

Performance of AI Chatbots in Adult Reconstruction Education: A Comparative Evaluation Using AAOS ResStudy Questions

TARISHI PARMAR, BS; AKSHAY KHANNA, BS; JONATHAN BERG, BS; ALEXANDER LINTON, MD; ALEC KELLISH, MD; LAUREN O'MARA, MD; STEPHANIE EBLE, MD; DAVID G. NAZARIAN, MD

ABSTRACT

BACKGROUND: Artificial intelligence (AI) chatbots are increasingly incorporated into orthopedic education, yet their ability to handle complex, clinically oriented questions remains unclear. This study evaluates the accuracy, response time, and citation reliability of contemporary AI chatbots when answering adult reconstruction questions from the AAOS ResStudy question bank.

METHODS: Fifty-five multiple-choice adult reconstruction questions were extracted and categorized by anatomic region (hip vs. knee), concept (clinical presentation, prognosis, management), and AAOS-defined difficulty. Seven chatbots (ChatGPT-3.5, ChatGPT-4, ChatGPT-4o, Gemini, Gemini Advanced, Claude, and Co-Pilot) were assessed. Accuracy, response time, and citation quality (existence and relevance) were recorded. Statistical analyses included chi-square testing, t-tests, Pearson correlation, one-way ANOVA, and Kruskal–Wallis testing ($p < 0.05$).

RESULTS: Overall chatbot accuracy was 67.5%. GPT-4o performed best (80%), outperforming AAOS members (75%), while Co-Pilot ranked second (70.9%). Accuracy varied by concept, with highest performance in Clinical Presentation (78.6%) and lowest in Prognosis (63.1%). Difficulty strongly influenced performance, with accuracy dropping from 84.7% on very easy questions to 45.4% on difficult ones. Average response time was 3.6 seconds, and no significant correlation existed between speed and accuracy. Among citation-producing chatbots, Co-Pilot achieved 100% citation existence and relevance.

CONCLUSION: Modern AI chatbots demonstrate moderate overall accuracy on adult reconstruction questions, with GPT-4o exhibiting the strongest performance. While chatbots excel at descriptive orthopedic knowledge, they continue to struggle with applied reasoning tasks such as prognosis and management. These findings support their use as supplementary, rather than primary, learning tools. Improving citation reliability and reasoning capabilities may enhance future educational applications.

LEVEL OF EVIDENCE: IV, Cross-Sectional Comparative Analysis

KEYWORDS: chatbots; orthopedic education; adult reconstruction; artificial intelligence; medical learning tools

INTRODUCTION

In recent years, the use of chatbots in education has surged, driven by their versatility and user-friendly design.¹ These AI-powered programs interact with users through dynamic, conversational queries.² Their integration into medical education has further expanded their popularity, highlighting their potential to enhance learning experiences. From automating tasks like summarizing and evaluating medical literature to developing interactive simulations, AI chatbots have diverse applications within medical education.³ Orthopedics, in specific, has shown a staggering tenfold increase in the use of interactive AI chatbots since 2010.⁴ From boasting about 80% accuracy when answering orthopedic trauma questions to having about 63.3% accuracy in answering hand surgery questions, several AI chatbots have shown promise as didactic tools in orthopedic education.^{5,6} On the other end of the spectrum, however, studies have been performed that render the accuracy of these bots' orthopedic knowledge questionable.⁷⁻¹⁰ One study indicated that while these AI bots have approximately the knowledge level of a PGY-2 orthopedic resident, they still fall short of passing the American Board of Orthopedic Surgery (ABOS) written exam, thus highlighting the need to be wary of blindly relying on them for orthopedic questions.⁸

As of late, the efficiency and accuracy of AI chatbots in orthopedics has garnered significant attention, becoming a prominent and rapidly expanding area of research.^{4-6,9,11-18} Some articles have specifically focused on the application of chatbots in hip and knee reconstruction.^{9,11-16,18,19} However, the majority of these articles concentrate on how AI chatbots can enhance patient understanding of hip and knee pathologies.^{9,11,12,14-16,18} Less is known about chatbots ability to answer clinical questions posed to medical professionals. After a comprehensive review of the current literature, we found no studies that evaluate the accuracy of different chatbots when answering adult reconstruction questions designed for orthopedic trainees.

This study therefore seeks to assess the performance of various AI chatbot models (ChatGPT-3.5, ChatGPT-4, ChatGPT-4o, Gemini, Gemini Advanced, Claude (Anthropic), and Co-Pilot) in answering adult reconstruction questions from the American Academy of Orthopaedic Surgeons (AAOS) question bank, ResStudy. The purpose of this investigation is to critically assess the performance of different

modern chatbots tasked with answering adult reconstruction questions and to evaluate their potential as a learning tool for orthopedic trainees.

