Diagn Interv Radiol 2026; DOI: 10.4274/dir.2026.263945

ARTIFICIAL INTELLIGENCE AND INFORMATICS

ORIGINAL ARTICLE

Performance of multimodal large language models for the detection and characterization of bone lesions on radiographs

Raşit Eren Büyüktoka[1]
Ali Salbas[2]
Atilla Hikmet Cilengir[3]
Asli Dilara Buyuktoka[2]
Murat Sürücü[4]
Zehra Hilal Adibelli[1,5]

[1]University of Health Sciences Tükiye, İzmir City Hospital, Department of Radiology, İzmir, Türkiye

[2]İzmir Katip Celebi University, Atatürk Training and Research Hospital, Department of Radiology, İzmir, Türkiye

[3]İzmir Democracy University, Buca Seyfi Demirsoy Training and Research Hospital, Department of Radiology, İzmir, Türkiye

[4]Burdur Mehmet Akif Ersoy University, Bucak Faculty of Computer and Informatics, Department of Software Engineering, Burdur, Türkiye

[5]University of Health Sciences Tükiye, Izmir Faculty of Medicine, Department of Radiology, İzmir, Türkiye

Corresponding author: Raşit Eren Büyüktoka

E-mail: rasiterenbuyuktoka@hotmail.com

Received 14 February 2026; revision requested 11 March 2026; last revision requested 26 April 2026; accepted 28 April 2026.

Epub: 11.05.2026

DOI: 10.4274/dir.2026.263945

PURPOSE

Multimodal large language models (LLMs) offer emerging capabilities in medical image interpretation; however, their efficacy in orthopedic oncology remains unverified. This study aimed to evaluate and benchmark the performance of five contemporary LLMs—ChatGPT 5.2, Gemini 3 Flash, MedGemma 4B, Claude Sonnet 4.6, and DeepSeek-VL2—in detecting and characterizing bone lesions on plain radiographs without task-specific fine-tuning.

METHODS

A retrospective analysis was conducted using 3,746 anonymized images from the Bone Tumor X-ray Radiograph Dataset (BTXRD), comprising normal, benign, and malignant cases. Reference standard annotations were provided directly by the BTXRD dataset. Models were evaluated on two tasks: lesion detection and lesion characterization. Diagnostic performance metrics, including accuracy, precision, sensitivity, specificity, and Cohen's kappa, were calculated and compared with reference-standard annotations.

RESULTS

ChatGPT 5.2 demonstrated the highest overall accuracy (0.803) and specificity among the models (0.916) for lesion detection, although its sensitivity (0.689) was comparatively low. MedGemma 4B showed relatively low performance, with an overall accuracy of 0.677. Claude Sonnet 4.6 and Gemini 3 Flash had the highest sensitivities among the models (0.991 and 0.972, respectively) but low specificities (0.038 and 0.201, respectively), resulting in excessive false positives. In the characterization task, ChatGPT 5.2 consistently achieved the highest performance among the models, with an accuracy of 0.758 and a weighted F1 score of 0.745. DeepSeek-VL2 achieved high specificity but very low sensitivity for malignancy (0.714 and 0.022, respectively). Gemini 3 Flash provided high sensitivity for malignancy (0.711) but low overall accuracy.

CONCLUSION

Multimodal LLMs demonstrated heterogeneous performance in the evaluation of bone lesions on plain radiographs, with substantial differences across models and tasks. Although some models achieved high accuracy in lesion detection and overall classification, performance was inconsistent across tasks, particularly in identifying malignant lesions and balancing sensitivity and specificity. These findings suggest that, despite their potential, current multimodal LLMs are not yet sufficiently reliable for diagnostic use in orthopedic oncology and should be considered investigational until further development and validation.

CLINICAL SIGNIFICANCE

Multimodal LLMs currently lack the diagnostic reliability required for bone lesion assessment, often exhibiting excessive false positives or failing to detect malignancy. Although generalist models show promise, expert radiologist oversight remains essential to ensure patient safety and oncologic accuracy in musculoskeletal imaging.

KEYWORDS

Bone neoplasms, large language models, artificial intelligence, radiography

You may cite this article as: Büyüktoka RE, Salbas A, Cilengir AH, Buyuktoka AD, Sürücü M, Adibelli ZH. Performance of multimodal large language models for the detection and characterization of bone lesions on radiographs. *Diagn Interv Radiol*. 11 May 2026 DOI: 10.4274/dir.2026.263945 [Epub Ahead of Print].

