

Accuracy and limitations of artificial intelligence chatbots in answering patient questions on scoliosis surgery

Precisión y limitaciones de los chatbots de IA para responder a las preguntas de los pacientes sobre la cirugía de escoliosis

Güngör Alibakan¹* and Yusuf Sülek²

¹Department of Orthopedics and Traumatology, University of Health Sciences Şişli Hamidiye Etfal Training and Research Hospital; ²Ministry of Health, Esenyurt Necmi Kadioğlu State Hospital . Istanbul, Turkey

Abstract

Objective: The aim of this study was to analyze the accuracy, reliability, and quality of the content of responses provided by artificial intelligence (AI)-based chatbots to frequently asked questions related to scoliosis surgery. **Methods:** A set of 25 questions related to diagnosis, treatment options, surgical risks, and recovery was developed based on clinical experience. The responses were anonymized and evaluated by two orthopedic surgeons using a 4-point scale to rate accuracy, clarity, and consistency. Cohen's kappa coefficient was used for interrater reliability, and the Kruskal-Wallis test was used for statistical analysis. **Results:** ChatGPT 4o performed best, with 60% of responses rated as excellent. Gemini Advanced 1.5 Pro followed with 56%, while Microsoft Copilot performed poorly, with 76% of responses requiring moderate clarification. The differences between the chatbots were statistically significant ($p < 0.001$). ChatGPT 4o and Gemini did not differ significantly from each other ($p = 0.850$), but both outperformed Copilot ($p < 0.001$). **Conclusions:** AI chatbots show great potential for patient education in scoliosis surgery. ChatGPT 4o was the most effective, although complex queries still require medical supervision.

Keywords: Scoliosis surgery. Artificial intelligence. Chatbot accuracy. Clinical decision support. Patient education. Surgical outcomes.

Resumen

Objetivo: Analizar la precisión, la fiabilidad y la calidad del contenido de las respuestas proporcionadas por los chatbots de inteligencia artificial a preguntas frecuentes relacionadas con la cirugía de escoliosis. **Métodos:** Se elaboró un conjunto de 25 preguntas relacionadas con el diagnóstico, las opciones de tratamiento, los riesgos quirúrgicos y la recuperación, basadas en la experiencia clínica. Las respuestas fueron anonimizadas y evaluadas por dos cirujanos ortopédicos utilizando una escala de 4 puntos para calificar la precisión, la claridad y la coherencia. Se utilizaron el coeficiente kappa de Cohen para la fiabilidad interevaluador y la prueba de Kruskal-Wallis para el análisis estadístico. **Resultados:** ChatGPT 4o obtuvo el mejor desempeño, con un 60% de respuestas calificadas como excelentes; Gemini Advanced 1.5 Pro le siguió con un 56%, mientras que Microsoft Copilot mostró un rendimiento inferior, con un 76% de respuestas que requirieron aclaraciones moderadas. Las diferencias entre los chatbots fueron estadísticamente significativas ($p < 0.001$). ChatGPT 4o y Gemini no difirieron significativamente entre sí ($p = 0.850$), pero ambos superaron a Copilot ($p < 0.001$). **Conclusiones:** Los chatbots de IA muestran un gran potencial para la educación del paciente en cirugía de escoliosis. ChatGPT 4o fue el más eficaz, aunque las consultas complejas aún requieren supervisión médica.

Palabras clave: Cirugía de escoliosis. Inteligencia artificial. Precisión de los chatbots. Apoyo a la toma de decisiones clínicas. Educación del paciente. Resultados quirúrgicos.

