

Feasibility of a multimodal large language model in interpreting plain radiographs of bone tumors: a pilot study

Samet Genez*, Hamza Özer

Department of Radiology, Faculty of Medicine, Bolu Abant İzzet Baysal University, Bolu, Türkiye

Received: 05.02.2026

Accepted: 04.03.2026

Published: 09.03.2026

Cite this article: Genez S, Özer H. Feasibility of a multimodal large language model in interpreting plain radiographs of bone tumors: a pilot study. *J Radiol Med.* 2026;3(1):5-8.

*Corresponding Author: Samet Genez, samet.genez@ibu.edu.tr

ABSTRACT

Aims: To evaluate the diagnostic accuracy and clinical feasibility of a multimodal large language model (ChatGPT-5) in interpreting plain radiographs of bone tumors and differentiating between benign and malignant lesions.

Methods: This retrospective pilot study utilized 50 verified bone tumor cases (27 benign and 23 malignant) sourced from the Radiopaedia database. Anonymized radiographs were processed by ChatGPT-5 using a standardized zero-shot prompt in independent sessions to prevent contextual bias. Model performance was assessed based on the accuracy of the most likely diagnosis, the inclusion of correct diagnoses within the top three differentials, and benign–malignant classification metrics. Statistical analysis included the Clopper–Pearson binomial method for confidence intervals and McNemar’s exact test to evaluate improvements in diagnostic accuracy and potential systematic error asymmetry.

Results: The model achieved 100% accuracy in identifying the imaging modality and the affected bone. The accuracy for the single most likely diagnosis was 56.0% (95% CI: 41.3-70.0), which significantly increased to 70.0% (95% CI: 55.4-82.1) when two differential diagnoses were included ($p=0.016$). For benign–malignant classification, the model demonstrated an overall accuracy of 76.0%, with a high specificity of 96.3% but a notably limited sensitivity for malignancy at 52.2%. A statistically significant error asymmetry indicated a systematic tendency toward benign classification ($p=0.006$).

Conclusion: While ChatGPT-5 demonstrates proficiency in foundational radiographic identification, its low sensitivity for malignancy remains a critical limitation for independent clinical use. The results suggest that while multimodal LLMs may serve as promising educational or triage aids, they currently require rigorous human expert oversight to maintain diagnostic safety in image interpretation.

Keywords: Bone tumors, large language models, ChatGPT-5, artificial intelligence, radiology

INTRODUCTION

Multimodal large language models (LLMs) have rapidly moved from text-only assistants to general-purpose systems that can increasingly support clinical tasks.¹⁻³ With the emergence of LLMs capable of processing images, interest has grown regarding their potential role as assistive tools in image interpretation workflows.^{4,5}

Radiographs are the first-line modality for suspected bone tumors and can provide high-yield clues such as lesion location, matrix, pattern of bone destruction, periosteal reaction, and soft-tissue extension;⁶ however, in many cases the most clinically relevant output is a plausible ranked differential diagnosis and correct benign-malignant triage rather than a single definitive histologic label.⁷

Early evaluations of LLMs in radiology largely emphasized text-based applications, including report rewriting and exam-style question answering.⁸⁻¹⁰ As these models have begun to be tested on image-based tasks, performance has appeared more variable, particularly in image-only settings, highlighting the importance of rigorous validation designs and clinically meaningful endpoints.¹¹⁻¹⁴

Despite increasing interest in general-purpose Artificial Intelligence (AI) for radiology, evidence specifically addressing bone tumor differential diagnosis on plain radiographs remains limited. Therefore, this study aimed to evaluate the diagnostic performance of ChatGPT-5 on plain radiographs of bone tumors using open-access cases with reference diagnoses.



METHODS

Ethics

Ethics committee approval was waived as the study was based on anonymized, open-access data from a public database, involving no direct patient contact or identifiable information.