METHODS

Question Selection and Categorization

Fifty-five multiple-choice questions on adult reconstruction topics were extracted from the AAOS Resident Question Bank. Questions containing images were excluded a priori, as most chatbot platforms lack the capability to reliably interpret visual data. From the remaining pool of eligible text-based questions, a random sample of 55 questions was generated for analysis. Random sampling was employed to minimize selection bias and enhance the representativeness of the dataset across a broad range of topics and difficulty levels within adult reconstruction. This approach also ensured methodological feasibility while maintaining consistency with prior studies evaluating AI chatbot performance in orthopedic education that have utilized similarly sized question sets.

They were then categorized by: 1) hip or knee-related content and 2) concept (clinical presentation, prognosis, and management). Concept categories were defined a priori by one investigator (AL) based on standard medical education frameworks: clinical presentation (questions focused on signs, symptoms, and diagnostic findings), prognosis (questions addressing disease course, outcomes, and risk factors), and management (questions concerning treatment decisions and interventions). Each question was assigned to a concept category by the same investigator, with categorization reviewed for consistency.

Difficulty was determined based on AAOS members' average performance: very easy (>90% correct), easy (80–90%), moderate (50–80%), and difficult (<50%).

Chatbot Testing and Data Collection

The questions were presented to ChatGPT-3.5, ChatGPT-4, ChatGPT-4o, Gemini, Gemini Advanced, Co-Pilot, and Claude. Accuracy and response times were recorded, with response time measured manually from prompt submission to the chatbot's first response.

While ChatGPT models could not provide citations, Gemini, Claude, and Co-Pilot did. Citation analysis was performed for all correct responses that included references. Three investigators (T.P., A.K., J.B.) independently verified citation existence (whether the cited reference could be located in standard medical databases) and citation relevance (defined as whether the referenced source was related to the clinical topic of the question being addressed). Discrepancies in citation assessment were resolved through consensus discussion among the three reviewers.

Statistical Analysis

Descriptive statistics analyzed chatbot performance by anatomic location, concept, difficulty, and response time. Chi-square tests compared accuracy, correlation tests examined accuracy vs. response speed, and t-tests compared knee vs. hip questions. One-way ANOVA assessed differences by difficulty and concept, while Kruskal-Wallis tested citation generation. A significance level of $p < 0.05$ was used, with all analyses conducted in IBM SPSS (2017). Given that all eligible non-image-based questions within the module were included, and the study was designed as an exploratory comparative analysis, a formal a priori power analysis was not performed.

RESULTS

Characteristics of Included Questions

Our study included 55 questions, with 36 focusing on the knee and 19 on the hip. Prognosis emerged as the most frequently assessed concept, encompassing 24 questions (43.6%), followed by management with 21 questions (38.2%), and clinical presentation with 10 questions (18.2%). Among knee-related questions, prognosis was the leading concept ($n=15$, 41.7%), while hip-related questions also predominantly explored prognosis ($n=9$, 47.4%). Regarding difficulty, the majority of the questions were categorized as moderate ($n=16$, 29.1%), while 14 questions (25.4%) were categorized as easy and another 14 (25.4%) as very easy, with 11 (20.0%) were classified as difficult.

Table 1 summarizes the distribution of the included questions by anatomic focus, conceptual emphasis, and difficulty level.

Table 1. Characteristics of Included Questions

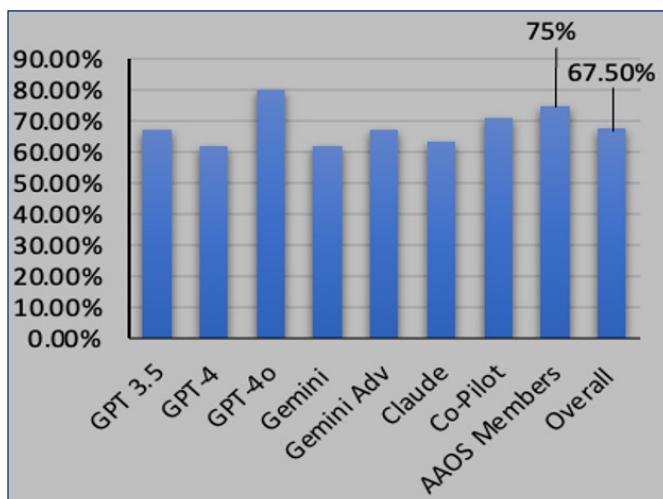
Characteristics of Questions (N=55) based on anatomic location, concept, and difficulty		
Anatomic Location	N	%
Knee	36	65.5
Hip	19	34.5
Concept	N	%
Clinical Presentation	10	18.2
Prognosis	24	43.6
Management	21	38.2
Difficulty	N	%
Very Easy	14	25.5
Easy	14	25.5
Moderate	16	29.1
Difficult	11	20.0

