Original Article Clinics in Orthopedic Surgery 2024;16:669-673 • https://doi.org/10.4055/cios23179

Performance of ChatGPT on Solving Orthopedic Board-Style Questions: A Comparative Analysis of ChatGPT 3.5 and ChatGPT 4

Sung Eun Kim, MD, Ji Han Lee, MD, Byung Sun Choi, MD, Hyuk-Soo Han, MD, Myung Chul Lee, MD, Du Hyun Ro, MD

Department of Orthopedic Surgery, Seoul National University Hospital, Seoul National University College of Medicine, Seoul, Korea

Background: The application of artificial intelligence and large language models in the medical field requires an evaluation of their accuracy in providing medical information. This study aimed to assess the performance of Chat Generative Pre-trained Transformer (ChatGPT) models 3.5 and 4 in solving orthopedic board-style questions.

Methods: A total of 160 text-only questions from the Orthopedic Surgery Department at Seoul National University Hospital, conforming to the format of the Korean Orthopedic Association board certification examinations, were input into the ChatGPT 3.5 and ChatGPT 4 programs. The questions were divided into 11 subcategories. The accuracy rates of the initial answers provided by Chat GPT 3.5 and ChatGPT 4 were analyzed. In addition, inconsistency rates of answers were evaluated by regenerating the responses.

Results: ChatGPT 3.5 answered 37.5% of the questions correctly, while ChatGPT 4 showed an accuracy rate of 60.0% ($p < 0.001$). ChatGPT 4 demonstrated superior performance across most subcategories, except for the tumor-related questions. The rates of inconsistency in answers were 47.5% for ChatGPT 3.5 and 9.4% for ChatGPT 4.

Conclusions: ChatGPT 4 showed the ability to pass orthopedic board-style examinations, outperforming ChatGPT 3.5 in accuracy rate. However, inconsistencies in response generation and instances of incorrect answers with misleading explanations require caution when applying ChatGPT in clinical settings or for educational purposes.

Keywords: *Artificial intelligence, Large language models, Orthopedic board examination*

The emergence of artificial intelligence (AI) and large language models (LLMs) has highlighted their potential in the medical field. As the application of AI and LLMs is expected to encompass throughout healthcare, it is important to assess its accuracy in providing medical information. AI programs have demonstrated competencies comparable to those of human medical specialists in diagnostic accuracy[1,2] and have even outperformed physicians in providing high-quality, empathetic responses to patients' questions.[3] Recent research reported that OpenAI's LLM, the Chat Generative Pre-trained Transformer (GPT), was able to pass the United States Medical Licensing Examination, achieving an accuracy rate of 60%.[4] Furthermore, it has surpassed the threshold in specialist board examinations in radiology and neurosurgery, reporting an accuracy rate of 81% and approximately 60%, respectively.[5-7] In light of the growing focus on evaluating ChatGPT's capabilities in the medical field, there is a surge in literature on this topic, examining whether ChatGPT possesses the qualification of a medical specialist.[8-10] This interest has also extended to the field of orthopedics. However, the volume of studies specifically evaluating the LLM's accuracy in addressing orthopedic problems remains limited.[11-13] In addition,

Received May 31, 2023; Revised January 29, 2024;
Accepted January 29, 2024
Correspondence to: Du Hyun Ro, MD
Department of Orthopedic Surgery, Seoul National University Hospital, 101 Daehak-ro, Jongno-gu, Seoul 03080, Korea
Tel: +82-2-2072-1995, Fax: +82-2-764-2718
E-mail: duhyunro@gmail.com

with 2 versions of ChatGPT available—model 3.5, which is freely accessible, and the more advanced model 4.0, a subscription-based version with enhanced accuracy and processing capabilities—a comparative analysis of their performances is essential.[14]

Therefore, the primary objective of this study was to evaluate the performance of ChatGPT in solving orthopedic problems by presenting it with a comprehensive set of orthopedic board-style questions. The accuracy of the generated responses was used as an indicator of the program's performance. Furthermore, we examined and compared the differences in accuracy between ChatGPT models 3.5 and 4. We hypothesized that ChatGPT would successfully pass the orthopedic board-style examination and that ChatGPT 4.0 would demonstrate superior performance compared to 3.5.

