Journal of Surgical Oncology

WILEY

RESEARCH ARTICLE OPEN ACCESS

Assessment of Artificial Intelligence Chatbot Responses to Common Patient Questions on Bone Sarcoma

Kameel Khabaz[1] | Nicole J. Newman-Hung[2] | Jennifer R. Kallini[2] | Joseph Kendal[3] | Alexander B. Christ[2] | Nicholas M. Bernthal[2] | Lauren E. Wessel[2]

[1]David Geffen School of Medicine at UCLA, Los Angeles, California, USA | [2]Department of Orthopaedic Surgery, University of California, Los Angeles, California, USA | [3]Department of Surgery, University of Calgary, Calgary, Alberta, Canada

Correspondence: Kameel Khabaz (kkhabaz@mednet.ucla.edu)

Received: 31 July 2024 | **Revised:** 4 October 2024 | **Accepted:** 12 October 2024

Funding: The authors received no specific funding for this work.

Keywords: artificial intelligence | bone sarcomas | chatbots | chondrosarcoma | Ewing sarcoma | osteosarcoma

ABSTRACT

Background and Objectives: The potential impacts of artificial intelligence (AI) chatbots on care for patients with bone sarcoma is poorly understood. Elucidating potential risks and benefits would allow surgeons to define appropriate roles for these tools in clinical care.

Methods: Eleven questions on bone sarcoma diagnosis, treatment, and recovery were inputted into three AI chatbots. Answers were assessed on a 5-point Likert scale for five clinical accuracy metrics: relevance to the question, balance and lack of bias, basis on established data, factual accuracy, and completeness in scope. Responses were quantitatively assessed for empathy and readability. The Patient Education Materials Assessment Tool (PEMAT) was assessed for understandability and actionability.

Results: Chatbots scored highly on relevance (4.24) and balance/lack of bias (4.09) but lower on basing responses on established data (3.77), completeness (3.68), and factual accuracy (3.66). Responses generally scored well on understandability (84.30%), while actionability scores were low for questions on treatment (64.58%) and recovery (60.64%). GPT-4 exhibited the highest empathy (4.12). Readability scores averaged between 10.28 for diagnosis questions to 11.65 for recovery questions.

Conclusions: While AI chatbots are promising tools, current limitations in factual accuracy and completeness, as well as concerns of inaccessibility to populations with lower health literacy, may significantly limit their clinical utility.

1 | Introduction

1.1 | Background

The advancement of large language models (LLMs) has driven the recent surge in AI chatbot adoption. LLMs leverage complex machine-learning architectures to understand and generate human-like text [1]. Despite growing excitement around applications of AI chatbots in healthcare, current adoption in orthopedic surgery remains limited [1–3]. Several studies have investigated using AI chatbots to augment counseling for patients with hip and knee osteoarthritis, shoulder instability, and carpal tunnel syndrome, demonstrating the potential to provide personalized and timely patient counseling [4–14]. However, the potential risks of using AI chatbots to counsel patients with complex, oncologic diagnoses have not been clearly elucidated [1, 15].

Bone sarcomas are a group of malignancies with an incidence of approximately 1.0 cases per 100 000 Americans per year, with significant clinical heterogeneity and complexity in long-term management [16–18]. Patients with bone sarcomas have become increasingly reliant on accessing online sources for clinical information, follow a longer pathway to diagnosis, and

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Author(s). *Journal of Surgical Oncology* published by Wiley Periodicals LLC.

experience anxiety about the evaluation and surveillance processes [19, 20]. Given these unique considerations for patients with bone sarcomas, there may be a need for adjunctive patient counseling tools to provide additional patient education and emotional support.

1.2 | Rationale

While applications of AI chatbots have been described in other orthopedic settings, the impact of chatbots on patient counseling for bone sarcoma remains unexplored [7, 10, 13]. This study evaluates the efficacy of three AI chatbots in answering common patient questions about bone sarcoma by measuring the clinical accuracy, understandability, actionability, and empathy of AI-generated responses. We hypothesized that while AI chatbots may provide coherent responses that are empathetic and understandable, they also may produce critical factual inaccuracies on bone sarcoma management. Based on the complexity of these oncologic diagnoses, factual inaccuracies may pose significant risks to patients seeking adjunctive counseling from these tools.

