Journal of Orthopaedic Surgery and Research

RESEARCH

Open Access

Is it a pediatric orthopaedic urgency or not? Can ChatGPT answer this question?

Sema Ertan Birsel[1*], Onur Oto[1], Baris Görgün[1], İrem Hazal İnan[2], Ali Şeker[3] and Muharrem İnan[1]

Abstract

Background Artificial intelligence (AI), particularly large language models (LLMs) such as ChatGPT, is increasingly studied in healthcare. This study evaluated the accuracy and reliability of the ChatGPT in guiding families on whether pediatric orthopaedic symptoms warrant emergency or outpatient care.

Methods Five common pediatric orthopaedic scenarios were developed, and ChatGPT was queried via a family-like language. For each scenario, two questions were asked: an initial general query and a follow-up query regarding the need for emergency versus outpatient care. ChatGPT's responses were evaluated for accuracy by analysing medical literature and online materials, completeness, and conciseness by two independent pediatric orthopaedic consultants, using a modified Likert scale. Responses were classified from "perfect" to "poor" based on total scores.

Results ChatGPT responded to 5 different scenarios commonly encountered in pediatric orthopaedics. The scores ranged from 8 to 10, with most responses requiring minimal clarification. While ChatGPT demonstrated strong diagnostic reasoning and actionable advice, occasional inaccuracies, such as recommending elevation for SCFE, highlighted areas for improvement.

Conclusion ChatGPT demonstrates potential as a supplemental tool for patient education and triage in pediatric orthopaedics, with generally accurate and accessible responses. These findings echo prior research on ChatGPT's potential and challenges in orthopaedics, emphasizing its role as a supplemental, not standalone, resource. While its strengths in providing accurate and accessible advice are evident, its limitations necessitate further refinement and cautious use under human supervision. Continued advancements in AI may further improve its safe integration into clinical care in pediartic orthopaedics.

Keywords Artificial intelligence, Pediatric orthopaedics, ChatGPT

*Correspondence:
Sema Ertan Birsel
drsemaertan@gmail.com
[1]Department of Orthopaedics and Traumatology, Ortopediatri Academy of Pediatric Otrhopaedics, Dikilitaş, Süleyman Seba Kompleksi, Hakkı Yeten Cd. 10/D, Beşiktaş, Istanbul 34349, Turkey
[2]Faculty of Medicine, Acibadem Mehmet Ali Aydinlar University, Ataşehir, Istanbul 34725, Turkey
[3]Cerrahpasa Faculty of Medicine, Department of Orthopaedics and Traumatology, Istanbul University- Cerrahpasa, Kocamustafapasa, Istanbul 34303, Turkey

Introduction

Artificial intelligence (AI) has rapidly emerged as a transformative technology across various domains, including healthcare. Among these advancements, large language models (LLMs) such as ChatGPT have gained widespread popularity due to their ability to generate human-like text responses to a wide range of queries. ChatGPT, developed by OpenAI, has been extensively used by patients and their families to obtain medical information, understand diagnoses, and make preliminary decisions regarding treatment options [1–3].

Families are beginning to explore large language models for health-related information. The increasing reliance on online medical information is not new. Studies have shown that over two-thirds of patients use the internet to seek answers to health-related questions before consulting a healthcare professional [1, 4]. However, the accuracy, reliability, and applicability of such information remain concerns [5]. ChatGPT stands out as a conversational AI tool that attempts to address these challenges by providing contextually relevant and evidence-based guidance. Previous research has demonstrated its potential to provide accurate and concise information for common orthopaedic conditions, including anterior cruciate ligament reconstruction and carpal tunnel syndrome [2, 6]. Despite these strengths, studies have also identified gaps in ChatGPT's ability to manage more complex or nuanced medical scenarios, underscoring the need for further evaluation [7, 8].

In pediatric orthopaedics, where timely and accurate decision-making is critical, the role of ChatGPT in guiding patients and families has yet to be fully explored. Shared decision-making plays a particularly important role in pediatric orthopaedics, where physicians must often collaborate not only with patients but also with parents or caregivers who are the primary decision-makers. Parents often face uncertainty in determining whether their child's symptoms warrant emergency care or can be addressed through outpatient consultation. This uncertainty can lead to delayed care in urgent cases or unnecessary utilization of emergency services. By assessing the ability of the ChatGPT to guide families in such scenarios, this study aims to determine its reliability as a supplemental tool in pediatric orthopaedic decision-making.

The aim of this study was to evaluate the ability of ChatGPT to guide patients (both in emergency and outpatient settings) by determining the accuracy and errors of its responses to common patient queries.

