Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Comparison of the Diagnostic Reliability of ChatGPT in Letournel–Judet Acetabulum Fracture Classification Based on Judet Radiographs According to Orthopaedic Residents

Muhammed Kilic

dr.kilic.06@gmail.com

Ankara Bilkent City Hospital

Ahmet Ozgur Yildirim

Ankara Bilkent City Hospital

Ibrahim Alper Yavuz

Ankara Bilkent City Hospital

Fatih Inci

Ankara Bilkent City Hospital

Erman Ceyhan

Ankara Bilkent City Hospital

Yakup Kahve

Ankara Bilkent City Hospital

Tahsin Aydin

Ankara Bilkent City Hospital

Ozdamar Fuad Oken

Ankara Bilkent City Hospital

Research Article

Keywords: pelvic injury, artificial intelligence, acetabular fracture, chatGPT, letournel–judet classification

Posted Date: July 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10168597/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background: Acetabular fractures are complex injuries requiring accurate radiographic evaluation, typically using the Letournel-Judet classification system. While large language models like ChatGPT have emerged as promising decision-support tools in medical practice, their efficacy in complex orthopaedic trauma remains unclear. This study aimed to evaluate the diagnostic reliability of ChatGPT-4o compared to orthopaedic residents in classifying acetabular fractures based on standard radiographs, and to assess the AI model's accuracy in identifying specific radiographic landmarks.

Methods: This retrospective observational study evaluated anonymized anteroposterior, iliac oblique, and obturator oblique radiographs from 184 patients with acetabular fractures. Cases were independently evaluated by ChatGPT-4o, two fourth-year orthopaedic residents, and an experienced trauma surgeon who served as the reference standard. A standardized prompt based on a systematic radiographic checklist was used to guide ChatGPT. Diagnostic agreement was assessed using Cohen's kappa coefficient, and McNemar's test was used to compare diagnostic performances.

Results: The overall diagnostic accuracy for the final Letournel-Judet classification was 17.9% for ChatGPT-4o, compared to 88.6% and 91.8% for the first and second orthopaedic residents, respectively. Pairwise agreement with the reference standard was almost perfect for both the first (kappa = 0.860) and second (kappa = 0.900) residents, whereas ChatGPT-4o demonstrated negligible agreement (kappa = -0.014). Interobserver agreement between the two residents was almost perfect (kappa = 0.946). Both residents performed significantly better than the AI model ($p < 0.001$), with no significant difference observed between the residents ($p = 0.070$). While ChatGPT-4o achieved moderate accuracy in identifying isolated landmarks like the iliac wing (82.1%), its overall checklist accuracy was 67.9%.

Conclusion: Current multimodal artificial intelligence models, specifically ChatGPT-4o, are fundamentally inadequate for classifying complex acetabular fractures using the Letournel-Judet system. Given its unacceptably low diagnostic accuracy and inability to reliably integrate multiple radiographic landmarks compared to clinical residents, the use of this AI model as a decision-support tool in pelvic trauma is definitively not recommended.

Trial Registration: ClinicalTrials.gov, NCT07673991, registered on June 2026, retrospectively registered.

Introduction

Acetabular fractures are complex injuries that require accurate radiographic evaluation for proper classification and surgical planning. The Letournel–Judet classification system remains the most widely used framework for describing acetabular fracture patterns and guiding treatment decisions [1-3]. Despite its widespread clinical use, the Letournel–Judet classification system demonstrates moderate interobserver variability, particularly in complex fracture patterns [3-5]. The lack of consensus among even experienced trauma surgeons when diagnosing these injuries highlights the persistent challenge of achieving diagnostic consistency in orthopedic trauma. Although the original classification relies on

plain radiographs, computed tomography (CT) with multiplanar reconstructions is currently considered the gold standard for the definitive evaluation and preoperative planning of these complex fractures [6-8]. Misclassification may result in an inappropriate surgical approach, inadequate reduction, and posttraumatic arthritis.

In recent years, the integration of Artificial Intelligence (AI) and machine learning into orthopedic surgery has gained significant momentum [9, 10]. AI systems have demonstrated remarkable success in the automated detection of fractures, bone age assessment, and the classification of relatively simpler injuries, such as hip and distal radius fractures [11]. Recent studies have shown that artificial intelligence can detect fractures with comparable accuracy to clinicians, and that AI-assisted classification has higher accuracy than clinicians alone [10-12]. AI models like ChatGPT demonstrated substantial agreement with clinicians in classifying basic fracture characteristics [13]. Large language models (LLMs), particularly ChatGPT-4o, have further expanded these boundaries with their multimodal capabilities, allowing for the direct analysis of complex radiographic images and clinical scenarios [12, 13].

