
	Total	38.0	66.5	33.0	35.2
2022	Foot and Ankle Surgery	1.0	4.0	−1.0	3.5
	Hand, Wrist, and Upper Extremity Surgery	3.0	3.0	0.0	3.2
	Adult Reconstructive Surgery	−4.0	−1.0	−1.0	3.8
	Spinal Surgery	1.0	4.0	3.0	2.2
	Orthopedic Oncology	2.0	6.0	4.0	2.2
	Pediatric Orthopedics	-2.0	-3.0	0.0	1.3
	Sports Traumatology, Arthroscopy, and Knee Surgery	1.5	5.5	−0.5	3.2
	Basic and General Orthopedics	3.0	4.0	−1.0	4.9
	Trauma	2.5	1.5	3.5	4.2
	Total	8.0	24.0	7.0	28.5
2023	Foot and Ankle Surgery	0.0	1.0	3.0	3.2
	Hand, Wrist, and Upper Extremity Surgery	1.0	1.0	2.0	3.0
	Adult Reconstructive Surgery	−0.5	2.5	−0.5	2.0
	Spinal Surgery	−1.0	6.0	−1.0	2.4
	Orthopedic Oncology	1.5	4.5	2.5	3.6
	Pediatric Orthopedics	1.5	0.5	4.5	3.8
	Sports Traumatology, Arthroscopy, and Knee Surgery	3.5	6.5	2.5	4.6
	Basic and General Orthopedics	4.0	6.0	5.0	4.6
	Trauma	−1.0	3.0	1.0	4.4
	Total	9.0	31.0	19.0	31.6
Total	Foot and Ankle Surgery	14.0	20.5	13.0	13.0
	Hand, Wrist, and Upper Extremity Surgery	14.0	18.0	5.0	17.0
	Adult Reconstructive Surgery	3.5	13.5	−2.5	14.9
	Spinal Surgery	6.0	16.0	4.0	15.4
	Orthopedic Oncology	8.5	16.5	15.5	15.0
	Pediatric Orthopedics	10.5	17.0	17.5	15.9
	Sports Traumatology, Arthroscopy, and Knee Surgery	21.5	22.5	2.5	14.9
	Basic and General Orthopedics	18.0	45.5	19.0	30.1
	Trauma	−4.0	19.5	2.0	21.7
	Total	92.0	189.0	76.0	157.9