2 | Materials and Methods

Eleven patient questions on bone sarcoma diagnosis, treatment, and recovery were generated from prior analysis of five online public discussion forums, containing a total of 570 posts from patients with osteosarcoma, Ewing sarcoma, and chondrosarcoma (Supporting Information S1: File 1) [20]. Using grounded theory, posts were thematically grouped into three selective categories of information support for diagnosis, treatment, and recovery [20]. A set of 15 preliminary questions were generated for each thematic group by feeding the posts to GPT-4, along with carefully engineered prompts asking for realistic patient questions that draw upon specific themes in the discussion forum posts. Eleven total patient questions were then manually selected for relevance and edited for clarity. Five questions pertained to bone sarcoma diagnosis, two pertained to treatment, and four pertained to recovery.

The patient questions were input into three different AI chatbots: GPT-4 (OpenAI, 2023, November 1, version), Bard (Google, 2023, November 1, version), and Claude 2 (Anthropic, accessed via Perplexity AI, 2023, November 1, version). These three models were developed by independent research teams using different neural network architectures, allowing for the sampling of a wider variety of tools. Each chatbot was used with default settings, and a separate user session was initiated for each question to prevent the model from using previous answers as context for new responses. Each question was accompanied by the following carefully worded prompt: "Answer the following question from a patient with bone sarcoma. Ensure that the answer is comprehensive in scope, thorough in detail, and empathetic in tone. The answer must be factually accurate, understandable, and actionable."

Answers from the chatbots were quantitatively assessed using a 5-point Likert scale for five clinical accuracy metrics: relevance to the question, balance and lack of bias, basis on established data, factual accuracy, and completeness in scope. Respondents were blinded to the chatbot used and were given the opportunity to provide comments on the chatbot's answers. Four fellowship-trained orthopedic oncologists (J.R.K., A.B.C., N.M.B., L.E.W.) independently scored all responses.

AI-generated answers were also assessed for understandability and actionability of information using the Patient Education Materials Assessment Tool (PEMAT). PEMAT is an internally consistent and reliable instrument, composed of a 24-question survey of yes-or-no questions, to assess the understandability and actionability of patient education materials [21]. Due to the text-based nature of AI-generated responses in this study, survey questions pertaining to visual aids, illustrations, tables, and checklists were excluded. Furthermore, one survey question on providing instructions for performing calculations was removed, resulting in a total of 15 PEMAT questions that were assessed in the study. In addition, empathy was assessed using a 5-point Likert scale. Three authors (K.K., N.J.N.H., J.R.K.) independently scored all responses. Readability was assessed with the Flesch–Kincaid grade level, which is one of the most commonly used readability scores corresponding to a US grade level and is based on sentence length and the number of syllables per word in a response [22, 23]. Statistical differences between response groups were assessed with two-sample independent t tests.

3 | Results

Thirty-three AI-generated answers to 11 patient questions on bone sarcoma were rated by four respondents for five clinical accuracy metrics: relevance to the question, balance and lack of bias, basis on established data, factual accuracy, and completeness in scope. This generated a total of 495 responses used in the analysis. The full set of patient questions and AI-generated answers can be found in Supporting Information S1: File 1. A sample question and answer generated by Bard is shown in Figure 1.

Overall, the average clinical accuracy score was 3.89 out of 5, and scores ranged from a minimum of 2 to a maximum of 5 for the clinical accuracy metrics: relevance to the question, balance and lack of bias, basis on established data, completeness in scope, and factual accuracy. As shown in Figure 2, GPT-4 and Claude 2 produced the highest-scoring answers for each clinical accuracy metric, with Bard producing lower-quality responses. Furthermore, while AI chatbots scored highly on relevance (4.23 ± 0.6) and balance/lack of bias (4.09 ± 0.7), they scored lower on basing responses on established data (3.77 ± 0.9) and ensuring the completeness (3.68 ± 0.8) and factual accuracy (3.66 ± 1.0) of generated text. Table 1 shows that AI chatbots exhibited more accuracy in answering questions about diagnosis (3.99 ± 0.8) than treatment (3.92 ± 1.0, $p = 0.494$) or recovery (3.75 ± 0.8, $p < 0.001$).