However, to the best of our knowledge, there are currently no studies in the literature specifically evaluating the diagnostic reliability and performance of large language models in the classification of acetabular fractures. While AI has been tested on various anatomical regions, its ability to navigate the sophisticated, stepwise radiographic integration required for the Letournel–Judet system remains unexplored.

Therefore the aim of this study is to evaluate the diagnostic accuracy and reliability of ChatGPT-4o in classifying acetabular fractures based on Judet radiographs. Furthermore, we aimed to compare the performance of this AI model with that of orthopaedic residents and an experienced trauma surgeon to determine its current feasibility as a clinical decision-support tool in complex pelvic trauma.

Materials and Methods

Study design

This study was designed as a retrospective single-arm cohort study. This study was conducted and reported in accordance with the CONSORT guidelines. Ethical approval was obtained from the institutional review board (approval number: TABED 2-26-1888), and the study was conducted in accordance with the Declaration of Helsinki. Patients who underwent surgery between May 2022 and December 2026 were included in the study A total of 184 patients aged older than 18 years with pelvic injury with acetabular fracture which had undergoing surgical treatment at our center. were included.

Patients with pelvic injuries but with non-acetabular fractures, those with incomplete imaging, those with poor radiographic quality or those who had previously undergone pelvic surgery or trauma were excluded.

Sample collection and processing

Imaging from cases presenting with pelvic injuries and undergoing surgical treatment at our center was selected from the institution's radiology archive. Standard anteroposterior pelvic, iliac oblique and obturator oblique radiographs were collected for each case.

AI Model Training and Structured Prompting Strategy

The article "Acetabular Fractures: Easier Classification with a Systematic Approach" by Brandser and Marsh was uploaded to Chat GPT [15]. The prompt used when evaluating the cases was as follows: "In these AP pelvic, iliac oblique, and obturator oblique radiographs, identify the acetabulum fracture as defined by Judet, using the systematic approach in the provided article and answering each question on the radiological checklist individually." Each case was evaluated using the same prompt we had previously defined. Chat GPT was asked to answer each question individually based on the systematic approach defined in the article and ultimately indicate the type of fracture.These questions were designed to evaluate key radiographic landmarks and fracture components including:

1. Is there a fracture of the obturator ring?
2. Is the ilioischial line disrupted?
3. Is the iliopectineal line disrupted?
4. Is there a fracture of the iliac wing?
5. Is there a fracture of posterior wall?
6. Does the fracture divide the acetabulum into top and bottom halves or front and back halves?
7. Is there a spur sign?

Based on the responses to these questions, the final fracture type was determined according to the Letournel–Judet acetabular fracture classification system.

Outcome Measures

Cases included in the study were independently classified by two fourth-year orthopaedic residents. To ensure objectivity, assessments were based solely on anteroposterior (AP), obturator oblique, and iliac oblique radiographs, without access to CT scans or intraoperative findings. Also a trauma surgeon with over a decade of experience in pelvic trauma surgery independently classified cases. To establish the true fracture pattern, the experienced trauma surgeon serving as the reference standard had full access to multiplanar computed tomography (CT) scans and intraoperative records. In contrast, the AI model and the orthopaedic residents were completely blinded to any CT data, basing their evaluations solely on the plain Judet radiographs.

Statistical Analysis

Descriptive statistics were used to summarize patient demographics and fracture characteristics. Continuous variables (e.g., patient age) were expressed as mean ± standard deviation (SD) and range, whereas categorical variables (e.g., sex and fracture types) were presented as frequencies and percentages.

The primary outcome measure was the diagnostic accuracy of ChatGPT-4o, orthopaedic residents, and the trauma surgeon in classifying acetabular fractures according to the Letournel–Judet classification system. The reference standard was established using three-dimensional pelvic CT findings together with the final intraoperative assessment of the experienced trauma surgeon. Accuracy was calculated as the proportion of correctly classified fractures relative to the reference standard.

Interobserver agreement between evaluators and the reference standard was assessed using Cohen's kappa (κ) coefficient for pairwise comparisons. Kappa values were interpreted according to the criteria described by Landis and Koch: <0, poor agreement; 0.01–0.20, slight agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and 0.81–1.00, almost perfect agreement.

