JOURNAL OF BALTALİMANI https://dergipark.org.tr/en/pub/job

*Research Article*

# Evaluation of ChatGPT's Performance in Residency Training Progress Exams and Competency Exams in Orthopedics and Traumatology

Yaşar Mahsut DİNÇEL[1] Gündüz Ercan KUTLUAY[2] Hadi SASANİ[3] Mehmet Ali ŞİMŞEK[4] Murat EREM[5]

[1] *Department of Orthopedics and Traumatology, Faculty of Medicine, Tekirdağ Namık Kemal University, Tekirdağ, Türkiye*
[2] *ÇOSB Kapaklı Devlet Hastahanesi, Tekirdağ, Türkiye*
[3] *Department of Radiology, Faculty of Medicine, Tekirdağ Namık Kemal University, Tekirdağ, Türkiye*
[4] *Department of Computer Programming, Vocational School of Technical Sciences, Tekirdağ Namık Kemal University, Tekirdağ, Türkiye*
[5] *Department of Orthopedics and Traumatology, Faculty of Medicine, Trakya University, Edirne, Türkiye*

ARTICLE INFO

***Article history:***
Submitted February 6

Accepted March 4

Publication March 30

***Keywords:***

ChatGPT

Board Examination

Orthopedics

Traumatology

Artificial Intelligence

***ORCID iDs of the Corresponding:***
Gündüz Ercan KUTLUAY, MD
0000-0002-1077-4945

**ABSTRACT**

**Background:** Artificial intelligence (AI) technologies have rapidly expanded into the field of medical education, offering innovative tools for training and assessment. This study aimed to evaluate the performance of the ChatGPT-3.5 language model in the "Residency Training Progress Examination" (UEGS) and the "Competency Examination" administered by the Turkish Society of Orthopedics and Traumatology (TOTBID). The objective was to determine whether ChatGPT performs comparably to orthopedic residents and whether it can achieve a passing score in the Competency Exam..

**Methods:** A total of 2,000 UEGS questions (2012–2023, excluding 2020) and 1,000 Competency Examination questions (2014–2023) were presented to ChatGPT-3.5 using standardized prompts designed within the Role–Goals–Context (RGC) framework. The model's responses were statistically compared with annual aggregate resident performance data using the Mann–Whitney U test. Bonferroni correction was applied for nine UEGS subcategory comparisons (adjusted significance threshold: $p < 0.0056$). Effect sizes (r) were calculated, and 95% confidence intervals for the primary comparison were estimated using bootstrap resampling.

**Results:** ChatGPT achieved the highest accuracy in the General Orthopedics category (62%) and the lowest in Adult Reconstructive Surgery (40%). In comparisons with residents, overall accuracy did not differ significantly. Although unadjusted analyses suggested differences in certain subcategories, none remained statistically significant after Bonferroni correction for multiple comparisons. In the Competency Exams, ChatGPT passed four of ten exams.

**Conclusion:** ChatGPT-3.5 demonstrated limited reliability and accuracy in orthopedic examinations and should be used cautiously as an educational support tool.

Future studies involving newer multimodal versions of large language models may clarify their potential role in medicaleducation and assessment.

**Doi:**
**10.5281/zenodo.18998621**
**Co-Author ORCID IDs**:

**Yaşar Mahsut DİNÇEL, MD**
0000-0001-6576-1802

**Hadi SASANİ ,MD**
0000-0001-6236-4123

**Mehmet Ali ŞİMŞEK ,MD**
0000-0002-6127-2195

**Murat EREM, MD**
0000-0002-9743-5515

**Corresponding author**:
Gündüz Ercan Kutluay, MD
gnduzercankutulay@gmail.com



## Introduction

Artificial intelligence (AI) has advanced rapidly in recent years, driven by increased computational power and the availability of large-scale data. From early theoretical neural network models, AI has evolved into modern deep learning systems capable of complex reasoning and language generation (Minh et al., 2022; Haenlein & Kaplan, 2019). The development of large language models such as GPT has further expanded the application of AI in healthcare and medical education (Ollivier et al., 2023).

