ORIGINAL ARTICLE

DIAGNOSIS AND MANAGEMENT DECISIONS OF LARGE LANGUAGE MODELS IN PEDIATRIC SUPRACONDYLAR HUMERUS FRACTURES: AN AAOS GUIDELINE-BASED COMPARATIVE STUDY

Pediatrik Suprakondiler Humerus Kırıklarında Büyük Dil Modellerinin Tanı ve Yönetim Kararları: AAOS Kılavuzuna Dayalı Karşılaştırmalı Çalışma

Abdullah Alper SAHİN[1], Murat ALPARSLAN[1]

[1] Ordu University, Faculty of Medicine, Department of Orthopedics and Traumatology, Ordu, Türkiye.

Abdullah Alper ŞAHİN, Assoc. Prof. Dr.
(https://orcid.org/0000-0002-8973-2050)
dr.a.alpersahin@gmail.com

Murat ALPARSLAN, Res. Asst.
(https://orcid.org/0009-0005-9491-6427)
muratalparslan61@gmail.com

Contact:
Assoc. Prof. Dr. Abdullah Alper ŞAHİN, Ordu University, Faculty of Medicine, Department of Orthopedics and Traumatology, Ordu, Türkiye.

ABSTRACT

Objective: This study aimed to evaluate the diagnostic performance and clinical safety of large language models (ChatGPT 5.2 Thinking, Gemini 3.1 Pro, and Claude Sonnet 4.6) in detecting pediatric supracondylar humerus (SCH) fractures, classifying them according to the Gartland system, and providing treatment recommendations based on American Academy of Orthopedic Surgeons (AAOS) clinical practice guidelines.

Material and Methods: In this retrospective observational study, radiographs and clinical data from 128 pediatric patients presenting with elbow trauma (84 fractures and 44 non-fractures) were analyzed. The reference standard was established through consensus evaluation by orthopedic specialists according to AAOS guidelines. Anonymized AP and lateral elbow radiographs, along with age, sex, and brief clinical history, were provided to each model. Diagnostic performance was assessed using sensitivity, specificity, and overall accuracy. Agreement in Gartland classification was evaluated using Cohen's kappa (κ). Clinical safety was examined by analyzing under-treatment and over-treatment risks in surgical decision recommendations.

Results: Gemini 3.1 demonstrated the highest sensitivity for fracture detection (96%) but had low specificity (41%). ChatGPT 5.2 showed a more balanced profile with both sensitivity and specificity of 64%, whereas Claude 4.6 demonstrated the lowest overall diagnostic accuracy (55%). In Gartland classification, Gemini 3.1 achieved the highest agreement (κ=0.60), while Claude 4.6 showed the lowest (κ=0.33). Regarding treatment recommendations, Gemini 3.1 avoided under-treatment of surgical cases but produced substantial over-treatment, whereas ChatGPT 5.2 missed 40% of surgically indicated cases.

Conclusion: Although LLMs may offer supportive insights in radiographic interpretation, their current performance demonstrates considerable risks of false-positive and false-negative decisions. These findings indicate that LLMs are not yet sufficiently reliable for independent clinical decision-making in pediatric SCH fractures, and expert oversight remains essential.

Keywords: *Artificial Intelligence; Large Language Models; Supracondylar Humerus Fractures; Radiography*

ÖZET

Amaç: Bu çalışma, büyük dil modellerinin (ChatGPT 5.2 Thinking, Gemini 3.1 Pro ve Claude Sonnet 4.6) pediatrik suprakondiler humerus (SCH) kırıklarının tespitinde, Gartland sistemine göre sınıflandırılmasında ve Amerikan Ortopedi Cerrahları Akademisi (AAOS) klinik uygulama kılavuzlarına dayalı tedavi önerileri sunmada tanısal performansını ve klinik güvenliğini değerlendirmek amacıyla yapılmıştır.