Study Design and Case Selection

This retrospective study was designed to evaluate the diagnostic performance of a large multimodal model in the interpretation of bone tumors on plain radiographs. A total of 50 cases of bone tumors with verified diagnoses were retrospectively identified and sourced from Radiopaedia (<https://radiopaedia.org>), a peer-reviewed open-source radiology database. Cases were selectively curated rather than consecutively sampled to form a pilot set of benign and malignant entities with diagnostic-quality radiographs and a clearly stated reference diagnosis. Rare tumors were not intentionally oversampled, although the educational nature of the source may favor more typical presentations. No formal case-difficulty grading was applied. The cohort was curated to include 27 benign and 23 malignant lesions. Inclusion criteria were: (1) availability of a diagnostic-quality conventional radiograph (X-ray) in at least one standard projection, (2) a definitive reference diagnosis confirmed by expert consensus or histopathology as provided by the database, and (3) clear visualization of the tumorous lesion. Cases with significant hardware, suboptimal positioning, or confounding secondary pathologies (e.g., advanced osteomyelitis mimicking a tumor) were excluded to ensure a focused assessment of tumor-specific features.

Image Acquisition and Reference Standard

For each case, the most representative plain radiograph was exported from the database as an anonymized 8-bit grayscale JPEG file. To ensure the model's assessment was purely image-based, all original annotations, markers, or clinical captions embedded in the images were removed. The reference standard consisted of the verified diagnosis provided by the Radiopaedia platform. For binary performance analysis, each diagnosis was further classified as either benign or malignant.

AI-Based Assessment Using ChatGPT-5

The diagnostic assessment was performed using ChatGPT-5.2 (Pro) (OpenAI, San Francisco, CA), a multimodal large language model, accessed via the web-based interface (<https://chatgpt.com>). All evaluations were conducted in February 2026 using the default vision-capable model settings without any task-specific fine-tuning or prior training.

For each case, the anonymized radiograph was uploaded into a new, independent chat session to prevent carryover effects or contextual bias from previous evaluations. A standardized zero-shot prompt was entered for every case:

“Interpret the provided plain X-ray image of a suspected bone tumor. Provide a brief radiographic assessment, then state the most likely diagnosis and 2 differential diagnoses”

The standardized prompting workflow is illustrated in **Figure**. The model was not provided with any patient demographics, clinical history, or laboratory findings; thus, this evaluation reflects an image-only scenario rather than real-world clinical

decision-making. The model's free-text responses, including its primary diagnosis and any mentioned differential diagnoses, were recorded for subsequent analysis.



Interpret the provided plain X-ray image of a suspected bone tumor. Provide a brief radiographic assessment, then state the most likely diagnosis and 2 differential diagnosis

Radiographic assessment (plain AP knee X-ray):

- Distal femoral metaphysis, eccentric cortical-based multilobulated lucent lesion along the lateral cortex
- Well-defined margins with a thin sclerotic rim, no cortical breakthrough
- No aggressive periosteal reaction, no soft-tissue mass, joint preserved

Most likely diagnosis:

Fibrous cortical defect / Non-ossifying fibroma (NOF spectrum lesion)

Differential diagnoses:

1. Osteofibrous dysplasia
2. Brodie abscess (subacute osteomyelitis)

Why NOF fits best:

Metaphyseal, eccentric, cortically based, geographic lytic lesion with sclerotic border and non-aggressive appearance in a long bone around the knee is classic — essentially a textbook benign developmental lesion.

📄 📄 📄 📄 📄 ...

Figure. Standardized prompt-and-response workflow used for ChatGPT-5 interpretation of bone tumor radiographs

Performance Metrics and Outcome Assessment

Model performance was assessed based on the accuracy of the most likely diagnosis, differential diagnosis (two differential diagnoses), and benign–malignant classification. Diagnostic correctness was assessed at the general diagnosis level; subtype-qualified labels (e.g., ‘conventional osteosarcoma’) were considered correct if they referred to the same underlying diagnostic entity (e.g., ‘osteosarcoma’). Additionally, the model's ability to correctly identify the imaging modality and the specific bone involved was recorded for each case.

