eISSN 2255-0569

ORIGINAL

How does ChatGPT perform on the European Board of Orthopedics and Traumatology examination? A comparative study

¿Cuál es el rendimiento de ChatGPT en el examen del Consejo Europeo de Ortopedia y Traumatología? Estudio comparativo

Sait Anıl Ulus

1. Dicle University, Faculty of Medicine, Department of Orthopedics and Traumatology, Diyarbakır, Turkey.

Corresponding author
Sait Anıl Ulus
E-mail: anil_ulus@hotmail.com

Received: 5 - VI - 2023
Accepted: 8 - VII - 2023

doi: 10.3306/AJHS.2023.38.06.43

Abstract

Objective: The objective of this investigation was to comprehensively compare the precision and responsiveness of GPT-3.5 and GPT-4 within the domain of Orthopedics and Traumatology. Specifically, our aim was to evaluate their capacity to provide accurate answers to a series of sample questions derived from the European Board of Orthopedics and Traumatology (EBOT) exam.
Methods: The study was conducted over the period from 10th May 2023 to 15th May 2023. It involved a comparative analysis of two AI language models, namely GPT-3.5 and GPT-4, specifically in the field of Orthopedics and Traumatology and with a focus on sample questions extracted from the EBOT exam. Two separate sets, each containing 80 sample questions (totaling 160 questions), were compiled from the pool of available EBOT sample questions.
Results: A total of 160 questions were included in the study, 80 questions from each group. In the field of General Orthopedics, GPT-4 demonstrated a higher success rate (75%) compared to GPT-3.5 (45%) (p=0.053). In the Traumatology domain, GPT-4 delivered a notable success rate of 80%, compared to GPT-3.5's (p=0.010). For Oncological Orthopedic Surgery, both models showed a similar trend (P=0.057). Overall, GPT-4 exhibited superior performance across all domains, with a cumulative success rate of 75% as compared to GPT-3.5's 43.75% (p<0.001). When considering the overall responses, the Odds Ratio between GPT-4 and GPT-3.5 was determined to be 3.8.
Conclusions: Based on the findings of this comparative study, it can be firmly concluded that GPT-4 demonstrates a remarkable superiority over GPT-3.5 in effectively addressing the EBOT exam sample questions.

Key words: ChatGPT, Orthopedics, Traumatology, EBOT, exam.

Resumen

Objetivo: El objetivo de esta investigación fue comparar exhaustivamente la precisión y capacidad de respuesta de la GPT-3.5 y la GPT-4 en el ámbito de la Ortopedia y la Traumatología. En concreto, nuestro objetivo era evaluar su capacidad para proporcionar respuestas precisas a una serie de preguntas de muestra derivadas del examen del European Board of Orthopedics and Traumatology (EBOT).
Métodos: El estudio se llevó a cabo durante el periodo comprendido entre el 10 de mayo de 2023 y el 15 de mayo de 2023. Consistió en un análisis comparativo de dos modelos lingüísticos de IA, a saber, GPT-3.5 y GPT-4, específicamente en el campo de la ortopedia y la traumatología y centrándose en preguntas de muestra extraídas del examen EBOT. Se recopilaron dos conjuntos distintos, cada uno de los cuales contenía 80 preguntas de muestra (en total 160 preguntas), a partir del conjunto de preguntas de muestra disponibles del EBOT.
Resultados: Se incluyeron en el estudio un total de 160 preguntas, 80 preguntas de cada grupo. En el ámbito de la Ortopedia general, la GPT-4 demostró una mayor tasa de aciertos (75%) en comparación con la GPT-3.5 (45%) (p=0,053). En el ámbito de la Traumatología, la GPT-4 obtuvo un notable índice de éxito del 80%, en comparación con la GPT-3.5 (p=0,010). En Cirugía Ortopédica Oncológica, ambos modelos mostraron una tendencia similar (p=0,057). En general, GPT-4 mostró un rendimiento superior en todos los dominios, con una tasa de éxito acumulada del 75% en comparación con el 43,75% de GPT-3.5 (p<0,001). Al considerar las respuestas globales, se determinó que la Odds Ratio entre la GPT-4 y la GPT-3.5 era de 3,8.
Conclusiones: Sobre la base de los resultados de este estudio comparativo, se puede concluir firmemente que la GPT-4 demuestra una notable superioridad sobre la GPT-3.5 a la hora de abordar eficazmente las preguntas de muestra del examen EBOT.