Overall Performance

The overall chatbot accuracy was 67.5%, with 37 out of 55 questions answered correctly. GPT-4o led (80.0%), followed by Co-Pilot (70.9%). ChatGPT 3.5 and Gemini Advanced each scored 67.3%, Claude 63.6%, and GPT-4 and Gemini had the lowest accuracy (61.8%). These accuracy differences were not statistically significant ($p=0.429$). AAOS members outperformed all bots except GPT-4o, averaging 75.0% accuracy.

The average response time was 3.6 seconds. Gemini Advanced was the slowest (5.8s), while GPT-3.5 was the fastest (2.5s). A weak, non-significant positive correlation was found between speed and accuracy ($r=0.218$, $p=0.639$) [Figure 1].

Figure 1. Accuracy Comparison of Chatbots and AAOS members



Performance According to Knee Versus Hip Questions

For knee-related questions, GPT-4o had the highest accuracy (83.3%), followed by GPT-3.5 and Gemini Advanced (72.2%). The lowest performers were GPT-4 and Gemini (63.9%). Co-Pilot and Claude scored 69.4% and 66.7%, respectively. Overall, chatbots answered 70.24% correctly, while AAOS members outperformed all except GPT-4o, averaging 73.6%.

For hip-related questions, Co-Pilot led (78.9%), followed by GPT-4o (73.7%). GPT-3.5, GPT-4, Gemini, and Gemini Advanced each scored 57.9%, while Claude had the lowest accuracy (52.6%). AAOS members again performed better, with 77.5% accuracy compared to the chatbots' 62.4%.

No significant differences were found in chatbot accuracy for pooled knee and hip questions ($p=0.112$; MD: 7.83) or response times ($p=0.898$; MD: 0.08). Gemini Advanced was the slowest (6.1s for knee, 5.3s for hip), while GPT-3.5 was the fastest (2.5s for knee, 2.6s for hip).

Performance According to Concept

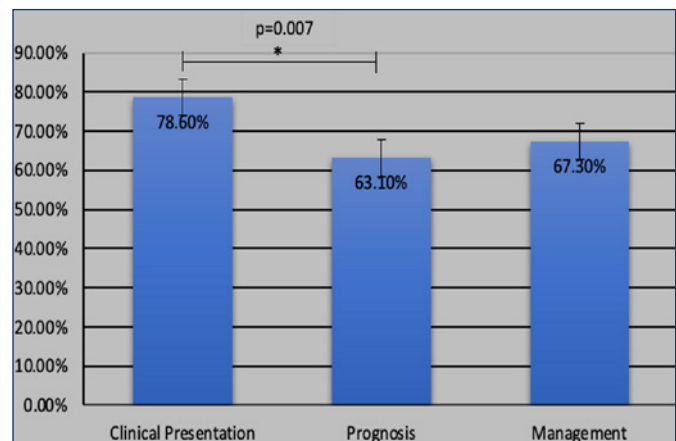
Mean accuracy was 78.6% for "Clinical Presentation," 63.1% for "Prognosis," and 67.3% for "Management."

For "Clinical Presentation," GPT-4o and Co-Pilot performed best (90%), while Gemini had the lowest accuracy (60%). In "Prognosis," GPT-4o led (79.2%), and GPT-4 performed worst (50%). For "Management," GPT-4 and GPT-4o scored highest (76.2%), while Claude and Co-Pilot had the lowest accuracy (57.1%). No significant differences were found among chatbot performances ($p=0.535$).

ANOVA showed a significant difference in chatbot accuracy across concepts ($p=0.007$). Performance on "Clinical Presentation" was significantly higher than "Prognosis" (95% C.I. = -29.07, -5.36), but no significant differences were observed between "Management" and the other two concepts.

No significant differences were found in response times across concepts ($p=0.432$). Average times were 3.7 sec for "Clinical Presentation," 3.6 sec for "Prognosis," and 3.9 sec for "Management." Gemini Advanced was the slowest overall, GPT-3.5 was fastest for "Prognosis" and Gemini was quickest for "Prognosis" and "Clinical Presentation" [Figure 2].