Gereç ve Yöntem: Bu retrospektif gözlemsel çalışmada, dirsek travması (84 kırık ve 44 kırık olmayan) ile başvuran 128 pediatrik hastanın radyografileri ve klinik verileri analiz edildi. Referans standardı, AAOS kılavuzlarına göre ortopedi uzmanlarının konsensüs değerlendirmesi ile belirlendi. Anonimleştirilmiş AP ve lateral dirsek radyografileri, yaş, cinsiyet ve kısa klinik öykü ile birlikte her modele sağlandı. Tanı performansı duyarlılık, özgüllük ve genel doğruluk kullanılarak değerlendirildi. Gartland sınıflandırmasındaki uyum Cohen'in kappa (κ) katsayısı kullanılarak değerlendirildi. Klinik güvenlik, cerrahi karar önerilerindeki yetersiz tedavi ve aşırı tedavi riskleri analiz edilerek incelendi.

Bulgular: Gemini 3.1, kırık tespiti için en yüksek duyarlılığı (%96) gösterdi, ancak özgüllüğü düşüktü (41%). ChatGPT 5.2, hem duyarlılık hem de özgüllük açısından %64 ile daha dengeli bir profil sergilerken, Claude 4.6 en düşük genel tanı doğruluğunu (%55) gösterdi. Gartland sınıflandırmasında Gemini 3.1 en yüksek uyumu (κ=0,60) elde ederken, Claude 4.6 en düşük uyumu (κ=0,33) gösterdi. Tedavi önerileri açısından Gemini 3.1, cerrahi vakaların yetersiz tedavisini önledi, ancak önemli ölçüde aşırı tedaviye yol açarken, ChatGPT 5.2 cerrahi olarak endike olan vakaların %40'ını kaçırdı.

Sonuç: LLM'ler radyografik yorumlamada destekleyici bilgiler sunsa da, mevcut performansları yanlış pozitif ve yanlış negatif kararların alınması riskini önemli ölçüde artırmaktadır. Bu bulgular, LLM'lerin pediatrik SCH kırıklarında bağımsız klinik karar verme için henüz yeterince güvenilir olmadığını ve uzman gözetiminin hala gerekli olduğunu göstermektedir.

Anahtar Kelimeler: *Yapay Zeka; Büyük Dil Modelleri; Suprakondiler Humerus Kırıkları; Radyografi*

Received/Geliş tarihi: 13.03.2026
Accepted/Kabul tarihi: 11.05.2026
DOI: 10.16919/bozoktip.1909287

Bozok Tıp Derg 2026;16(2):303-311
Bozok Med J 2026;16(2):303-311

Cite this aritcle as: Şahin AA, Alparslan M. Diagnosis and Management Decisions of Large Language Models in Pediatric Supracondylar Humerus Fractures: An AAOS Guideline-Based Comparative Study. Bozok Med J 2026;16(2):303-311
DOI: https://doi.org/10.16919/bozoktip.1909287

INTRODUCTION

The introduction of widely accessible large language model platforms-including ChatGPT (OpenAI, San Francisco, CA, USA), Gemini (formerly Bard; Google, Mountain View, CA, USA), and Claude (Anthropic, San Francisco, CA, USA)-has markedly increased the reach, usability, and clinical visibility of artificial intelligence (AI) systems. These large language model (LLM)-based artificial intelligence chatbots enable users to access vast amounts of information across multiple disciplines, including medicine, almost instantaneously through a single interface (1). LLMs are increasingly being used in the healthcare field to improve patient information, optimize clinician-patient interaction, support scientific research, and enhance the quality of medical education (2). Research in the literature shows an increasing interest in evaluating whether orthopedic and traumatology surgeons can contribute to the clinical decision-making mechanisms of LLMs, update treatment guidelines by synthesizing existing scientific data, and improve patient satisfaction (3). Research shows that approximately two-thirds of orthopedic patients obtain information about their conditions from online sources, but only half of this group considers discussing the information they have obtained with their physicians (4). With the increased accessibility and expanded use of LLMs, patients will increasingly turn to these digital resources to learn about their conditions and treatment options. Supracondylar humerus fractures seen in childhood most commonly occur in the 5-7 age group following a fall onto an outstretched hand and are the most common type of elbow fracture in the pediatric population (5,6). In 2011, the American Academy of Orthopedic Surgeons (AAOS) published clinical practice guidelines (CPG) for pediatric supracondylar humerus fractures, offering 6 evidence-based recommendations ranging from moderate to limited and 8 non-definitive statements (7). The frequent occurrence of pediatric supracondylar humerus fractures in emergency departments and the need for surgical treatment in a significant proportion of cases leads parents to seek information about their child's injury process, treatment alternatives, and potential complications. Given the increasing accessibility of AI, families are likely to use LLM-based chatbots to meet this information need. Therefore, it is necessary to evaluate the compliance and accuracy levels of these systems with evidence-based clinical guidelines. A review of the current literature reveals no studies analyzing LLM responses to patient radiographs in the diagnosis and treatment of pediatric supracondylar humerus fractures in comparison with current guideline recommendations. Therefore, our study aimed to evaluate the ability of ChatGPT, Claude, and Gemini to generate responses based on the AAOS guidelines for pediatric supracondylar humerus fractures.

