Performance of ChatGPT-5 in Diagnosing Fractures on Proximal Humerus and Intertrochanteric Femur X-Rays

I Ibad Sha[1], Ibrahim S Majeed[2], Riju R[3]

Learning Point of the Article:

ChatGPT-5 can assist with fracture screening, but its limitations mean it cannot replace expert radiographic interpretation.

Abstract

Introduction: Large language models (LLMs) such as ChatGPT-5 offer new possibilities for interpreting medical images, but their effectiveness in orthopedic radiograph analysis remains largely unexplored.

Objective: To evaluate the diagnostic performance of ChatGPT-5 in detecting and classifying fractures on shoulder and hip X-rays, specifically proximal humerus and intertrochanteric (IT) femur fractures.

Materials and Methods: A retrospective study of 120 anonymized anteroposterior (AP) radiographs (60 shoulder and 60 hip) was conducted. Each case was independently reviewed by orthopedic experts, establishing a reference standard. ChatGPT-5 analyzed the same images using structured prompts and was assessed for fracture detection accuracy, sensitivity, specificity, and agreement on detailed fracture features.

Results: ChatGPT-5 achieved 87.5% sensitivity and 100% specificity in detecting proximal humerus fractures ($\kappa = 0.74$), and 100% sensitivity but only 16.7% specificity in IT femur fractures ($\kappa = 0.24$). While it identified major fracture patterns and comminution reliably, it frequently hallucinated fractures in normal hip X-rays and missed fine details such as lesser tuberosity fragments and dislocations.

Conclusion: ChatGPT-5 shows high sensitivity for orthopedic fracture detection and produces coherent, structured reports. However, limitations in specificity and fine-detail recognition restrict its autonomous clinical use. It may serve as a triage or educational tool with human oversight or be integrated into hybrid artificial intelligence workflows.

Keywords: ChatGPT-5, artificial intelligence radiology, fracture detection, proximal humerus, Intertrochanteric femur.

Introduction

The emergence of artificial intelligence (AI) in musculoskeletal (MSK) radiology has provided powerful tools to improve fracture detection and classification on imaging [1]. Deep learning models, particularly convolutional neural networks (CNNs), have achieved high accuracy in identifying fractures on radiographs [2]. For example, a CNN trained on shoulder X-rays distinguished proximal humerus fractures from normal images with approximately 99% sensitivity and 97% specificity, performing on par with specialized orthopedic radiologists [2]. AI-assisted interpretation can also reduce human error in detecting subtle or complex fractures that might otherwise be missed due to fatigue or inexperience [3]. Recent meta-analyses confirm that modern AI tools are non-inferior to clinicians for overall fracture detection on X-rays [4]. Timely diagnosis of fractures is critical to guide management and prevent

Access this article online

Website:
www.jocr.co.in

DOI:
https://doi.org/10.13107/jocr.2026.v16.i01.6702

Author's Photo Gallery

Dr. I Ibad Sha

Dr. Ibrahim S Majeed

Dr. Riju R

[1]Department of Orthopaedic Surgery, Government Medical College, Thiruvananthapuram, India,
[2]Department of Orthopaedic Surgery, Mount Zion Medical College, Adoor, Kerala, India.
[3]Department of Orthopaedic Surgery, Lifeline Hospital, Adoor, Kerala, India.

Address of Correspondence:
Dr. I Ibad Sha,
Department of Orthopaedic Surgery, Government Medical College, Thiruvananthapuram, India.
E-mail: ibadshah47@gmail.com

Submitted: 13/10/2025; Review: 17/11/2025; Accepted: December 2025; Published: January 2026

DOI: https://doi.org/10.13107/jocr.2026.v16.i01.6702

complications from missed injuries.

Large language models (LLMs) such as ChatGPT have so far been applied mainly to text-based medical tasks, including generating reports and answering clinical questions. The newest iterations, GPT-4 and beyond, incorporate vision capabilities, raising the question of whether a general AI can interpret medical images [5]. Early investigations into GPT-4 Vision's radiology performance showed mixed results: The model answered text-only radiology examination queries with approximately 81% accuracy but correctly solved only about 48% of image-based questions [6]. This gap suggests that while LLMs possess broad medical knowledge, extracting precise diagnostic information from images remains challenging. Nevertheless, if an advanced model such as ChatGPT-5 can reliably analyze radiographs, it could assist in clinical triage or act as a second-reader system to enhance radiologist confidence.

Fracture radiographs present a useful test domain for such AI because of their high volume and well-defined diagnostic targets. The present study focused on two common yet distinct fracture types: Proximal humerus fractures and intertrochanteric (IT) hip fractures. These injuries often have subtle radiographic features that influence management. This study evaluated ChatGPT-5's ability to interpret shoulder and hip radiographs, including fracture detection accuracy and agreement with experts on key fracture characteristics (e.g., displacement, comminution, and alignment), as well as IT fracture stability. Diagnoses by orthopedic specialists served as the gold standard. The goal was to establish an initial benchmark of ChatGPT-5's diagnostic reliability on real-world fracture cases and to explore its potential role in clinical radiographic interpretation workflows.