Comparisons of diagnostic accuracy rates between ChatGPT-4o and human evaluators were performed using McNemar's test for paired categorical data. A p value of <0.05 was considered statistically significant. All statistical analyses were conducted using IBM SPSS Statistics version 29.0 (IBM Corp., Armonk, NY, USA).

Results

A total of 222 patients were included in the study. However, 26 (11.7%) patients were excluded from the study due to non-acetabular fractures (such as sacrum, pubic rami, or isolated iliac wing fractures) and 12 (5.4%) were excluded because the defined pelvic radiographs (AP, iliac oblique, obturator oblique) were incomplete or the image quality was poor. Radiographs of the remaining 184 patients (82.9%) were anonymized and randomly ordered before evaluation, and then uploaded to ChatGPT, which was asked to answer a set of questions individually. (Table 1)

A total of 184 acetabular fracture cases were included in the final analysis. ChatGPT-4o correctly classified 33 of 184 cases, corresponding to an overall diagnostic accuracy of 17.9%. (Table 2) In the structured seven-question Judet radiographic checklist analysis, ChatGPT-4o provided 875 correct responses out of 1288 individual checklist items, yielding an overall checklist accuracy of 67.9%. (Table 3) Among the individual checklist items, the highest accuracy was observed for spur sign identification (159/184, 86.4%) and iliac wing involvement (151/184, 82.1%). In contrast, the lowest accuracy was observed for posterior wall fracture detection (97/184, 52.7%) and assessment of whether the fracture divided the acetabulum into front–back or top–bottom components (103/184, 56.0%). The accuracy rates for obturator foramen disruption, ilioischial line disruption, and iliopectineal line disruption were 68.5%, 66.8%, and 63.0%, respectively.(Table 3)

The orthopaedic residents demonstrated substantially higher diagnostic accuracy than ChatGPT-4o. The 4th-year first resident correctly classified 163 of 184 cases, corresponding to an accuracy of 88.5%, whereas the 4th-year second resident correctly classified 169 of 184 cases, corresponding to an accuracy of 91.8%. (Table 2) Agreement with the reference standard was almost perfect for the first resident (κ = 0.860) and almost perfect for the second resident (κ = 0.900). Interobserver agreement between the two residents was also almost perfect (κ = 0.946). (Table 4)

When ChatGPT-4o was compared with the orthopaedic residents, both residents significantly outperformed the AI model. McNemar analysis demonstrated that the diagnostic accuracy of both the second resident and the first resident was significantly higher than that of ChatGPT-4o ($p < 0.001$ for both comparisons). There was no statistically significant difference in diagnostic accuracy between the two residents ($p = 0.070$). Pairwise agreement between ChatGPT-4o and the residents was low, with only slight agreement between ChatGPT-4o and the second resident (κ = 0.095) and between ChatGPT-4o and the first resident (κ = 0.112). Agreement between ChatGPT-4o and the reference standard was poor, with a Cohen's kappa value of κ = 0.052. (Table 5)

Confusion matrix analysis revealed a marked tendency of ChatGPT-4o to misclassify acetabular fractures as both-column fractures. This overclassification pattern was particularly evident in posterior wall, posterior column, anterior wall–column, T-type, and transverse plus posterior wall fracture patterns. Among true posterior wall fractures, ChatGPT-4o correctly classified only 5 of 53 cases. Most posterior wall fractures were misclassified as both-column fractures, anterior wall fractures, anterior column fractures, posterior column fractures, or anterior wall–column fractures. Among true both-column fractures, ChatGPT-4o correctly classified 6 of 43 cases, while the most frequent incorrect classification was transverse fracture. True posterior column fractures and anterior wall–column fractures were not correctly classified by ChatGPT-4o in any case. The highest subtype-specific performance was observed for transverse plus posterior wall fractures, for which ChatGPT-4o correctly classified 10 of 20 cases. (Table 6)

Overall, these findings indicate that although ChatGPT-4o showed moderate performance in identifying certain individual radiographic landmarks, particularly spur sign and iliac wing involvement, its final Letournel–Judet fracture classification accuracy was low. The model appeared to recognize some isolated radiographic features but showed substantial difficulty integrating these findings into the correct global fracture pattern, especially in complex acetabular injuries requiring spatial interpretation of anterior and posterior column involvement.