MATERIAL AND METHODS

This retrospective observational study was conducted to evaluate the diagnostic performance and clinical decision-making safety of multimodal large language models (LLMs) in the assessment of pediatric supracondylar humerus fractures using radiographic images. The study focused on the ability of LLMs to detect fractures, classify fractures, and recommend appropriate treatment according to established clinical guidelines. Prior to the study, approval was obtained from the local ethics committee (Date:07.01.2026, No:2026/14). Radiographs and clinical records of pediatric patients presenting with elbow trauma and followed in the Department of Orthopedics and Traumatology between January 1, 2022, and December 1, 2025 were retrospectively reviewed.

A total of 128 pediatric elbow radiographic examinations obtained from patients presenting with elbow trauma were included in the study. Each case consisted of standard anteroposterior and lateral elbow radiographs. According to the reference evaluation, 84 cases were diagnosed with supracondylar humerus fractures, while 44 cases showed no evidence of fracture. All cases were retrospectively selected from the institutional radiographic archive.

Patients were included in the study if they met the following criteria:

- Pediatric patients aged 0-12 years.
- Radiographs of sufficient technical quality, including complete anteroposterior (AP) and lateral elbow views.
- Patients presenting with elbow trauma.
- For fracture-positive cases, the modified Gartland classification and treatment strategy, including

Bozok Tıp Derg 2026;16(2):303-311
Bozok Med J 2026;16(2):303-311

conservative management or closed reduction and percutaneous pinning, were clearly documented in the clinical records. For fracture-negative cases, the absence of fracture was confirmed by orthopedic specialist consensus based on radiographic evaluation.

- Complete and accessible medical records available for retrospective review through the Hospital Information Management System (HBYS) and Picture Archiving and Communication System (PACS).
- Patients were excluded if any of the following conditions were present:
- Patients older than 12 years or younger than newborn age.
- Missing AP or lateral radiographs, or images with insufficient diagnostic quality.
- Pathological fractures (e.g., tumor-related fractures, bone dysplasia, osteogenesis imperfecta).
- History of previous surgical intervention in the elbow region (e.g., osteotomy, internal fixation).
- Radiographs that could not be adequately evaluated due to severe polytrauma involving multiple anatomical regions.
- Incomplete or unclear clinical documentation, including treatment strategy, surgical details, or radiographic evaluation records.
- Other types of elbow fractures not included in the supracondylar fracture classification (e.g., lateral condyle fractures, medial epicondyle fractures, olecranon fractures).

The reference standard for fracture diagnosis, classification, and treatment recommendation was established by orthopedic specialists using American Academy of Orthopedic Surgeons (AAOS) clinical practice guidelines for the management of pediatric supracondylar humerus fractures (7). All cases were independently evaluated by two experienced orthopedic surgeons. In cases of disagreement between the two reviewers, a third orthopedic specialist reviewed the case, and consensus was reached. This consensus assessment served as the reference standard for subsequent comparisons.

Three widely used multimodal large language models capable of interpreting radiographic images were evaluated: ChatGPT 5.2 Thinking (OpenAI, San Francisco, CA, USA), Gemini 3.1 Pro (Google, Mountain View, CA, USA), Claude Sonnet 4.6 (Anthropic, San Francisco, CA, USA). All models were accessed through their official web-based interfaces. Evaluations were conducted on three separate days - February 21, February 24, and February 27, 2026 - corresponding to three independent sessions per case, to assess response consistency and reduce session-related bias. For each case, the models were provided with standardized clinical and radiographic information in separate evaluation sessions. The following data were presented to each model:

- Patient age and sex
- A brief clinical history describing the trauma mechanism
- Anteroposterior and lateral elbow radiographs
- Each model was then asked to answer the following clinical questions:
- “Is there a supracondylar humerus fracture in this pediatric patient?”
- “If present, please specify the fracture type according to the modified Gartland classification.”
- “What is the most appropriate treatment approach for this fracture?”