Materials & methods

In this study, five different scenarios were created to reflect common reasons for presentation to pediatric orthopaedics. These scenarios were developed by a study team consisting of two pediatric orthopedic consultants and one senior orthopedic surgeon, based on their clinical experience and frequent consultation reasons observed in both emergency and outpatient settings. The selection of scenarios was further informed by institutional data and previous literature describing the most common musculoskeletal complaints in children [9, 10]. Each scenario was formulated in a way that would realistically reflect how a patient's family might describe the situation when seeking advice. The phrasing of the questions was informed by commonly used expressions and concerns encountered during in-person consultations in both emergency and outpatient pediatric orthopaedic settings. The ChatGPT model used in this study was GPT-4o-mini, accessed via the ChatGPT app (OpenAI). The ChatGPT was asked about these scenarios in sentences that could be asked by the patient's family and asked to answer whether they should visit the emergency department or the outpatient department for these situations. For each scenario, two questions and/or a follow-up input either in the form of a question or an additional clinical detail were asked (Table 1). The first objective is to determine the general history of the condition. A more descriptive second question was formed on the basis of the answer given by the system to the first question. The responses were analyzed for accuracy according to

Table 1 Questions posed to ChatGPT as if asked by patients' families

Question 1	***A***	*My child has a pain in his right knee that started today. A patient had an upper respiratory tract infection before. Should I take my son to the emergency or outpatient clinic?*
	B	*He has moderate pain without redness and warmth around the knee.*
Question 2	***A***	*My child has a pain in his right knee that started today. Should I go to the emergency or outpatient clinic for this?*
	B	He has moderate pain, redness and warmth on his knee.
Question 3	***A***	My child has pain in the elbow. There is no history of any fall. I grabbed his arm and lifted him while getting off the stairs and the pain started after that. In this case, should I go to the emergency or outpatient clinic?
	B	He has mild-modarate pain and could not move his elbow
Question 4	***A***	My child fell in the playground. He has pain and swelling in his elbow. Should I take him to the emergency or outpatient clinic?
	B	He has moderate pain and could not move his elbow. There is not any numbless and tingling on his extremity.
Question 5	***A***	I have a 13-year-old overweight son. Starting 3 weeks ago, he had mild pain in the thigh area. However, the pain gradually increased. Should I take my son to the emergency or outpatient clinic?
	B	So what should I do and what should I not do until I make an appointment and go to the outpatient clinic?

Table 2 Classification of answers according to scores

Perfect answer	10–12 points
Good answer	8–9 points
Moderate answer	6–7 points
Poor answer	3–5 points

the literature and online resources such as; OrthoKids, Orthobullets, Pediatric Orthopaedic Society of North America (POSNA) and European Pediatric Orthopaedic Society (EPOS) websites.

Responses were classified on the basis of the study by Casey et al. using Likert scales assessing accuracy, completeness, and conciseness [2]. Accuracy was scored on a 12-point scale divided into six categories, completeness on a 6-point scale divided into three categories, and conciseness on a 6-point scale divided into three categories (Table 2). For simplicity in reporting, the overall accuracy scores were grouped into four categorical levels: "perfect" (10–12 points), "good" (8–9 points), "moderate" (6–7 points), and "poor" (3–5 points).

The informative text provided by ChatGPT was recorded for this study by a non-observer senior orthopaedic surgeon. Two pediatric orthopaedic consultants independently reviewed the responses to each question and given scores for them.

In addition to the Likert scale scores, responses were qualitatively categorized based on the degree of clarification needed. An "excellent response" required no further clarification, while a "satisfactory response" indicated that only minimal clarification was necessary. These qualitative categories helped provide a clearer interpretation of ChatGPT's performance.

Results

Two authors rated the responses with an inter-rater reliability (Cohen's kappa) of 0.71 (95% CI 0.39–1.00), which is representative of substantial agreement.

Question 1-a *My child has a pain in his right knee that started today. A patient had an upper respiratory tract infection before. Should I take my son to the emergency or outpatient clinic?*

Question 1-b *He has moderate pain without redness and warmth around the knee.*

Answer Appendix file.

Likert score 10.

Analysis Excellent response requiring no further clarification.
ChatGPT suggested that the patient schedule an appointment at the outpatient clinic for this scenario which is suitable for reactive arthritis. ChatGPT provided an excellent summary according to the patient's complaints and gave appropriate information about the details necessary for more accurate guidance. In response to the second question, to which details were added, appropriate guidance and general follow-up recommendations were appropriately indicated.

