Journal of the Pediatric Orthopaedic Society of North America 11 (2025) 100164

Contents lists available at ScienceDirect

Journal of the Pediatric Orthopaedic Society of North America

journal homepage: www.jposna.com

Original Research

Performance of Artificial Intelligence in Addressing Questions Regarding the Management of Pediatric Supracondylar Humerus Fractures

John D. Milner, MD[1,*]; Matthew S. Quinn, MD[1]; Phillip Schmitt, BS[1]; Ashley Knebel, BA[1]; Jeffrey Henstenburg, MD[2]; Adam Nasreddine, MD[2]; Alexandre R. Boulos, MD, MPH[1]; Jonathan R. Schiller, MD[1]; Craig P. Eberson, MD[1]; Aristides I. Cruz Jr., MD, MBA[2]

[1] *Department of Orthopaedic Surgery, Brown University, Warren Alpert Medical School, Providence, RI, USA*
[2] *Division of Sports Medicine, Boston Children's Hospital, Boston, MA, USA*

ARTICLE INFO

Keywords:
Pediatric elbow
ChatGPT
Artificial intelligence
Supracondylar humerus
Fracture

ABSTRACT

Background: The vast accessibility of artificial intelligence (AI) has enabled its utilization in medicine to improve patient education, augment patient–physician communications, support research efforts, and enhance medical student education. However, there is significant concern that these models may provide responses that are incorrect, biased, or lacking in the required nuance and complexity of best practice clinical decision-making. Currently, there is a paucity of literature comparing the quality and reliability of AI-generated responses. The purpose of this study was to assess the ability of ChatGPT and Gemini to generate reponses to the 2022 American Academy of Orthopaedic Surgeons' (AAOS) current practice guidlines on pediatric supracondylar humerus fractures. We hypothesized that both ChatGPT and Gemini would demonstrate high-quality, evidence-based responses with no significant difference between the models across evaluation criteria.

Methods: The responses from ChatGPT and Gemini to responses based on the 14 AAOS guidelines were evaluated by seven fellowship-trained pediatric orthopaedic surgeons using a questionnaire to assess five key characteristics on a scale from 1 to 5. The prompts were categorized into nonoperative or preoperative management and diagnosis, surgical timing and technique, and rehabilitation and prevention. Statistical analysis included mean scoring, standard deviation, and two-sided t-tests to compare the performance between ChatGPT and Gemini. Scores were then evaluated for inter-rater reliability.

Results: ChatGPT and Gemini demonstrated consistent performance across the criteria, with high mean scores across all criteria except for evidence-based responses. Mean scores were highest for clarity (ChatGPT: 3.745 ± 0.237, Gemini 4.388 ± 0.154) and lowest for evidence-based responses (ChatGPT: 1.816 ± 0.181, Gemini: 3.765 ± 0.229). There were notable statistically significant differences across all criteria, with Gemini having higher mean scores in each criterion ($P < .001$). Gemini achieved statistically higher ratings in the relevance ($P = .03$) and evidence-based ($P < .001$) criteria. Both large language models (LLMs) performed comparably in the accuracy, clarity, and completeness criteria ($P > .05$).

Conclusions: ChatGPT and Gemini produced responses aligned with the 2022 AAOS current guideline practices for pediatric supracondylar humerus fractures. Gemini outperformed ChatGPT across all criteria, with the greatest difference in scores seen in the evidence-based category. This study emphasizes the potential for LLMs, particularly Gemini, to provide pertinent clinical information for managing pediatric supracondylar humerus fractures.

Key Concepts:

(1) The accessibility of artificial intelligence has enabled its utilization in medicine to improve patient education, support research efforts, enhance medical student education, and augment patient–physician communications.
(2) There is a significant concern that artificial intelligence may provide responses that are incorrect, biased, or lacking in the required nuance and complexity of best practice clinical decision-making.
(3) There is a paucity of literature comparing the quality and reliability of AI-generated responses regarding management of pediatric supracondylar humerus fractures.