All radiographic images were presented in anonymized PNG format. To evaluate response consistency and reduce session-related bias, each case was evaluated by each model in three independent sessions conducted on three separate days. Every evaluation was performed in a completely new chat session, ensuring that the models had no access to prior responses or contextual information from previous cases. This repeated evaluation protocol was designed to reduce the influence of session-specific response variability and to derive a single final classification for each case. Responses from three independent sessions for each case were consolidated into one final classification using a predefined majority rule.

The responses generated by each large language model were systematically reviewed and converted into a structured numerical coding system to enable direct comparison with the reference standard. Each model output was independently examined and categorized according to three parameters:

- **Fracture presence:** 1=No fracture detected, 2=Supracondylar humerus fracture present
- Modified Gartland fracture classification (8): 1=Type I, 2=Type II, 3=Type III, 4=Type IV
- **Treatment recommendation:** 1=Conservative management, 2=Surgical treatment (closed reduction and percutaneous pinning)
- Thus, each model response was transformed into a three-component numerical code (e.g., 2-3-2) representing fracture presence, modified Gartland classification, and treatment recommendation. All coded responses were entered into a structured dataset and compared with the reference standard established by orthopedic specialists. In cases where the model responses contained ambiguous or non-standard wording, the interpretation was performed according to predefined coding rules to maintain consistency across evaluations. This standardized coding framework allowed quantitative comparison of model performance in terms of fracture detection accuracy, classification agreement, and treatment recommendation safety.
- **The performance of the LLMs was evaluated across three domains:**
- **Fracture Detection:** Diagnostic performance for fracture detection was evaluated across all cases (n=128). The metrics were calculated: Sensitivity, Specificity, Accuracy, Area under the receiver operating characteristic curve (AUC).
- **Modified Gartland Classification Agreement:** Agreement between the model-generated classifications and the reference classification was evaluated in fracture cases only (n=84) using quadratic weighted Cohen's kappa coefficient (κ).
- **Clinical Safety of Treatment Recommendations:** Treatment recommendations were evaluated within fracture cases (n=84). Two clinically relevant misclassification types were defined:

◊ **Undertreatment:** Cases requiring surgical management according to the reference standard but incorrectly recommended as conservative treatment by the model.

◊ **Overtreatment:** Cases suitable for conservative treatment but incorrectly recommended for surgical intervention.

Statistical Analysis

All statistical analyses were performed using IBM SPSS Statistics software (version 24.0; IBM Corp., Armonk, NY, USA). Diagnostic performance for fracture detection was evaluated by calculating sensitivity, specificity, accuracy, and area under the receiver operating characteristic curve (AUC) based on confusion matrices comparing model outputs with the reference standard. Agreement between the modified Gartland classifications generated by the large language models and the reference classification was assessed using the quadratic weighted Cohen's kappa coefficient (κ). For treatment recommendations, surgical sensitivity and specificity were calculated relative to the reference standard. Clinically relevant misclassification patterns were evaluated by determining the frequency of undertreatment (surgical cases incorrectly classified as conservative) and overtreatment (conservative cases incorrectly recommended for surgery). Results were summarized using descriptive statistics.

RESULTS

A total of 128 pediatric cases were included in the analysis. According to the guideline-based reference standard, 84 patients (65.6%) had supracondylar humerus fractures, while 44 patients (34.4%) had no fracture. Among the fracture cases (n=84), 30 patients (35.7%) required surgical treatment, and 54 patients (64.3%) were managed conservatively.

All patients (n=128) were evaluated in this analysis. Differences were observed among the models in detecting the presence of fracture. Gemini demonstrated the highest sensitivity (0.96), missing only 3 fracture cases (FN=3). However, its specificity was low (0.41), with 26 false-positive classifications (FP=26) in non-fracture cases. ChatGPT showed a more balanced performance (Sensitivity 0.64; Specificity 0.64) but missed 30 fracture cases (FN=30). Claude exhibited the lowest diagnostic performance (sensitivity 0.58; specificity 0.50; FN=35). Overall accuracy and discriminative capacity (AUC) were moderate across all models: ChatGPT: Accuracy 0.64; AUC 0.64, Gemini: Accuracy 0.77; AUC 0.69, Claude: Accuracy 0.55; AUC 0.54. None of the models achieved high combined sensitivity and specificity (Table 1).

