Journal of Experimental and Clinical Medicine
https://dergipark.org.tr/omujecm

Research Article

J Exp Clin Med
2025; 42(1): 40-42
doi: 10.52142/omujecm.42.1.8

Comparison of ChatGPT-3.5 and Google Bard Performance on Turkish Orthopaedics and Traumatology National Board Examination

Murat KORKMAZ *, Abdullah KAHRAMAN

Department of Orthopedics and Traumatology, Istanbul Faculty of Medicine, Istanbul University, Istanbul, Türkiye

Received: 21.11.2024 • **Accepted/Published Online:** 29.11.2024 • **Final Version:** 28.03.2025

Abstract

This study ia a cross-sectional study to evaluate and compare the responses of two chatbots to compare the performance of ChatGPT-3.5 and Google Bard on the Turkish Orthopaedics and Traumatology National Board Examination. The questions of the Turkish Orthopaedics and Traumatology National Board Examination were asked to the chatbots one by one to have them indicate what the correct answer was and determine the difficulty level of the questions. The examination consists of 100 questions; 92 were included in the study. It was found that ChatGPT-3.5 answered 54.3% of the questions correctly, while Google Bard answered 45.7% of the questions correctly. When the correlation of difficulty and accuracy between the two AI models was evaluated, it was found that both were poorly correlated between the two different AI models (r=0.290 and p=0.005 for difficulty; r=0.314 and p=0.002 for accuracy). Both language models showed about 50% success on the Turkish Orthopaedics and Traumatology National Board Examination. Both found similar levels of difficulty in the questions.

Keywords: accuracy, Bard, ChatGPT-3.5, difficulty, orthopedics

1. Introduction

The scope of artificial intelligence (AI) models in medicine is gradually increasing (1-3). These models, such as ChatGPT and Google Bard, are supported by studies that show success in many areas, such as clinical decision-making, disease diagnosis, imaging of complex conditions, and medical planning (4). In the field of orthopaedics, AI models have several functions, such as suggesting medical treatment, analysing surgical cases, and assisting in teaching (5). For example, predicting early mortality in patients with critical fractures (6), analysing treatment effects in disc herniations using CT images based on AI algorithms (7), and using AI as a learning aid in orthopaedic education for residents (8) are some of the uses of these technologies. Studies have investigated the ability to answer questions correctly in assistant-level board examinations in various fields of medicine (1, 9). Assessing the performance of these AI models in specialised board-style examination questions is very important for understanding and evaluating their clinical utility (1).

This study analysed the answers and comments of ChatGPT 3.5 and Google Bard, both AI models, to the questions of the 1st phase of the Turkish National Board Examination of 2024, which measures national competence in the field of orthopaedics and traumatology, and aimed to compare their performance.

2. Materials and Methods

This study is a comparative, cross-sectional study that evaluates and compares the performance of ChatGPT 3.5 and Google Bard, two AI speech models, on the 2024 Turkish National Board Examination in Orthopedics and Traumatology. The exam consists of 100 questions assessing general orthopaedic knowledge. Eight questions were excluded from the study because they contained photographs. The exam questions were presented individually to both models by two different people. The following introductory sentence was added before each question: "The following question is a national board-level exam question in the field of orthopaedics and traumatology. You are expected to read the question and rate its difficulty as "easy, medium, difficult" and give the correct answer." The performance of each AI model was evaluated by comparing the proportion of correct answers they gave to the questions, their accuracy rates and the level of difficulty they recognised.

2.1. Statistical analysis

Analyses were performed using SPSS v.26 (IBM Corp.,

*Correspondence: muratkorkmaz.md@gmail.com

Armonk, NY, USA). The chi-square and Fisher's exact tests were used for categorical variables. Independent samples t-test was used for analysis between groups. Pearson's correlation coefficient was used for correlation analysis. The degree of correlation was evaluated according to the coefficient values: r=0.81-1.0 means 'excellent', r=0.61-0.80 means 'very good', r=0.41-0.60 means 'good', r=0.21-0.40 means 'moderate', and r=0.0-0.20 means 'poor' (10, 11). Statistical significance was accepted as $p \leq 0.05$ in all tests.