Palabras clave: ChatGPT, Ortopedia, Traumatología, EBOT, examen.

Cite as: Ulus SA. How does ChatGPT perform on the European Board of Orthopedics and Traumatology examination? A comparative study. *Academic Journal of Health Sciences 2023*; 38 (6):43-6 doi: 10.3306/AJHS.2023.38.06.43

Introduction

The ever-evolving field of Artificial Intelligence (AI) has profoundly impacted numerous domains, manifesting an extraordinary ability to innovate and redefine the way we assimilate and apply knowledge[1]. Among these, the Chat Generative Pre-trained Transformer (ChatGPT) holds a distinctive place as a leading-edge tool in the realm of Natural Language Processing (NLP). This AI-driven model, trained on copious amounts of data, has displayed unparalleled proficiency in generating remarkably human-like responses to conversational text prompts[2,3].

Originally released to the public in November 2022, ChatGPT, initially powered by the GPT-3.5 language model, garnered widespread recognition and interest due to its impressive capabilities[4]. Merely months later, in March 2023, the introduction of an upgraded iteration known as GPT-4 promised enhanced precision, further fueling the excitement around this technology. Nevertheless, amidst this enthusiasm, a critical debate persists regarding the suitability and effectiveness of deploying ChatGPT within the field of Medicine[5].

The infusion of AI into medical practice, and more specifically into Orthopedics and Traumatology, presents a cutting-edge avenue to probe, interpret, and devise solutions to intricate surgical dilemmas. OpenAI's language models, GPT-3.5 and GPT-4, have already shown promising potential across various disciplines, including medicine. However, a comprehensive evaluation of their proficiency in responding accurately to professional, domain-specific inquiries is a research area yet to be extensively explored[6,7].

This study endeavors to undertake a meticulous examination of the performance of these two AI models within the context of Orthopedics and Traumatology surgery. We aim to gauge their accuracy and responsiveness by assessing their ability to correctly respond to a series of sample questions typically presented in the European Board of Orthopedics and Traumatology (EBOT) exam questions.

Materials and Methods

In this comparative study, we sought to evaluate the performance of two language models, GPT-3.5 and GPT-4, in the field of Orthopedics and Traumatology. Our analysis focused on assessing the correctness of the models' responses to a series of questions related to four distinct domains: General Orthopedics, Traumatology, Oncological Orthopedic Surgery, and Pediatric Orthopedic Surgery.

A total of 160 questions were included in this comparative analysis, split evenly between the two models with 80 questions assigned randomly to each (Sample questions: **figure 1** and **figure 2**). The choice of questions aimed to cover a diverse array of topics within each domain to provide a comprehensive examination of the models' understanding and knowledge. The replies generated by both AI models were meticulously evaluated by specialists in the field of Orthopedics and Traumatology. Their assessments were grounded on the accuracy and pertinence of each response to the posed question.

After the thorough evaluation and scoring procedure, the count of correct responses from each model was tallied. These cumulative correct answers were subsequently converted into a percentage of the total questions, thereby yielding an accuracy score for each model. By comparing the accuracy scores of GPT-3.5 and GPT-4, we aimed to identify the model that exhibited superior performance in addressing the sample questions from the EBOT exams.

Figure 1: Sample question-1.