Materials and Methods

$$n = \frac{Z^2\, S(1-S)}{L^2\, P}$$

A total of 120 anonymized radiographs (60 proximal humerus and 60 IT femur) were included. Sample size was determined using Buderer's method for estimating the precision of sensitivity in diagnostic studies. Buderer's formula is:

$$n = \frac{(1.96)^2 \times 0.85 \times 0.15}{(0.10)^2 \times 0.80} \approx 61 \text{radiographics.}$$

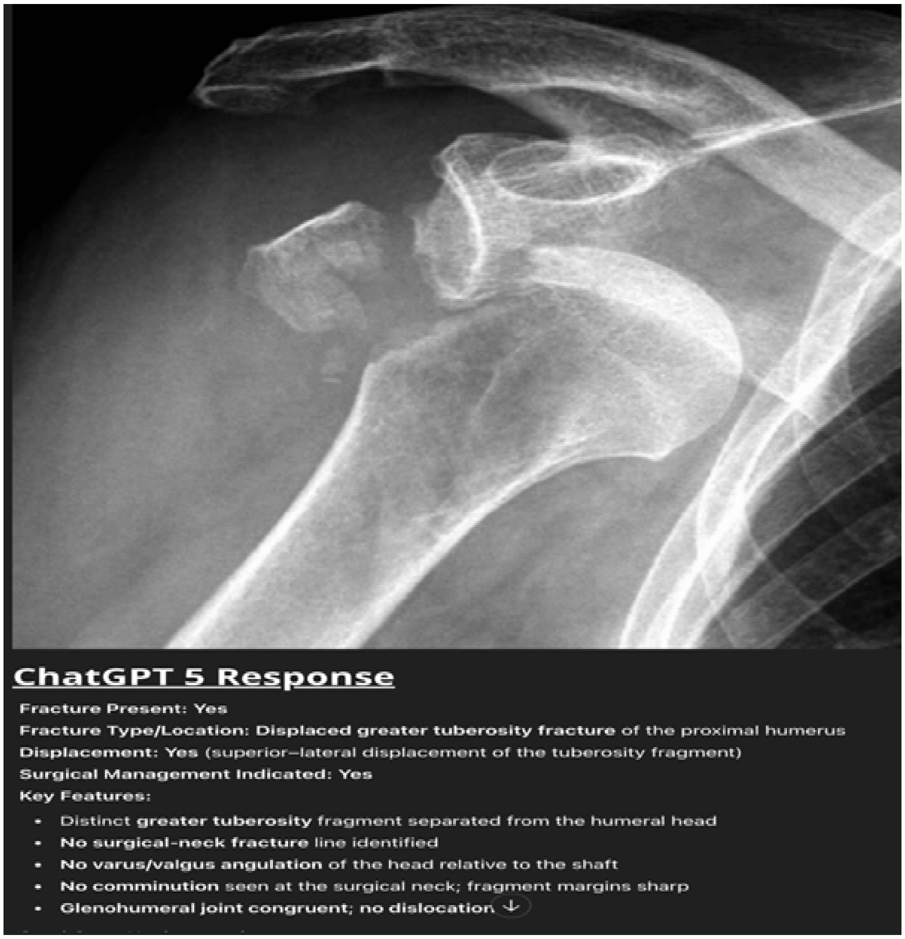
Where SSS = anticipated sensitivity, LLL = desired half-width of the confidence interval (CI) (precision), PPP = prevalence of the condition in the sample, and ZZZ = 1.96 for a 95% confidence level.

Using an expected sensitivity S = 0.85 S = 0.85S = 0.85 (based on prior fracture-detection studies), precision L = 0.10L = 0.10L = 0.10 and prevalence P =0.80 P = 0.80 P = 0.80, the required sample size for estimating sensitivity was:

To allow for heterogeneity and potential exclusions, we selected 60 radiographs per anatomical region (48 fractures and 12 normals), giving an empirical sample close to the calculated minimum and sufficient to estimate sensitivity with ~±10% precision (95% CI). Two anatomical groups were studied: 60 shoulder radiographs for proximal humerus injuries, and 60 hip/pelvic radiographs for IT femur injuries. Each case consisted of a single anteroposterior (AP) X-ray demonstrating either a fracture or a normal study.

For the fracture cases, we included acute isolated fractures of the proximal humerus or IT region confirmed by clinical and radiologic assessment. Cases spanned a range of fracture patterns (from non-displaced to complex comminuted fractures) to challenge the model's classification abilities. For the proximal humerus, all Neer classification types (one-part through four-part fractures, including fracture-dislocations) were represented. For IT femur, the spectrum included both stable (e.g., minimal comminution and intact lateral wall) and unstable patterns (comminuted, lateral wall compromised, reverse obliquity, etc.). We also included a subset of normal radiographs (n = 12 in each group) with no fracture, to evaluate specificity. Each radiograph was de-identified, and only a single view (AP) per case was used to match the typical inputs that might be given to a general AI model. We excluded pediatric cases, pathological fractures, post-operative films, or cases with implants, as well as cases with insufficient imaging quality.