3. Results

The responses of two different AI models to the Turkish Orthopedics and Traumatology National Board Examination were evaluated. It was found that ChatGPT 3.5 answered 54.3% of the questions correctly, while Bard answered 45.7% of the questions correctly. There was no significant difference between the two groups in the accuracy of the AI models' answers to the questions (p=0.241). When assessing the difficulty of the questions posed by the AI models, ChatGPT 3.5 reported that 3.3% of the questions were easy (n=3), 88% were medium, and 8.7% were difficult. Bard, on the other hand, reported that 3.3% of the questions were easy (n=3), 90.2% were medium, and 6.5% were hard. No significant difference was found between the two AI models in determining the level of difficulty (p=0.654). When the relationship between the accuracy of the answers given and the difficulty of the questions was evaluated within the group, no statistically significant difference was observed between the results of ChatGPT and Bard (p=0.541 and 0.611, respectively) (Table 1). When the correlation between difficulty level and accuracy rate was evaluated between the two AI models, it was found that both were correlated at a low level between two different AI models (r=0.290 and p=0.005 for difficulty level; r=0.314 and p=0.002 for accuracy rate) (Table 2).

Table 1. Assessment of initial artificial intelligence responses by difficulty level as determined by the authors

	Chat GPT 3.5			**Gemini**			***pa***
	Incorrect n(%)	*Correct n(%)*	p^c	*Incorrect n(%)*	*Correct n(%)*	p^c	
Easy	2 (66.7%)	1 (33.3%)		2 (66.7%)	1 (33.3%)		
Medium	36 (44.4%)	45 (55.6%)	0.541	45 (54.2%)	38 (45.8%)	0.611	0.654
Hard	4 (50.0%)	4 (50.0%)		3 (50.0%)	3 (50.0%)		
***p**b*	0.241						

p^a: independent samples t-test for difficulty levels of the questions between two groups, p^b: independent samples t-test for accuracy of the questions between two groups, p^c: chi-square test for analyzing correct answer rate by difficulty category

Table 2. Correlations of difficulty levels and accuracy rates between ChatGPT and Gemini answers

	Difficulty levels (Gemini)	Accuracy rates (Gemini)
Difficulty levels (ChatGPT) *r* *p*	0.290 0.005*	0.146 0.165
Accuracy rates (ChatGPT) *r* *p*	0.185 0.078	0.314 0.002*

4. Discussion

To the best of our knowledge, this is the first study to compare the performance of ChatGPT-3.5 and Google Bard on national board-level questions in Orthopedics and Traumatology. Both models performed similarly, but chatGPT-3.5 led with a success rate of 54.3%. Both AI language models also found similar levels of difficulty in the questions. Moderate correlations were found between the accuracy rates of the two AI models, as well as between their difficulty levels.

ChatGPT-3.5 and ChatGPT-4 are both known to be successful in important tests. Studies have shown that they successfully pass MBBS and the United States Medical Licencing Exam (USMLE) Steps 1 and 2 (12, 13).

Lum compared the performance of the chatbot and orthopaedic residents on the American Board of Orthopedic Surgery exam and found that ChatGPT answered 47% of the questions correctly. When the response rate was compared according to the duration of orthopaedic specialisation, it was found to be similar for the first-year residents. As the difficulty of the questions increased, the ability to give correct answers decreased (14). Sparks et al. assessed the orthopaedic knowledge of ChatGPT-3.5 on orthopaedic board-style questions using the Orthobullets dataset and found a pass rate of 55.9%. It was reported that the performance of ChatGPT-3.5 on this exam was between the average performance of an intern and a second-year resident (15). Similarly, our study found that ChatGPT answered about half of the national exam questions correctly. This shows that it was not very successful in answering the questions. Studies have reported that chatgpt's low level of judgment and limited logical reasoning ability are the reasons for its inability to choose the appropriate response in clinical scenarios. (1, 16)

Traoré et al. (17) evaluated ChatGPT-3.5's answers to the European Board of Hand Surgery (EBHS) diploma exam, while Thibaut et al. (18) evaluated Google Bard's answers to

the same questions; both studies found that neither ChatGPT nor Bard could pass the first part of the EBHS diploma exam. In our study, the success rate of both AI models was similar, although Bard gave a lower rate of correct answers. The low performance of chatbots in board exams can be explained by the lack of modelling by engineers to train bots in orthopaedics and even medicine, and the lack of resources.

The study has some limitations. One of them is the use of ChatGPT-3.5 instead of ChatGPT-4, which is a more recent version. However, ChatGPT-4 is limited in use because it is paid, and ChatGPT-3.5 is more easily accessible to everyone. Additionally, this study can be considered as a preliminary study to show the differences between GPT-3.5 and GPT-4 by evaluating the performance of GPT-4 in future studies. Secondly, these chatbots cannot analyse videos or images; therefore, the questions with images were not included in this study.