The Generative Pre-trained Transformer (GPT) model, introduced by OpenAI in 2018, is a deep learning–based system trained on massive text datasets to generate human-like responses. The ChatGPT interface, launched in 2022, made this technology accessible to the general public. Newer versions of large language models, such as GPT-4 and subsequent multimodal iterations, have demonstrated improved accuracy and reasoning capabilities (Ollivier et al., 2023). In healthcare, AI applications now contribute to diagnosis, image interpretation, and patient management with increased accuracy and reduced workload for clinicians. In education, AI-based learning systems are being tested for their potential to improve comprehension and self-directed learning (Liu et al., 2021; Wu et al., 2020; Yang & Shulruf, 2019).

In this study, questions from the Residency Training and Progress Examination (UEGS) and the Competency Examination, organized by the Turkish Society of Orthopedics and Traumatology (TOTBID) across various years, were presented to the ChatGPT-3.5 model. The results were compared with participant outcomes to evaluate whether ChatGPT performs better than resident physicians in the UEGS and whether it can achieve a passing score in the Competency Examination. The findings aim to clarify whether ChatGPT can serve as a supplementary educational tool in residency training.

## Materials and Methods

### Study design and data sources

This was a retrospective, descriptive, and comparative study that evaluated ChatGPT-3.5's performance in the Residency Training Progress Examination (UEGS) and the Competency Examination administered by the Turkish Society of Orthopedics and Traumatology (TOTBID).

Both exams are organized annually and publicly accessible on the official TOTBID website (Gönen, 2013; Türk Ortopedi ve Travmatoloji Birliği Derneği, n.d.). The UEGS is a national examination designed to assess the progress of orthopedic and traumatology residents in Türkiye. It has been held annually since 2009 and includes theoretical questions from all subfields of orthopedics. The exam initially contained 100 questions but expanded to 200 questions in 2014. The present study used UEGS exams from 2012 to 2023, excluding 2020, as those questions were unavailable online.

Resident performance data were extracted from the publicly available annual TOTBID reports / website summaries. The number of resident participants by year was: 2012 (n=713), 2013 (n=725), 2014 (n=725), 2015 (n=718), 2016 (n=742), 2017 (n=755), 2018 (n=846), 2019 (n=884), and 2021 (n=1031). Participant-level raw data (individual correct / incorrect per person) were not accessible; only annual aggregate summaries were available. No comparable resident summaries could be retrieved for 2020, 2022, and 2023; therefore, resident comparisons were limited to overlapping years with available TOTBID summaries.

The Competency Examination, prepared by the Turkish Orthopedics and Traumatology Education Council (TOTEK), has been conducted since 2003 and consists of two stages. Only the first stage, comprising 100 multiple-choice questions, was analyzed in this study (Benli & Acaroğlu, n.d.; Acaroğlu et al., 2014). Participants who answer at least 60 questions correctly are considered successful. Competency Exam questions from 2014 to 2023 were included.

**ChatGPT interaction and data collection procedure**

All questions from the UEGS and Competency Exams were presented to the ChatGPT-3.5 model (OpenAI, web interface; tested in March 2023). The free version of ChatGPT-3.5 web interface, which represents the most widely used configuration during the study period, was employed. The model was instructed to answer in Turkish to ensure linguistic compatibility with the original exam format.

Each examination year (e.g., UEGS 2012, UEGS 2013, etc.) was conducted in a separate independent chat session. Within each session, all questions belonging to that specific examination were entered sequentially. Each question was queried a single time, and no regeneration or repeated sampling was performed. Responses were recorded exactly as generated.

The same standardized prompt structure and wording were used consistently for all questions to ensure reproducibility. No feedback or corrections were provided to the model during data entry.
Example of standardized prompt used for Competency Exam: "You are an orthopedic resident preparing for the national examination. Read the following question carefully and select the most appropriate answer among the options. Then, provide a brief (one-sentence) explanation for your choice. Answer in Turkish."

In the official UEGS format, participants are allowed to respond using three options: "Correct," "Incorrect," or "I do not know (blank)." The examination applies a negative marking system in which each incorrect answer invalidates one correct answer. Therefore, leaving a question blank may represent a strategic choice.

For the UEGS analysis, ChatGPT-3.5 was explicitly instructed to follow the same response format. The standardized prompt used was:
"You are an orthopedic resident taking the official UEGS (Residency Training Progress Examination). Each question must be evaluated according to the UEGS response format. You may respond using only one of the following options: 'Correct,' 'Incorrect,' or 'I do not know (blank).' If you are uncertain, select 'I do not know (blank)' rather than guessing. Respond using only one of the three options. Then, provide a brief (one-sentence) explanation for your choice. Answer in Turkish."