Figure 2. Overall Accuracy Comparison of Chatbots Based on Concept



Performance According to Difficulty

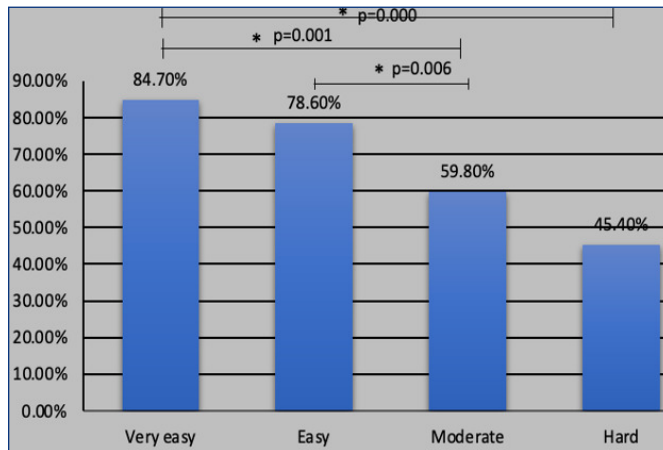
The average accuracy was 84.7% for very easy, 78.6% for easy, 59.8% for moderate, and 45.4% for hard questions.

For very easy questions, GPT-4o scored highest (100%), while Claude was lowest (71.4%). In easy questions, Co-Pilot performed best (92.9%), while Gemini scored lowest (64.3%). For moderate questions, GPT-4o, Claude, and Co-Pilot tied (68.8%), with Gemini lowest (43.7%). In hard questions, GPT-4o and Gemini led (63.6%), while GPT-4 was lowest (27.3%).

ANOVA showed no significant difference among chatbots ($p=0.097$). However, performance on very easy questions was significantly better than on hard ($p=0.000$) and moderate ($p=0.001$) ones. Easy questions also had significantly higher scores than moderate ones ($p=0.006$).

Chatbots averaged 3.6 seconds per question, with Gemini Advanced the slowest and GPT-3.5 the fastest [Figure 3].

Figure 3. Overall Accuracy Comparison of Chatbots Based on Question Difficulty



Literature Search Ability

Among the four bots providing citations, Gemini, Gemini Advanced, Claude, and Co-Pilot, the average citation existence was 84.5%, with an average relevance of 90.7%. Co-Pilot performed best, achieving 100% in both categories,

followed by Gemini Advanced (85.3% existence, 82.8% relevance). Gemini had 79.4% existence but 100% relevance, while Claude had the lowest existence (73.5%) and 80% relevance.

Kruskal-Wallis analysis showed no significant differences in citation existence ($H=3.00$, $p=0.392$) or relevance ($H=2.70$, $p=0.440$) across bots [Figure 4, Table 2].