Discussion

The classification of acetabular fractures is still controversial and there is no gold standard protocol (3-5). The classification of complex acetabular fractures remains inherently difficult, often leading to interobserver variability even among highly experienced orthopaedic trauma surgeons. Despite this persistent clinical uncertainty, the diagnostic potential of artificial intelligence in navigating such

complex fracture patterns remains unexplored. Recent literature has increasingly highlighted the success of artificial intelligence in medical imaging, with several studies suggesting that AI models demonstrate diagnostic performance comparable to, or even exceeding, that of human experts in evaluating radiographs. Machine learning algorithms and deep neural networks have shown radiologist-level accuracy in fundamental tasks such as simple fracture detection, skeletal age assessment, and basic triage [11,13,16]. However, our findings present a stark contrast to this general optimism; when faced with a complex, multi-step anatomical evaluation like the Letournel-Judet classification, the multimodal large language model completely failed to match the diagnostic accuracy of clinical residents.

To the best of our knowledge, the most important feature that distinguishes the present study from previous literature is that it represents the first investigation to evaluate the diagnostic reliability and performance of a multimodal artificial intelligence model in the complex classification of acetabular fractures. When we look at the studies that have evaluating the use of artificial intelligence models for fracture classification, Kuo et al. (11) demonstrated through a systematic review and meta-analysis that artificial intelligence models can achieve high diagnostic accuracy in fracture detection and serve as effective decision-support tools, particularly in triage processes. However, most of these studies have primarily focused on basic fracture detection and simpler injury patterns. Our findings bridge this gap in the literature by demonstrating that the current performance of artificial intelligence is not as reliable in complex acetabular fracture classification, which requires the highly sophisticated, multi-step Letournel–Judet classification system. Cha et al. (12) conducted a systematic review highlighting that while artificial intelligence and machine learning models have shown significant potential in the diagnosis and classification of hip fractures, the majority of current research is limited to basic fracture identification. Their review underscores a clinical need for models that can handle more complex fracture patterns. Our study contributes to this area of research by evaluating the application of a multimodal AI model in the context of the Letournel–Judet system, demonstrating that such models currently struggle to achieve the reliability required for managing the complexities of acetabular fractures. Olczak et al. (13) demonstrated that artificial intelligence models could successfully differentiate between various orthopaedic fracture types using standard trauma radiographs, highlighting the potential for AI to serve as an objective screening tool. However, their research emphasized that these models are trained on specific, often limited datasets. Our study expands upon this by testing a multimodal model (ChatGPT-4o) in a complex, multi-step classification environment (Letournel–Judet) rather than a basic identification task, revealing that the transition from simple fracture detection to hierarchical classification remains a significant challenge for contemporary AI. Recently, Walker et al. (14) reported on the interrater reliability of ChatGPT-4 in radiographic fracture classification, providing essential evidence regarding the model's performance in clinical orthopaedic settings. While their findings established a baseline for AI's diagnostic potential in trauma, our study further explores this domain by focusing specifically on the Letournel–Judet classification of acetabular fractures. Unlike Walker et al., who may have addressed broader trauma categories, our investigation highlights the specific difficulties AI faces when tasked with the multi-step hierarchical synthesis required for complex pelvic fractures.

Collectively, these studies confirm that while large language models show promise as supportive tools, their current limitations in spatial interpretation and structural classification underscore the necessity for expert clinical oversight before their integration into trauma workflows.

Although ChatGPT demonstrated good performance in assessing the iliac wing, obturator foramen continuity, and the presence of the spur sign, and moderate performance in evaluating column continuity, its performance in assessing posterior wall integrity was markedly limited. Furthermore, the model showed poor accuracy in determining the final fracture classification. The spur sign requires recognition of a specific geometric relationship between the iliac wing fragment and the acetabular dome. The model's tendency to generate frequent false positives suggests misidentification of other radiographic features as spur signs—a form of visual hallucination consistent with known LLM failure modes in complex radiographic interpretation tasks. These findings indicate that, although Chat GPT can recognize certain individual radiographic features, it is unable to effectively integrate this information to produce an accurate Letournel–Judet classification of acetabular fractures. This dissociation indicates that the failure is not primarily one of individual landmark recognition but of hierarchical decision synthesis: the model cannot reliably integrate sequentially acquired binary landmark information into a coherent classification that adheres to the Letournel–Judet decision tree. The consistent fair agreement for view may arise from a limitation in the visual interpretation of 2 dimensional radiographic image.Higher accuracy rates are likely to be achieved with the analysis of 3D CT sections. Its ability to accurately identify anatomical landmarks individually in radiographs is remarkable. In a short time, LLMs like ChatGPT have made significant progress. Future research should explore hybrid human–AI workflows in which AI landmark annotation supports rather than replaces human classification. Despite these advancements, applications in highly specialized multi-step tasks such as detailed fracture classification may require further refinement, specialized training, or the use of optimized guidance techniques.