*Correspondence:

Güngör Alibakan
E-mail: gngr_alibakan@hotmail.com

Date of reception: 03-07-2025

Date of acceptance: 19-07-2025

DOI: 10.24875/CIRU.25000347

Cir Cir. (ahead of print)

Contents available at PubMed

www.cirugiaycirujanos.com

0009-7411/© 2025 Academia Mexicana de Cirugía. Published by Permanyer. This is an open access article under the terms of the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Introduction

Healthcare is undergoing a significant transformation with the rapid advancement of digital technologies. Artificial intelligence (AI) plays an important role, with algorithms and tools supporting diagnostic processes in different branches, offering remarkable advantages in terms of accuracy and speed.¹⁻⁶ In particular, AI-based chatbots are increasingly being used in the area of patient information and support from health professionals.^{6,7}

The bots have been slowly leading the way in the emergence of a primary source of health-related information due to their accessibility and user-friendliness for people searching for such data. However, the way in which these technologies are useful with regard to accuracy, reliability, and coverage of the information they provide, compared to health professionals' information, is an open question.⁷⁻¹⁴ The content of chatbot responses is of great importance in the quality of answers given to a patient, especially in specialized fields like orthopedics, particularly in the treatment processes of complex cases such as scoliosis surgery.

Scoliosis surgery is an area where detailed information sharing between the patient and the physician is necessary. Accurate and intelligible information about diagnostic procedures, surgical and non-surgical treatment options, possible complications, and recovery processes can not only enhance patient satisfaction but also potentially affect the treatment outcome.^{15,16} AI-based chatbots may be a good tool for fast access to such information. However, there are limited studies on the compatibility of the information provided by these bots with expert opinions and the adequacy of their answers to complex questions.^{7-14,17,18} In this respect, the possible and not-so-possible facets of chatbots while providing information on scoliosis surgery should be well understood for the development of future use cases. The objective of this study is to analyze the accuracy, reliability, and content of the responses of three AI-based chatbots, ChatGPT 4o (OpenAI), Gemini Advanced 1.5 Pro (Google), and Microsoft Copilot (Microsoft), to common patient questions regarding scoliosis surgery. Our hypothesis is that these chatbots can provide high accuracy in general information about scoliosis surgery but may show significant deficiencies in detailed or specialized topics. Accordingly, we aim to provide concrete data to

evaluate the role of chatbots in patient education and clinical practice.

Methods

Study design

This is a cross-sectional comparative study that evaluates the accuracy, reliability, and content of AI-based chatbot responses to questions related to scoliosis surgery.

Setting and participants

ChatGPT 4o (OpenAI), Gemini Advanced 1.5 Pro (Google), and Microsoft Copilot were selected for analysis as AI-based chatbot models. The answers from the above-mentioned chatbots were reviewed and evaluated by two orthopedic spine surgeons independently. Both surgeons have more than 5 years of clinical experience in spine surgery.

Question development

Based on clinical experience and common questions patients ask about scoliosis surgery, a set of 25 questions was developed (Table 1). The questions were designed to cover a wide range of topics, including diagnostic criteria for scoliosis, available surgical and non-surgical treatment options and associated risk factors, possible complications, and recovery expectations.

Data collection

Each of the 25 questions was posed to all three chatbots (ChatGPT 4o [OpenAI], Gemini Advanced 1.5 Pro (Google) and Microsoft Copilot) by independent evaluators under the same conditions to ensure standardization. Responses generated by the chatbots were collected and anonymized to maintain platform blindness prior to evaluation. To assess response consistency, each question was asked twice at different intervals to assess whether the chatbot's responses remained constant over time.

Evaluation metrics

Responses were scored using a 4-point scale based on a rating system adapted from Mika et al.⁸ (Table 2),

Table 1. Common patient questions regarding scoliosis surgery

S. No.	Question
1	What is scoliosis and how is it diagnosed?
2	What are the available surgical and non-surgical treatment options for scoliosis?
3	What are the risks of scoliosis surgery?
4	Is scoliosis surgery safe?
5	Who is a suitable candidate for scoliosis surgery?
6	What is the best age for scoliosis surgery?
7	How long does it take to recover after scoliosis surgery?
8	Which techniques are used during scoliosis surgery?
9	When can I return to my daily activities after scoliosis surgery?
10	What is the success rate of scoliosis surgery?
11	What is the possibility of recurrence of scoliosis after surgery?
12	How long does scoliosis surgery take?
13	Is physiotherapy necessary after scoliosis surgery?
14	How long does the pain last after scoliosis surgery?
15	Can I do sports after scoliosis surgery?
16	How much does scoliosis surgery cost?
17	What is the lifespan of the implants used during scoliosis surgery?
18	What are the complication rates after scoliosis surgery?
19	What happens if scoliosis is not treated?
20	Does my spine completely recover after scoliosis surgery?
21	What preparations should I make before scoliosis surgery?
22	Are there different surgical approaches for scoliosis surgery? Which technique is better?
23	How long will I stay in hospital after scoliosis surgery?
24	How common is infection after scoliosis surgery?
25	How much spinal mobility is preserved after scoliosis surgery?