Bozok Tıp Derg 2026;16(2):303-311
Bozok Med J 2026;16(2):303-311

Table 1. Diagnostic performance of large language models for fracture detection

Model	TP	FP	TN	FN	Sensitivity	Specificity	Accuracy	AUC
ChatGPT	54	16	28	30	0.64	0.64	0.64	0.64
Gemini	81	26	18	3	0.96	0.41	0.77	0.69
Claude	49	22	22	35	0.58	0.50	0.55	0.54

TP: True positive, FP: False positive, TN: True negative, FN: False negative, AUC: Area under the curve

Only patients with fractures (n=84) were evaluated in this analysis. Agreement with the reference classification was evaluated using quadratic weighted Cohen's kappa. Gemini: κ=0.60, ChatGPT: κ=0.47, Claude: κ=0.33. Gemini demonstrated the highest agreement, approaching good concordance. ChatGPT showed moderate agreement, whereas Claude exhibited lower concordance. None of the models achieved high-level classification reliability (Table 2).

Table 2. Classification agreement and clinical safety profile in fracture cases

Model	Gartland κ (weighted)	Surgical sensitivity	Surgical specificity	Undertreatment (n)	Overtreatment (n)
ChatGPT	0.47	0.60	0.78	12	12
Gemini	0.60	1.00	0.37	0	34
Claude	0.33	0.60	0.67	12	18

Surgical Sensitivity (Undertreatment Risk): Among the 30 surgically indicated cases:

- Gemini correctly identified all cases (Sensitivity 1.00; FN=0).
- ChatGPT misclassified 12 surgical cases as conservative (FN=12; Sensitivity 0.60).
- Claude also misclassified 12 surgical cases (FN=12; Sensitivity 0.60).
- Thus, Gemini showed the lowest undertreatment risk (Table 2).
- **Overtreatment Risk (False Surgical Recommendation):** Among conservatively managed cases (n=54):
- Gemini recommended surgery in 34 cases (Specificity 0.37).
- Claude misclassified 18 cases.
- ChatGPT misclassified 12 cases (Specificity 0.78).

ChatGPT demonstrated the most balanced surgical decision profile.

Clinical Safety Interpretation: The models demonstrated distinct safety patterns. Gemini minimized fracture and surgical misses but showed substantial overtreatment tendency. ChatGPT presented a more balanced specificity profile but carried a notable risk of missing surgically indicated cases. Claude showed lower reliability across diagnostic, classification, and management domains. Importantly, none of the evaluated large language models achieved a safety profile compatible with independent clinical use, as no model simultaneously maintained high sensitivity and high specificity in both diagnostic and treatment decisions (Figure 1 and Figure 2).

Bozok Tıp Derg 2026;16(2):303-311
Bozok Med J 2026;16(2):303-311

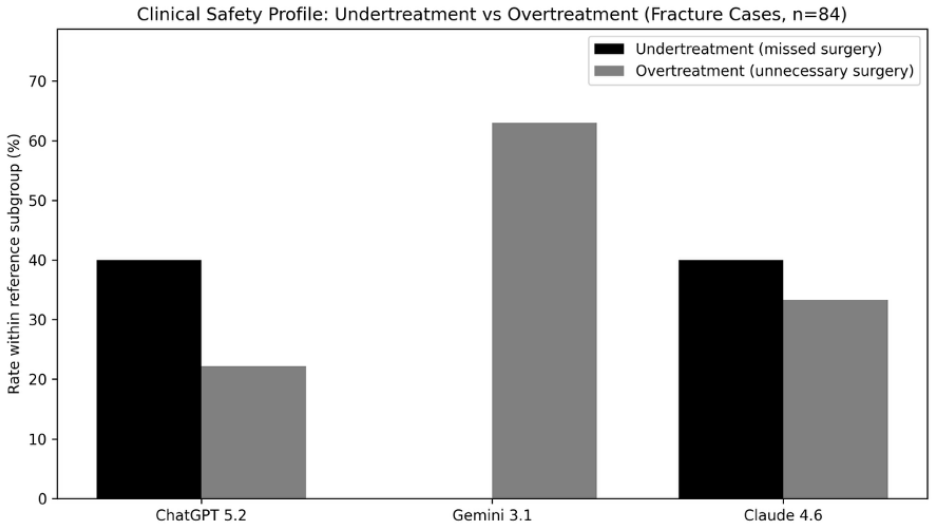

Figure 1. Clinical safety profile of large language models in pediatric supracondylar humerus fractures