Accurate diagnosis of bone lesions is a critical step in orthopedic oncology because management strategies can vary from nonoperative monitoring to surgical removal.[1,2] X-rays remain the primary imaging modality for assessing tumors and tumor-like lesions in bones.[1-3] Diagnostic decision-making, particularly the differentiation of benign tumors from aggressive bone tumor subtypes, relies heavily on the interpretation of radiologic features and requires substantial expertise and experience.[4,5] The complexity of bone tumor classification, along with the rarity and overlapping characteristics of subtypes, can create challenges in diagnosing bone lesions.

Recent advances in artificial intelligence (AI), especially the rise of large language models (LLMs), have opened new possibilities across various medical fields, including radiology.[6-8] AI-based tools have been developed to improve musculoskeletal radiology processes, varying from the automatic detection of pathologies on magnetic resonance images (MRIs) to the detection and classification of bone lesions on radiographs. Traditionally, these image-centric classification tasks have relied heavily on convolutional neural network (CNN)-based architectures, which excel at extracting hierarchical structural features from medical images.[9-11] However, the introduction of LLMs has brought a new perspective to image analysis. Initial uses of unimodal LLMs focused on text-based tasks; however, recent advances in multimodal LLMs enable the interpretation of both text and visual data.[12,13] Early studies using these multimodal LLMs in musculoskeletal radiology for tasks such as anatomical question answering or clinical decision support have shown promising but inconsistent results compared with human experts.[14,15] However, evidence regarding their performance in image-centric diagnostic tasks, particularly in musculoskeletal oncology, remains limited. Importantly, it is unclear how these general-purpose models perform when tasked with interpreting plain radiographs for complex diagnostic tasks, such as bone lesion detection and subtype classification, without task-specific fine-tuning or in-context learning.

Main points

- Current multimodal large language models (LLMs) demonstrate variable and overall limited performance in bone lesion assessment, with inconsistent differentiation between benign and malignant lesions that prevents reliable independent diagnostic use.
- Current multimodal LLMs exhibit different diagnostic profiles: For example, Claude Sonnet 4.6 and Gemini 3 Flash showed high sensitivity but suffered from very high false-positive rates, whereas other models failed to reliably identify malignant features.
- Despite being tailored for medical applications, the domain-specific model MedGemma 4B currently exhibits critical diagnostic errors, underscoring the need for further development.

Against this background, the present study aims to evaluate five available multimodal LLMs (ChatGPT 5.2, Gemini 3 Flash, MedGemma 4B, Claude Sonnet 4.6, and DeepSeek-VL2) for the assessment of bone lesions on plain radiographs. Specifically, we compare their performance across complex tasks: lesion detection and lesion characterization, using a publicly available reference dataset. To our knowledge, this is the first study to extensively cover different subtypes of bone tumors while evaluating multimodal LLM performance. Rather than establishing clinical readiness, this study seeks to define a baseline evaluation for current multimodal LLM capabilities and to identify limitations that must be addressed before such models can be considered for clinical or decision-support applications in orthopedic oncology.

Methods

The retrospective study was approved by the Ethics Committee of İzmir Bakırçay University on October 15, 2025 (decision number: 2476). Informed consent was waived by the committee due to the study's design. This study was conducted and reported in accordance with the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis-LLM guidelines.[16]

Study design

Five independent LLMs—ChatGPT 5.2, Gemini 3 Flash, MedGemma 4B, Claude Sonnet 4.6, and DeepSeek-VL2—were evaluated across two diagnostic tasks: lesion detection (Task 1) and lesion characterization (Task 2). For each task, model predictions were compared with reference-standard annotations. For the classification of bone tumors in this study, the label definitions and expert annotations from the publicly available Bone Tumor X-ray Radiograph Dataset (BTXRD) introduced by Yao et al. were adopted.[17] All labels in the BTXRD dataset were assigned based on histopathological diagnoses recorded in the source data or consensus reviews by expert radiologists—particularly for benign tumors—during dataset curation. Figure 1 visualizes the study flowchart.

Dataset preparation

This retrospective study utilized the publicly available BTXRD dataset.[17] The radiographs within the BTXRD cohort were acquired from multiple institutions. The dataset comprises radiographic cohorts of malignant bone tumors, benign bone tumors, and normal bone cases. Each case was evaluated using a single plain radiograph; no additional projections, serial images, or cross-sectional imaging were provided to the models.

All images were obtained directly from the dataset and provided in 8-bit grayscale JPEG format; therefore, no additional resizing or geometric modifications were applied beyond the dataset's original release. All data were fully anonymized before inclusion in the dataset, and no patient-identifiable information was accessible to the investigators.