* Corresponding author: Department of Orthopaedic Surgery, Brown University Warren Alpert Medical School, 593 Eddy Street, Providence, RI 02903, USA.
E-mail address: john_milner@brown.edu (J.D. Milner).

https://doi.org/10.1016/j.jposna.2025.100164
Received 17 September 2024; Received in revised form 4 January 2025; Accepted 8 January 2025
Available online 9 March 2025

(4) In our study, both ChatGPT and Gemini produced responses that were well aligned with the AAOS current guideline practices for pediatric supracondylar humerus fractures; however, Gemini outperformed ChatGPT across all criteria, with the greatest difference in scores seen in the evidence-based category.

Level of Evidence: Level II

Introduction

The accessibility and utilization of artificial intelligence (AI) has increased dramatically with the introduction of platforms like ChatGPT (OpenAI, San Francisco, CA, USA) and Gemini (formerly Bard) (Google, Mountain View, CA, USA) [1,2]. ChatGPT gained 100 million users in the first two months of its launch, making it one of the fastest-growing consumer applications [3]. These large language model (LLM)–based AI chatbots give users near immediate access to massive amounts of information across various disciplines, including medicine, with the click of a button [4].

LLMs in medicine have already been used to improve patient education, augment patient–physician communications, support research efforts, and enhance medical student education [5]. Similar trends are also found in orthopaedics, with prior research demonstrating an interest amongst orthopaedic surgeons in exploring how LLMs can support clinical decision-making, update treatment guidelines through the synthesis of medical literature, and enhance patient satisfaction [6]. With this potential in mind, there is significant concern that these models may provide incorrect, biased responses or lack the required nuance and complexity of best practice clinical decision-making [3,6]. Furthermore, these responses may include literature citations that do not exist, which calls into question the reliability of these responses [7]. The risks of providing patients with error-prone information before and during treatment cannot be understated. Nearly two-thirds of orthopaedic patients report using the internet to obtain orthopaedic information related to their condition; however, only half plan to discuss the information with their physician [8]. As LLMs become more accessible and continue to grow in popularity, patients will likely increasingly use these online resources to learn about their conditions and treatment options. With increasing use and inconsistent discussion of findings with their surgeons, information obtained through LLMs may obscure patients' understanding of their condition and negatively impact the shared decision-making process.

Due to the relatively recent introduction of publicly accessible AI chatbots, few studies have attempted to describe the accuracy and appropriateness of LLM responses to questions about orthopaedic conditions and treatments. Furthermore, the available research has focused primarily on ChatGPT as the sole LLM in the study [9–11]. In pediatric supracondylar humerus fracture management, to our knowledge, no study has compared LLM responses with current guidelines on managing these fractures. This study aimed to assess ChatGPT and Gemini's ability to generate responses based on the 2022 American Academy of Orthopaedic Surgeons (AAOS) guidelines on pediatric supracondylar humerus fractures [12]. We hypothesized that both LLMs responses would achieve high ratings across quality criteria, with no significant difference between the models across evaluation criteria.

Methods

Study design

The methodology for this study was adapted from the previous literature on LLM response consistency with AAOS guidelines [10]. First, the 14 guidelines were reviewed and LLM prompts were generated by the authors (Table 1). The prompts were subclassified into three categories: (1) nonoperative or preoperative evaluation and diagnosis, (2) surgical timing and technique, (3) rehabilitation and prevention. The guidelines had varying recommendation strength, with two supported by moderate evidence, two by limited evidence, eight by inconclusive evidence, and two by consensus opinion. Each prompt was inputted into ChatGPT (version 4.0), and Gemini (September 04, 2024 update) initiated each time with a new chat session to eliminate bias from previous prompts. ChatGPT and Gemini responses to each prompt were transferred to a survey. The fellowship-trained pediatric orthopaedic surgeons were asked to evaluate the LLM responses using a previously established questionnaire [10]. Institutional review board review was not required since no protected health information was collected.

Assessment of outputs

Surgeons were asked to grade LLM responses on a 1–5 scale based on six categories: relevance, accuracy, clarity, completeness, evidence-based, and consistency. These categories were derived from previous LLM quality assessment methodology and the literature on information quality [10,13]. The specific questions asked were the following:

- "Does the answer directly address the question asked?" (relevance).
- "Is the information provided in the answer correct and up to date, compared with current AAOS clinical practice guidelines?" (accuracy).
- "Is the answer easy to understand and well-organized?" (clarity).
- "Does the answer provide all the information necessary to fully answer the question?" (completeness).
- "Does the answer provide relevant research and data to support the information provided?" (evidence-based).