Statistical Analysis

The data analyses were performed using R software (version 4.3.1). The primary endpoint was the accuracy of the initial most likely diagnosis. Proportions were reported as n/N with 95% confidence intervals (CIs) calculated using the exact (Clopper-Pearson) binomial method.

To assess the clinical value of differential diagnoses, the improvement from the initial diagnosis accuracy to the differential diagnosis was evaluated using McNemar's exact test. For benign–malignant classification, a 2×2 confusion matrix was used to calculate sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).

Balanced accuracy was reported as a point estimate (mean of sensitivity and specificity); 95% CIs were provided for sensitivity and specificity only. Potential systematic bias (e.g., a tendency toward benign classification) was evaluated by analyzing error asymmetry (false negatives vs. false positives) using McNemar's test. A p-value <0.05 was considered statistically significant.

RESULTS

A total of 50 plain radiographs of bone tumors were evaluated (reference labels: 23 malignant and 27 benign lesions). ChatGPT-5 provided responses for all cases, with no response refusals observed (0/50; 0.0% [95% CI: 0.0-7.1]). Additionally, the model correctly labeled the radiograph modality and identified the affected bone in 100% of cases (50/50; 95% CI: 92.9-100). Overall diagnostic accuracy—defined as an exact match between the model's single most likely diagnosis and the reference diagnosis—was 56.0% (28/50; 95% CI: 41.3-70.0). When the model was allowed to provide up to two differential diagnoses, the correct-diagnosis rate increased to 70.0% (35/50; 95% CI: 55.4-82.1), when the reference diagnosis was present in any of the three reported diagnoses (primary+two differentials) representing a statistically significant improvement over single-diagnosis performance (McNemar's exact test, $p=0.016$).

For benign–malignant classification, overall accuracy was 76.0% (38/50; 95% CI: 61.8-86.9; **Table 1**). Sensitivity for malignancy was 52.2% (12/23; 95% CI: 30.6-73.2) and specificity was 96.3% (26/27; 95% CI: 81.0-99.9). The positive predictive value was 92.3% (12/13; 95% CI: 64.0-99.8) and the negative predictive value was 70.3% (26/37; 95% CI: 53.0-84.1), yielding a balanced accuracy of 74.3% (**Table 2**). The model predicted malignancy less frequently than the reference prevalence (26.0% vs 46.0%), suggesting a tendency toward benign classification; this error asymmetry was supported by McNemar's exact test ($p=0.006$; **Table 2**).

Table 1. Confusion matrix summarizing the multimodal LLM's benign-malignant predictions for bone tumor radiographs

	Reference malignant (n=23)	Reference benign (n=27)
Predicted malignant (n=13)	TP=12	FP=1
Predicted benign (n=37)	FN=11	TN=26

LLM: Large language model, TP: True positive, FP: False positive, FN: False negative, TN: True negative

Table 2. Performance metrics for benign-malignant classification of bone tumors on plain radiographs by the multimodal LLM

Metric	Estimate	95% CI (exact)
Accuracy	76.0% (38/50)	61.8-86.9
Sensitivity	52.2% (12/23)	30.6-73.2
Specificity	96.3% (26/27)	81.0-99.9
PPV	92.3% (12/13)	64.0-99.8
NPV	70.3% (26/37)	53.0-84.1
Balanced accuracy	74.3%	—

LLM: Large language model, CI: Confidence interval, PPV: Positive predictive value, NPV: Negative predictive value

DISCUSSION

This pilot study provides critical insights into the diagnostic capabilities of ChatGPT-5 in the radiographic interpretation

of bone tumors. Our results demonstrate that while the model excels in basic descriptive tasks such as identifying imaging modalities and anatomical locations, its performance in definitive diagnostic reasoning and malignancy detection remains limited.