The AI-generated responses showed a high level of understandability but mixed actionability depending on the question type. As shown in Table 2, PEMAT understandability scores were high for patient questions about diagnosis ($84.0\% \pm 36.7\%$), treatment ($82.8\% \pm 37.8\%$), and recovery ($85.3\% \pm 35.4\%$) across

Question: *I understand that surgery is a common treatment option for bone sarcoma. What are the surgical options, and what are the potential risks and complications?*

Bard: Bone sarcoma surgery aims to remove the cancerous tumor while preserving as much of the affected bone and surrounding tissue as possible. The specific surgical approach depends on the location and size of the tumor, the type of bone involved, and your overall health.

Types of Limb-Sparing Surgery

Wide Excision: The most common type of limb-sparing surgery involves removing the tumor and a wide margin of healthy tissue around it.
Intramedullary Rodding: For tumors located inside the bone marrow, an intramedullary rod is inserted into the bone to replace the missing bone and provide support.
Bone Grafting: If a significant portion of bone is removed, bone grafting may be used to fill the defect. Bone grafts can be taken from another part of the patient's body or from a donor.
Endoprosthetic Reconstruction: In some cases, an endoprosthesis, a metal or plastic implant, may be used to replace the missing bone and restore function.

Amputation
Amputation is considered when limb-sparing surgery is not feasible, either due to the tumor's location or extent or because of concerns about the limb's function or stability. Amputation involves removing the affected limb, either at the joint or through the bone. Types of amputation include above-knee amputation (AKA), below-knee amputation (BKA), above-elbow amputation (AEA), and below-elbow amputation (BEA).

FIGURE 1 | Sample patient question with AI-generated answer. The response, generated by Bard, is edited for brevity. While demonstrating the capacity for AI chatbots to generate coherent, understandable answers for patients' questions, this response also highlights the risks of relying on AI chatbots to provide accurate counseling for diagnoses with complex treatment options. Raters gave this answer an average score of 2.5 out of 5 for factual accuracy and 2.75 out of 5 for basis on established data. Raters also commented that the response incorrectly cites intramedullary rodding as an appropriate treatment for a primary bone tumor. This finding highlights the potential for AI chatbots to produce false information, which has been previously described [1].

all AI chatbots. However, PEMAT actionability scores were higher for questions on diagnosis (80.2% ± 40.1%) than treatment (64.6% ± 48.3%) and recovery (60.6% ± 49.1%). GPT-4 responses to questions on diagnosis exhibited higher actionability (96.3% ± 19.2%) than any other subgroup (Table 3).

The AI-generated responses incorporated varying levels of empathy. GPT-4 exhibited the highest level of empathy across all question types, with an average rating of 4.12 versus 3.55 for Bard and 3.64 for Claude 2 (Table 4). Responses generated by Bard scored the lowest for diagnosis questions (3.20 ± 1.3), while those generated by Claude 2 scored the lowest in empathy for treatment (3.67 ± 1.2) and recovery (3.58 ± 1.5) questions.

GPT-4 produced the longest responses, with an average word count of 450, while Claude 2 produced the shortest responses at an average word count of 329. As shown in Table 5, Flesch–Kincaid Grade Level scores ranged between 10.28 ± 1.9 for diagnosis questions to 11.65 ± 1.8 for recovery questions and approached statistical significance ($p = 0.067$). For each question type, GPT-4 (11.56 ± 1.8) and Bard (11.49 ± 1.6) produced more complex text than Claude 2 (9.82 ± 1.7).

4 | Discussion

This study evaluated the responses of AI chatbots to common patient questions regarding diagnosis, treatment, and management of bone sarcoma. The three models (GPT-4, Bard, and Claude 2) generated answers of variable overall quality. The responses were empathetic, easily understandable, and relevant to the questions being asked, aligning with prior findings [7, 10, 13]. As in previous papers, chatbots frequently suggested discussing questions with the healthcare team, which is important considering the supplementary nature of medical AI chatbots [10, 11]. Furthermore, all three models provided balanced and unbiased responses. However, biases in training data (e.g., underrepresentation of certain demographics or variables) are known to impact the outcomes of AI models, warranting further investigation of how AI chatbots may perpetuate or mitigate certain biases in orthopedic applications [1, 24].