After both the ChatGPT and Gemini responses had been presented to the surgeon, a final question was asked:

- "How similar are the two different responses for each question?" (consistency).

Statistical analysis

Mean and standard deviation were calculated for each category. Additionally, inter-rater reliability was calculated. Data analysis was performed using Microsoft Excel.

Results

ChatGPT and Gemini responses received mean scores greater than 3/5 across all criteria except for evidence-based responses (Table 2).

For prompts categorized under diagnosis and nonoperative/preoperative management, both LLMs achieved high scores in clarity (ChatGPT: 4.107 ± 0.382, Gemini: 4.464 ± 0.294) and low scores in evidence-based responses (ChatGPT: 1.857 ± 0.4, Gemini: 3.357 ± 0.516). Gemini achieved statistically higher ratings in the relevance ($P = .03$) and evidence-based ($P < .001$) criteria. Both LLMs performed comparably in the accuracy, clarity, and completeness criteria ($P > .05$).

In the surgical timing and technique category, the highest scores for both LLMs were in the clarity criterion (ChatGPT: 3.592 ± 0.379, Gemini: 4.388 ± 0.234). The lowest score for ChatGPT was in the evidence-based criterion (1.776 ± 0.244), while the lowest score for Gemini was in the accuracy criterion (3.959 ± 0.347). Across all criteria, Gemini scored higher than ChatGPT with a statistically significant difference ($P < .001$).

The rehabilitation and prevention category showed the highest mean score for ChatGPT in completeness (3.964 ± 0.356) and for Gemini in

Table 1.
List of 2022 AAOS Guidelines on Pediatric Supracondylar Humerus fracture, recommendation strength, and author-generated LLM prompt.

Recommendation strength	AAOS guideline	LLM prompt
Moderate	We suggest nonsurgical immobilization of the injured limb for patients with acute (e.g. Gartland type I) or non-displaced pediatric supracondylar fractures of the humerus or posterior fat pad sign	In pediatric patients with acute (Gartland type I) non-displaced supracondylar fractures of the humerus or who have a posterior fat pad sign, what type of elbow immobilization should be utilized?
Moderate	We suggest closed reduction with pin fixation for patients with displaced (Gartland type II and III, and displaced flexion) pediatric supracondylar fractures of the humerus.	What is the appropriate treatment in pediatric patients with displaced (Gartland type II and III and displaced flexion) supracondylar fractures of the humerus?
Limited	The practitioner might use two or three laterally introduced pins to stabilize the reduction of displaced pediatric supracondylar fractures of the humerus. Considerations of potential harm indicate that the physician might avoid the use of a medial pin.	In pediatric patients with displaced supracondylar fractures of the humerus, should pins be used to stabilize the reduction, and if so, how many pins should be used?
Inconclusive	We cannot recommend for or against using an open incision to introduce a medial pin to stabilize the reduction of displaced pediatric supracondylar fractures of the humerus.	In pediatric patients with displaced supracondylar fractures of the humerus, how should a medial pin be used to stabilize the reduction?
Inconclusive	We are unable to recommend for or against a time threshold for reduction of displaced pediatric supracondylar fractures of the humerus without neurovascular injury.	In pediatric patients with displaced supracondylar fractures of the humerus without neurovascular injury, is there a recommended time threshold for reduction?
Limited	The practitioner might perform open reduction for displaced pediatric supracondylar fractures of the humerus with varus or other malposition after closed reduction.	How should varus or other malposition of displaced supracondylar humerus fractures in pediatric patients be addressed?
Consensus	In the absence of reliable evidence, the opinion of the work group is that emergent closed reduction of displaced pediatric supracondylar humerus fractures be performed in patients with decreased perfusion of the hand.	In pediatric patients with displaced supracondylar fractures of the humerus with decreased perfusion to the hand, should emergent closed reduction be performed?
Consensus	In the absence of reliable evidence, the opinion of the work group is that open exploration of the antecubital fossa be performed in patients who have absent wrist pulses and are underperfused after reduction and pinning of displaced pediatric supracondylar humerus fractures.	In pediatric patients with displaced supracondylar fractures of the humerus who have absent wrist pulses and are underperfused after reduction and pinning, should open exploration of the antecubital fossa be performed?
Inconclusive	We cannot recommend for or against open exploration of the antecubital fossa in patients with absent wrist pulses but with a perfused hand after reduction of displaced pediatric supracondylar humerus fractures.	In pediatric patients with displaced supracondylar fractures of the humerus with absent wrist pulses but with a perfused hand after reduction, should open exploration of the antecubital fossa be performed?
Inconclusive	We are unable to recommend an optimal time for removal of pins and mobilization in patients with displaced pediatric supracondylar fractures of the humerus.	In pediatric patients with displaced supracondylar fractures of the humerus, is there an optimal time for removal of pins and mobilization?
Inconclusive	We are unable to recommend for or against routine supervised physical or occupational therapy for patients with pediatric supracondylar fractures of the humerus.	In pediatric patients with supracondylar fractures of the humerus, should supervised physical or occupational therapy be routinely utilized?
Inconclusive	We are unable to recommend an optimal time for allowing unrestricted activity after injury in patients with healed pediatric supracondylar fractures of the humerus.	In pediatric patients with healed supracondylar fractures of the humerus, is there an optimal time for allowing unrestricted activity after injury?
Inconclusive	We are unable to recommend optimal timing of or indications for electrodiagnostic studies or nerve exploration in patients with nerve injuries associated with pediatric supracondylar fractures of the humerus.	In pediatric patients with nerve injuries associated with supracondylar fractures of the humerus, what is the optimal time or indications for electrodiagnostic studies or nerve exploration?
Inconclusive	We are unable to recommend for or against open reduction and stable fixation for adolescent patients with supracondylar fractures of the humerus.	In adolescent patients with supracondylar fractures of the humerus, should open reduction and stable fixation be utilized?