Although the model was given the option to leave questions blank, ChatGPT-3.5 provided a response of either "Correct" or "Incorrect" for

all items and did not select the blank option for any question. In contrast, the mean proportion of blank responses among resident participants in the 2016, 2017, 2018, 2019, and 2021 examinations was 26.5% (8).
UEGS performance comparisons were based on net scores calculated according to the official negative marking system, ensuring that both ChatGPT and resident participants were evaluated using the same scoring rules.
Each interaction was designed following the Role–Goals–Context (RGC) Framework, which defines the AI's role (exam participant), objective (select the most accurate answer), and contextual boundaries (question content and available options) (Tabatabaian, 2024).

**Subcategorization of questions**

To allow domain-specific performance analysis, all questions were classified into subcategories representing the major divisions of orthopedics and traumatology.
For the UEGS, nine subcategories were defined: general orthopedics, trauma, pediatric orthopedics, spinal surgery, hand and upper extremity surgery, foot and ankle surgery, sports trauma and knee arthroscopy, orthopedic oncology, and adult reconstructive surgery
For the Competency Exam, ten subcategories were used: basic sciences, pediatric orthopedics, pediatric trauma, adult trauma, upper extremity and hand surgery, lower extremity and foot surgery, arthroscopic and sports surgery, adult reconstructive surgery and arthroplasty, spinal surgery, and infections and tumors (Table 1).

**Table 1.** Distribution of questions by subcategories in Competency Exam and UEGS.

| Competency Exam Subcategories | Number of Questions (n) | UEGS Subcategories | Number of Questions (n) |
|---|---|---|---|
| Basic Sciences | 127 | General Orthopedics | 285 |
| Pediatric Orthopedics | 151 | Trauma | 257 |
| Pediatric Trauma | 110 | Pediatric Orthopedics | 218 |
| Adult Trauma | 130 | Spinal Surgery | 212 |
| Upper Extremity and Hand Surgery | 88 | Hand, Wrist, and Upper Extremity Surgery | 202 |
| Lower Extremity and Foot Surgery | 52 | Foot and Ankle Surgery | 180 |
| Arthroscopic Surgery and Sports Traumatology | 102 | Sports Trauma, Arthroscopy, and Knee Surgery | 247 |
| Adult Reconstructive Surgery and Arthroplasty | 96 | Orthopedic Oncology | 186 |
| Spinal Surgery | 80 | Adult Reconstructive Surgery | 213 |
| Infections and Tumors | 64 | | |
| Total | 1000 | Total | 2000 |

**Limitations of AI interaction**

Because ChatGPT-3.5 does not support image interpretation, the model was unable to answer questions containing radiographs, clinical photographs, or other visual material. In total, 56 such questions were skipped: 7 (2023), 8 (2022), 17 (2021), 5 (2020), 5 (2019), 4 (2018), 1 (2017), 5 (2016), 3 (2015), and 1 (2014). These were excluded from accuracy calculations.

All remaining UEGS questions (n = 2000) and non-visual Competency Exam questions (n = 944) received valid text-based responses.

**Statistical analysis**

All statistical analyses were performed using PASW Statistics for Windows, version 18.0 (SPSS Inc., Chicago, IL, USA). Because the unit of analysis consisted of annual aggregate values and the sample size was small, non-parametric methods were used for group comparisons. Descriptive statistics were reported as means ± standard deviation (SD) and frequencies. Accuracy was calculated as the number of correct answers divided by the total number of exam questions (including incorrect and unanswered items).

The Mann–Whitney U test was applied to compare ChatGPT and resident groups. For UEGS subcategory analyses (nine comparisons), Bonferroni correction was used to adjust for multiple testing, and the adjusted significance threshold was set at $p < 0.0056$.

Effect sizes (r) were calculated for Mann–Whitney U tests. For the primary comparison of total correct answers, 95% confidence intervals were estimated using bootstrap resampling.
For resident comparisons, statistical testing was performed only for examination years with available resident summary data (2012–2019 and 2021). In Table 3, the unit of analysis was exam-year (annual mean values), because resident data were available only as yearly aggregates. Therefore, the resident comparison represents an aggregate-level (ecological) comparison, and results should be interpreted with this limitation in mind.