METHODS

As this study did not involve patient data and medical records, it was conducted without the approval from the Institutional Review Board. The questions used in the study were sourced from the Department of Orthopedic Surgery at Seoul National University Hospital, which have been validated by board-certified orthopedic surgeons and adapted to align with the format of the Korean Orthopedic Association board certification examinations. The orthopedic surgery board certification examination consists of 160 text-only questions and 100 image-based questions. However, for the purposes of this analysis, only the 160 text-only questions were included due to the limitation of ChatGPT 3.5, which does not support image interpretation. As ChatGPT can generate varying responses when prompted to regenerate answers, the initial answer provided by ChatGPT was considered for primary evaluation. To further assess the consistency of ChatGPT, we used the 'regenerate' function, prompting the model to provide a second answer to the same question, and the number of changed responses was evaluated.

The threshold for passing the board-style examination was set at 60%, in alignment with the passing threshold for orthopedic specialist certification examinations in Korea. Questions were input in Korean, which was automatically translated into English by ChatGPT, with the authors ensuring the accuracy of the translations and making manual corrections as necessary. The study then compared the accuracy rates between ChatGPT 3.5 and ChatGPT 4. The examination questions consisted of multiple-choice items with 5 options, requiring the selection of 1 correct answer. The distribution of questions across orthopedic subcategories was as follows: hip (n = 17), knee (n = 18), ankle and foot (n = 15), spine (n = 18), shoulder (n = 15), hand (n = 18), pediatrics (n = 19), tumor (n = 10), general trauma (n = 13), infection and metabolism (n = 8), and

Table 1. Accuracy Rates of ChatGPT 3.5 and ChatGPT 4 According to Subcategories

Variable	Number of correct answers (%)		*p*-value
	GPT 3.5	GPT 4	
Hip (n = 17)	3 (17.6)	6 (35.3)	0.243
Knee (n = 18)	6 (33.3)	11 (61.1)	0.095
Ankle and foot (n = 15)	7 (46.7)	10 (66.7)	0.269
Spine (n = 18)	8 (44.4)	14 (77.8)	0.040*
Shoulder (n = 15)	6 (40.0)	12 (80.0)	0.025*
Hand (n = 18)	7 (38.9)	12 (66.7)	0.071
Pediatrics (n = 19)	9 (47.4)	14 (73.7)	0.097
Tumor (n = 10)	4 (40.0)	3 (30.0)	0.639
General trauma (n = 13)	4 (30.8)	5 (31.3)	0.680
Infection and metabolism (n = 8)	3 (37.5)	5 (62.5)	0.317
Basic science and rehabilitation (n = 9)	3 (33.3)	6 (66.7)	0.157
Total (n = 160)	60 (37.5)	96 (60.0)	< 0.001*

*Statistically significant at $p < 0.05$.

basic science and rehabilitation (n = 9), resulting in a total of 160 questions.

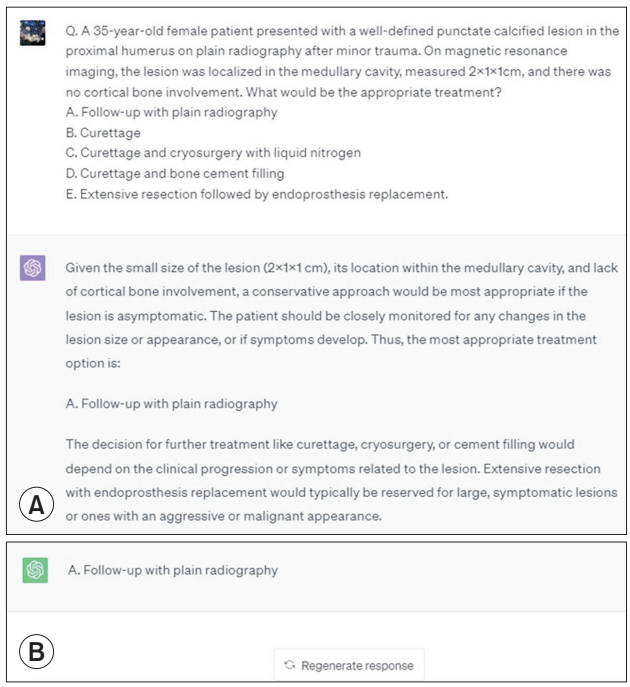

Fig. 1. Example of a question answered by ChatGPT 4 (A) and ChatGPT 3.5 (B). The correct answer is option "A. Follow-up with plain radiography." ChatGPT 4 provided an appropriate explanation, whereas ChatGPT 3.5 presented only the answer without any explanation.