This study has several limitations. The prompts used in this study represent a significant limitation. Even though the same prompt was used in each evaluation, the chosen prompt reflects only one of many possible options that could be used to guide ChatGPT. Linear learning models (LLMs) like ChatGPT are highly sensitive to the wording, form, and context of prompts, leading to variability that can affect the results. Prompt dependency makes it difficult to replicate and generalize these findings across different LLMs. Another significant limitation is that only two orthopedic residents from the same institution evaluated the cases, further reducing generalizability. Additionally, only 184 cases were evaluated, which can be considered a small sample size. Using a larger and more diverse case series would strengthen the study's findings. Furthermore, since only ChatGPT was used in this study, comparing it with other AI models could provide a better understanding of AI's role in fracture classification.

Conclusion

In conclusion, current multimodal artificial intelligence models, specifically ChatGPT-4o, are fundamentally inadequate for classifying complex acetabular fractures using the Letournel-Judet

system. Given its unacceptably low diagnostic accuracy (17.9%) and inherent inability to integrate multiple radiographic landmarks compared to orthopaedic residents, we strongly advise against the use of this AI model as a diagnostic or decision-support tool in clinical practice for pelvic trauma. Until these systems evolve to master complex spatial and anatomical synthesis, relying on them for acetabular fracture classification is definitively not recommended. These rapidly developing tools promise wide applications in orthopedic surgery in the future, and appear to have significant potential benefits, particularly in residents training. Further studies in the literature are needed to guide the development of these tools.

Declarations

Ethics approval and consent to participate: Ethics approval for this study was obtained from the Institutional Review Board of the Ethics Committee of the Ankara Bilkent City Hospital (Approval No. TABED 2-26-1888) Written informed consent was obtained from all participants to take part in the study. The study was conducted in accordance with the Good Clinical Practice guidelines and the recommendations of the Declaration of Helsinki.

Consent for publication: Not applicable.

Availability of data and materials: The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests: The authors declare that they have no competing interests.

Funding: No funding was received for this study.

Authors' contributions: M.K conceived and designed the study. O.F.O, A.O.Y, F.I, E.C, I.A.Y, Y.K, T.A performed the surgical procedures. M.K collected and analysed the data. M.K prepared the figures, M.K drafted the manuscript. All authors read and approved the final manuscript.

References

1. Rowe CR, Lowell JD. Prognosis of fractures of the acetabulum. J Bone Joint Surg Am. 1961;43:30–59.
2. Judet R, Judet J, Letournel E. Fractures of the acetabulum: classification and surgical approaches for open reduction. J Bone Joint Surg Am. 1964;46:1615–46.
3. Letournel E. Acetabulum fractures: classification and management. Clin Orthop. 1980;151:81–106.
4. Beaulé PE, Dorey FJ, Matta JM. Letournel classification for acetabular fractures: assessment of interobserver reliability. J Bone Joint Surg Am. 2003;85:1704–9.
5. Petrisor BA, Bhandari M, Orr RD, Mandel S, Kwok DC, Schemitsch EH. Variance in the classification of acetabular fractures. J Orthop Trauma. 2003;17(4):256–9.