The reference standard was established through independent review by two board-certified orthopedic surgeons, with a third senior reviewer resolving any discrepancies. The experts were provided the radiographs and asked to record: (1) whether a fracture was present, and (2) detailed fracture characteristics as defined below. For proximal humerus fractures, experts recorded the Neer classification (number of parts) and specific features: whether any fragment was displaced (per Neer's criteria of >1 cm or >45° angulation), whether the greater tuberosity (GT) fragment was present and displaced >5 mm, whether the lesser tuberosity was involved, presence of varus angulation of the humeral head, presence of calcar region comminution (medial hinge disruption), and whether an associated glenohumeral (GH) dislocation was present [7]. For IT femur fractures, experts recorded: fracture stability classification (designating "unstable" if classic risk factors such as a broken lateral wall, large comminution, reverse obliquity or subtrochanteric extension were present, vs. "stable" if not), fracture displacement (whether the fracture was non-displaced/impacted vs. displaced); lateral wall integrity (intact vs. fractured lateral femoral cortex, a key determinant of



ChatGPT 5 Response

Fracture Present: Yes

Fracture Type/Location: Displaced greater tuberosity fracture of the proximal humerus

Displacement: Yes (superior–lateral displacement of the tuberosity fragment)

Surgical Management Indicated: Yes

Key Features:

- Distinct **greater tuberosity** fragment separated from the humeral head
- **No surgical-neck fracture** line identified
- **No varus/valgus angulation** of the head relative to the shaft
- **No comminution** seen at the surgical neck; fragment margins sharp
- **Glenohumeral joint congruent; no dislocation**

Figure 1: Anteroposterior shoulder radiograph of the shoulder showing a displaced fracture of the greater tuberosity with clear anterior glenohumeral dislocation. ChatGPT-5 correctly identified the fracture but failed to recognize the associated dislocation.

stability); presence of comminution (particularly of the posteromedial cortex); presence of a reverse obliquity fracture line; extension of the fracture into the subtrochanteric region; and presence of a separate lesser trochanter (LT) fragment [8]. These specific features were chosen because they are clinically important descriptors that affect management. The expert interpretations served as the gold standard for all analyses in the study.

ChatGPT-5 model and prompting

We accessed the ChatGPT-5 model (developer: OpenAI) through its multimodal interface in June 2025 [9]. ChatGPT-5 is an advanced successor to GPT-4, purported to have enhanced vision capabilities and context length. Although ChatGPT-5 is not explicitly trained on radiology datasets or our institutional radiographs, it is a general-purpose multimodal LLM with vision capabilities trained on broad image–text corpora. Because such models are increasingly used informally by clinicians and trainees for preliminary interpretation, evaluating their real-world diagnostic performance and failure patterns is clinically relevant. This study, therefore, aimed to assess the model's baseline capabilities and limitations when applied to routine orthopedic radiographs.

We interacted with ChatGPT-5 in a standardized manner for each case. The radiograph image was uploaded to the chat interface, and we entered a fixed prompt instructing the model to analyze the image and provide a structured report. The prompt was formulated to mirror a radiologist's approach and was kept consistent for all cases. An example prompt was: "Analyze the attached X-ray. Answer: (a) Is there a fracture? (yes/no); (b) If yes, describe the fracture including: for a proximal humerus, the Neer classification and whether there is displacement, GT fragment (>5 mm), lesser tuberosity fragment, varus malalignment, calcar comminution, or shoulder dislocation; for an IT femur, state if it is stable or unstable, if displaced, if lateral wall is intact, if comminuted, if reverse obliquity pattern, if subtrochanteric extension, and if a separate LT fragment is present." The model was asked to be concise and use the same structured format for each case. We did not allow the model to see any patient history or the expert's answers. Each case was processed independently in a new chat session to avoid any carryover of information. The output from ChatGPT-5 was then recorded, including the binary fracture presence call and each feature reported (yes/no or the classification).