Statistical Analysis

The chi-square test was used to compare the accuracy rates between ChatGPT 3.5 and ChatGPT 4. Statistical significance was set at a p-value of less than 0.05. All statistical analyses were conducted using IBM SPSS statistics ver. 25 (IBM Corp.).

RESULTS

The overall accuracy rate and accuracy rates for each subcategory are presented in Table 1. Of the total 160 questions, ChatGPT 3.5 correctly answered 60 questions (37.5%), while ChatGPT 4 provided 96 correct answers (60.0%), showing a statistically significant difference ($p < 0.001$). Among the subcategories, ChatGPT 4 demonstrated significantly higher accuracy rates in the areas of spine and shoulder ($p = 0.040$ and $p = 0.025$, respectively). While ChatGPT 4 generally showed superior accuracy rates in most subcategories relative to ChatGPT 3.5, the accuracy rate for ChatGPT 3.5 in the tumor subcategory (40.0%) exceeded that of ChatGPT 4 (30.0%), although this difference was not statistically significant. Figs. 1-3 show examples of the questions presented to ChatGPT and the corresponding responses generated, which have been provided in English for illustrative purposes. Concerning the measure of response consistency, ChatGPT 3.5 exhibited a notable alteration in 76 of its initial responses (47.5%),

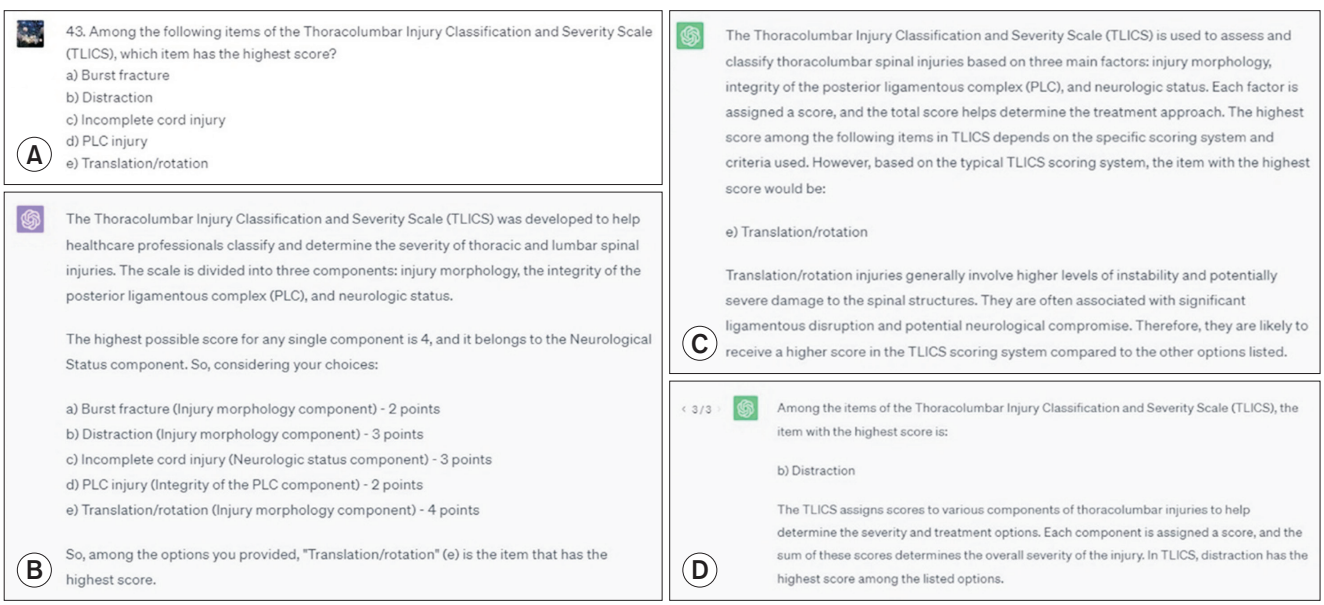

Fig. 2. Example of a simple, knowledge-based question incorrectly answered by both ChatGPT 4 and ChatGPT 3.5. (A) The correct answer is option "b) Distraction." According to the Thoracolumbar Injury Classification and Severity Scale (TLICS) scoring system, burst fracture is assigned 2 points, distraction is assigned 4 points, incomplete cord injury is assigned 3 points, posterior ligamentous complex (PLC) injury is assigned 3 points, and translation/rotation is assigned 3 points. (B) Initial answer provided by ChatGPT 4. (C) Initial answer provided ChatGPT 3.5. (D) Corrected answer provided by ChatGPT 3.5 through the "Regenerate response" function.