across four main measures: accuracy, clarity, completeness, and consistency. Accuracy referred to the degree that the responses were factually accurate and adhered to current medical guidelines. Clarity measured the simplicity and accessibility of the language used, ensuring that it was understandable to patients. Completeness depended on whether all critical aspects of the question were addressed, and consistency referred to the similarity of the responses over multiple attempts. Scores range from 1 to 4, where 1

Table 2. Chatbot response rating system

Response accuracy score	Response accuracy description
1	Excellent response not requiring clarification
2	Satisfactory requiring minimal clarification
3	Satisfactory requiring moderate clarification
4	Unsatisfactory requiring substantial clarification

meant very good answers which did not require an explanation, and 4 meant inadequate answers which were wrong or incomplete.

Scoring criteria

Scoring of chatbot responses was from 1 to 4 on the basis of the information quality and accuracy. A score of 1 represented an excellent response-accuracy, complete, no explanation was necessary. A score of 2 indicated a satisfactory response that was largely accurate, but required a small amount of explanation for full understanding. Responses scored a 3 either because missing information or partially correct information was present in the response; such responses required moderate explanation. Finally, a score of 4 was awarded to responses that were unsatisfactory, inaccurate, or incomplete, and indicated significant gaps in the information provided. These criteria provided a standardized and comprehensive assessment of chatbot performance.

Statistical analysis

For statistical analysis, categorical statistics were used to summarize chatbot performance across the grading criteria. The Kruskal–Wallis Test was used to compare chatbot response scores, and significance was set at $p < 0.05$. In addition, Bonferroni correction was applied to further investigate the distribution of response accuracy in different categories. Inter-rater reliability between reviewers was measured using Cohen's Kappa.

Ethical considerations

Ethical standards were complied with by excluding patient-identifiable data in all chatbot interactions. The study focused only on the responses generated by the

publicly available AI, and no human subjects were directly involved in the research. This ensured an ethical evaluation of chatbot capabilities in addressing scoliosis-related questions.

Results

The performance of three AI-based chatbots (ChatGPT 4o, Gemini Advanced 1.5 Pro, and Microsoft Copilot) was evaluated based on responses to 25 frequently asked questions about scoliosis surgery. Detailed questions and full responses are provided in supplementary material 1. Responses were graded using a 4-point scale assessing accuracy, clarity, completeness, and consistency (Table 3). ChatGPT 4o provided excellent answers (1 point) in 60% of cases, while the remaining 40% were rated as satisfactory answers (2 points) requiring minor clarification. Gemini Advanced 1.5 Pro performed similarly, rating 56% of responses as excellent and 44% as satisfactory. In contrast, Microsoft Copilot performed the poorest, with 76% of responses requiring moderate explanation (3 points) and only 20% rated satisfactory. None of the Copilot responses were rated as excellent.

Statistical analysis revealed a significant difference between the performances of the three chatbots ($p < 0.001$). While pairwise comparisons showed no significant difference between ChatGPT 4o and Gemini Advanced 1.5 Pro ($p = 0.850$), both significantly outperformed Microsoft Copilot ($p < 0.001$) (Table 4). Inter-rater reliability assessed using Cohen's kappa showed high agreement between raters, with kappa values of 0.833 for ChatGPT 4o, 0.759 for Gemini Advanced 1.5 Pro, and 0.736 for Microsoft Copilot (Fig. 1).