Figure 4. Comparison of Chatbots Based on Citation Generating Ability

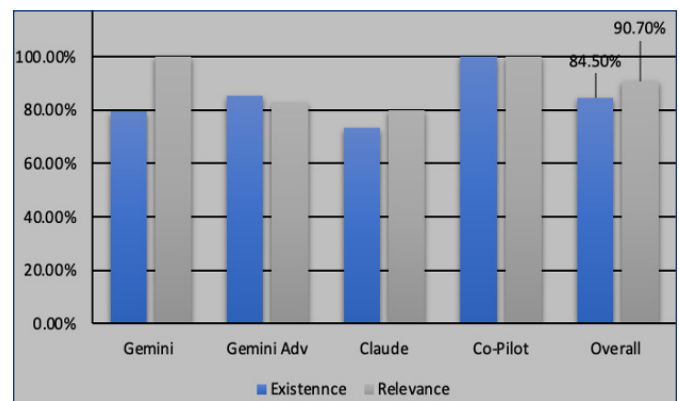


Table 2. Summary of Chatbot Performance

Category	Measure	Best Performer	Worst Performer	Mean
Overall Performance	Accuracy	80.0% (GPT-4o)	61.8% (GPT-4, Gemini)	67.5%
	Response Time	2.5s (GPT-3.5)	5.8s (Gemini Advanced)	3.6s
Knee vs Hip Questions	Knee Accuracy	83.3% (GPT-4o)	63.9% (GPT-4, Gemini)	70.24%
	Hip Accuracy	78.9% (Co-Pilot)	52.6% (Claude)	62.4%
	Knee Response Time	2.5s (GPT-3.5)	6.1s (Gemini Advanced)	3.6s
	Hip Response Time	2.6s (GPT-3.5)	5.3s (Gemini Advanced)	3.6s
Concept Performance	Clinical Presentation Accuracy	90% (GPT-4o, Co-Pilot)	60.0% (Gemini)	78.6%
	Prognosis Accuracy	79.2% (GPT-4o)	50.0% (GPT-4)	63.1%
	Management Accuracy	76.2% (GPT-4, GPT-4o)	57.1% (Claude, Co-Pilot)	67.3%
	Clinical Presentation Response Time	2.4s (Gemini)	5.2s (Gemini Advanced)	3.7s
	Prognosis Response Time	2.6s (GPT-3.5)	5.7s (Gemini Advanced)	3.6s
	Management Response Time	2.4s (Gemini)	6.3s (Gemini Advanced)	3.9s
Difficulty Performance	Very Easy Accuracy	100% (GPT-4o)	71.4% (Claude)	84.7%
	Easy Accuracy	92.9% (Co-Pilot)	64.3% (Gemini)	78.6%
	Moderate Accuracy	68.8% (GPT-4o, Claude, Co-Pilot)	43.7% (Gemini)	59.8%
	Hard Accuracy	63.6% (GPT-4o, Gemini)	27.3% (GPT-4)	45.4%
	Very Easy Response Time	2.5s (GPT-3.5)	6.2s (Gemini Advanced)	3.6s
	Easy Response Time	2.5s (GPT-3.5)	5.6s (Gemini Advanced)	3.6s
	Moderate Response Time	2.6s (GPT-3.5)	5.7s (Gemini Advanced)	3.6s
	Hard Response Time	2.5s (GPT-3.5)	5.9s (Gemini Advanced)	3.6s
Literature Search Ability	Citation Existence	100% (Co-Pilot)	73.5% (Claude)	84.5%
	Citation Relevance	100% (Gemini, Co-Pilot)	80% (Claude)	90.7%

DISCUSSION

Our findings show an overall chatbot accuracy of 67.5% for adult reconstruction questions, with GPT-4o leading at 80% and Co-Pilot at 70.9%. GPT-4 and Gemini scored lowest at 61.8%. While differences in performance were observed across models, these did not reach statistical significance, and therefore should be interpreted cautiously. Chatbots struggled with difficult questions, averaging only 45.4% accuracy. They performed best on “Clinical Presentation” (78.6%), compared to “Prognosis” (63.1%) and “Management” (67.3%), suggesting stronger descriptive knowledge than applied reasoning. Citation accuracy also varied, Co-Pilot achieved 100% for both existence and relevance, while Gemini Advanced had the lowest citation existence (73.5%) and 80% relevance. The overall moderate performance of 67.5% achieved by the chatbots in our study aligns with the current literature.^{9,20-25} GPT-4o performed the best, aligning with prior research that identifies the GPT4o architecture as being more robust in medical question-answering tasks compared to its predecessors.^{26,27} Surprisingly, ChatGPT-3.5 (67.3%) outperformed GPT-4 (61.8%) in our study, particularly on prognosis and management-related questions. This contradicts prior studies, which consistently report GPT-4 as superior.^{10,21,28,29} One plausible explanation is that GPT-4’s enhanced reasoning capabilities may overcomplicate questions requiring straightforward application of knowledge, as noted by Knoedler et al.²³ This suggests that GPT-3.5’s simpler reasoning mechanisms may have better aligned with the design of our dataset. However, given the lack of statistically significant differences, these findings should not be interpreted as definitive evidence. The performance of Gemini, formerly iterated as Google Bard, has also been a topic of contention in the literature.³⁰ Our study found Gemini’s accuracy (61.8%) to be at the lower end of the spectrum, which contrasts starkly with Vaishya et al, who reported an exceptional 100% accuracy for Google Bard.²⁹ Instead our findings align with Lubitz and Latario’s study, where Google Bard demonstrated a lackluster performance with 49.8% accuracy.²⁴ Such discrepancies could stem from differences in study design, including the type and complexity of questions posed. No significant differences, however, were found in chatbot accuracy between knee and hip questions.

The 75.0% average accuracy of the AAOS member respondents surpassed that of all chatbots, except for ChatGPT 4o (80.0%). This finding aligns with existing literature indicating that chatbots have consistently struggled to pass orthopedic examinations across various countries and in different formats.^{6,25,28,29} Similar to human respondents, the chatbots performed well on easier questions but struggled with more difficult ones. Interestingly, Maraqa et al found that ChatGPT 4o achieved scores equal to or higher than those of residents for the French orthopedic and trauma surgery exam (DES) exam.²⁶ The superior proficiency of GPT-4o across evaluations highlights its potential as a tool for orthopedic

trainees. However, further research into its evolving capabilities and limitations is essential for fully understanding its role in supplementing or enhancing traditional training methods.