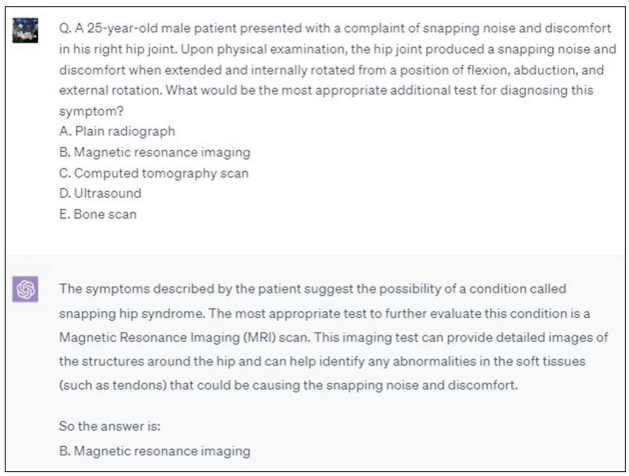

Fig. 3. Example of a question requiring inference incorrectly answered by ChatGPT 4. In this case, internal snapping of the hip is suspected. The most appropriate diagnostic test for the presented case would be option "D. Ultrasound." However, ChatGPT provides "B. Magnetic resonance imaging" as the correct answer, despite the individual sentence in its explanation being correct.

whereas ChatGPT 4 displayed a more consistent performance, with only 15 of its initial responses (9.4%) undergoing modification.

DISCUSSION

The most important finding of this study is that ChatGPT 4 showed a problem-solving proficiency to pass the orthopedic specialist examination, whereas ChatGPT 3.5 did not reach the passing standard. Previous studies reported that the accuracy rates of ChatGPT 4 exceeded the passing threshold of medical board certification examinations,[5-7] which is comparable to the results of this study. In terms of orthopedic question performance, accuracy has varied from 29% to 73.6% in existing research.[11-13] In our study, we attribute the improved performance of ChatGPT 4 (with an accuracy of 60%) over its predecessor, ChatGPT 3.5 (with an accuracy rate of 37.5%), to the continuous enhancement of LLM models.[15] Interestingly, despite the language of the input questions being Korean, ChatGPT translated them into English and generated responses efficiently, showing its multilingual proficiency.

In this study, there were instances where both ChatGPT 4 and 3.5 provided incorrect answers to simple, knowledge-based questions (Fig. 2). The "regenerate response" function occasionally changed incorrect answers to correct ones, and vice versa, without providing an explanation for these changes. This inconsistency was especially pronounced in ChatGPT 3.5, where 47.5% of answers underwent changes upon regeneration, whereas ChatGPT 4 exhibited greater consistency, with only 9.4% of responses being altered. In addition to the inaccuracies shown by both ChatGPT 4 and 3.5, these inconsistency rates raise questions about their ability to reliably provide information. Furthermore, as ChatGPT 4 displayed a lower accuracy rate in the tumor subcategory, the superiority of ChatGPT 4 over ChatGPT 3.5 may not be generalized across all subject areas.

When solving problems requiring inference from given scenarios, ChatGPT demonstrated quick reasoning abilities. However, there were instances where ChatGPT's answers and explanations contained incorrect information, despite the individual sentences being mostly accurate (Fig. 3). In addition, incorrect explanations were generated with confidence, a known limitation of LLMs.[16] In some cases, ChatGPT generated nonexistent references to support its answers, known as "hallucination."[17] This finding implies the necessity for the expertise and review of a medical professional for verification. Therefore, although ChatGPT's performance proved sufficient to pass an orthopedic board-style examination, it should be used with caution in clinical settings or for educational purposes.