AAOS, American Academy of Orthopaedic Surgeons; LLM, large language model.

relevance (4.071 ± 0.362). The lowest mean scores were in the evidence-based criteria for ChatGPT (1.893 ± 0.339) and accuracy for Gemini (2.714 ± 0.542). Gemini had higher mean scores that were statistically significant in the clarity ($P = .013$) and evidence-based ($P < .001$) criteria, but other criteria had comparable ratings ($P > .05$).

The consistency of responses between the two LLMs was rated at a mean of 2.571 ± 0.213 overall. Consistency was highest for the diagnosis and nonoperative/preoperative management category (3.036 ± 0.409) and lowest for the surgical timing and technique category (2.224 ± 0.282). The inter-rater reliability (IRR) ranged from 0.44 to 0.5 and had a mean of 0.47 (Table 3).

Discussion

The present study's results demonstrate ChatGPT and Gemini's effectiveness in generating responses aligned with the 2022 AAOS guidelines for pediatric supracondylar humerus fracture management. For both LLMs, mean scores were highest in clarity and lowest for evidence-based responses. The generally low scores in evidence-based criteria are likely due to limited citations to support the responses. Gemini provided citations more often than ChatGPT, which never offered a bibliography; however, the number of citations in each reaction was usually limited. Additionally, the citations from Gemini were not always from peer-reviewed articles or reputable journals. Overall, Gemini had a statistically significant higher mean score in each criterion. Specifically, Gemini scored statistically higher in relevance and evidence-based criteria for prompts under diagnosis and management, in all requirements for prompts under surgical timing and technique, and in clarity and evidence-based criteria for prompts under rehabilitation and prevention. The results of this study emphasize the potential for gathering clinically relevant and accurate answers to questions regarding the management of pediatric supracondylar humerus fractures from LLMs

Table 2.
Average ratings for artificial intelligence-generated responses.