**Ethical considerations**

No human participants or patient data were involved in this study. All exam data were publicly available on the official TOTBID website and analyzed in aggregate form. The artificial intelligence algorithm we used in our study is an "open access" artificial intelligence platform. Therefore, formal ethics committee approval was not required.

## Results

**Performance of ChatGPT in the UEGS**

In the UEGS examinations conducted between 2012 and 2023 (excluding 2020), the total number of analyzed questions was 2,000, distributed across nine orthopedic subcategories as shown in Table 1. ChatGPT achieved an overall accuracy rate of 52%, providing 1,043 correct and 957 incorrect responses. The highest accuracy was recorded in the General Orthopedics category (62%), followed by Trauma (57%) and Orthopedic Oncology (57%). The lowest accuracy was found in Adult Reconstructive Surgery (40%) and Foot and Ankle Surgery (45%). Detailed accuracy (correct answers / total exam questions) rates by subcategory are presented in Table 2.

When compared statistically with resident physicians' annual summary scores obtained from the TOTBID database, ChatGPT's total accuracy did not differ significantly ($p = 0.895$, $r = 0.03$). In unadjusted subcategory analyses, differences were observed in Spinal Surgery (mean = 10.64 ± 2.54 vs. 7.54 ± 2.72, $p = 0.034$) and Adult Reconstructive Surgery (mean = 7.27 ± 3.20 vs. 8.71 ± 1.43, $p = 0.015$); however, these differences did not remain statistically significant after Bonferroni correction for nine comparisons (adjusted $p < 0.0056$). Although overall correct-answer rates were comparable, analysis of the annual mean incorrect-answer counts presented in Table 3 showed that ChatGPT had higher mean incorrect responses than residents in all 9 UEGS subcategories. After Bonferroni correction (adjusted $p < 0.0056$), statistically significant differences in incorrect responses persisted in Hand, Wrist, and Upper Extremity Surgery ($p = 0.003$), Sports Trauma and Knee Surgery ($p = 0.001$), and Adult Reconstructive Surgery ($p < 0.001$).

**Table 2.** Accuracy of ChatGPT-3.5 by subcategories in the UEGS

| Subcategory | Number of Questions (n) | Correct Answers (n) | Accuracy (%) |
|---|---|---|---|
| General Orthopedics | 285 | 176 | 62 |
| Trauma | 257 | 146 | 57 |
| Pediatric Orthopedics | 218 | 112 | 51 |
| Spinal Surgery | 212 | 117 | 55 |
| Hand, Wrist, and Upper Extremity Surgery | 202 | 93 | 46 |
| Foot and Ankle Surgery | 180 | 81 | 45 |
| Sports Trauma, Arthroscopy, and Knee Surgery | 247 | 127 | 51 |
| Orthopedic Oncology | 186 | 106 | 57 |
| Adult Reconstructive Surgery | 213 | 85 | 40 |
| **Total** | **2,000** | **1,043** | **52** |