Outcome measures and statistical analysis

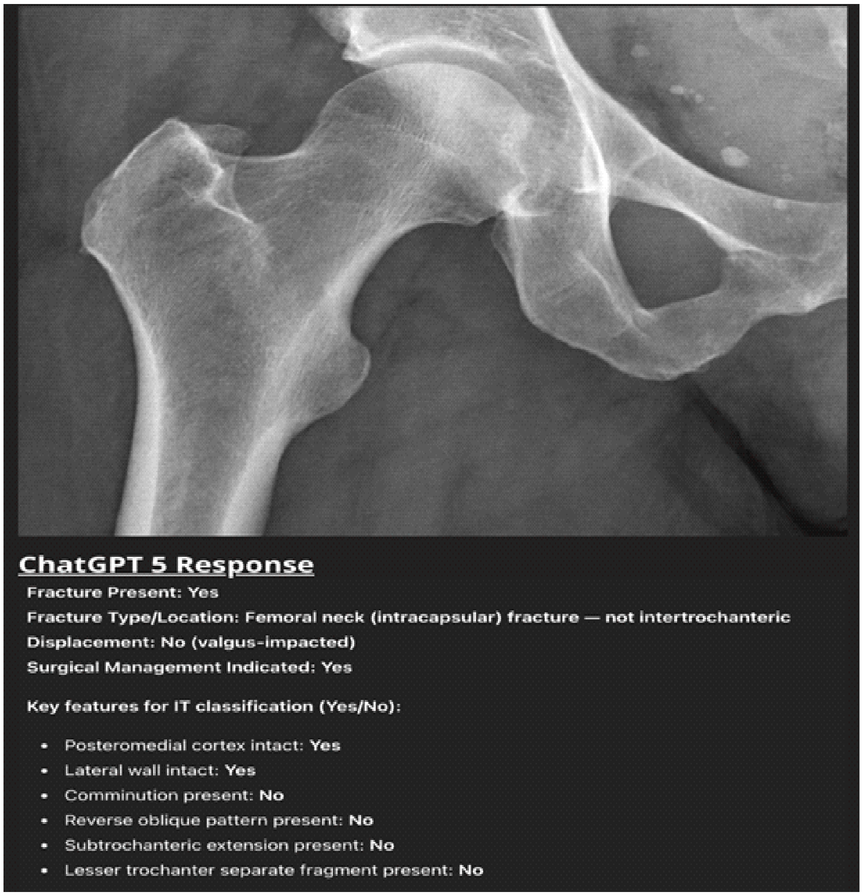
The primary outcome was the diagnostic agreement between ChatGPT-5 and the gold standard for fracture detection (presence/absence). Secondary outcomes were the performance metrics for each feature in the structured report (e.g., correctly identifying displacement, classification, etc.). For each binary outcome (fracture presence and each feature), we calculated sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and overall accuracy of ChatGPT-5's responses, using standard definitions. Cohen's kappa coefficient was calculated to assess agreement beyond chance for each feature (treating multi-class classifications as categorical for kappa calculation). For the Neer classification (a multi-class variable with 1-part, 2-part, 3-part, and 4-part categories), we evaluated the proportion of cases where ChatGPT-5's stated classification exactly matched the expert classification (overall accuracy), and we computed a kappa for multicategory agreement. Similarly, for IT fracture stability (binary stable vs. unstable classification), we analyzed agreement and kappa. In cases where ChatGPT-5 missed a fracture entirely (false negative), its feature descriptions were considered "no" by default for that case (since the model did not describe any features if it said no fracture). This penalized the model appropriately for failing to identify features when the fracture was missed. Conversely, if the model hallucinated a fracture on a normal case (false positive), any features it described were counted as false positive feature identifications.

All analyses were performed in IBM Statistical Package for the Social Sciences (SPSS) Statistics, Version 26 (IBM Corp., Armonk, NY, USA). Diagnostic test characteristics – sensitivity, specificity, PPV, NPV, and overall accuracy – were calculated with 95% CIs using exact (Clopper–Pearson) methods for proportions. Agreement between ChatGPT-5 and the reference standard was assessed with Cohen's kappa (κ) with 95% CIs; κ was interpreted as: <0.20 poor, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, >0.80 almost perfect. For multicategory variables (e.g., Neer classes), we report exact agreement (%) and weighted κ (linear weights). Continuous variables, where applicable, are presented as mean ± SD or median (IQR) based on distribution. Two-sided α = 0.05 was adopted; no between-group inferential comparisons were planned, given the descriptive validation design. Figures (bar charts of sensitivity/specificity and κ) were produced in SPSS. Missing data were not present; each case contributed one observation per endpoint. This retrospective analysis used fully de-identified images and was exempt from institutional review board review per local policy.

Results

ChatGPT-5 demonstrated strong performance in detecting the presence of fractures, though with variability between anatomical sites. In the proximal humerus group (n = 60; 48 fractures, 12 normal), the model identified 42 of 48 fractures and correctly labeled all 12 normals, resulting in sensitivity 87.5%, specificity 100%, and κ = 0.74 (substantial agreement). False negatives were predominantly non-displaced or minimally displaced surgical neck fractures. For example, in a case with a fracture dislocation of the proximal humerus, ChatGPT-5 correctly identified a displaced GT fracture but failed to detect an associated anterior GH dislocation, despite clear humeral head displacement relative to the glenoid, highlighting its blind spot for associated dislocations critical for surgical planning (Fig. 1).