Criteria	Diagnosis and nonoperative/preoperative management			Surgical timing and technique			Rehabilitation and prevention			*Overall (all categories)*		
	ChatGPT	Gemini	*P*-value	ChatGPT	Gemini	*P*-value	ChatGPT	Gemini	*P* value	ChatGPT	Gemini	*P* value
Relevance	3.821 ± 0.473	4.464 ± 0.276	**.003***	3.49 ± 0.344	4.265 ± 0.279	**<.001***	3.821 ± 0.335	4.071 ± 0.362	.183	3.643 ± 0.227	4.296 ± 0.17	**<.001***
Accuracy	3.679 ± 0.484	4.071 ± 0.39	.054	2.898 ± 0.363	3.959 ± 0.347	**<.001***	2.821 ± 0.515	2.714 ± 0.542	.621	3.051 ± 0.26	3.714 ± 0.258	**<.001***
Clarity	4.107 ± 0.382	4.464 ± 0.294	.067	3.592 ± 0.379	4.388 ± 0.234	**<.001***	3.786 ± 0.408	4.25 ± 0.313	**.013***	3.745 ± 0.237	4.388 ± 0.154	**<.001***
Completeness	4.036 ± 0.433	4.429 ± 0.293	.078	3.367 ± 0.369	4.265 ± 0.267	**<.001***	3.964 ± 0.356	4.036 ± 0.37	.646	3.663 ± 0.238	4.276 ± 0.171	**<.001***
Evidence-based	1.857 ± 0.4	3.357 ± 0.516	**<.001***	1.776 ± 0.244	4.265 ± 0.227	**<.001***	1.893 ± 0.339	3.393 ± 0.455	**<.001***	1.816 ± 0.181	3.765 ± 0.229	**<.001***
Consistency	3.036 ± 0.409		n/a	2.224 ± 0.282		n/a	2.643 ± 0.406		n/a	2.571 ± 0.213		n/a

Mean scores were highest for clarity (ChatGPT: 3.745 ± 0.237, Gemini 4.388 ± 0.154) and lowest for evidence-based responses (ChatGPT: 1.816 ± 0.181, Gemini: 3.765 ± 0.229). There were notable statistically significant differences across all criteria, with Gemini having higher mean scores in each criterion ($P < .001$).
Bold means $P < 0.05$.

and also suggest that Gemini may be a better model for obtaining this information.

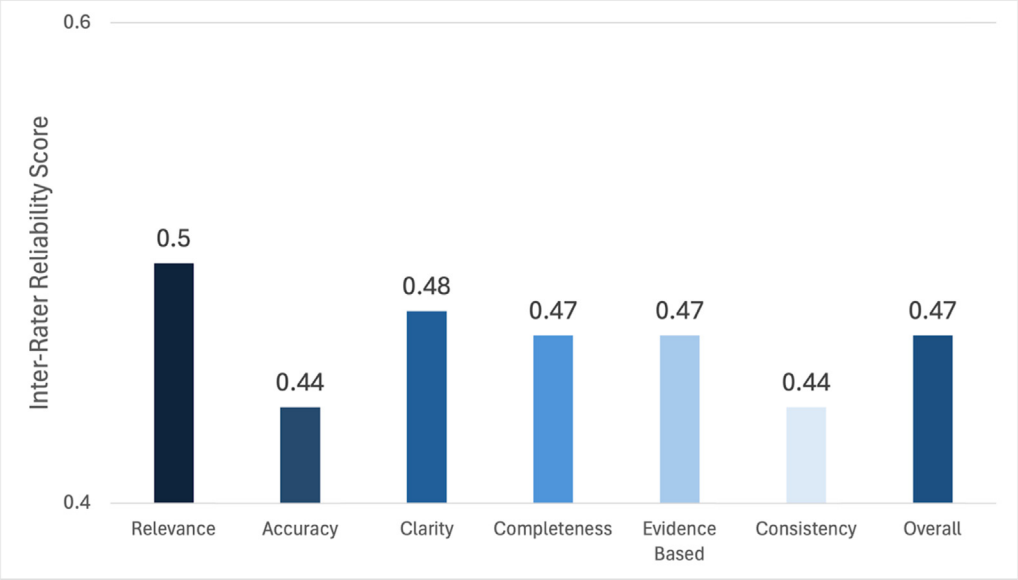
There is growing interest in understanding the role that LLMs, such as ChatGPT and Gemini, can play in providing information to patients regarding health concerns and diagnoses. ChatGPT can help patients access support groups, provide information about pharmaceutical side effects, track medications, and set reminders to take medications [14]. Several studies have highlighted the potential benefits of LLMs, specifically in orthopaedic surgery, such as in surgical planning, making a diagnosis, provider education, data collection, and patient communication [15–19]. While there are many potential positives, it is also essential to acknowledge the severe consequences LLMs can have on patients if the responses provided are inaccurate or inadequate. Given the increasing accessibility of LLMs and the significant percentage of patients who already use online resources for health information before clinic visits, physicians must recognize patient usage of these resources and how the strengths and limitations of a platform may ultimately impact patients.