**Table 3.** Comparison of ChatGPT-3.5 and resident participants in the

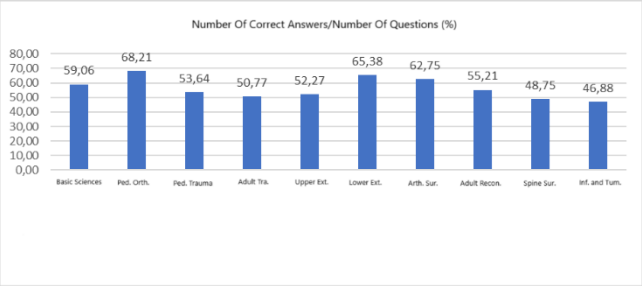
| Subcategory | ChatGPT Mean ± SD (Correct) | Residents Mean ± SD (Correct) | p-value | ChatGPT Mean ± SD (Incorrect) | Residents Mean ± SD (Incorrect) | p-value |
|---|---|---|---|---|---|---|
| General Orthopedics | 16.00 ± 6.48 | 13.17 ± 6.41 | 0.426 | 9.91 ± 6.04 | 5.92 ± 2.82 | 0.070 |
| Trauma | 13.27 ± 4.34 | 13.29 ± 1.69 | 0.929 | 10.10 ± 4.48 | 5.77 ± 0.94 | 0.038 |
| Pediatric Orthopedics | 10.18 ± 3.63 | 10.29 ± 1.22 | 0.965 | 9.64 ± 3.93 | 5.14 ± 1.47 | 0.007 |
| Spinal Surgery | 10.64 ± 2.54 | 7.54 ± 2.72 | 0.034 | 8.64 ± 4.13 | 4.23 ± 1.02 | 0.070 |
| Hand, Wrist, and Upper Extremity Surgery | 8.45 ± 3.53 | 10.93 ± 3.47 | 0.070 | 9.91 ± 4.18 | 5.16 ± 1.42 | **0.003** |
| Foot and Ankle Surgery | 7.36 ± 4.92 | 10.75 ± 4.95 | 0.085 | 9.00 ± 5.72 | 5.72 ± 2.30 | 0.085 |
| Sports Trauma, Arthroscopy, and Knee Surgery | 11.45 ± 4.39 | 11.49 ± 2.47 | 0.757 | 10.91 ± 3.70 | 6.51 ± 1.35 | **0.001** |
| Orthopedic Oncology | 9.37 ± 4.86 | 7.90 ± 3.51 | 0.566 | 7.27 ± 3.90 | 3.96 ± 1.85 | 0.057 |
| Adult Reconstructive Surgery | 7.27 ± 3.20 | 8.71 ± 1.43 | 0.015 | 11.64 ± 3.44 | 4.31 ± 0.75 | **<0.001** |
| **Total** | 94.10 ± 24.84 | 94.58 ± 8.33 | 0.895 | 87.73 ± 28.46 | 46.44 ± 5.07 | 0.010 |

** Statistical comparisons were performed using the Mann–Whitney U test. For UEGS subcategory analyses (nine comparisons), Bonferroni correction was applied, and statistical significance was set at $p < 0.0056$. Values are presented as mean ± standard deviation (SD) of annual aggregate scores. "Correct" and "Incorrect" indicate the annual mean number of correct and incorrect answers per examination year. Resident data represent publicly available annual summary averages rather than individual participant-level scores.*

**Performance of ChatGPT in the Competency Exams**

A total of 1,000 Competency Exam questions from 2014–2023 were analyzed. After excluding 56 image-based questions that the model could not interpret, 944 questions were evaluated. ChatGPT achieved its highest accuracy (correct answers / total exam questions) in Pediatric Orthopedics (68.2%), followed by Lower Extremity and Foot Surgery (65.4%), and its lowest accuracy in Infections and Tumors (38%). Figure 1 illustrates response percentages across subcategories.

**Figure 1.** Percentages of correct answers by ChatGPT-3.5 in the Competency Exam subcategories.




In the ten annual Competency Exams analyzed, ChatGPT passed four exams (years 2016, 2017, 2019, and 2022) by achieving ≥ 60% correct answers (Table 4).
Its best performance occurred in 2019 (72% accuracy), while the lowest score was in 2023 (44%).

**Table 4**. ChatGPT-3.5 performance in the Competency Exams (2014–2023)

| Year | Unanswered Questions (n) | Correct Answers (n) | Incorrect Answers (n) | Result |
|---|---|---|---|---|
| 2023 | 7 | 44 | 49 | Failed |
| 2022 | 8 | 61 | 31 | Passed |
| 2021 | 17 | 45 | 38 | Failed |
| 2020 | 5 | 47 | 48 | Failed |
| 2019 | 5 | 72 | 23 | Passed |
| 2018 | 4 | 59 | 37 | Failed |
| 2017 | 1 | 70 | 29 | Passed |
| 2016 | 5 | 60 | 35 | Passed |
| 2015 | 3 | 57 | 40 | Failed |
| 2014 | 1 | 54 | 45 | Failed |
| **Total** | **56** | **569** | **375** | **4 / 10 Passed** |

## Discussion

This study examined the performance of the ChatGPT-3.5 artificial intelligence model in Türkiye's Residency Training and Progress Examination (UEGS) and Competency Examination, both organized by the Turkish Society of Orthopedics and Traumatology (TOTBID).
ChatGPT achieved its highest accuracy in the General Orthopedics category (62%) and its lowest accuracy in Adult Reconstructive Surgery (40%). In unadjusted analyses, higher mean correct answers were observed for ChatGPT in the Spine Surgery category; however, this difference did not remain statistically significant after correction for multiple comparisons. Notably, Spine Surgery was also among the subcategories in which resident physicians demonstrated relatively lower accuracy. These findings may offer perspective for those involved in the evaluation and refinement of orthopedic and traumatology residency curricula, particularly in subspecialty areas such as spine surgery (Acaroğlu et al., 2014).