Consistent with prior literature, ChatGPT-5 achieved 100% accuracy in identifying the imaging modality and the affected bone. This aligns closely with the work of Hiredesai et al.,¹⁵ who reported a 98.9% accuracy rate for modality identification in upper extremity pathologies. Such high performance suggests that LLMs have successfully integrated foundational medical imaging characteristics into their visual processing frameworks. However, a significant gap exists between “recognizing” a radiograph and “interpreting” its complex pathological features.¹⁶

In our study, ChatGPT-5 reached a diagnostic accuracy of 56% for the most likely diagnosis, which increased to 70% when considering the top two differential diagnoses. This is notably higher than the 19.90% “image-only” accuracy reported by Atakır et al.¹⁷ in a general radiology dataset. This discrepancy may be attributed to the highly structured nature of bone tumor patterns—such as lesional margins and matrix mineralization—which may be more conducive to the model's pattern recognition than the heterogeneous cases used in broader studies. Nevertheless, a 56% diagnostic accuracy is unlikely to support unsupervised clinical use, suggesting that LLMs currently serve best as assistive tools under radiologist oversight.^{16,17}

A critical finding of our study is a tendency toward benign classification, with a sensitivity of only 52.2% for malignancy. This benign-leaning asymmetry ($p=0.006$) is clinically consequential, as missed malignant bone tumors may delay appropriate referral and definitive management. In an image-only workflow, these results suggest that a general-purpose LLMs may have limited sensitivity for malignancy, reinforcing the need for radiologist oversight and task-specific validation before clinical deployment.¹⁶ Furthermore, research by Atakır et al.¹⁷ suggests that LLMs are heavily text-biased; their diagnostic consistency and accuracy improve dramatically when textual clinical context is provided, yet the addition of the image itself often yields diminishing marginal returns.

The limitations observed in ChatGPT-5's visual reasoning are likely due to its architecture, which was primarily optimized for natural language processing rather than specialized medical computer vision.¹⁸ While the latest iterations show “numerical improvements” over previous versions in medical assessments, they still exhibit hallucinations and struggle with subtle findings that human experts easily identify.⁴ For instance, Hiredesai et al.¹⁵ noted that ChatGPT 4.0 often declined to provide a diagnosis or provided generalized information instead of specific findings.

Limitations

This study has several limitations. First, the use of static, single-view radiographs does not reflect the dynamic nature of clinical practice, where radiologists often utilize multiple views and longitudinal data. Second, cases were drawn from an open-access educational repository and

were curated primarily for teaching purposes, which may introduce selection/spectrum bias toward well-demonstrated archetypal examples and may not reflect the prevalence or full spectrum of presentations in the general clinical population. Third, our evaluation focused on a “zero-shot” prompting approach, prompt engineering and optimized instructions can significantly reduce information loss and improve the clarity of AI outputs. Future research should explore “few-shot” learning or the integration of LLMs with specialized Convolutional Neural Networks (CNNs), which have historically outperformed general-purpose LLMs in fracture and tumor detection.

CONCLUSION

ChatGPT-5 demonstrates a promising but still insufficient capacity for the independent interpretation of bone tumor radiographs. While it can serve as a valuable triage or educational aid, potentially enhancing the performance of residents in structured scenarios-its low sensitivity for malignancy underscores the indispensable role of human expert oversight. As these models evolve, their integration into radiology must be governed by rigorous validation to ensure they supplement, rather than compromise, diagnostic safety.

ETHICAL DECLARATIONS

Ethics Committee Approval

Since the study involved no human or animal subjects, clinical interventions, or identifiable patient data, ethics committee approval was not required.

Informed Consent

Since the study did not involve human or animal subjects, clinical interventions, or identifiable patient data, informed consent was not required.

Peer Review Process

This manuscript was subject to external peer review.

Conflict of Interest

The authors declare no conflicts of interest related to this study.

Financial Disclosure

The authors received no financial support for the conduct or publication of this research.

Author Contributions

Project Conceptualization and Study Design: SG; Data Collection: SG; Statistical Analysis: SG, HÖ; Manuscript Drafting: SG, HÖ; Review of the Final Manuscript Submitted for Publication: SG, HÖ.