AI chatbots did not produce highly actionable responses. Actionability is the quality that patients with varying levels of health literacy can identify what they can do based on presented information [21]. The lack of actionability, especially in answers to questions on bone sarcoma treatment and recovery, complicates the usage of AI chatbots to guide patients in how to take control of their care. This challenge may be worsened by the complexity of AI-generated answers, which can worsen disparities for patients with low health literacy. As demonstrated in previous studies, the AI chatbots evaluated in our study consistently wrote at a ninth-to-twelfth-grade reading level, above national recommendations of sixth and eighth-grade reading levels from the American Medical Association and the National Institute of Health [7, 23, 25, 26]. While chatbot responses were not explicitly biased in their clinical evaluation

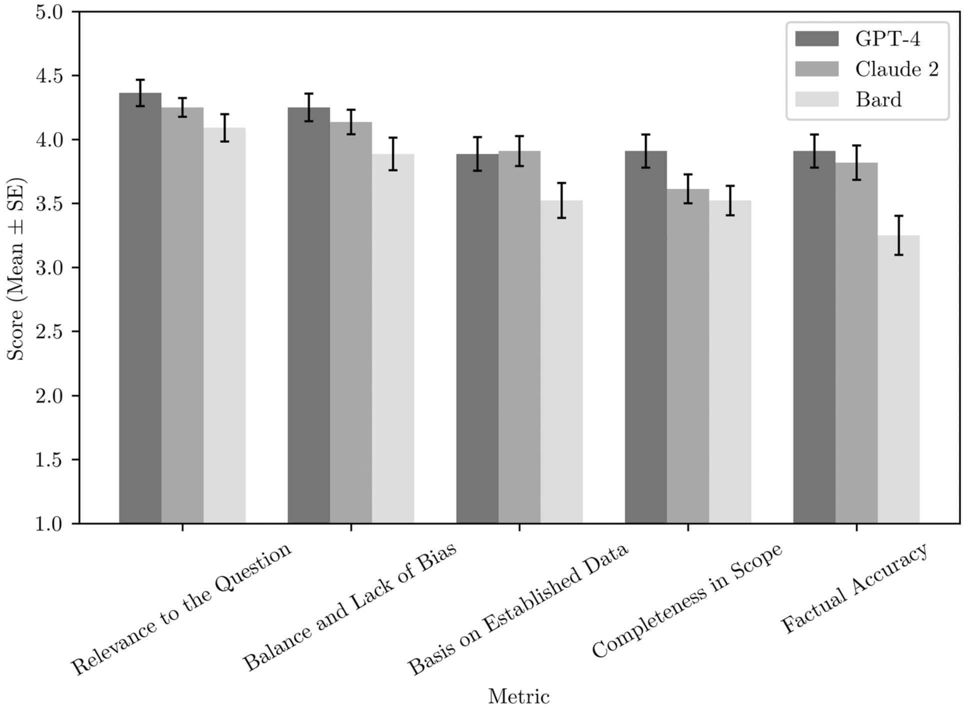

FIGURE 2 | Clinical evaluation of AI Chatbot responses to bone sarcoma patient questions. On average, AI chatbot-generated responses scored highest on relevance (4.23 ± 0.6), followed by balance/lack of bias (4.09 ± 0.7), basing responses on established data (3.77 ± 0.9), completeness in scope (3.68 ± 0.8), and factual accuracy (3.66 ± 1.0). GPT-4 and Claude 2 produced the highest-scoring answers for each clinical accuracy metrics, while Bard produced lower-quality responses.

TABLE 1 | Clinical evaluation of AI Chatbot responses by question type.

	Mean (standard deviation)			
Question type	**GPT-4 (OpenAI)**	**Bard (Google)**	**Claude 2 (Perplexity)**	**All**
Diagnosis	4.16 (0.8)	3.78 (0.8)	4.02 (0.7)	3.99 (0.8)
Treatment	4.28 (0.9)	3.60 (1.1)	3.90 (0.8)	3.92 (1.0)
Recovery	3.84 (0.8)	3.52 (0.9)	3.88 (0.8)	3.75 (0.8)

TABLE 2 | PEMAT understandability of AI Chatbot responses by question type.

	Mean (standard deviation)			
Question type	**GPT-4 (OpenAI)**	**Bard (Google)**	**Claude 2 (Perplexity)**	**All**
Diagnosis	87.82 (32.8)	87.90 (32.7)	76.43 (42.6)	84.04 (36.7)
Treatment	85.48 (35.5)	88.71 (31.9)	74.19 (44.1)	82.80 (37.8)
Recovery	79.23 (40.7)	87.97 (32.7)	88.64 (31.9)	85.32 (35.4)

TABLE 3 | PEMAT actionability of AI Chatbot responses by question type.