When analyzed against specific evaluation criteria, ChatGPT 4o achieved the highest accuracy (60%), followed by Gemini Advanced 1.5 Pro (56%), while Microsoft Copilot scored 0% for accuracy, reflecting its inability to provide fully accurate responses (Fig. 2). While both ChatGPT 4o and Gemini Advanced 1.5 Pro exhibited comparable clarity and consistency, Microsoft Copilot exhibited the lowest consistency, with noticeable variability in responses to the same questions over time. In terms of completeness, ChatGPT 4o provided the most comprehensive answers, while Microsoft Copilot frequently omitted critical information.

The evaluation of temporal consistency revealed that ChatGPT 4o and Gemini Advanced 1.5 Pro provided consistent responses over repeated queries,

Table 3. Response accuracy scores (%) across ChatGPT 4o, Gemini advanced 1.5 Pro, and Microsoft copilot

Response accuracy score	ChatGPT 4o (%)	Gemini Advanced 1.5 Pro (%)	Microsoft copilot (%)
1	60	56	0
2	40	44	20
3	0	0	76
4	0	0	4

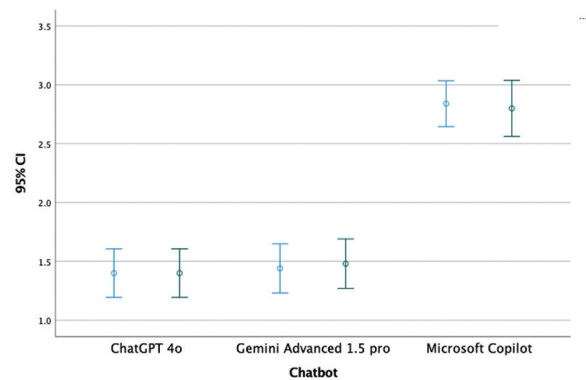


Figure 1. Mean response accuracy scores with 95% confidence intervals for response accuracy scores of ChatGPT 4o, Gemini Advanced 1.5 Pro, and Microsoft Copilot.

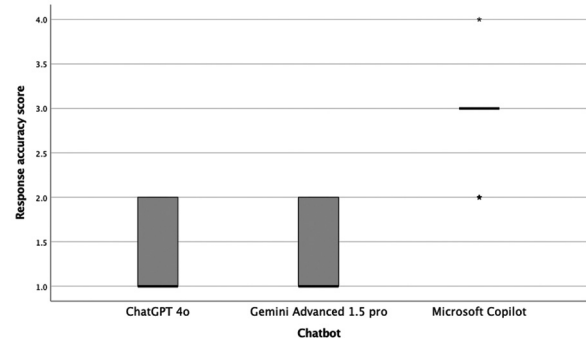


Figure 2. Response accuracy scores across ChatGPT 4o, Gemini Advanced 1.5 Pro, and Microsoft Copilot.

while Microsoft Copilot's responses varied significantly, raising concerns about its reliability. In all, ChatGPT 4o was the most capable chatbot in providing full and accurate information on scoliosis surgery, followed by Gemini Advanced 1.5 Pro. Microsoft Copilot showed very poor performance for complex and specific questions, thus limiting its clinical application.

Table 4. Pairwise comparisons of chatbot performance

Comparison	Test statistic	Standard error	Standard test statistic	Sig.
ChatGPT 4o-Gemini Advanced 1.5 pro	-1.100	5.797	-0.190	0.850
ChatGPT 4o-Microsoft Copilot	-34.900	5.797	-6.020	< 0.001
Gemini Advanced 1.5 pro-Microsoft Copilot	-33.800	5.797	-5.830	< 0.001