When comparing the results of ChatGPT and residents in the UEGS, no statistically significant difference was found in the total number of correct answers ($p = 0.895$, $r = 0.03$), indicating a negligible effect size. The bootstrap 95% confidence interval for the primary comparison included zero, further supporting the absence of a meaningful difference between groups. However, analysis of the annual mean incorrect-answer counts (Table 3) showed that ChatGPT demonstrated higher mean incorrect responses than residents across all nine subcategories. After Bonferroni adjustment ($p < 0.0056$) statistically significant differences in incorrect responses persisted only in selected domains. Taken together, these findings indicate that, despite similar overall correct-answer rates, ChatGPT-3.5 showed less stable response patterns in certain subspecialties compared with orthopedic and traumatology residents in the UEGS.

In the Competency Examination, ChatGPT passed four out of ten exams by correctly answering at least 60 questions out of 100. In the remaining six exams, the model failed to reach the required passing score. Among the 1,000 questions analyzed, ChatGPT-3.5 was unable to answer 56 image-based questions due to its lack of visual interpretation capability. With newer versions such as ChatGPT-4, which can interpret images, more questions could be answered, potentially resulting in higher success rates. In a comparative study using UEGS questions, ChatGPT-4 demonstrated significantly higher accuracy than ChatGPT-3.5 (Ayik et al., 2025).

Likewise, a recent study using questions from the Turkish Competency Exam reported that ChatGPT-4o achieved higher accuracy than human participants across all orthopedic subdomains (Yağar et al., 2025). This finding aligns with the overall trend of improved performance observed in newer AI systems.

Our findings suggest that ChatGPT-3.5 did not demonstrate superior performance compared with resident physicians and showed less consistent response patterns in certain domains. Similar studies have demonstrated that ChatGPT performs well in some exams but poorly in others (Ruksakulpiwat et al., 2023; Khan et al., 2023; Oztermeli & Oztermeli, 2023; Sumbal et al., 2024; Wang et al., 2023; Aljindan et al., 2023). In certain cases, ChatGPT outperformed participants, whereas in others, as in our study, participants achieved better results.

ChatGPT became one of the fastest-growing computer programs in history, reaching 100 million active users within two months of its public release (Alessandri Bonetti et al., 2024). Like other AI models, it draws from a wide range of data sources, including peer-reviewed journal articles, textbooks, and online content (Massey et al., 2023). As new versions are released, both the technical capabilities of the model and the size of its knowledge base expand. Thus, it is expected that future AI models will continue to improve their ability to evaluate and answer questions.

This study has several limitations. First, resident and specialist performance data were obtained from publicly available annual summaries on the official TOTBID website rather than participant-level raw datasets. Therefore, comparisons between ChatGPT and residents were conducted using annual aggregate values, and subcategory analyses represent ecological-level comparisons rather than individual-level inferential statistics. This limitation reduces the precision of statistical comparisons and may obscure within-year variability among participants. Furthermore, although net scores were calculated according to the official negative marking system, detailed subcategory-level data on blank or unanswered questions were not available in the publicly accessible TOTBID summaries. While overall blank response proportions were reported for selected years, the distribution of blank responses across specific subspecialties could not be determined. Because residents may strategically omit more difficult questions in certain domains to avoid penalty, this factor may have influenced subspecialty-level comparisons. Therefore, comparisons between ChatGPT and resident performance should be interpreted cautiously in light of this limitation.

Additionally, the negative marking structure of the UEGS may introduce a response-threshold bias. Because incorrect answers cancel correct ones, resident participants may strategically leave uncertain questions unanswered to minimize score penalties. In contrast, although ChatGPT was allowed to select the blank option, the model generated a definitive response ("Correct" or "Incorrect") for all items. This difference in response behavior may create a systematic bias when directly comparing performance.