GPT-3.5 had the fastest response time (2.5 sec), while Gemini Advanced was the slowest (5.8 sec), possibly reflecting a trade-off between speed and response complexity. Faster times in GPT-3.5 may indicate simpler processing, while longer times could suggest more intricate reasoning. A weak, non-significant correlation was observed between slower response times and higher accuracy. Importantly, however, these differences are unlikely to be clinically or educationally meaningful, and response time should not be considered a primary determinant of chatbot utility in this context.

The variation in chatbot accuracy across concepts, Clinical Presentation (78.6%), Management (67.3%), and Prognosis (63.1%), suggests differing strengths. GPT-4o and Co-Pilot excelled in Clinical Presentation questions, achieving 90% accuracy. This is consistent with a study by Jain et al which highlights the superior ability of the bots in answering Basic Science questions, emphasizing their ability to output rote facts.³¹ The comparatively lower performance in Prognosis and Management aligns with challenges documented in the literature, where chatbots struggle with predictive reasoning tasks.²²

Citation generation is a critical capability for chatbots in medical education. However, a major challenge is their tendency to generate ‘hallucinated’ citations, non-existent references, or citations that are irrelevant to the topic.^{32,33} Among the bots evaluated, all versions of ChatGPT reported their inability to generate real citations. In contrast, Co-Pilot excelled, achieving 100% accuracy in citation existence and relevance. Gemini Advanced and Claude also demonstrated high relevance (100% and 82.8%, respectively). These findings highlight that while many chatbots can generate citations, the quality and reliability vary widely. Based on our findings, Co-Pilot appears most adept at referencing the relevant literature for adult reconstruction questions. Consistent with existing studies, we observed that the bots can reference scientifically relevant articles, but manual fact-checking is strongly recommended.³²

Our study is one of the first to compare multiple chatbots’ abilities in answering specific orthopedic questions. Future investigations could focus on refining chatbot algorithms to improve performance on more complex, reasoning-based questions such as in prognosis and management. Enhancing citation generation accuracy and integrating up-to-date medical literature would also improve their utility in medical education. Lastly, exploring cultural and demographic variations in chatbot performance across regions and specialties could help optimize their application in diverse educational environments.

Strengths and Limitations

This study should be interpreted in the context of several limitations. The analysis was restricted to a single AAOS ResStudy Adult Reconstruction module and included only non-image-based questions, which may limit generalizability, particularly for topics requiring radiographic or intraoperative image interpretation. Although the dataset represents the complete set of eligible questions within the selected module, the overall number of questions resulted in relatively small subgroup sizes after stratification by concept, anatomic location, and difficulty, which may limit statistical power for subgroup comparisons. Response time was measured manually, introducing potential measurement bias and limiting precision. In addition, while citation analysis was performed by multiple independent reviewers with discrepancies resolved through consensus, formal inter-rater reliability metrics were not calculated. Finally, chatbot performance was assessed using a standardized, single-prompt approach, which may not fully reflect real-world use where iterative prompting and user interaction can influence responses. Despite these considerations, this study provides a structured and clinically relevant evaluation of chatbot performance on standardized orthopedic educational content. Additional strengths include the use of a standardized and clinically relevant question bank with a comprehensive evaluation framework. The use of real-world, trainee-level questions and systematic categorization by concept and difficulty further enhances the practical relevance of the findings.

CONCLUSION

This study evaluates AI chatbots' performance on adult reconstruction questions, focusing on accuracy, speed, and citation capabilities. GPT-4o led with 80.0% accuracy, followed by Co-Pilot (70.9%), while GPT-4 and Gemini scored lower (61.8%). The chatbots excelled in Clinical Presentation (78.6%) but struggled with Prognosis (63.1%) and Management (67.3%). Faster response times didn't always correlate with higher accuracy. Despite promising potential for orthopedic education, chatbots remain supplementary due to inconsistent performance and citation issues. Future research should integrate specific databases and cultural factors to improve their utility in medical education.