Meanwhile, it is important to note that ChatGPT was not designed to replace medical professionals as a diagnostic tool, but rather to assist users by providing information and supporting various tasks.[15] Additionally, it is pertinent to acknowledge that ChatGPT's training corpus only extends up to 2021 (latest version when the authors conducted the analysis), warranting consideration when utilizing the program for medical inquiries in current scenarios. Moreover, as the responses of ChatGPT are generated from large-scale text patterns, limitations in its judgement capabilities should be considered.[18]

Limitations of this study include the use of text-only questions without incorporating imaging tests commonly used in actual clinical settings. Additionally, considering that ChatGPT is predominantly trained in English, there could be a difference in accuracy when inputting and solving problems in a non-English language. Although ChatGPT automatically translated the questions into English, and these translations were subsequently verified by the authors, the translation process itself may have impacted the performance. Furthermore, the responses provided by ChatGPT lacked solid evidence and did not include concrete, reliable references. Lastly, as the evaluation was conducted using a question bank, and actual specialist examination questions are not publicly available, there might have been differences in the difficulty level.

In conclusion, ChatGPT demonstrated the potential to pass orthopedic board-style examinations, with

ChatGPT 4 achieving a significantly higher accuracy rate compared to ChatGPT 3.5. However, there were instances of incorrect answers and changes in responses. Therefore, caution is required when applying ChatGPT in clinical settings or for educational purposes.

CONFLICT OF INTEREST

No potential conflict of interest relevant to this article was reported.

ORCID

Sung Eun Kim https://orcid.org/0000-0002-1954-9875
Ji Han Lee https://orcid.org/0000-0002-6363-0574
Byung Sun Choi https://orcid.org/0000-0002-4492-4358
Hyuk-Soo Han https://orcid.org/0000-0003-1229-8863
Myung Chul Lee https://orcid.org/0000-0002-8150-1573
Du Hyun Ro https://orcid.org/0000-0001-6199-908X

REFERENCES

1. Zhu J, Shen B, Abbasi A, Hoshmand-Kochi M, Li H, Duong TQ. Deep transfer learning artificial intelligence accurately stages COVID-19 lung disease severity on portable chest radiographs. PLoS One. 2020;15(7):e0236621.
2. Krusche M, Callhoff J, Knitza J, Ruffer N. Diagnostic accuracy of a large language model in rheumatology: comparison of physician and ChatGPT-4. Rheumatol Int. 2024;44(2):303-6.
3. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183(6):589-96.
4. 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.
5. Bhayana R, Bleakney RR, Krishna S. GPT-4 in radiology: improvements in advanced reasoning. Radiology. 2023;307(5):e230987.
6. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. Radiology. 2023;307(5):e230582.
7. Hopkins BS, Nguyen VN, Dallas J, et al. ChatGPT versus the neurosurgical written boards: a comparative analysis of artificial intelligence/machine learning performance on neurosurgical board-style questions. J Neurosurg. 2023;139(3):904-11.
8. Barbour AB, Barbour TA. A radiation oncology board exam of ChatGPT. Cureus. 2023;15(9):e44541.
9. Cheong RC, Pang KP, Unadkat S, et al. Performance of artificial intelligence chatbots in sleep medicine certification board exams: ChatGPT versus Google Bard. Eur Arch Otorhinolaryngol. 2023 Dec 20 [Epub]. https://doi.org/10.1007/s00405-023-08381-3
10. Sakai D, Maeda T, Ozaki A, Kanda GN, Kurimoto Y, Takahashi M. Performance of ChatGPT in board examinations for specialists in the Japanese Ophthalmology Society. Cureus. 2023;15(12):e49903.
11. 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(3):e23.00056.
12. 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(23):1173-9.
13. Lum ZC. Can artificial intelligence pass the American board of orthopaedic surgery examination?: orthopaedic residents versus ChatGPT. Clin Orthop Relat Res. 2023;481(8):1623-30.
14. Wu T, He S, Liu J, et al. A brief overview of ChatGPT: the history, status quo and potential future development. IEEE/CAA J Autom Sin. 2023;10(5):1122-36.
15. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023;388(13):1233-9.
16. Xiao Y, Wang WY. On hallucination and predictive uncertainty in conditional language generation. arXiv [Preprint]. 2021 [cited 2024 Jan 23]. Available from: https://doi.org/10.48550/arXiv.2103.15025
17. Beutel G, Geerits E, Kielstein JT. Artificial hallucination: GPT on LSD? Crit Care. 2023;27(1):148.
18. Alberts IL, Mercolli L, Pyka T, et al. Large language models (LLM) and ChatGPT: what will the impact on nuclear medicine be? Eur J Nucl Med Mol Imaging. 2023;50(6):1549-52.