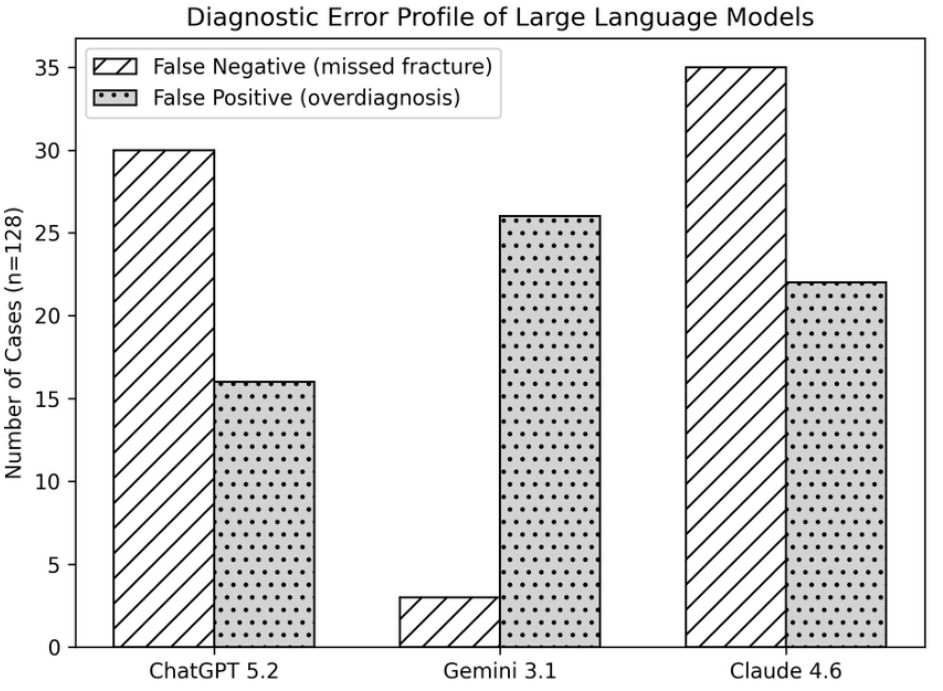

Figure 2. Diagnostic error profile of large language models in detecting pediatric supracondylar humerus fractures

Bozok Tıp Derg 2026;16(2):303-311
Bozok Med J 2026;16(2):303-311

DISCUSSION

In this study, we evaluated the diagnostic accuracy and clinical decision-making safety of ChatGPT 5.2 Thinking, Gemini 3.1 Pro, and Claude Sonnet 4.6-the three most widely used multimodal large language models worldwide-in assessing pediatric supracondylar humerus fractures using radiographic images. Our results demonstrated that while LLMs showed moderate performance in fracture detection and classification, substantial variability existed in their clinical decision-making profiles. Gemini 3.1 achieved the highest sensitivity for fracture detection and did not miss any surgically indicated cases, indicating a lower risk of undertreatment. However, this model exhibited a marked tendency toward overtreatment by recommending surgical intervention in a considerable proportion of conservatively manageable fractures. In contrast, ChatGPT 5.2 demonstrated a more balanced specificity profile but missed a notable number of surgically indicated fractures, reflecting a potential risk of undertreatment. Claude 4.6 showed comparatively lower agreement in fracture detection, classification, and treatment recommendations. Taken together, these findings suggest that although large language models may provide supportive information in the interpretation of pediatric elbow radiographs, their current performance profiles do not yet support safe independent clinical decision-making.