References

- Labadze L, Grigolia M, Machaidze L. Role of AI chatbots in education: systematic literature review. *Int J Educ Technol High Educ*. 2023;20(1):56. doi:10.1186/s41239-023-00426-1
- Adamopoulou E, Moussiades L. Chatbots: History, technology, and applications. *Mach Learn Appl*. 2020;2:100006. doi:10.1016/j.mlwa.2020.100006
- Lee H. The rise of ChatGPT: Exploring its potential in medical education. *Anat Sci Educ*. 2024;17(5):926-931. doi:10.1002/ase.2270
- Fayed AM, Mansur NSB, de Carvalho KA, Behrens A, D'Hooghe P, de Cesar Netto C. Artificial intelligence and ChatGPT in Orthopaedics and sports medicine. *J Exp Orthop*. 2023;10(1):74. doi:10.1186/s40634-023-00642-8
- Ulus SA. How does ChatGPT perform on the European Board of Orthopedics and Traumatology examination? A comparative study. *Acad J Health Sci Med Balear*. 2023;38(6):43-46.
- Thibaut G, Dabbagh A, Liverneux P. Does Google's Bard Chatbot perform better than ChatGPT on the European hand surgery exam? *Int Orthop*. 2024;48(1):151-158. doi:10.1007/s00264-023-06034-y
- Lum ZC. Can Artificial Intelligence Pass the American Board of Orthopaedic Surgery Examination? Orthopaedic Residents Versus ChatGPT. *Clin Orthop*. 2023;481(8):1623-1630. doi:10.1097/CORR.0000000000002704
- Lum ZC, Collins DP, Dennison S, et al. Generative Artificial Intelligence Performs at a Second-Year Orthopedic Resident Level. *Cureus*. 16(3):e56104. doi:10.7759/cureus.56104
- Magruder ML, Rodriguez AN, Wong JCJ, et al. Assessing Ability for ChatGPT to Answer Total Knee Arthroplasty-Related Questions. *J Arthroplasty*. 2024;39(8):2022-2027. doi:10.1016/j.arth.2024.02.023
- Agharia S, Szatkowski J, Fraival A, Stevens J, Zhou Y. The ability of artificial intelligence tools to formulate orthopaedic clinical decisions in comparison to human clinicians: An analysis of ChatGPT 3.5, ChatGPT 4, and Bard. *J Orthop*. 2024;50:1-7. doi:10.1016/j.jor.2023.11.063
- Assessing ChatGPT Responses to Common Patient Questions Regarding Total Hip Arthroplasty | Request PDF. Accessed August 8, 2024. https://www.researchgate.net/publication/372417465_Assessing_ChartGPT_Responses_to_Common_Patient_Questions_Regarding_Total_Hip_Arthroplasty
- Mika AP, Mulvey HE, Engstrom SM, Polkowski GG, Martin JR, Wilson JM. Can ChatGPT Answer Patient Questions Regarding Total Knee Arthroplasty? *J Knee Surg*. 2024;37(9):664-673. doi:10.1055/s-0044-1782233
- Yang J, Ardavanis KS, Slack KE, Fernando ND, Della Valle CJ, Hernandez NM. Chat Generative Pretrained Transformer (ChatGPT) and Bard: Artificial Intelligence Does not yet Provide Clinically Supported Answers for Hip and Knee Osteoarthritis. *J Arthroplasty*. 2024;39(5):1184-1190. doi:10.1016/j.arth.2024.01.029
- AlShehri Y, McConkey M, Lodhia P. ChatGPT Provides Satisfactory but Occasionally Inaccurate Answers to Common Patient Hip Arthroscopy Questions. *Arthrosc J Arthrosc Relat Surg*. Published online June 22, 2024. doi:10.1016/j.arthro.2024.06.017
- Johns WL, Martinazzi BJ, Miltenberg B, Nam HH, Hammoud S. ChatGPT Provides Unsatisfactory Responses to Frequently Asked Questions Regarding Anterior Cruciate Ligament Reconstruction. *Arthrosc J Arthrosc Relat Surg*. 2024;40(7):2067-2079. e1. doi:10.1016/j.arthro.2024.01.017
- Evaluating the accuracy and relevance of ChatGPT responses to frequently asked questions regarding total knee replacement | Knee Surgery & Related Research. Accessed August 8, 2024. <https://link.springer.com/article/10.1186/s43019-024-00218-5>
- Parekh AS, McCahon JAS, Nghe A, Pedowitz DI, Daniel JN, Parekh SG. Foot and Ankle Patient Education Materials and Artificial Intelligence Chatbots: A Comparative Analysis. *Foot Ankle Spec*. Published online March 19, 2024:19386400241235834. doi:10.1177/19386400241235834
- Wright BM, Bodnar MS, Moore AD, et al. Is ChatGPT a trusted source of information for total hip and knee arthroplasty patients? *Bone Jt Open*. 2024;5(2):139-146. doi:10.1302/2633-1462.52.BJO-2023-0113.R1
- Ridha A, Ahmed I, Raj S, et al. No Time for ChitChat? Artificial intelligence Chatbots as a tool to identify research priorities in hip and knee arthroplasty. Published online October 19, 2023. doi:10.21203/rs.3.rs-3339904/v1