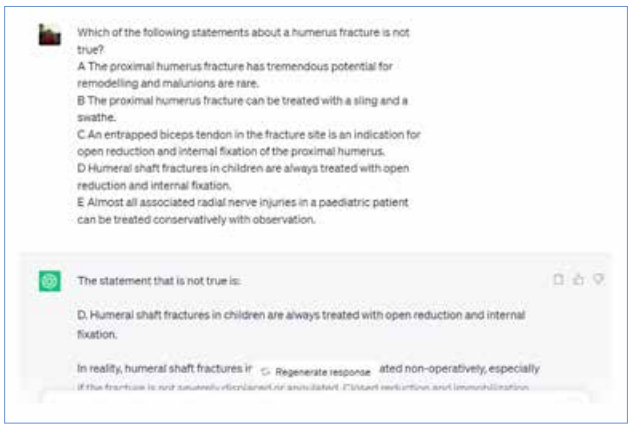


Figure 2: Sample question-2.

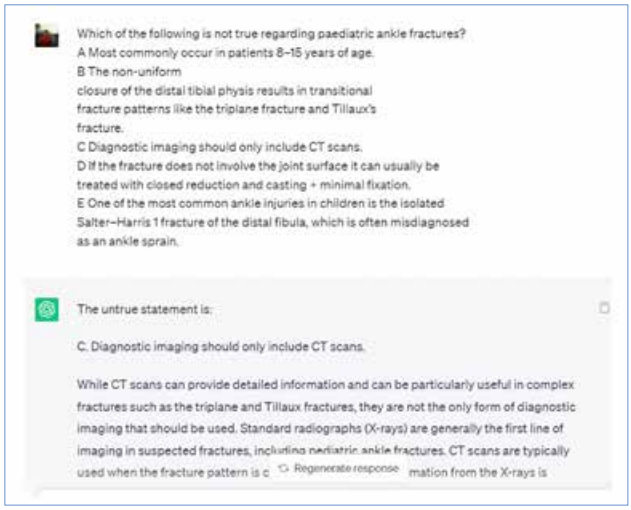
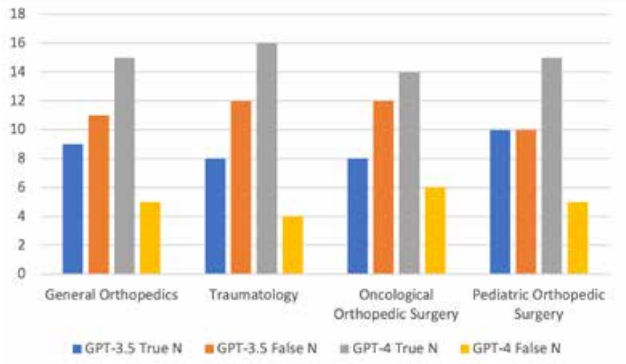


Table I: Comparison of GPT-3.5 and GPT-5 in terms of answers number.

	GPT-3.5				GPT-4				
	True		False		True		False		
	N	%	N	%	N	%	N	%	*p*-value
General Orthopedics	9	45,0	11	55,0	15	75,0	5	25,0	0.053
Traumatology	8	40,0	12	60,0	16	80,0	4	20,0	0.010
Oncological Orthopedic Surgery	8	40,0	12	60,0	14	70,0	6	30,0	0.057
Pediatric Orthopedic Surgery	10	50,0	10	50,0	15	75,0	5	25,0	0.102
Total	35	43,75	45	56,25	60	75,00	20	25,00	<0.001

Figure 3: Analyzing the relative frequency of the answers.



For the inclusion criteria, we incorporated a selection of questions from the EBOT study questions, chosen randomly. On the other hand, questions that involved images were excluded from the study due to the limitations of the GPT program, which does not have the capability to process visual content.

Statistical analysis

Descriptive statistics, including frequencies and other relevant characteristics, were employed to analyze the data for all items. Categorical variables were evaluated using the Chi-square test, and in certain instances, the Fisher exact test was utilized. The statistical analyses were conducted using IBM SPSS Statistics for Windows, Version 24.0 (IBM Corp., Armonk, NY, USA). Two-sided p-values were calculated, and statistical significance was determined by considering p-values below 0.05.