In contrast, for the IT femur group (n = 60; 48 fractures, 12 normal), ChatGPT-5 detected all 48 fractures (100% sensitivity) but overcalled 10 of 12 normals as fractured, yielding specificity 16.7%, accuracy 83.3%, and κ = 0.24 (fair agreement). The model's tendency to hallucinate was highlighted by a case in which it identified a "displaced femoral neck fracture with varus angulation" on a normal hip X-ray despite no visible fracture line or trabecular disruption, exemplifying the model's propensity for hallucination (Fig. 2).



ChatGPT 5 Response

Fracture Present: Yes

Fracture Type/Location: Femoral neck (intracapsular) fracture — not intertrochanteric

Displacement: No (valgus-impacted)

Surgical Management Indicated: Yes

Key features for IT classification (Yes/No):

- Posteromedial cortex intact: Yes
- Lateral wall intact: Yes
- Comminution present: No
- Reverse oblique pattern present: No
- Subtrochanteric extension present: No
- Lesser trochanter separate fragment present: No

Figure 2: Normal anteroposterior hip radiograph that ChatGPT-5 misinterpreted as a "valgus-impacted femoral-neck fracture." Despite intact cortical outlines and an undisturbed trabecular pattern, the model reported a displaced intracapsular fracture requiring surgery, illustrating its tendency toward false-positive fracture detection and poor specificity in the intertrochanteric group.

Feature-level performance for proximal humerus fractures is presented in Table 1.

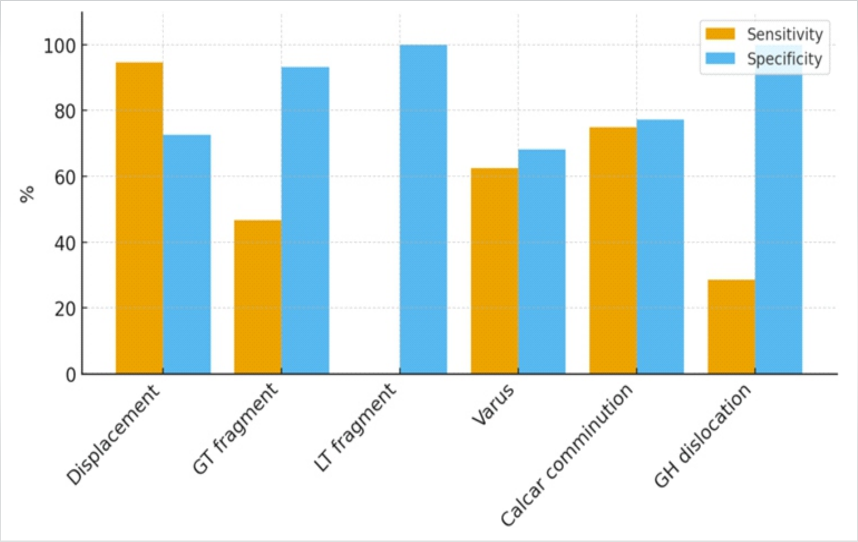


Figure 3: Sensitivity and specificity for each feature recognition in proximal humerus fractures.

ChatGPT-5 performed well in detecting displacement (sensitivity 94.7%, κ = 0.70) and calcar comminution (sensitivity 75.0%, κ = 0.47). GT displacement was recognized in fewer than half of cases (46.7% sensitivity) but with high specificity (93.3%). Lesser tuberosity involvement was missed in all cases (0% sensitivity, κ = 0.00). GH dislocations were identified in only 2 of 7 cases (28.6% sensitivity), as seen in the aforementioned case where the model overlooked a clear dislocation. Varus malalignment was inconsistently assessed, with only fair agreement (κ = 0.26). Overall, the model was reliable for detecting gross displacement but had substantial blind spots for smaller fragments and associated dislocations. These findings are depicted in Fig. 3, where sensitivity and specificity are contrasted for each fracture feature.

In the IT fracture group, ChatGPT-5 showed high agreement with experts for displacement (sensitivity 86.4%, specificity 100%, κ = 0.69) and reverse obliquity (sensitivity 100%, specificity 80.8%, κ = 0.53). Subtrochanteric extension was detected with moderate sensitivity (50.0%) but high specificity (88.9%), producing κ = 0.41. However, lateral wall integrity, a key determinant of stability, was poorly assessed (sensitivity 18.2%, specificity 52.6%, κ = −0.29). Similarly, comminution and LT fragments were detected in most true cases (78.6% sensitivity) but also overcalled in 50% of stable cases, reflecting the model's bias toward overestimating instability, as evidenced by the misidentification of a normal hip X-ray as a femoral neck fracture. These findings are summarized in Table 2 and shown graphically in Fig. 4, which highlights the overcalling of instability markers despite strong sensitivity for reverse obliquity.

Discussion

This study evaluated ChatGPT-5, a state-of-the-art LLM, for radiographic interpretation of proximal humerus and IT femur fractures, representing the first benchmark of a GPT-series model against experienced clinicians for detailed orthopedic imaging analysis. The results highlight a dichotomy: ChatGPT-5 achieved near-human sensitivity in fracture detection but struggled with specificity and detailed characterization, limiting its standalone clinical utility.