	Mean (standard deviation)			
Question type	**GPT-4 (OpenAI)**	**Bard (Google)**	**Claude 2 (Perplexity)**	**All**
Diagnosis	96.30 (19.2)	66.67 (48.0)	77.78 (42.4)	80.25 (40.1)
Treatment	72.22 (46.1)	68.75 (47.9)	50.00 (51.9)	64.58 (48.3)
Recovery	63.89 (48.7)	53.33 (50.7)	64.29 (48.8)	60.64 (49.1)

TABLE 4 | Empathy of AI Chatbot responses by question type.

Question type	Mean (standard deviation)			
	GPT-4 (OpenAI)	Bard (Google)	Claude 2 (Perplexity)	All
Diagnosis	4.27 (1.1)	3.20 (1.3)	3.67 (1.7)	3.71 (1.4)
Treatment	4.00 (1.3)	4.17 (0.8)	3.67 (1.2)	3.94 (1.1)
Recovery	4.00 (1.0)	3.67 (1.6)	3.58 (1.5)	3.75 (1.4)

TABLE 5 | Mean Flesch–Kincaid grade level of AI Chatbot responses by question type.

Question type	Mean (standard deviation)			
	GPT-4 (OpenAI)	Bard (Google)	Claude 2 (Perplexity)	All
Diagnosis	10.51 (1.3)	10.93 (1.8)	9.38 (2.4)	10.28 (1.9)
Treatment	11.94 (2.2)	11.43 (1.0)	10.42 (1.4)	11.26 (1.4)
Recovery	12.49 (1.4)	12.40 (2.2)	10.08 (0.6)	11.65 (1.8)

scores, their poor readability may effectively limit their utility for those with lower health literacy.

Perhaps the most significant concern with AI chatbots is providing misinformation through "hallucination," the generation of fabricated information that is presented as factually correct [1]. All three models scored an average of less than a 4 out of 5 in factual accuracy and completeness in scope, with Bard scoring significantly lower than GPT-4 and Claude 2. In one case presented in Section 3, a model gave critically inaccurate operative treatment recommendations by providing intramedullary nailing as a surgical option for a primary bone sarcoma. In another example, models produced overly optimistic responses to a question on recovery timelines and expected postoperative outcomes. These results underscore existing concerns that AI chatbots may provide misleading information through hallucination [10, 13].

Hallucinations may have been exacerbated by the fact that responses did not reference scientific literature, which is a known limitation of current models. There is significant evidence of fabrication in AI-generated bibliographic citations due to the inherently probabilistic nature of LLMs [7, 13, 27]. Furthermore, the broad training data and general-use capabilities of existing publicly available LLMs are misaligned with the highly specialized knowledge of medical subspecialties like orthopedic oncology, where diagnosis, treatment, and recovery require multidisciplinary expert-level care [28]. Without the development of purpose-built LLMs, there may be significant limitations in implementing AI chatbots as adjunctive tools. These findings may conflict with prior literature examining chatbot use in counseling patients with hip osteoarthritis, where diagnosis and treatment are generally less nuanced, suggesting that further training on disease-specific data is necessary before widespread use for all patients with musculoskeletal disease [10]. Finally, given LLMs do not yet possess complex reasoning skills, further development is necessary before adjunctive use for counseling patients with musculoskeletal disease [29, 30].

5 | Limitations

This study was limited to a small set of representative patient questions that were synthesized from text in online, anonymous bone sarcoma discussion forums and thus were not directly validated or phrased by patients. Three general-use AI chatbots were assessed; however, large-scale performance of LLMs may be inferior to LLMs that are fine-tuned for domain-specific applications, such as medical questions [31, 32]. Due to the probabilistic nature of LLMs, AI chatbots produce different responses when processing the same query multiple times and produce answers that are highly dependent on the inputted prompt; these additional sources of variability must be investigated. Also, chatbots can engage in conversation that responds to follow-up questions and may provide additional context or clarification. While responses generated by AI chatbots were quantitatively rated by the authors, they were not evaluated by patients or compared to answers that physicians would give. Further clinical investigation into the application of AI chatbots for patient counseling through patient evaluation of chatbot compared to physician responses is warranted.