ChatGPT-3.5 and Google Bard had similar performances in answering the Turkish Orthopaedics and Traumatology National Board Examination, but chatGPT-3.5 led with a success rate of 54.3%. The two AI language models also found similar levels of difficulty in the questions. Moderate correlations were found between the accuracy rates of the two AI models, as well as between their difficulty levels.

Conflict of interest

We certify that there is no conflict of interest with any financial organization regarding the manuscript.

Funding

None.

Acknowledgments

I thank to Assoc. Prof. Merve Damla Korkmaz for her expertise in statistical analysis.

Authors' contributions

Concept: M.K., A.K., Design: M.K., A.K., Data Collection or Processing: M.K., A.K., Analysis or Interpretation: M.K., Literature Search: M.K., A.K., Writing: M.K., A.K.

References

1. Menekşeoğlu AK, İş EE. Comparative performance of artificial ıntelligence models in physical medicine and rehabilitation board-level questions. Rev Assoc Med Bras (1992). 2024;70(7):e20240241.
2. Mejia MR, Arroyave JS, Saturno M, Ndjonko LCM, Zaidat B, Rajjoub R, et al. Use of ChatGPT for Determining Clinical and Surgical Treatment of Lumbar Disc Herniation With Radiculopathy: A North American Spine Society Guideline Comparison. Neurospine. 2024;21(1):149-58.
3. Chang MC. Use of artificial intelligence in the field of pain medicine. World J Clin Cases. 2024;12(2):236-9.
4. Sancheti P, Bijlani N, Shyam A, Yerudkar A, Lunawat R. ORTHO AI : World's First ARTIFICIAL INTELLIGENCE IN ORTHOPAEDICS. J Orthop Case Rep. 2023;13(12):178-9.
5. Chatterjee S, Bhattacharya M, Pal S, Lee SS, Chakraborty C. ChatGPT and large language models in orthopedics: from education and surgery to research. J Exp Orthop. 2023;10(1):128.
6. Han T, Xiong F, Sun B, Zhong L, Han Z, Lei M. Development and validation of an artificial intelligence mobile application for predicting 30-day mortality in critically ill patients with orthopaedic trauma. Int J Med Inform. 2024 Apr;184:105383.
7. Fan X, Qiao X, Wang Z, Jiang L, Liu Y, Sun Q. Artificial Intelligence-Based CT Imaging on Diagnosis of Patients with Lumbar Disc Herniation by Scalpel Treatment. Comput Intell Neurosci. 2022 May 27;2022:3688630.
8. Gan W, Ouyang J, Li H, Xue Z, Zhang Y, Dong Q, Huang J, Zheng X, Zhang Y. Integrating ChatGPT in Orthopedic Education for Medical Undergraduates: Randomized Controlled Trial. J Med Internet Res. 2024 Aug 20;26:e57037.
9. Khan AA, Yunus R, Sohail M, Rehman TA, Saeed S, Bu Y, et al. Artificial Intelligence for Anesthesiology Board-Style Examination Questions: Role of Large Language Models. J Cardiothorac Vasc Anesth. 2024;38(5):1251-9.
10. Fayers PM, Machin D. Quality of life: The assessment, analysis and reporting of patient-reported outcomes: John Wiley & Sons; 2015.
11. Korkmaz MD, Korkmaz M, Altın YF, Akgül T. Adaptation and validation of the Turkish version of the Quality of Life Profile for Spinal Deformities in idiopathic scoliosis. Acta Orthop Traumatol Turc. 2024;58(3):182-6.
12. Subramani M, Jaleel I, Krishna Mohan S. Evaluating the performance of ChatGPT in medical physiology university examination of phase I MBBS. Adv Physiol Educ. 2023;47(2):270-1.
13. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023;9:e45312.
14. 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.
15. Sparks CA, Kraeutler MJ, Chester GA, Contrada EV, Zhu E, Fasulo SM, et al. Inadequate Performance of ChatGPT on Orthopedic Board-Style Written Exams. Cureus. 2024;16(6):e62643.
16. Cuthbert R, Simpson AI. Artificial intelligence in orthopaedics: can Chat Generative Pre-trained Transformer (ChatGPT) pass Section 1 of the Fellowship of the Royal College of Surgeons (Trauma & Orthopaedics) examination? Postgrad Med J. 2023;99(1176):1110–1114.
17. Traoré SY, Goetsch T, Muller B, Dabbagh A, Liverneaux PA. Is ChatGPT able to pass the first part of the European Board of Hand Surgery diploma examination? Hand Surg Rehabil. 2023;42(4):362-4.
18. 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):151-8.