Recent studies have increasingly explored the potential role of large language models in clinical decision support across various medical specialties, including orthopedics (9,10). Marcaccini et al. evaluated the clinical performance of three different large language models (ChatGPT-4o, DeepSeek-V3, and Gemini 1.5) in the diagnosis and treatment management of hand fractures (9). They presented the clinical details of 58 cases with hand fractures to the models and compared the management plans generated by the models with the actual clinical decisions made by experienced surgeons. This study reported that ChatGPT-4o achieved a high level of accuracy, while Gemini 1.5 showed significantly lower performance, emphasizing that AI responses are not yet reliable enough to replace human expertise, especially in complex trauma management. In our study, three different LLMs were evaluated for their performance in diagnosis, modified Gartland classification, and treatment recommendation for pediatric supracondylar humerus fractures by presenting them with the short clinical history of the cases along with anonymized AP and lateral elbow radiographs. Our results showed significant performance heterogeneity among the models; the highest accuracy was found in Gemini (0.77), while the lowest accuracy was found in Claude (0.55). Similarly, sensitivity in fracture detection was highest in Gemini (0.96), while Claude (0.58) and ChatGPT (0.64) showed lower sensitivity values. Although Marcaccini et al.'s study on hand fractures indicated that ChatGPT-4o suggested surgical plans with high accuracy, our data shows that ChatGPT 5.2 incorrectly classified 40% of surgical cases as conservative, proving that these models cannot yet replace human expertise in complex pediatric trauma management.

Prior literature evaluating multimodal large language models in medical image interpretation has demonstrated that performance varies across tasks and clinical contexts. In musculoskeletal radiology, Horiuchi et al. reported a diagnostic accuracy of 43% for GPT-4-based ChatGPT, and Güneş et al. reported an accuracy of 45% for GPT-4o in visual neuroanatomy tasks (11,12). Bulut et al. evaluated ChatGPT-4o, Gemini 2.0, and Claude 3.5 for scaphoid fracture detection and assessment of surgical indications, using radiographs presented on three different days; Claude 3.5 achieved the highest sensitivity (57.1%), followed by Gemini 2.0 (18.2%) and ChatGPT-4o (9.1%) (10). In the present study, which focused on pediatric supracondylar humerus fractures, Gemini 3.1 demonstrated a fracture-detection sensitivity of 96% and did not miss any surgically indicated cases (0 undertreatment). However, surgical specificity was low (37%), indicating a high rate of surgical recommendations among conservatively managed cases (overtreatment). ChatGPT 5.2 showed a relatively

Bozok Tıp Derg 2026;16(2):303-311
Bozok Med J 2026;16(2):303-311

higher specificity profile but missed 12 surgically indicated cases (surgical sensitivity 60%). Overall, these findings indicate that model performance and error patterns differ across anatomical regions and clinical tasks; notably, whereas Claude 3.5 showed the highest sensitivity in scaphoid fracture detection in Bulut et al., Gemini 3.1 demonstrated the highest sensitivity for supracondylar humerus fracture detection in our cohort (10). In our study, a similar inconsistency is also noticeable when examining the bone fracture classification capabilities of the models, which differs from others. In the modified Gartland classification, Gemini 3.1 (κ=0.60) showed a near-good agreement, while Claude 4.6 (κ=0.33) remained at a very low level of reliability. These results reveal that the models did not demonstrate consistent and reliable performance in the classification step and that classification accuracy should be additionally evaluated in clinical decision-making processes.

Noda et al. demonstrated how the diagnostic accuracy of large language models (LLMs) varies across different anatomical regions and clinical tasks when integrated into orthopedic radiology (13). Noda et al. found that ChatGPT-4 demonstrated moderate agreement (κ=0.420) in classifying adult pertrochanteric fractures as “stable” or “unstable” according to AO/OTA standards and performed at a level comparable to surgeons. In contrast, in our study examining pediatric supracondylar humerus fractures, the performance of the models was found to be much more heterogeneous in terms of clinical safety thresholds (13). In pediatric cases, the Gemini 3.1 model demonstrated a very high sensitivity of 96% and stronger agreement in the modified Gartland classification (κ=0.60), but it showed a serious tendency toward “overtreatment” due to low specificity (41%). While Noda et al.'s study found that ChatGPT-4's accuracy (0.714) and sensitivity (0.700) metrics, our study shows that ChatGPT 5.2 missed 40% of cases requiring surgery (60% surgical sensitivity), indicating that the models carry a risk of “undertreatment” in pediatric trauma management. Noda et al. argue that ChatGPT is reasonably reliable, yielding results comparable to those of expert surgeons; however, our study emphasizes that, in their current state, these technologies are not yet safe for independent clinical decision-making processes, as no model can simultaneously maintain high sensitivity and specificity. In conclusion, while both studies agree that LLMs provide a valuable “second opinion” in radiological image interpretation, the models' limitations in capturing specific anatomical details (such as lateral wall integrity or pediatric elbow contours) necessitate expert oversight.