6 | Conclusion

AI-powered chatbots are promising tools given their potential to supplement traditional patient counseling for patients with bone sarcoma. While chatbots may produce understandable and empathetic responses, patients and surgeons should approach using these tools with caution, as there may be significant limitations in the factual accuracy and actionability of responses. Further investigation into the limitations of AI chatbots, in addition to advancement in capabilities for patient education, is necessary before widespread incorporation into clinical use, particularly for patients with complex musculoskeletal diagnoses.

Acknowledgments

The authors have nothing to report.

Ethics Statement

University of California Los Angeles's Human Subjects Research Review Board granted an exemption of this study from further review.

Conflicts of Interest

Alexander B. Christ reports an honorarium with Enovis and the following consultancies: ZimmerBiomet, Onkos, Smith & Nephew, Daiichi Sankyo, and Globus Medical. Nicholas M. Bernthal reports funding from Deciphera Pharm, LLC, unrelated to this work. The other authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. B. Meskó and E. J. Topol, "The Imperative for Regulatory Oversight of Large Language Models (or Generative AI) in Healthcare," *npj Digital Medicine* 6 (2023): 120.

2. G. M. Iannantuono, D. Bracken-Clarke, C. S. Floudas, M. Roselli, J. L. Gulley, and F. Karzai, "Applications of Large Language Models in Cancer Care: Current Evidence and Future Perspectives," *Frontiers in Oncology* 13 (2023): 1268915.

3. R. Yang, T. F. Tan, W. Lu, A. J. Thirunavukarasu, D. S. W. Ting, and N. Liu, "Large Language Models in Health Care: Development, Applications, and Challenges," *Health Care Science* 2 (2023): 255–263.

4. S. Chatterjee, M. Bhattacharya, S. Pal, S.-S. Lee, and C. Chakraborty, "ChatGPT and Large Language Models in Orthopedics: From Education and Surgery to Research," *Journal of Experimental Orthopaedics* 10 (2023): 128.

5. H. Decker, K. Trang, J. Ramirez, et al., "Large Language Model−Based Chatbot vs Surgeon-Generated Informed Consent Documentation for Common Procedures," *JAMA Network Open* 6 (2023): e2336997.

6. H. L. Hofmann, G. A. Guerra, J. L. Le, et al., "The Rapid Development of Artificial Intelligence: GPT-4's Performance on Orthopedic Surgery Board Questions," *Orthopedics* 47, no. 2 (2023): 1–5.

7. E. T. Hurley, B. S. Crook, S. G. Lorentz, et al., "Evaluation High-Quality of Information from ChatGPT (Artificial Intelligence—Large Language Model) Artificial Intelligence on Shoulder Stabilization Surgery," *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 40, no. 3 (2023): 726–731.e6.

8. J. C. Kvedar, A. L. Fogel, E. Elenko, and D. Zohar, "Digital Medicine's March on Chronic Disease," *Nature Biotechnology* 34 (2016): 239–246.

9. L. A. Merrell, N. D. Fisher, and K. A. Egol, "Large Language Models in Orthopaedic Trauma: A Cutting-Edge Technology to Enhance the Field," *Journal of Bone and Joint Surgery* 105 (2023): 1383–1387.

10. A. P. Mika, J. R. Martin, S. M. Engstrom, G. G. Polkowski, and J. M. Wilson, "Assessing ChatGPT Responses to Common Patient Questions Regarding Total Hip Arthroplasty," *Journal of Bone and Joint Surgery* 105 (2023): 1519–1526.

11. S. Pagano, S. Holzapfel, T. Kappenschneider, et al., "Arthrosis Diagnosis and Treatment Recommendations in Clinical Practice: An Exploratory Investigation With the Generative AI Model GPT-4," *Journal of Orthopaedics and Traumatology* 24 (2023): 61.

12. M. G. Rizzo, N. Cai, and D. Constantinescu, "The Performance of ChatGPT on Orthopaedic in-Service Training Exams: A Comparative Study of the GPT-3.5 Turbo and GPT-4 Models in Orthopaedic Education," *Journal of Orthopaedics* 50 (2024): 70–75.

13. I. Seth, Y. Xie, A. Rodwell, et al., "Exploring the Role of a Large Language Model on Carpal Tunnel Syndrome Management: An Observation Study of ChatGPT," *The Journal of Hand Surgery* 48 (2023): 1025–1033.