Therefore, the findings may reflect AI accuracy under a de facto forced-response condition rather than true net-score performance under authentic examination strategy conditions.

Second, complete resident data were not available for all examination years. Specifically, participant numbers and detailed performance summaries for 2020, 2022, and 2023 could not be retrieved from the TOTBID website. Consequently, statistical comparisons were restricted to overlapping years with accessible resident summaries, which may have influenced representativeness.

Third, multiple comparisons across subcategories increase the risk of type I error. Although Bonferroni correction was applied to adjust the significance threshold, this conservative approach may have reduced statistical power and masked potentially meaningful differences.

Finally, the stochastic nature of large language models represents an additional methodological limitation. Each examination question was queried only once, and repeated independent runs were not performed. Although separate chat sessions were used for each examination year, intra-session variability was not evaluated. Therefore, the findings reflect a single-run performance snapshot rather than a reproducibility estimate across multiple iterations.

Finally, the use of chatbots in medical education is an emerging trend supported by many educators and medical professionals. OpenAI's ChatGPT offers several potential advantages for both students and teachers (Huang et al., 2023; Moritz et al., 2023). Recent reviews have highlighted that generative AI tools hold promise for orthopedic education and training but also pose challenges related to reliability, ethical use, and integration into curricula. It is important to remember that these systems are still evolving and have not yet reached perfection. There remain significant gaps in both theoretical and practical orthopedic education that AI tools cannot yet fill (Atik, 2024).

## Conclusion

ChatGPT-3.5 demonstrated variable accuracy across orthopedic examination subcategories but did not show a statistically significant advantage over residents after adjustment for multiple comparisons. Although overall performance levels were comparable in certain domains, the model generated a higher number of incorrect responses across most categories.

While ChatGPT may offer supportive value as an adjunct educational tool, its current performance does not support independent or unsupervised use in orthopedics and traumatology training. As artificial intelligence systems continue to evolve, future multimodal models with enhanced reasoning capabilities may achieve greater reliability and educational applicability in medical assessment settings.

## Ethics Approval and Consent to Participate

Not Applicable

## Consent for Publication

Not Applicable

## Data Availability

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

## Competing Interests

The authors declare that they have no competing interests

## Funding

Not applicable

## Clinical Trial Number

Not applicable

## Authors' Contributions

YMD and GEK wrote the main manuscript.
HS and ME asked questions to ChatGPT.
MAŞ performed the statistical work.

## Acknowledgements

The authors would like to express their sincere gratitude to all doctors who helped improve the orthopedic examination for a better education.

## References

Acaroğlu, E., Kahraman, S., Senköylü, A., Berk, H., Caner, H., Özkan, S., ... (2014). Core curriculum (CC) of spinal surgery: A step forward in defining our profession. Acta Orthopaedica et Traumatologica Turcica, 48(5), 475–478.

Alessandri Bonetti, M., Giorgino, R., Gallo Afflitto, G., De Lorenzi, F., & Egro, F. M. (2024). How does ChatGPT perform on the Italian Residency Admission National Exam compared to 15,869 medical graduates? Annals of Biomedical Engineering, 52(4), 745–749.

Aljindan, F. K., Al Qurashi, A. A., Albalawi, I. A. S., Alanazi, A. M. M., Aljuhani, H. A. M., Falah Almutairi, F., ... (2023). ChatGPT conquers the Saudi Medical Licensing Exam: Exploring the accuracy of artificial intelligence in medical knowledge assessment and implications for modern medical education. Cureus, 15(9), Article e45043.

Atik, O. Ş. (2024). Artificial intelligence: Who must have autonomy the machine or the human? Joint Diseases and Related Surgery, 35(1), 1–2.
Ayik, G., Kolac, U. C., Aksoy, T., Yilmaz, A., Sili, M. V., Tokgozoglu, M., ... (2025). Exploring the role of artificial intelligence in Turkish orthopedic progression exams. Acta Orthopaedica et Traumatologica Turcica, 59(1), 18–26.