Discussion

The findings from this study revealed a profound variation in the performances of ChatGPT 4o, Gemini Advanced 1.5 Pro, and Microsoft Copilot. ChatGPT 4o showed consistency in providing scoliosis surgical information and remained ahead of the pack on the same question, especially on aspects of accuracy and content. On some metrics, Gemini Advanced 1.5 Pro did not differ much but came out a bit behind in consistency. Microsoft Copilot had poor comprehensiveness in its answers or was not that accurate, therefore limiting its applicability in the clinical and patient education settings. To the best of our knowledge, this is the first study in the literature regarding responses from AI-based chatbots to questions about scoliosis surgery from both a patient and physician perspective. The superiority of ChatGPT 4o can be attributed to its high-performance natural language processing and larger training data that allow the chatbot to respond accurately and in a well-structured manner with facts. Moreover, the kappa score of 0.833 signifies great rater agreement on reliability in giving consistent and accurate information. Gemini Advanced 1.5 Pro showed reasonable accuracy, with a kappa of 0.759, but it tended to be less accurate more often and often required further explanation, especially with more complex topics of surgical technique and post-operative recovery. Microsoft Copilot had the lowest kappa value, 0.736, which represents limited reliability in approaching complex clinical questions. These large differences in chatbot performance underpin both the potential and limitations of AI in healthcare. While ChatGPT 4o and Gemini Advanced 1.5 Pro show promise for supporting patient education and clinical decision-making, they were not perfect in most cases. For example, no single chatbot scored full marks in all categories, thus indicating areas in which their use can be further optimized. On the contrary, incomplete responses from Microsoft Copilot give an indication that this tool is not yet ready for clinical applications,

especially in situations where complex and specific knowledge is needed. These results align with prior literature emphasizing the role of AI-driven chatbots in improving patient education and optimizing clinical workflows.⁷⁻¹⁴ However, the limitations seen, particularly in handling very specialized queries, raise the need to consider these tools in addition to expert supervision when using them in clinical practice. For example, although ChatGPT 4o could answer general and moderately complex questions, nuanced or context-specific questions were often less clear and less detailed than might be expected from a clinical expert. This points to how continuous development or improvement in AI algorithms is indispensable for the complexity of medical knowledge. These results have, from a clinical perspective, implications for patient and physician use: Patients can use these tools as sources of reliable, accessible information, particularly for common questions about scoliosis surgery, recovery, and risks involved with it. As for the doctors, they would also serve as additional complementary decision support tools while exercising caution against such guidance in complex cases.

This study has several limitations. First, the evaluation was limited to responses from three dedicated chatbots, and the results may not be generalizable to other AI systems. Second, although the raters were blind to the source of the responses, the subjective nature of the scoring may introduce bias. In addition, the study focused on 25 predefined questions, which, while comprehensive, may not cover the full range of potential queries encountered in clinical practice. By examining a wider range of questions in future research, a larger pool of chatbots and their real-time use in clinical settings can be evaluated to provide more accurate insights into practical applications.

Conclusions

AI chatbots, particularly ChatGPT 4o and Gemini Advanced 1.5 Pro, show significant potential in

providing accurate and reliable information about scoliosis surgery. However, their limitations highlight the need for human oversight and additional refinement to improve their use in clinical settings. As these technologies advance, it is crucial to integrate AI tools into healthcare workflows carefully, ensuring they support rather than substitute expert clinical judgment.

Funding

The authors declare that they have not received funding.

Conflicts of interest

The authors declare no conflicts of interest.

Ethical considerations

Protection of humans and animals. The authors declare that no experiments on humans or animals were performed for this research.

Confidentiality, informed consent, and ethical approval. This study did not involve human participants or patient data; it evaluated publicly available chatbot responses. Therefore, ethics committee approval and informed consent were not required.

Declaration on the use of artificial intelligence. Generative AI was used only for language editing; it was not used to generate text, figures, or tables.

Supplementary data

Supplementary data are available at DOI: 10.24875/CIRU.25000347. These data are provided by the corresponding author and published online for the benefit of the reader. The content of supplementary data is the sole responsibility of the authors.