20. Gan W, Ouyang J, Li H, et al. Integrating ChatGPT in Orthopedic Education for Medical Undergraduates: Randomized Controlled Trial. *J Med Internet Res*. 2024;26:e57037. doi:10.2196/57037
21. Kim SE, Lee JH, Choi BS, Han HS, Lee MC, Ro DH. Performance of ChatGPT on Solving Orthopedic Board-Style Questions: A Comparative Analysis of ChatGPT 3.5 and ChatGPT 4. *Clin Orthop Surg*. 2024;16(4):669-673. doi:10.4055/cios23179
22. Chow R, Hasan S, Zheng A, et al. The Accuracy of Artificial Intelligence ChatGPT in Oncology Examination Questions. *J Am Coll Radiol*. 2024;21(11):1800-1804. doi:10.1016/j.jacr.2024.07.011
23. Knoedler L, Alfertshofer M, Knoedler S, et al. Pure Wisdom or Potemkin Villages? A Comparison of ChatGPT 3.5 and ChatGPT 4 on USMLE Step 3 Style Questions: Quantitative Analysis. *JMIR Med Educ*. 2024;10:e51148. doi:10.2196/51148
24. Lubitz M, Latario L. Performance of Two Artificial Intelligence Generative Language Models on the Orthopaedic In-Training Examination. *Orthopedics*. 2024;47(3):e146-e150. doi:10.3928/01477447-20240304-02
25. Sparks CA, Kraeutler MJ, Chester GA, et al. Inadequate Performance of ChatGPT on Orthopedic Board-Style Written Exams. *Cureus*. 2024;16(6):e62643. doi:10.7759/cureus.62643
26. Maraqa N, Samargandi R, Poichotte A, Berhouet J, Benhenneda R. Comparing performances of french orthopaedic surgery residents with the artificial intelligence ChatGPT-4/4o in the French diploma exams of orthopaedic and trauma surgery. *Orthop Traumatol Surg Res OTSR*. Published online December 4, 2024;104080. doi:10.1016/j.otsr.2024.104080
27. Ayık G, Ercan N, Demirtaş Y, Yıldırım T, Çakmak G. Evaluation of ChatGPT-4o's answers to questions about hip arthroscopy from the patient perspective. *It Dis Relat Surg*. 2025;36(1):193-199. doi:10.52312/jdrs.2025.1961
28. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT-3.5, ChatGPT-4, and Orthopaedic Resident Performance on Orthopaedic Assessment Examinations. *JAAOS-J Am Acad Orthop Surg*. 2023;31(23):1173. doi:10.5435/JAAOS-D-23-00396
29. Vaishya R, Iyengar KP, Patralekh MK, et al. Effectiveness of AI-powered Chatbots in responding to orthopaedic postgraduate exam questions—an observational study. *Int Orthop*. 2024;48(8):1963-1969. doi:10.1007/s00264-024-06182-9
30. Rayhan S. *The Impact of Changing the Name of Google AI from Bard to Gemini*. 2024. doi:10.13140/RG.2.2.30294.83520
31. Jain N, Gottlich C, Fisher J, Campano D, Winston T. Assessing ChatGPT's orthopedic in-service training exam performance and applicability in the field. *J Orthop Surg*. 2024;19(1):27. doi:10.1186/s13018-023-04467-0
32. Chen A, Chen DO. Accuracy of Chatbots in Citing Journal Articles. *JAMA Netw Open*. 2023;6(8):e2327647. doi:10.1001/jama-networkopen.2023.27647
33. Li W, Li J, Ma W, Liu Y. Citation-Enhanced Generation for LLM-based Chatbots. In: Ku LW, Martins A, Srikumar V, eds. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics; 2024:1451-1466. doi:10.18653/v1/2024.acl-long.79

Authors

Tarishi Parmar, BS, Sidney Kimmel Medical College, Thomas Jefferson University, Philadelphia, PA.
 Akshay Khanna, BS, Sidney Kimmel Medical College, Thomas Jefferson University, Philadelphia, PA.
 Jonathan Berg, BS, Sidney Kimmel Medical College, Thomas Jefferson University, Philadelphia, PA.
 Alexander Linton, MD, Rothman Orthopaedic Institute, Philadelphia, PA.
 Alec Kellish, MD, Rothman Orthopaedic Institute, Philadelphia, PA.
 Lauren O'Mara, MD, Rothman Orthopaedic Institute, Philadelphia, PA.
 Stephanie Eble, MD, Rothman Orthopaedic Institute, Philadelphia, PA.
 David G. Nazarian, MD, Rothman Orthopaedic Institute, Philadelphia, PA.

Disclosures

Conflicts of Interest: None

Acknowledgments: None

Funding: None

Correspondence

Tarishi Parmar, BS
 tarishiparmar20051@gmail.com