Study Limitations

This study has several limitations that should be acknowledged. First, the study has a single-center, retrospective design; this may limit its generalizability in terms of the technical quality of radiographs and the patient population. Second, the large language models (LLMs) evaluated are highly dynamic systems, and their performance depends on the specific versions used (ChatGPT 5.2, Gemini 3.1, Claude 4.6) and updates at the time of evaluation; the evolution of models over time may affect the long-term validity of the results. Third, since the content and sources of the datasets used to train the models are not disclosed by the companies, the potential biases of these models cannot be fully estimated. Furthermore, although the study is based on AAOS guidelines, there is debate in the literature that these guidelines (e.g., the 2011 version) do not fully reflect current practice. Finally, our study did not analyze the readability levels of the models or the complexity of the natural language questions (patient-generated queries) that patients might direct to these models; this constitutes a limitation in measuring the actual success of artificial intelligence in patient education.

CONCLUSION

In conclusion, this study demonstrates that multimodal large language models hold significant technological potential for the diagnosis and management of pediatric supracondylar humerus fractures; however,

their current performance profiles do not yet meet the safety standards required for independent clinical decision-making processes. While the technological potential offered by LLMs provides a valuable "second opinion" in radiological image interpretation, the high rates of false positives and false negatives indicate that more specific training is needed for the full integration of these systems in a manner that is fully compliant with and safe for use in AAOS guidelines.

Acknowledgements

The authors declare that there are no conflicts of interest. No financial support was received for the conduct of this study.

REFERENCES

1. 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.
2. 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.
3. 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.
4. 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.
5. Kolac UC, Oral M, Sili MV, Ibik S, Aydinoglu HS, Bakircioglu S, et al. Identifying risk factors for open reduction in pediatric supracondylar humerus fractures. J Pediatr Orthop. 2024;44:573-8.
6. Yilmaz T, Dur IH, Kabakci T, Bulut MA, Akgok B, Kolac UC, et al. Effect of fracture level on optimal Kirschner wire configuration in pediatric supracondylar humerus fractures: A finite element analysis. Jt Dis Relat Surg. 2025;36(3):648-58.
7. Heggeness MH, Sanders JO, Murray J, Pezold R, Sevarino KS. Management of Pediatric Supracondylar Humerus Fractures. J Am Acad Orthop Surg. 2015;23(10):e49-51.
8. Wilkins KE. Fractures and dislocations of the elbow region. In: Rockwood CA Jr, Wilkins KE, King RE, eds. Fractures in children. Vol 3, 4th ed. Philadelphia: Lippincott-Raven, 1996. p. 680-1.
9. Marcaccini G, Seth I, Xie Y, Susini P, Pozzi M, Cuomo R, et al. Breaking Bones, Breaking Barriers: ChatGPT, DeepSeek, and Gemini in Hand Fracture Management. J Clin Med. 2025;14(6):1983.
10. Bulut B, Yortanlı M, Gür A, Akkan Öz M, Mutlu H. Diagnostic capabilities of large language models in the detection of scaphoid fractures in the emergency department. Ulus Travma Acil Cerrahi Derg. 2025;31(10):987-94.
11. Horiuchi D, Tatekawa H, Oura T, Shimono T, Walston SL, Takita H, et al. ChatGPT's diagnostic performance based on textual vs. visual information compared to radiologists' diagnostic performance in musculoskeletal radiology. Eur Radiol. 2025;35:506-16.
12. Güneş YC, Ülkir M. Comparative Performance Evaluation of Multimodal Large Language Models, Radiologist, and Anatomist in Visual Neuroanatomy Questions. Uludağ Üniversitesi Tıp Fakültesi Dergisi. 2024;50:551-6.
13. Noda M, Takahara S, Hayashi S, Inui A, Oe K, Matsushita T. Evaluating ChatGPT's Performance in Classifying Pertrochanteric Fractures Based on Arbeitsgemeinschaft für Osteosynthesefragen/Orthopedic Trauma Association (AO/OTA) Standards. Cureus. 2025;17:e78068.