References

1. Soe NN, Yu Z, Latt PM, Lee D, Ong JJ, Ge Z, et al. Evaluation of artificial intelligence-powered screening for sexually transmitted infections-related skin lesions using clinical images and metadata. *BMC Med.* 2024;22:296.
2. Kanar M, Olçar AH, Sülek Y, Alibakan G, Gök B, Armağan R. Developing a fast computer vision model for diagnosing and classifying hip fractures. *Eur Arch Med. Res.* 2024;40:214-20.
3. Xu D, Sui L, Zhang C, Xiong J, Wang VY, Zhou Y, et al. The clinical value of artificial intelligence in assisting junior radiologists in thyroid ultrasound: a multicenter prospective study from real clinical practice. *BMC Med.* 2024;22:293.
4. Hidaka R, Matsuda K, Igari T, Takeuchi S, Imoto Y, Yagi S, et al. Development and accuracy of an artificial intelligence model for predicting the progression of hip osteoarthritis using plain radiographs and clinical data: a retrospective study. *BMC Musculoskelet Disord.* 2024;25:893.
5. Jeon YD, Jung KH, Kim MS, Kim H, Yoon DK, Park KB. Clinical validation of artificial intelligence-based preoperative virtual reduction for near 3- or 4-part proximal humerus fractures. *BMC Musculoskelet Disord.* 2024;25:669.
6. Jiang W, Liu T, Sun B, Zhong L, Han Z, Lu M, et al. An artificial intelligence application to predict prolonged dependence on mechanical ventilation among patients with critical orthopaedic trauma: an establishment and validation study. *BMC Musculoskelet Disord.* 2024;25:1089.
7. Bafor A, Strub D, Kold S, Iobst CA, Tulchin-Francis K. Chatbots in limb lengthening and reconstruction surgery: how accurate are the responses? *J Pediatr Orthop.* 2025;45:33-6.
8. Mika AP, Martin JR, Engstrom SM, Polkowski GG, Wilson JM. Assessing ChatGPT responses to common patient questions regarding total hip arthroplasty. *J Bone Joint Surg.* 2023;105:1519-26.
9. Hu X, Niemann M, Kienzie A, Braun K, Back DA, Gwinner C, et al. Evaluating ChatGPT responses to frequently asked patient questions regarding periprosthetic joint infection after total hip and knee arthroplasty. *Digit Health.* 2024;10:20552076241272620.
10. Taylor WL 4th, Cheng R, Weinblatt AI, Bergstein V, Long WJ. An artificial intelligence chatbot is an accurate and useful online patient resource prior to total knee arthroplasty. *J Arthroplasty.* 2024;39:S358-62.
11. Casciato D, Mateen S, Cooperman S, Pesavento D, Brandao RA. Evaluation of online AI-generated foot and ankle surgery information. *J Foot Ankle Surg.* 2024;63:680-3.
12. Amen TB, Torabian KA, Subramanian T, Yang BW, Liimakka A, Fufa D. Quality of ChatGPT responses to frequently asked questions in carpal tunnel release surgery. *Plast Reconstr Surg Glob Open.* 2024;12:e5822.
13. Lack BT, Mouhawasse E, Childers JT, Jackson GR, Daji SV, Yerke-Hansen P, et al. Can ChatGPT answer patient questions regarding reverse shoulder arthroplasty? *J ISAKOS.* 2024;9:100323.
14. White CA, Masturov YA, Haunschild E, Michaelson E, Shukla DR, Cagle PJ. Can ChatGPT reliably answer the most common patient questions regarding total shoulder arthroplasty? *J Shoulder Elbow Surg.* 2025;34:e254-64.
15. Trobisch PD, Ricart PA, Verma K. Clinical and demographic characteristics of online community members—does it represent reality? *Spine J.* 2011;11 9:826-31.
16. Yeramosu T, Dominy CL, Arvind V, Isleem UN, Cho SK. Scoliosis surgery: a social media analysis of content, tone, and perspective. *J Am Acad Orthop Surg.* 2023;31:26-33.
17. Blasco JM, Díaz-Díaz B, Igual-Camacho C, Pérez-Maletzki J, Hernández-Guilén D, Roig-Casasús S. Effectiveness of using a chatbot to promote adherence to home physiotherapy after total knee replacement, rationale and design of a randomized clinical trial. *BMC Musculoskelet Disord.* 2023;24:491.
18. Asfuroğlu ZM, Yağar H, Gümüşoğlu E. High accuracy but limited readability of large language model-generated responses to frequently asked questions about Kienböck's disease. *BMC Musculoskelet Disord.* 2024;25:879.