Few prior studies have attempted to assess the accuracy of LLMs in providing answers to common orthopaedic conditions, and most of these studies have focused exclusively on ChatGPT and adult conditions [11,20,21]. Recently, Amaral et al. evaluated ChatGPT's ability to answer common questions regarding pediatric in-toeing and found that 90% of responses were satisfactory or better with moderate consistency when asked the same question on different occasions [22]. Adelstein et al. similarly found that ChatGPT could provide adequate answers to frequent questions about slipped capital femoral epiphysis and consistently reiterated the importance of professional medical evaluation [23]. In line with prior research, our study demonstrated that ChatGPT provides satisfactory responses that were aligned with AAOS guidelines for the management of pediatric supracondylar fractures. Notably, previous studies have not assessed Gemini, which, in our research, outperformed ChatGPT in all categories, emphasizing a potential gap in the previously published literature. Initial findings suggest that LLMs may benefit providers and parents in answering questions about pediatric orthopaedic conditions. Given the paucity of the present literature, further research should be conducted to assess the consistency and accuracy of ChatGPT and Gemini responses, with particular attention paid to responses regarding more urgent conditions.

IRR is a quantitative measure of the agreement amongst evaluators in assessing a given variable and functions as a potential indicator of consistency during subjective evaluations [24]. Overall IRR in this study was 0.47 with the highest IRR was 0.5 in the relevance category. Using a similar methodology, Magruder et al. also reported a mean IRR of 0.33 and a maximum IRR of 0.43 in assessing the responses of ChatGPT to total knee arthroplasty questions based on AAOS guidelines [10]. Lower IRR scores on these two highly different topics may potentially highlight the difficulty in objectively assessing LLM responses and the challenge in controlling for inherent subjectivity in the evaluation process. One explanation for the lower IRR scores in these studies may be that in a numeric survey, any discrepancy between ratings is treated equivalently. For example, two surgeons providing ratings of 5 and 4 decrease the IRR by the same amount as ratings of 5 and 1. Disadvantageously, this may conceal a general, if not exact, agreement among authors regarding the performance of the LLMs. An additional explanation for decreased IRR in this study is that the highest recommendation strength of these 14 AAOS' clinical practice guidelines was "moderate", suggesting a lack of substantial evidence to support one management decision over another and resulting in divided opinions among the surgeons. Given that this is a new area of research, it is essential to maintain consistency, including calculating an IRR, to facilitate comparison between studies.

The value of platforms like ChatGPT and Gemini is contingent upon their accurate and pertinent information in an accessible and convenient format. While Amaral et al. reported that ChatGPT could give satisfactory answers 90% of the time, they also found that ChatGPT's response had a collegiate readability level despite initially being prompted with questions appropriate for an elementary reading level

Table 3.
Inter-rater reliability scores for each survey question characteristic.



The highest IRR was in the relevance category (0.5), and the lowest was in the accuracy and consistency categories (0.44).

[22]. The average American adult reads at an eighth-grade level, and the AMA and NIH generally recommend that patient education materials not exceed the sixth-grade level to ensure comprehensibility [25]. Therefore, most of the responses provided by ChatGPT may be excessively challenging or even incomprehensible for the average user. Gomez-Caballo et al. reported that Gemini provided significantly more readable responses than ChatGPT regarding questions about post-operative plastic surgery care; however, in both cases, most responses were at a college reading level [26]. Users may be able to prompt ChatGPT and Gemini to deliver more comprehensible responses, but this could lead to oversimplification and omission of clinically relevant information. The potential for confusion and variation in the quality of information supplied by LLMs emphasizes the importance of office visits in which medical providers can provide clinically relevant information delivered in a patient-specific context.