Question 2-a *My child has a pain in his right knee that started today. Should I go to the emergency or outpatient clinic for this?*

Question 2-b He has moderate pain, redness and warmth on his knee.

Answer Appendix file.

Likert score 9.5.

Analysis Satisfactory response requiring minimal clarification.
ChatGPT suggested that the patients visit emergency department as soon as possible for this scenario, which is appropriate for septic arthritis. ChatGPT provided a satisfactory summary according to the patient's complaints and provided appropriate information about the details needed for more accurate advice. In response to the second question, to which details were added, appropriate guidance and possible diagnosis was provided by ChatGPT. Furthermore, it also makes it easier for families by listing what they need to take with them when they are traveling to the emergency room.

Question 3-a My child has pain in the elbow. There is no history of any fall. I grabbed his arm and lifted him while getting off the stairs and the pain started after that. In this case, should I go to the emergency or outpatient clinic?

Question 3-b He has mild moderate pain and could not move his elbow.

Answer Appendix file.

Likert score 9.5.

Analysis Satisfactory response requiring minimal clarification.
ChatGPT suggested that this patient had nursemaid's elbow condition. In the first question, without giving details, it stated that there is no need to go to the emergency room with the sentence 'Unless the pain is severe, your child is very distressed or the arm does look deformed, you do not necessarily need to go to the emergency room'. However, in response to the second

question, where details were given, it was suggested to go to the emergency room or outpatient clinic as soon as possible.

Question 4-a My child fell in the playground. He has pain and swelling in his elbow. Should I take him to the emergency or outpatient clinic?

Question 4-b He has moderate pain and could not move his elbow. There is not any numbness and tingling on his extremity.

Answer Appendix file.

Likert score 8.

Analysis Satisfactory response requiring minimal clarification.
ChatGPT gave suggestions for more information in the first question where the details were not given in full. In the first question without giving details, it stated that there is no need to go to the emergency room with the sentence 'some ability to move the elbow and arm, even if limited'. After receiving the details, the patient was recommended to go to the emergency department for possible elbow fracture or dislocation. It also gave the things to be done until the first treatment and these were appropriate suggestions.

Question 5-a I have a 13-year-old overweight son. Starting 3 weeks ago, he had mild pain in the thigh area. However, the pain gradually increased. Should I take my son to the emergency or outpatient clinic?

Question 5-b So what should I do and what should I not do until I make an appointment and go to the outpatient clinic?

Answer Appendix file.

Likert score 8.5.

Analysis Satisfactory response requiring minimal clarification.
The ChatGPT made a preliminary diagnosis of SCFE in accordance with the complaints and recommended that the patient be referred to an outpatient clinic. When asked what to do and what not to do until the outpatient clinic referral, it gave generally appropriate recommendations except for elevation, which has no place in the treatment of SCFE.

Discussion

Artificial intelligence (AI), specifically large language models (LLMs) such as ChatGPT, is increasingly utilized in healthcare for tasks ranging from patient education to clinical decision support. This study assessed the accuracy and completeness of ChatGPT's responses to pediatric orthopedic scenarios, providing insights into its potential applications in clinical practice. Within the limited scope of this study, ChatGPT provided generally accurate and appropriate guidance in the selected pediatric orthopedic scenarios. Our findings are based on a limited set of scenarios and single queries, so broader clinical applicability of ChatGPT should be interpreted with caution and warrants further study.

In scenarios involving common conditions like reactive arthritis and septic arthritis, ChatGPT demonstrated a high level of diagnostic reasoning and offered actionable advice, which was consistent with best practices. Similarly, for conditions like nursemaid's elbow and slipped capital femoral epiphysis (SCFE), its guidance was broadly accurate but occasionally included recommendations (e.g., limb elevation) that lacked evidence-based support. These observations align with broader evaluations of ChatGPT in orthopaedics, such as studies by Sparks et al. and Smith et al., which highlighted the chatbot's strengths in basic medical reasoning while acknowledging limitations in nuanced scenarios [1, 7, 11].

ChatGPT's performance in this study mirrors findings from recent literature. For example, Gaudiani et al. demonstrated that ChatGPT-4 outperformed Google in providing accurate and complete responses to patient inquiries about anterior cruciate ligament reconstruction, emphasizing its potential as a patient education tool [6]. Similarly, Hlavinka et al. reported that ChatGPT responses to be more detailed but less readable compared to Google, highlighting the trade-off between comprehensiveness and accessibility [4]. This duality was also evident in our study, where ChatGPT's responses were comprehensive but occasionally included unnecessary or incorrect details, underscoring the need for careful oversight.