Results

The results of the study comparing the performance of GPT-3.5 and GPT-4 are presented in **table I**. The data includes the number and percentage of true and false responses for each domain, namely General Orthopedics, Traumatology, Oncological Orthopedic Surgery, and Pediatric Orthopedic Surgery. The total number and percentage of true and false responses across all domains are also provided.

A total of 160 questions were included in the study, 80 questions from each group. In the field of General Orthopedics, GPT-4 demonstrated a higher success rate (75%) compared to GPT-3.5 (45%). Although this indicates a substantial improvement, the *p*-value of 0.053 suggests that this difference might have occurred by chance and was marginally above the commonly accepted statistical significance level ($p<0.05$). In the Traumatology domain, GPT-4 delivered a notable success rate of 80%, compared to GPT-3.5's 40% - a twofold increase ($p=0.010$). For Oncological Orthopedic Surgery, both models showed a similar trend, with GPT-4 achieving a success rate of 70% compared to GPT-3.5's 40% ($P=0.057$). Pediatric Orthopedic Surgery saw another considerable performance boost with GPT-4, with a success rate of 75% against GPT-3.5's 50% ($p=0.102$). Overall, GPT-4 exhibited superior performance across all domains, with a cumulative success rate of 75% as compared to GPT-3.5's 43.75%. This overall difference was found to be highly statistically significant with a *p*-value of less than 0.001 (**Table I**). The graph related to the comparison of GPT-3.5 and GPT-4 for each item's responses is provided in **figure 3**.

When considering the overall responses, the Odds Ratio between GPT-4 and GPT-3.5 was determined to be 3.8. This infers that GPT-4 had a 3.8-fold higher probability of correctly answering the queries related to Orthopedics and Traumatology compared to its predecessor, GPT-3.5.

Discussion

To the best of our knowledge, this study represents the first comprehensive comparison between GPT-3.5 and GPT-4 in their ability to respond to EBOT examination questions in the field of Orthopedics and Traumatology. The findings from this research highlight notable disparities in the accuracy and responsiveness of GPT-3.5 and GPT-4 when it comes to addressing domain-specific inquiries within the Orthopedics and Traumatology domain. Specifically, GPT-4 exhibited a statistically significant improvement in its ability to answer EBOT-related questions compared to its predecessor, GPT-3.5.

Previous studies in the field of medical question-answering research have often focused on more specific tasks, aiming to enhance model performance at the expense of generalizability[1,3-7]. For example, Jin et al.[8] achieved a 68.1% accuracy with their model that responded to yes-or-no questions, based on information available in the PubMed abstract corpus. However, the pursuit of more

versatile models has encountered several challenges. Jin et al.[9] achieved an accuracy of 36.7% on a dataset of 12,723 questions from Chinese medical licensing exams. In another study, ChatGPT demonstrated an accuracy of over 50% in all examinations, surpassing 60% in certain analyses[10]. The USMLE pass threshold, which may vary by year, is approximately 60%[10]. As a result, ChatGPT now approaches the range required for a passing score. Similarly, Ha et al.[11] reported a mere 29% accuracy on 454 USMLE Step 1 and Step 2 questions in 2019. Gilson and colleagues found that ChatGPT had a correct response rate of 58% when presented with questions from the United States Medical Licensing Examination Step 1 and Step 2 exams[12]. However, in our study, in the field of General Orthopedics, GPT-4 demonstrated a higher success rate (75%) compared to GPT-3.5 (45%) (p=0.053). In the Traumatology domain, GPT-4 delivered a notable success rate of 80%, compared to GPT-3.5's (p=0.010). For Oncological Orthopedic Surgery, both models showed a similar trend (P=0.057). Overall, GPT-4 exhibited superior performance across all domains, with a cumulative success rate of 75% as compared to GPT-3.5's 43.75% (p<0.001). When considering the overall responses, the Odds Ratio between GPT-4 and GPT-3.5 was determined to be 3.8.