REFERENCES

1. Lecler A, Duron L, Soyer P. Revolutionizing radiology with GPT-based models: current applications, future possibilities and limitations of ChatGPT. *Diagn Interv Imaging*. 2023;104(6):269-274. doi:10.1016/j.diii.2023.02.003
2. Genez S, Özer H, Buz Yaşar A, et al. Evaluation of ChatGPT-5 for automated ASPECTS assessment on non-contrast CT in acute ischemic stroke. *Diagnostics (Basel)*. 2025;15(24):3160. doi:10.3390/diagnostics15243160
3. Nguyen D, Rao A, Mazumder A, Succi MD. Exploring the accuracy of embedded ChatGPT-4 and ChatGPT-4o in generating BI-RADS scores: a pilot study in radiologic clinical support. *Clin Imaging*. 2025; 117:110335. doi:10.1016/j.clinimag.2024.110335
4. Wu Z, Li S, Zhao X. The application of ChatGPT in medical education: prospects and challenges. *Int J Surg*. 2025;111(1):1652-1653. doi:10.1097/JS9.0000000000001887
5. Liu M, Okuhara T, Chang X, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. *J Med Internet Res*. 2024; 26:e60807. doi:10.2196/60807
6. Costelloe CM, Madewell JE. Radiography in the initial diagnosis of primary bone tumors. *AJR Am J Roentgenol*. 2013;200(1):3-7. doi:10.2214/AJR.12.8488
7. Priolo F, Cerase A. The current role of radiography in the assessment of skeletal tumors and tumor-like lesions. *Eur J Radiol*. 1998;27 Suppl 1: S77-S85. doi:10.1016/s0720-048x(98)00047-3.
8. Kitamura FC. ChatGPT Is shaping the future of medical writing but still requires human judgment. *Radiology*. 2023;307(2):e230171. doi:10.1148/radiol.230171
9. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198
10. Liao W, Liu Z, Dai H, et al. Differentiating ChatGPT-generated and human-written medical texts: quantitative study. *JMIR Med Educ*. 2023;9:e48904. doi:10.2196/48904
11. 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(3):e240153. doi:10.1148/radiol.240153
12. Handa P, Chhabra D, Goel N, Krishnan S. Exploring the role of ChatGPT in medical image analysis. *Biomed Signal Process Control*. 2023;86:105292. doi:10.1016/j.bspc.2023.105292.
13. Dehdab R, Brendlin A, Werner S, et al. Evaluating ChatGPT-4V in chest CT diagnostics: a critical image interpretation assessment. *Jpn J Radiol*. 2024;42(10):1168-1177. doi:10.1007/s11604-024-01606-3
14. Lacaita PG, Galijasevic M, Swoboda M, et al. The accuracy of ChatGPT-4o in interpreting chest and abdominal X-Ray images. *J Pers Med*. 2025; 15(5):194. doi:10.3390/jpm15050194
15. Hiredesai AN, Martinez CJ, Anderson ML, Howlett CP, Unadkat KD, Noland SS. Is artificial intelligence the future of radiology? Accuracy of ChatGPT in radiologic diagnosis of upper extremity bony pathology. *Hand (N Y)*. 2026;21(1):73-80. doi:10.1177/15589447241298982
16. Yang X, Chen W. The performance of ChatGPT on medical image-based assessments and implications for medical education. *BMC Med Educ*. 2025;25(1):1192. doi:10.1186/s12909-025-07752-0
17. Atakır K, Işın K, Taş A, Önder H. Diagnostic accuracy and consistency of ChatGPT-4o in radiology: influence of image, clinical data, and answer options on performance. *Diagn Interv Radiol*. 2025. doi:10.4274/dir.2025.253460
18. Lyu Q, Tan J, Zapadka ME, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. *Vis Comput Ind Biomed Art*. 2023; 6(1):9. doi:10.1186/s42492-023-00136-5