This study has several potential limitations. Given that this is a developing research area, limited studies are available, so we could not calculate a power analysis to determine the appropriate sample size. To prompt ChatGPT and Gemini responses, we translated the AAOS clinical practice guidelines into questions, which may have introduced bias in the process. While the prompts were designed to gather relevant information from the LLM without introducing bias, they deliberately avoided referencing the AAOS guidelines. As a result, this study does not assess the LLMs' access to the guidelines but evaluates how well their responses align with them. Additionally, surgeon bias may have impacted response grading; however, surgeons were blinded to the source of each response to best control this. The number of reviewers was comparable to prior studies and should reduce the impact of individual biases [10,27]. Investigative categories, such as clarity, are also somewhat subjective, and scores could vary significantly based on the perspective of physicians versus patients. Potential bias and the general subjectivity of the categories likely play a role in the generally low inter-rater reliability across categories. The grading system used to score responses in this study is based on previously published studies but has not been formally validated [10]. Finally, while this study provided important insights into the quality of information provided by LLM responses, it did not consider other factors, such as readability or accessibility for different groups (patients with disabilities, children, or those with limited English proficiency). These should be further investigated to understand the actual utility of LLMs in the context of patient education.

Conclusion

In conclusion, ChatGPT and Gemini produced responses aligned with the 2022 AAOS current guideline practices for pediatric supracondylar humerus fractures. Gemini outperformed ChatGPT across all criteria, with the greatest difference in scores seen in the evidence-based category. Gemini also performed particularly well with questions about surgical timing and technique relative to ChatGPT. This study emphasizes the potential for LLMs, particularly Gemini, to provide pertinent clinical information for managing pediatric supracondylar humerus fractures.

Informed patient consent

The author(s) declare that no patient consent was necessary as no images or identifying information are included in the article.

Author contributions

John D. Milner: Writing – review & editing, Writing – original draft, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Matthew S. Quinn:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Phillip Schmitt:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Ashley Knebel:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jeffrey Henstenburg:** Writing – review & editing, Writing – original draft, Methodology, Investigation,

Formal analysis, Conceptualization. **Adam Nasreddine:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Alexandre R. Boulos:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jonathan R. Schiller:** Writing – review & editing, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Craig P. Eberson:** Writing – review & editing, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Aristides I. Cruz:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Funding

No funding was received for this project.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank Janine Molino, PhD, for her expertise as a biostatistician and reviewer of our statistical analysis.

Supplementary material

Supplementary material related to this article can be found at https://doi.org/10.1016/j.jposna.2025.100164.