The study by Casey et al. further supports ChatGPT's utility, demonstrating that it offers concise, accurate guidance for common conditions like carpal tunnel syndrome, comparable to top-tier online medical resources [2]. However, our findings echo concerns raised by Kim et al. about ChatGPT's inconsistencies and occasional inaccuracies, which may hinder its application in more specialized or high-stakes clinical scenarios [8].

Additionally, Zusman et al. evaluated ChatGPT's appropriateness in addressing parental concerns for pediatric orthopaedic conditions compared to OrthoKids, a specialty-governed resource by the Pediatric Orthopaedic Society of North America (POSNA) [12]. Their

findings revealed that ChatGPT's responses aligned with OrthoKids' information in 93% of cases, although inconsistencies in recommending specialist consultation and unprompted treatment recommendations were noted. These results underscore ChatGPT's potential as a reliable supplemental tool, while highlighting areas for improvement in ensuring accurate referrals and minimizing unsolicited advice [3].

ChatGPT's ability to provide fast, evidence-informed responses has significant implications for patient education and triage. For pediatric orthopaedic conditions, it can bridge informational gaps for families, facilitating more informed decision-making. Studies, such as those by Yüce et al. and Artamonov et al., have emphasized its potential to complement traditional medical consultations, particularly in scenarios requiring detailed explanations of surgical procedures or diagnostic pathways [7, 13] .

Nevertheless, caution is warranted. ChatGPT's inability to interpret clinical images or perform individualized assessments limits its utility in contexts requiring precise diagnostic input. As Sparks et al. highlighted, ChatGPT's underperformance on orthopaedic board-style questions underscores the importance of relying on expert oversight in clinical decision-making [1].

Improving ChatGPT's integration of evidence-based guidelines could enhance its reliability and application in specialized fields like pediatric orthopaedics. The development of healthcare-specific LLMs, trained on validated medical literature, may address current gaps in accuracy and context sensitivity. Additionally, the incorporation of real-time clinical data and diagnostic imaging capabilities could expand its utility.

This study has several limitations. One limitation of our study is that the second prompt for each scenario was dynamically generated based on ChatGPT's response to the first question. While this approach was intended to mimic real-life conversational flow between patients' families and an AI tool, it introduced a degree of subjectivity and limited standardization. As such, variability in the phrasing of follow-up inputs was not formally assessed for intra- or inter-rater reliability, which may affect reproducibility. Second limitation of this study is that the questions were developed based on clinical experience and institutional data rather than being directly extracted from standardized educational resources such as OrthoKids, OrthoBullets, POSNA, and EPOS. Incorporating questions directly derived from these sources may enhance methodological difficulty in future research. Another limitation of this study is the relatively small number of queries used, which contributed to a wide confidence interval in the inter-rater reliability analysis. Future research with a larger query set is needed to improve the precision and power of reliability estimates. Lastly, the evaluations by the two pediatric orthopedic consultants were conducted in a single session, and intra-rater reliability was not assessed. Therefore, the consistency of individual reviewers' ratings over time could not be evaluated. Furthermore, we did not assess response consistency through repeated queries, nor did we control for ChatGPT versioning and update-related variability, which may impact reproducibility. The questions were not validated for readability or language clarity, and it remains uncertain whether they reflect real-world parental phrasing. In addition, the questions provided to ChatGPT were structured and clinically focused, which may not reflect the fragmented or non-specific information typically conveyed by families in real-world consultations.

Conclusion

This study reinforces the promise of the ChatGPT as a supplemental tool for patient education and preliminary guidance in pediatric orthopaedic scenarios. However, its limitations necessitate cautious application, with human oversight remaining essential. As AI continues to evolve, its role in healthcare will likely expand, underscoring the need for ongoing evaluation and refinement.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13018-025-05981-z.

Supplementary Material 1

Author contributions

All authors contributed to the study conception and design. Material preparation and data collection were performed by Sema Ertan Birsel and İrem Hazal İnan. Data analysis were performed by Onur Oto and Barış Görgün. The manuscript was written by Sema Ertan Birsel and Ali Şeker. Concept was designed by Muharrem İnan. Supervisor of thestudy was Muharrem İnan. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding

None of the authors, their immediate families, and any research foundation with which they are affiliated received any financial payments or other benefits from any commercial entity related to the subject of this article. This work did not receive funding from any organization.

Data availability

No datasets were generated or analysed during the current study.

Declarations

Competing interests

The authors declare no competing interests.