This study has certain limitations that should be acknowledged. The ChatGPT algorithm was initially trained on a dataset compiled from information generated on or before 2021. As a result, the model's prompts can only incorporate knowledge available up until that point. Due to the closed nature of the model and the absence of a public application programming interface (API), we were unable to fine-tune it with task-specific data and explore the level of inherent stochasticity it possesses. Nevertheless, these limitations did not impede our analysis as the focus of this research was on evaluating ChatGPT's performance in the context of the EBOT exam. Thirdly, ChatGPT receives regular updates that incorporate user-provided inputs. The version of ChatGPT utilized in this study was an older iteration compared to the version released at the time of the study's completion. Considering all factors, it is reasonable to hypothesize that the model's performance will not significantly decline with each new iteration and may even improve when applied to the outlined task.

Conclusions

Based on the findings of this comparative study, it can be firmly concluded that GPT-4 demonstrates a remarkable superiority over GPT-3.5 in effectively addressing the EBOT exam sample questions with a successful rate (75%).

Funding
No

Conflict of interest
No

Acknowledgment
No

References

1. OpenAI. Introducing ChatGPT. Accessed from: https://openai.com/blog/chatgpt, Accessed May 12, 2023.

2. Parsa A, Ebrahimzadeh MH. ChatGPT in Medicine; a Disruptive Innovation or Just One Step Forward? Arch Bone Jt Surg. 2023;11(4):225-226. doi: 10.22038/abjs.2023.22042. PMID: 37180295; PMCID: PMC10167532.

3. Lee TC, Staller K, Botoman V, Pathipati MP, Varma S, Kuo B. ChatGPT Answers Common Patient Questions About Colonoscopy. Gastroenterology. 2023 May 5:S0016-5085(23)00704-7. doi: 10.1053/j.gastro.2023.04.033. Epub ahead of print. PMID: 37150470.

4. Shue E, Liu L, Li B, Feng Z, Li X, Hu G. Empowering Beginners in Bioinformatics with ChatGPT. bioRxiv [Preprint]. 2023 Mar 8:2023.03.07.531414. doi: 10.1101/2023.03.07.531414. PMID: 36945641; PMCID: PMC10028953.

5. Strong E, DiGiammarino A, Weng Y, Basaviah P, Hosamani P, Kumar A, et al. Performance of ChatGPT on free-response, clinical reasoning exams. medRxiv [Preprint]. 2023 Mar 29:2023.03.24.23287731. doi: 10.1101/2023.03.24.23287731. PMID: 37034742; PMCID: PMC10081420.

6. Hügle T. The wide range of opportunities for large language models such as ChatGPT in rheumatology. RMD Open. 2023 Apr;9(2):e003105. doi: 10.1136/rmdopen-2023-003105. PMID: 37116985; PMCID: PMC10151992.

7. Temsah MH, Jamal A, Aljamaan F, Al-Tawfiq JA, Al-Eyadhy A. ChatGPT-4 and the Global Burden of Disease Study: Advancing Personalized Healthcare Through Artificial Intelligence in Clinical and Translational Medicine. Cureus. 2023 May 23;15(5):e39384. doi: 10.7759/cureus.39384. PMID: 37223340; PMCID: PMC10204616.

8. Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. PubMedQA: a dataset for biomedical research question answering. arXiv doi: 10.48550/arXiv.1909.06146. Preprint posted online on September 13, 2019

9. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Applied Sciences. 2021;11(14):6421.

10. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198.

11. Ha LA, Yaneva V. Automatic question answering for medical MCQs: can it go further than information retrieval?. Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019); RANLP 2019; September 2-4, 2019; Varna, Bulgaria. 2019. pp. 418-2.

12. 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? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023;9:e45312.