References

[1] ChatGPT. Accessed June 19, 2024. https://chatgpt.com/.
[2] Gemini - chat to supercharge your ideas. Accessed June 19, 2024. https://gemini.google.com/.
[3] Homolak J. Opportunities and risks of ChatGPT in medicine, science, and academic publishing: a modern Promethean dilemma. Croat Med J 2023;64(1):1–3. https://doi.org/10.3325/cmj.2023.64.1.
[4] Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023;6:1169595. https://doi.org/10.3389/frai.2023.1169595.
[5] Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large Language models in medicine: the potentials and pitfalls: a narrative review. Ann Intern Med 2024;177(2):210–20. https://doi.org/10.7326/M23-2772.
[6] Giorgino R, Alessandri-Bonetti M, Luca A, Migliorini F, Rossi N, Peretti G, et al. ChatGPT in orthopedics: a narrative review exploring the potential of artificial intelligence in orthopedic practice. Front Surg 2023;10:1284015. https://doi.org/10.3389/fsurg.2023.1284015.
[7] McGowan A, Gui Y, Dobbs M, Shuster S, Cotter M, Selloni A, et al. ChatGPT and Bard exhibit spontaneous citation fabrication during psychiatry literature search. Psychiatry Res 2023;326:115334. https://doi.org/10.1016/j.psychres.2023.115334.
[8] Tyrrell Burrus M, Werner BC, Starman JS, Kurkis GM, Pierre JM, Diduch DR, et al. Patient perceptions and current trends in Internet use by orthopedic outpatients. HSS J 2017;13(3):271–5. https://doi.org/10.1007/s11420-017-9568-2.
[9] Kuroiwa T, Sarcon A, Ibara T, Yamada E, Yamamoto A, Tsukamoto K, et al. The potential of ChatGPT as a self-diagnostic tool in common orthopedic diseases: exploratory study. J Med Internet Res 2023;25:e47621. https://doi.org/10.2196/47621.
[10] Magruder ML, Rodriguez AN, Wong JCJ, Erez O, Piuzzi NS, Scuderi GR, et al. Assessing ability for ChatGPT to answer total knee arthroplasty-related questions. J Arthroplast 2024. https://doi.org/10.1016/j.arth.2024.02.023. Published online February 14. S0883–5403(24)00122-0.
[11] 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 Am. 2023;105(19):1519–26. https://doi.org/10.2106/JBJS.23.00209.
[12] Abzug JM, Herman MJ. Management of supracondylar humerus fractures in children: current concepts. J Am Acad Orthop Surg 2012;20(2).
[13] Arazy O, Kopak R. On the measurability of information quality. J Am Soc Inf Sci Technol 2011;62(1):89–99. https://doi.org/10.1002/asi.21447.
[14] MdR Islam. Urmi TJ, Mosharrafa RA, Rahman MS, Kadir MF. Role of ChatGPT in health science and research: a correspondence addressing potential application. Health Sci Rep 2023;6(10):e1625. https://doi.org/10.1002/hsr2.1625.
[15] Chatterjee S, Bhattacharya M, Pal S, Lee SS, Chakraborty C. ChatGPT and large language models in orthopedics: from education and surgery to research. J Exp Orthop 2023;10:128. https://doi.org/10.1186/s40634-023-00700-1.
[16] Cheng K, Sun Z, He Y, Gu S, Wu H. The potential impact of ChatGPT/GPT-4 on surgery: will it topple the profession of surgeons? Int J Surg Lond Engl 2023;109(5):1545–7. https://doi.org/10.1097/JS9.0000000000000388.
[17] Morya VK, Lee HW, Shahid H, Magar AG, Lee JH, Kim JH, et al. Application of ChatGPT for orthopedic surgeries and patient care. Clin Orthop Surg 2024;16(3):347–56. https://doi.org/10.4055/cios23181.
[18] 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. https://doi.org/10.7759/cureus.58874.
[19] 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. https://doi.org/10.1186/s40634-023-00642-8.
[20] Wrenn SP, Mika AP, Ponce RB, Mitchell PM. Evaluating ChatGPT's ability to answer common patient questions regarding hip fracture. J Am Acad Orthop Surg 2024. https://doi.org/10.5435/JAAOS-D-23-00877. Published online May 14.
[21] Anastasio AT, Mills FB, Karavan MP, Adams SB. Evaluating the quality and usability of artificial intelligence-generated responses to common patient questions in foot and ankle surgery. Foot Ankle Orthop 2023;8(4):24730114231209919. https://doi.org/10.1177/24730114231209919.
[22] Amaral JZ, Schultz RJ, Martin BM, Taylor T, Touban B, McGraw-Heinrich J, et al. Evaluating chat generative pre-trained transformer responses to common pediatric in-toeing questions. J Pediatr Orthop 2024. https://doi.org/10.1097/BPO.0000000000002695. Published online April 30.
[23] Adelstein JM, Sinkler MA, Li LT, Mistovich RJ. ChatGPT responses to common questions about slipped capital femoral epiphysis: a reliable resource for parents? J Pediatr Orthop 2024;44(6):353–7. https://doi.org/10.1097/BPO.0000000000002681.
[24] McHugh ML. Interrater reliability: the kappa statistic. Biochem Med 2012;22(3):276–82.
[25] Eltorai AEM, Ghanian S, Adams CA, Born CT, Daniels AH. Readability of patient education materials on the American association for surgery of trauma website. Arch Trauma Res 2014;3(2):e18161. https://doi.org/10.5812/atr.18161.
[26] Gomez-Cabello CA, Borna S, Pressman SM, Haider SA, Sehgal A, Leibovich BC, et al. Artificial intelligence in postoperative care: assessing large language models for patient recommendations in plastic surgery. Healthc Basel Switz 2024;12(11):1083. https://doi.org/10.3390/healthcare12111083.
[27] Quinn M, Milner JD, Schmitt P, Morrissey P, Lemme N, Marcaccio S, et al. Artificial intelligence large language models address anterior cruciate ligament reconstruction: superior clarity and completeness by Gemini compared to ChatGPT-4 in response to American Academy of orthopedic surgeons clinical practice guidelines. Arthrosc J Arthrosc Relat Surg Off Publ Arthrosc Assoc N Am Int Arthrosc Assoc 2024. https://doi.org/10.1016/j.arthro.2024.09.020. Published online September 21. S0749–8063(24)00736-9.