Received: 7 April 2025 / Accepted: 28 May 2025
Published online: 04 June 2025

References

1. Sparks CA, Kraeutler MJ, Chester GA, Contrada EV, Zhu E, Fasulo SM, Scillia AJ. Inadequate performance of ChatGPT on orthopedic board-style written exams. Cureus. 2024;16(6):e62643. PMID: 39036109; PMCID: PMC11258215. https://doi.org/10.7759/cureus.62643.
2. Casey JC, Dworkin M, Winschel J, Molino J, Daher M, Katarincic JA, Gil JA, Akeman E. ChatGPT: a concise google alternative for people seeking accurate and comprehensive carpal tunnel syndrome information. Hand Surg Rehabil. 2024;101757. Epub ahead of print. PMID: 39103051. https://doi.org/10.1016/j.hansur.2024.101757.
3. Amaral JZ, Schultz RJ, Martin BM, Taylor T, Touban B, McGraw-Heinrich J, McKay SD, Rosenfeld SB, Smith BG. Evaluating chat generative pre-trained transformer responses to common pediatric in-toeing questions. J Pediatr Orthop. 2024;44(7):e592–7. Epub 2024 Apr 30. PMID: 38686934. https://doi.org/10.1097/BPO.0000000000002695.
4. Hlavinka WJ, Sontam TR, Gupta A, Croen BJ, Abdullah MS, Humbyrd CJ. Are large language models a useful resource to address common patient concerns on hallux valgus? A readability analysis. Foot Ankle Surg. 2024;6:S1268-7731(24)00181-4. Epub ahead of print. PMID: 39117535. https://doi.org/10.1016/j.fas.2024.08.002.
5. Pirkle S, Yang J, Blumberg TJ. Do ChatGPT and gemini provide appropriate recommendations for pediatric orthopaedic conditions?? J Pediatr Orthop. 2025;45(1):e66–71. Epub 2024 Aug 22. PMID: 39171426. https://doi.org/10.1097/BPO.0000000000002797.
6. Gaudiani MA, Castle JP, Abbas MJ, Pratt BA, Myles MD, Moutzouros V, Lynch TS. ChatGPT-4 generates more accurate and complete responses to common patient questions about anterior cruciate ligament reconstruction than google's search engine. Arthrosc Sports Med Rehabil. 2024;6(3):100939. PMID: 39006779; PMCID: PMC11240040. https://doi.org/10.1016/j.asmr.2024.100939.
7. Artamonov A, Bachar-Avnieli I, Klang E, Lubovsky O, Atoun E, Bermant A, Rosinsky PJ. Responses from ChatGPT-4 show limited correlation with expert consensus statement on anterior shoulder instability. Arthrosc Sports Med Rehabil. 2024;6(3):100923. PMID: 39006799; PMCID: PMC11240044. https://doi.org/10.1016/j.asmr.2024.100923.
8. 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):669–73. Epub 2024 Mar 7. PMID: 39092297; PMCID: PMC11262944. https://doi.org/10.4055/cios23179.
9. de Inocencio J, Carro MÁ, Flores M, Carpio C, Mesa S, Marín M. Epidemiology of musculoskeletal pain in a pediatric emergency department. Rheumatol Int. 2016;36(1):83–9. Epub 2015 Aug 11. PMID: 26259985. https://doi.org/10.1007/s00296-015-3335-9.
10. Schwend RM, Geiger J. Outpatient pediatric orthopedics. Common and important conditions. Pediatr Clin North Am. 1998;45(4):943–71. PMID: 9728195. https://doi.org/10.1016/s0031-3955(05)70054-7.
11. Smith AM, Jacquez EA, Argintar EH. Assessing the efficacy of an AI-Powered chatbot (ChatGPT) in providing information on orthopedic surgeries: a comparative study with expert opinion. Cureus. 2024;16(6):e63287. PMID: 39070516; PMCID: PMC11283313. https://doi.org/10.7759/cureus.63287.
12. Zusman NL, Bauer M, Mann J, Goldstein RY. AI = Appropriate insight? ChatGPT appropriately answers parents' questions for common pediatric orthopaedic conditions. J Pediatr Orthop Soc North Am. 2023;5(4):762–71. ISSN 2768–2765. https://doi.org/10.55275/JPOSNA-2023-762.
13. Yüce A, Erkurt N, Yerli M, Misir A. The potential of ChatGPT for high-quality information in patient education for sports surgery. Cureus. 2024;16(4):e58874. PMID: 38800159; PMCID: PMC11116739. https://doi.org/10.7759/cureus.58874.

Publisher's note