ChatGPT-5 demonstrated exceptional sensitivity, detecting 100% of hip fractures and 87.5% of proximal humerus fractures in our 120-case sample, aligning with AI trends favoring over-detection to minimize false negatives, critical in trauma triage [10]. This performance mirrors earlier deep learning models with high sensitivity for fracture screening [10]. Notably, proximal humerus results (87.5% sensitivity and 100% specificity) approached specialized CNNs, which report 97–99% sensitivity and specificity [11]. This is remarkable given ChatGPT-5's lack of radiology-specific training, suggesting its broad image and text training enabled generalizable pattern recognition. The model excelled in identifying distinct features, such as reverse obliquity in all IT fractures, likely due to its characteristic appearance, possibly learned from textbooks or online content. It also reliably detected comminution (with some false

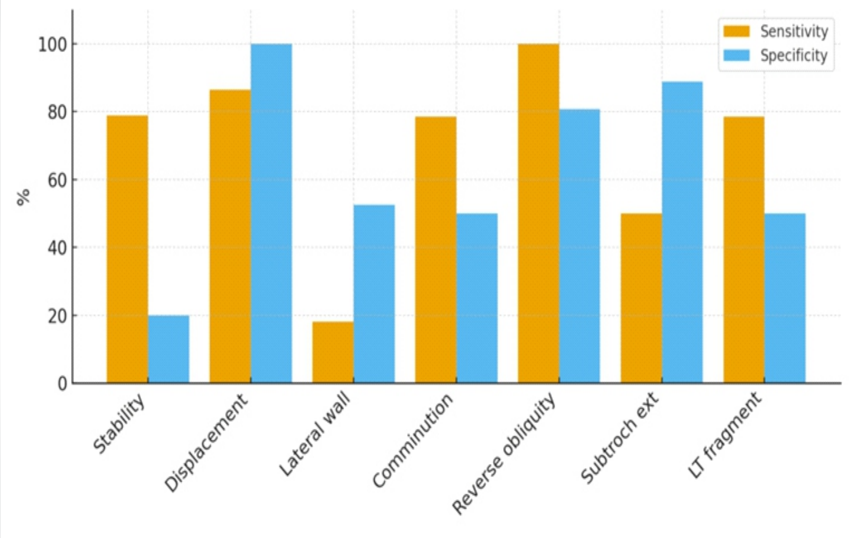


Figure 4: Sensitivity and specificity of ChatGPT-5 for each feature in intertrochanteric femur fractures.

Table 1: Diagnostic performance per feature-proximal humerus fractures

Feature	Sensitivity (%)	Specificity (%)
Fracture presence	87.5	100
Displacement present	94.7	72.7
GT fragment >5 mm	46.7	93.3
LT fragment involvement	0	100
Varus malalignment	62.5	68.2
Calcar comminution	75	77.3
GH dislocation	28.6	100

GT: Greater tuberosity, LT: Lesser trochanter, GH: Glenohumeral

positives) and gross displacement across both anatomies, indicating competence in interpreting basic bone fragment alignment, a fundamental radiographic skill.

The model's output was notably fluent and structured, often resembling radiology report impressions (e.g., "unstable comminuted IT fracture with subtrochanteric extension and a displaced lesser trochanter fragment"). This suggests potential for report generation and standardization, a role explored in prior studies converting free text to structured reports [12]. In clinical settings, ChatGPT-5 could draft reports, saving time for radiologists by pre-populating structured fields. While accuracy issues limit immediate utility, a hybrid workflow-where the model proposes drafts for human correction-could enhance efficiency if drafts are mostly accurate [12]. Existing radiology AI products use similar approaches for simpler tasks, and LLMs offer flexibility for varied descriptions.

Despite its strengths, ChatGPT-5's limitations preclude clinical adoption. Its low specificity, with hallucinations of fractures in five normal hip X-rays and misidentification of femoral neck fractures in three, poses significant risks, as false positives could lead to unnecessary treatments [13]. Unlike task-specific AI, which flags ambiguous regions, ChatGPT-5's confident but incorrect assertions reflect known LLM issues [13]. A systematic review cautioned against unsupervised LLM use in medicine due to such pitfalls [14].

The model also struggled with fine-grained classifications. For proximal humerus fractures, it failed to identify lesser tuberosity fragments, critical for distinguishing 3-part from 4-part fractures, resulting in only 20% accuracy for 4-part fractures. In contrast, dedicated deep learning models achieve 71–90% accuracy for Neer classifications [15, 16]. For IT fractures, ChatGPT-5 labeled 95% as unstable (79% sensitivity and 20% specificity), far below vision AI performance (85–88% sensitivity and 95–99% specificity) [17, 18]. Misidentification of subtle markers, such as lateral wall integrity (50% accuracy), likely contributed, as this sign is crucial for stability assessment [17]. These shortcomings stem from the model's lack of specialized radiographic training, unlike CNNs optimized for such tasks.