14. D. Truhn, C. D. Weber, B. J. Braun, et al., "A Pilot Study on the Efficacy of GPT-4 in Providing Orthopedic Treatment Recommendations From Mri Reports," *Scientific Reports* 13 (2023): 20159.

15. J. Kim, Z. R. Cai, M. L. Chen, J. F. Simard, and E. Linos, "Assessing Biases in Medical Decisions via Clinician and AI Chatbot Responses to Patient Vignettes," *JAMA Network Open* 6 (2023): e2338050.

16. Z. Burningham, M. Hashibe, L. Spector, and J. D. Schiffman, "The Epidemiology of Sarcoma," *Clinical Sarcoma Research* 2 (2012): 14.

17. M.-P. Curado, B. Edwards, H. R. Shin, et al., *Cancer Incidence in Five Continents Volume IX* (Lyon: IARC, 2007), https://publications.iarc.fr/Book-And-Report-Series/Iarc-Scientific-Publications/Cancer-Incidence-In-Five-Continents-Volume-IX-2007.

18. K. Nakano, "Challenges of Systemic Therapy Investigations for Bone Sarcomas," *International Journal of Molecular Sciences* 23 (2022): 3540.

19. K. Castleton, T. Fong, A. Wang-Gillam, et al., "A Survey of Internet Utilization Among Patients With Cancer," *Supportive Care in Cancer* 19 (2011): 1183–1190.

20. A. E. Paulson, A. Stein, J. K. Kendal, N. M. Bernthal, and L. E. Wessel, "Most Patients With Bone Sarcomas Seek Emotional Support and Information About Other Patients' Experiences: A Thematic Analysis," *Clinical Orthopaedics & Related Research* 482 (2024): 161–171.

21. S. J. Shoemaker, M. S. Wolf, and C. Brach, "Development of the Patient Education Materials Assessment Tool (PEMAT): A New Measure of Understandability and Actionability for Print and Audiovisual Patient Information," *Patient Education and Counseling* 96 (2014): 395–403.

22. J. P. Kincaid, R. P. Fishburne Jr., L. C. Richard, and S. Brad, *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*, Report 56 (Millington, Tennessee: Institute for Simulation and Training, 1975), https://stars.library.ucf.edu/cgi/viewcontent.cgi?article=1055&context=istlibrary.

23. M. K. Rooney, G. Santiago, S. Perni, et al., "Readability of Patient Education Materials From High-Impact Medical Journals: A 20-Year Analysis," *Journal of Patient Experience* 8 (2021): 2374373521998847.

24. P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in Health and Medicine," *Nature Medicine* 28 (2022): 31–38.

25. S. Badarudeen and S. Sabharwal, "Assessing Readability of Patient Education Materials: Current Role in Orthopaedics," *Clinical Orthopaedics & Related Research* 468 (2010): 2572–2580.

26. H. Roberts, D. Zhang, and G. S. M. Dyer, "The Readability of AAOS Patient Education Materials: Evaluating the Progress Since 2008," *Journal of Bone and Joint Surgery* 98 (2016): e70.

27. W. H. Walters and E. I. Wilder, "Fabrication and Errors in the Bibliographic Citations Generated by ChatGPT," *Scientific Reports* 13 (2023): 14045.

28. H. Liu, W. Xue, Y. Chen, et al., "A Survey on Hallucination in Large Vision-Language Models," 2024, http://arxiv.org/abs/2402.00253.

29. J. Huang and K. C.-C. Chang, "Towards Reasoning in Large Language Models: A Survey," in *Findings of the Association for Computational Linguistics (ACL)*, (Association for Computational Linguistics, 2023), 1049–1065.

30. J. Huang, X. Chen, S. Mishra, et al., "Large Language Models Cannot Self-Correct Reasoning Yet," 2023, https://arxiv.org/abs/2310.01798.

31. H. W. Chung, L. Hou, S. Longpre, et al., "Scaling Instruction-Finetuned Language Models," 2022, http://arxiv.org/abs/2210.11416.

32. A. J. Thirunavukarasu, R. Hassan, S. Mahmood, et al., "Trialling a Large Language Model (ChatGPT) in General Practice With the Applied Knowledge Test: Observational Study Demonstrating Opportunities and Limitations in Primary Care," *JMIR Medical Education* 9 (2023): e46599.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.