6. O'Toole RV, Cox G, Shanmuganathan K, et al. Evaluation of computed tomography for determining the diagnosis of acetabular fractures. J Orthop Trauma. 2010;24(5):284–90.
7. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. Radiogr Rev Publ Radiol Soc N Am Inc. 2017;3 7(2):505–15.
8. Mack LA, Harley JD, Winquist RA. CT of acetabular fractures: analysis of fracture patterns. AJR Am J Roentgenol. 1982;138:407–12.
9. Yücens M, Aydemir AN, Demirkan AF. Assessment of interobserver reliability for the Letournel and Judet classification. Acta Ortop Bras. 2024;32(1):e265219.
10. OpenAI AJ, Adler S et al. GPT-4 technical report. Published online March 4, 2024. 10.48550/arXiv.2303.08774
11. Kuo RYL, Harrison C, Curran TA, et al. Artificial intelligence in fracture detection: a systematic review and meta-analysis. Radiology. 2022;304:50–62.
12. Cha Y, Kim JT, Park CH, et al. Artificial intelligence and machine learning on diagnosis and classification of hip fracture: systematic review. J Orthop Surg Res. 2022;17:520.
13. Olczak J, Fahlberg N, Maki A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs. Acta Orthop. 2017;88:581–6.
14. Aliyah N, Walker BS, a JB, Smith MD, b Samuel K, Simister MD, b Om Patel MBA, BS, MBS, a Soham Choudhary, BS, a Michael Seidu, BS, a David Dallas-Orr, MD, MBA, MTM, b Shannon Tse, BMBS, b Hania Shahzad, MD, b Patrick Wise, MD, b Michelle Scott, MD, b, Augustine M. Saiz, MD, b and Zachary C. Lum, DO. Assessing Interrater Reliability of ChatGPT-4 and Orthopaedic Clinicians in Radiographic Fracture Classification. J Orthop Trauma 2026;40:e27–e32.
15. Brandser E, Marsh JL. Acetabular fractures: easier classification with a systematic approach. AJR Am J Roentgenol. 1998;171:1217–28.
16. Lindsey RV, Daluiski A, Chopra S, et al. Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci U S A. 2018;115(45):11591–6.

Tables

Table 1. Demographic Characteristics of the Patients

Number of patients	184
Sex (F/M)	56 (30.4%) / 128 (69.6%)
Mean age (years)	39.8 ± 18.0 (range: 18–80)

Table 2. Diagnostic Accuracy of Orthopaedic Residents vs. ChatGPT-4o (n=184)

Measurement	1st resident	2nd resident	ChatGPT-4o
Correct classification (n)	163	169	33
Accuracy(%)	88.5%	91.8%	17.9%

Table 3. Accuracy of ChatGPT-4o in the Structured Radiographic Checklist Analysis (n=184)

Radiographic Checklist Item	Correct Responses (n)	Accuracy(%)
Spur sign identification	159	86.4%
Iliac wing involvement	151	82.1%
Obturator foramen disruption	126	68.5%
Ilioischial line disruption	123	66.8%
Iliopectineal line disruption	116	63.0%
Front-back or top-bottom component assessment	103	56.0%
Posterior wall fracture detection	97	52.7%
Overall Checklist Accuracy	875/1288	67.9%

Table 4. Interrater Reliability Assessment using Cohen's kappa coefficient

Comparison	Cohen's kappa	Strength of Agreement
4th-Year Resident (R2) vs. Reference	0.900	Almost perfect
4th-Year Resident (R1) vs. Reference	0.860	Almost perfect
ChatGPT-4o vs. Reference	-0.014	Poor / Negligible

Table 5. Comparative Performance and Pairwise Agreement Between ChatGPT-4o and Orthopaedic Residents

Comparison	Accuracy Difference	P-value*	Cohen's kappa
ChatGPT-4o vs. Resident 1	Significant	< 0.001	0.112
ChatGPT-4o vs. Resident 2	Significant	< 0.001	0.095
Resident 1 vs. Resident 2	Not Significant	0.070	-
ChatGPT-4o vs. Reference	-	-	0.052

**p-values derived from McNemar's test for diagnostic accuracy comparisons between pairs.*

Agreement strength based on Landis and Koch criteria: 0.01–0.20 (Slight), < 0.00 (Poor).

Table 6. Common Misclassification Patterns of ChatGPT-4o by Fracture Type

Actual Fracture Type (Reference)	Actual Fracture Type (Reference)
Posterior wall	Both-column, Anterior wall, Anterior wall–column
Both-column	Transverse
Anterior column	Both-column
Anterior wall	Transverse, Both-column, T-type
Posterior column	Both-column
Transverse + Posterior wall	Posterior wall, T-type, Transverse

Poor Diagnostic Sensitivity: ChatGPT-4o demonstrated a notably low diagnostic performance in identifying true posterior wall fractures (9.4%).

Complete Diagnostic Failure: The model failed to correctly classify any cases of posterior column and anterior wall–column fractures.

Overclassification Bias: A significant tendency for "overclassification" was observed; the model frequently misclassified various complex fracture patterns as "both-column" fractures, indicating a structural bias in its analytical framework.