Benli, İ., & Acaroğlu, E. (2011). Türk Ortopedi ve Travmatoloji Birliği Derneği (TOTBİD) Türk Ortopedi ve Travmatoloji Eğitim Konseyi Yeterlik Sınavları. Acta Orthopaedica et Traumatologica Turcica, 45(2). https://dergipark.org.tr/en/download/article-file/169969

Gönen, D. E. (2013). 2012-2013 TOTBİD-TOTEK Uzmanlık Eğitimi Gelişim Sınavı Raporu (UEGS). Türk Ortopedi ve Travmatoloji Birliği Derneği. https://totbid.org.tr/uploads/files/uegs_2013_rapor.pdf

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. California Management Review, 61(4), 5–14.

Huang, Y., Gomaa, A., Semrau, S., Haderlein, M., Lettmaier, S., Weissmann, T., ... (2023). Benchmarking ChatGPT-4 on a radiation oncology in-training exam and Red Journal Gray Zone cases: Potentials and challenges for AI-assisted medical education and decision making in radiation oncology. Frontiers in Oncology, 13, Article 1265024.

Khan, R. A., Jawaid, M., Khan, A. R., & Sajjad, M. (2023). ChatGPT - Reshaping medical education and clinical management. Pakistan Journal of Medical Sciences, 39(2). https://pjms.org.pk/index.php/pjms/article/view/7653

Liu, P. R., Lu, L., Zhang, J. Y., Huo, T. T., Liu, S. X., & Ye, Z. W. (2021). Application of artificial intelligence in medicine: An overview. Current Medical Science, 41(6), 1105–1115.

Massey, P. A., Montgomery, C., & Zhang, A. S. (2023). Comparison of ChatGPT–3.5, ChatGPT-4, and orthopaedic resident performance on orthopaedic assessment examinations. Journal of the American Academy of Orthopaedic Surgeons, 31(23), 1173.
Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: A comprehensive review. Artificial Intelligence

Review, 55(5), 3503–3568.

Moritz, S., Romeike, B., Stosch, C., & Tolks, D. (2023). Generative AI (gAI) in medical education: Chat-GPT and co. GMS Journal for Medical Education, 40(4), Article Doc54.

Ollivier, M., Pareek, A., Dahmen, J., Kayaalp, M. E., Winkler, P. W., Hirschmann, M. T., ... (2023). A deeper dive into ChatGPT: History, use and future perspectives for orthopaedic research. Knee Surgery, Sports Traumatology, Arthroscopy, 31(4), 1190–1192.

Oztermeli, A. D., & Oztermeli, A. (2023). ChatGPT performance in the medical specialty exam: An observational study. Medicine, 102(32), Article e34673.

Ruksakulpiwat, S., Kumar, A., & Ajibade, A. (2023). Using ChatGPT in medical research: Current status and future directions. Journal of Multidisciplinary Healthcare, 16, 1513–1520.

Sumbal, A., Sumbal, R., & Amir, A. (2024). Can ChatGPT-3.5 pass a medical exam? A systematic review of ChatGPT's performance in academic testing. Journal of Medical Education and Curricular Development, 11, Article 23821205241238641.

Tabatabaian, M. (2024). Prompt engineering using ChatGPT: Crafting effective interactions and building GPT apps. Mercury Learning and Information.

Türk Ortopedi ve Travmatoloji Birliği Derneği. (n.d.). TOTBİD resmi sitesi. https://totbid.org.tr/tr/

Wang, X., Gong, Z., Wang, G., Jia, J., Xu, Y., Zhao, J., ... (2023). ChatGPT performs on the Chinese National Medical Licensing Examination. Journal of Medical Systems, 47(1), Article 86.

Wu, D., Xiang, Y., Wu, X., Yu, T., Huang, X., Zou, Y., ... (2020). Artificial intelligence-tutoring problem-based learning in ophthalmology clerkship. Annals of Translational Medicine, 8(11), Article 700.

Yang, Y. Y., & Shulruf, B. (2019). Expert-led and artificial intelligence (AI) system-assisted tutoring course increase confidence of Chinese medical interns on suturing and ligature skills: Prospective pilot study. Journal of Educational Evaluation for Health Professions, 16, Article 7.

Yağar, H., Gümüşoğlu, E., & Mert Asfuroğlu, Z. (2025). Assessing the performance of ChatGPT-4o on the Turkish Orthopedics and Traumatology Board Examination. Joint Diseases and Related Surgery, 36(2), 304–310.