This study bridges a gap between general LLM capabilities and specialized medical imaging AI. Previous ChatGPT radiology studies focused on non-interpretive tasks, such as explaining reports in lay language or drafting impressions [18]. Diagnostic reasoning studies using text-based vignettes reported 50–60% accuracy for ChatGPT-4, consistent with our findings of partial reliability [18]. A study by Jiao et al. on GPT-4 for MSK magnetic resonance imaging appropriateness (text-based) concluded LLMs can assist but not replace clinicians [19]. For image interpretation, data are limited; one study found GPT-4 had ~20% accuracy for radiographic positioning errors, mirroring our complex feature accuracy [20]. These studies highlight LLMs' tendency to miss specifics, producing partially correct answers.

In fracture diagnosis, domain-specific training is critical. ChatGPT-5's high sensitivity but low specificity for hip fractures resembles early computer vision models before rigorous tuning. Modern CNNs, trained on thousands of labeled images, achieve balanced metrics (>90% sensitivity/specificity) [10, 18]. ChatGPT-5, as a generalist, behaves like a well-read layperson, listing relevant features but inconsistently applying them to images due to limited visual perception training. For example, it may know "lesser tuberosity fracture implies 4-part" from texts but cannot reliably identify such fragments on X-rays, unlike CNNs optimized for pixel patterns [16].

Table 2: Diagnostic performance per feature – intertrochanteric femur fractures

Feature	Sensitivity (%)	Specificity (%)
Fracture presence	100	16.7
Stability classification (unstable vs. stable)	78.9 (unstable)	20
Displacement present	86.4	100
Lateral wall intact	18.2	52.6
Comminution present	78.6	50
Reverse obliquity pattern	100	80.8
Subtrochanteric extension	50	88.9
LT fragment present	78.6	50

LT: Lesser trochanter

ChatGPT-5's sensitivity suggests immediate applications as a triage "safety net" in high-volume settings, flagging potential fractures for radiologist review [10, 18]. However, its low specificity necessitates CNN integration to filter false positives. In medical education, its rapid, human-like descriptions (e.g., "suggests varus collapse, likely unstable") could serve as a learning tool for trainees, provided outputs are vetted [20]. Studies have noted LLMs' effectiveness as radiology tutors [20].

Future improvements may involve integrating LLMs with vision models. A pipeline where a CNN identifies fractures and fragments, followed by an LLM generating refined reports, could leverage both strengths. Ongoing research into fine-tuning LLMs on medical image-report pairs aims to improve pixel-to-finding mapping. Confidence thresholding, absent in current LLMs, could reduce false positives by flagging uncertain outputs, an active area in AI safety research. ChatGPT-5's hallucinations and lack of reasoning transparency make standalone use unsafe. False positives/negatives risk patient harm (e.g., unnecessary surgery or missed dislocations), requiring human oversight. ChatGPT-5's kappa (0.08) and unstable bias for hip fractures are inadequate, though its fracture detection sensitivity rivals clinicians. Specificity remains a key shortfall, as humans rarely mistake normal anatomy for fractures.

The study's findings are subject to several limitations. The sample size of 120 cases is modest and may not represent the full spectrum of fracture variations. As data were collected from a single center, the results may not be generalizable to different hospital settings or imaging protocols. The process of coding the AI's free-text output into specific decisions introduced a degree of subjectivity. Furthermore, the evaluation was based on a hypothetical ChatGPT-5 model, and its assumed capabilities may not perfectly align with future LLMs. However, the observed trends are likely applicable to current multimodal GPTs. Finally, this was a retrospective study, meaning it did not assess the practical impact of the AI on clinical workflow, such as diagnostic speed or integration into patient care.

Conclusions

In this study, we found that ChatGPT-5 showed encouraging potential in orthopedic imaging, demonstrating high sensitivity for fracture detection and a remarkable ability to produce fluent, human-like reports. However, the model's performance was compromised by significant limitations, including a lack of specificity and a tendency to hallucinate fractures on normal radiographs. While the high sensitivity suggests a potential role as a triage or educational tool, its errors currently preclude it from serving as an autonomous diagnostic device. Therefore, we conclude that ChatGPT-5 is not yet ready for clinical use in fracture interpretation without the direct supervision of a human expert. Future work should focus on hybrid models that integrate the language capabilities of LLMs with the visual precision of specialized medical imaging AI.

Clinical Message

ChatGPT-5 may support clinicians by highlighting obvious fractures, but its limitations make it unsafe to use as a standalone diagnostic tool. No AI system – particularly general-purpose models – should be entrusted with decisions that could impact a patient's life.

Declaration of patient consent: The authors certify that they have obtained all appropriate patient consent forms. In the form, the patient has given the consent for his/ her images and other clinical information to be reported in the journal. The patient understands that his/ her names and initials will not be published and due efforts will be made to conceal their identity, but anonymity cannot be guaranteed.

Conflict of interest: Nil **Source of support:** None

References

1. Bhatnagar A, Kekatpure AL, Velagala VR, Kekatpure A. A review on the use of artificial intelligence in fracture detection. Cureus 2024;16:e58364.

2. Olczak J, Fahlberg N, Maki A, Razavian AS, Jilert A, Stark A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs. Acta Orthop 2017;88:581-6.

3. Gitto S, Serpi F, Albano D. AI applications in musculoskeletal imaging: A narrative review. Skelet Radiol 2024;53:361-72.

4. Jung J, Dai J, Liu B, Wu Q. Artificial intelligence in fracture detection with different image modalities and data types: A systematic review and meta-analysis. Radiology 2024;3(1):e0000438.

5. Yang Z, Yao Z, Tasmin M, Vashisht P, Jang WS, Ouyang F, et al. Unveiling GPT-4V's hidden challenges behind high accuracy on USMLE questions: Observational Study. J Med Internet Res 2025;27:e65146.

6. Hayden N, Gilbert S, Poisson LM, Griffith B, Klochko C. Performance of GPT-4 with vision on text- and image-based

ACR diagnostic radiology in-training examination questions. Radiology 2024;312:e240153.

7. Carofino BC, Leopold SS. Classifications in brief: The neer classification for proximal humerus fractures. Clin Orthop Relat Res 2013;471:39-43.

8. Lewis J, Macey R, Lewis J, Stokes J, Gill JR, Cook JA, et al. Surgical interventions for treating extracapsular hip fractures in older adults: A network meta-analysis. Cochrane Database Syst Rev 2022;10:CD013405.

9. OpenAI. Introducing GPT-5. OpenAI. Introducing GPT-5. San Francisco, CA: OpenAI; 2025. Available from: https://openai.com/index/introducing-gpt-5 [Last accessed on 2025 Aug 31].

10. Husarek J, Hess S, Razaeian S, Ruder TD, Sehmisch S, Müller M, et al. Artificial intelligence in commercial fracture detection products: A systematic review and meta-analysis of diagnostic test accuracy. Sci Rep 2024;14:23053.

11. Kuo RY, Harrison C, Curran TA, Jones B, Freethy A, Cussons D, et al. Artificial intelligence in fracture detection: A systematic review and meta-analysis. Radiology 2022;304:50-62.

12. Gertz RJ, Dratsch T, Bunck AC, Lennartz S, Iuga AI, Hellmich MG, et al. Potential of GPT-4 for detecting errors in radiology reports: Implications for reporting accuracy. Radiology 2024;311:232714.

13. Keshavarz P, Bagherieh S, Nabipoorashrafi SA, Chalian H, Rahsepar AA, Kim GH, et al. ChatGPT in radiology: A systematic review of performance, pitfalls, and future perspectives. Diagn Interv Imaging 2024;105:251-65.

14. Hirosawa T, Harada Y, Tokumasu K, Ito T, Suzuki T, Shimizu T. Evaluating ChatGPT-4's diagnostic accuracy: Impact of visual data integration. JMIR Med Inform 2024;12:e55627.

15. Chung SW, Han SS, Lee JW, Oh KS, Kim NR, Yoon JP, et al. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. Acta Orthop 2018;89:468-73.

16. Sharma S. Artificial intelligence for fracture diagnosis in orthopedic X-rays: Current developments and future potential. SICOT J 2023;9:21.

17. Liu XS, Nie R, Duan AW, Yang L, Li X, Zhang LT, et al. YOLOX-SwinT algorithm improves the accuracy of AO/OTA classification of intertrochanteric fractures by orthopedic trauma surgeons. Chin J Traumatol 2025;28:69-75.

18. Ueda D, Mitsuyama Y, Takita H, Horiuchi D, Walston SL, Tatekawa H, et al. ChatGPT's diagnostic performance from patient history and imaging findings on the diagnosis please quizzes. Radiology 2023;308:e231040.

19. Jiao W, Wang W, Huang JT, Wang X, Shi S, Tu Z. Is ChatGPT a Good Translator? Yes with GPT-4 as the Engine. Vol. 4. New York: Cornell University; 2023.

20. Arruzza ES, Evangelista CM, Chau M. The performance of ChatGPT-4.0 in medical imaging evaluation: A cross-sectional study. J Educ Eval Health Prof 2024;21:29.

Conflict of Interest: Nil
Source of Support: Nil

Consent: The authors confirm that informed consent was obtained from the patient for publication of this article

How to Cite this Article

ShaII, Majeed IS, Riju R. Performance of ChatGPT-5 in Diagnosing Fractures on Proximal Humerus and Intertrochanteric Femur X-Rays. Journal of Orthopaedic Case Reports 2026 January;16(01):366-373.