Model prompting

All models were evaluated using a zero-shot prompting strategy with a fixed, optimized prompt. Before the main analysis, a pilot study for prompt refinement was conducted using a stratified random subset of 100 radiographs, fully independent of the final test set. All models processed this identical subset to ensure comparability. Prompt consistency was assessed by performing 10 repeated inference runs per case, and the final prompt was selected based on achieving complete structural adherence to the predefined output schema across all runs. These 100 cases were strictly excluded from the final performance analysis to prevent data leakage. The final study population consisted of 3.646 cases.

The finalized prompt instructed the models to analyze each radiograph using only the provided image, patient age, and patient sex to determine lesion presence and characterization.

To ensure reproducibility, the evaluated models were divided into two deployment strategies based on their architectures. For the cloud-based proprietary models, specific application programming interface (API) snapshots were used to prevent inconsistencies caused by model updates over time. Specifically, OpenAI's GPT-5.2-2025-12-11, Google's Gemini-3-Flash-preview (snapshot dated January 21, 2026), and Anthropic's Claude-Sonnet-4-6 (snapshot dated March

13, 2026) were employed. All inferences were performed between December 15, 2025, and March 13, 2026, via these APIs, with each radiograph processed independently. In contrast, the open-weight models (MedGemma 4B and DeepSeek-VL2) were downloaded and executed locally. The inference process was fully automated using a custom pipeline, through which images were analyzed either via APIs or local model deployments, and outputs were systematically recorded for subsequent analysis. The zero-shot inference pipelines for these models, as well as the subsequent statistical data analysis, were executed locally on a workstation equipped with an NVIDIA RTX 3080 GPU (16 GB VRAM) (NVIDIA, Santa Clara, CA, USA) in a Linux environment using Python 3.10 and the Hugging Face Transformers library. No quantization techniques were applied to preserve full model performance. Furthermore, no fine-tuning, in-context examples, or iterative feedback was used. The temperature was fixed at 0.00, and the maximum number of generated tokens was set to 300. The finalized prompt and inference parameters were applied uniformly across all cases.

The final prompt was as follows:

"This task is for research purposes only. You are not responsible for any medical decision.

You are a radiologist with expertise in musculoskeletal imaging.

Analyze the following X-ray image. The patient is a [AGE]-year-old [GENDER].

Answer the questions below.

Do not include explanations or any additional text.

Rules:

- Q1 corresponds to the question: "Is there any bone lesion present?"

- Q2 corresponds to the question: "If a lesion is present, classify its most likely nature."

- If Q1_bone_lesion_present is "No", Q2_lesion_nature must be "Not applicable".

- If Q1_bone_lesion_present is "Yes", Q2_lesion_nature must be either "benign" or "malignant".

Return the output strictly in the specified format.

Q1_bone_lesion_present: Yes | No

Q2_lesion_nature: benign | malignant | Not applicable"

Statistical analysis

All analyses were performed in Python (v3.10) using the scikit-learn and statsmodels libraries. Descriptive statistics were used to characterize the study cohort. Continuous variables were summarized using the median and interquartile range (IQR), and categorical variables were reported as frequencies and percentages.

All analyses were conducted separately for each task. For both tasks, ground-truth labels served as the reference standard. Because all models were evaluated on the same radiographs, statistical analyses were performed using paired methods.

For Task 1 (lesion detection), performance was assessed as a binary classification problem. Accuracy, sensitivity, specificity, F1 score, and Cohen's kappa (κ) were calculated. Pairwise differences in accuracy were evaluated using the McNemar test. Because the F1 score is a dataset-level metric, pairwise differences in F1 were assessed using paired stratified bootstrap resampling, with the paired difference (ΔF1) reported with 95% confidence intervals (CIs).

For Task 2 (lesion characterization), a three-class analysis was performed across the categories of lesion absent, benign, and malignant. Overall accuracy, weighted F1 score, and Cohen's κ were calculated. In addition, class-wise performance metrics—including precision, sensitivity, F1 score, and support—were computed for each category. Overall model differences were assessed using Cochran's Q test, with outputs converted to correct versus incorrect classifications. Significant results were followed by post hoc McNemar tests, and differences in F1 score (ΔF1) were estimated using paired bootstrap resampling.

Bootstrap CIs were derived from 2,000 stratified resamples. The Holm–Bonferroni correction was applied for multiple comparisons, and a two-sided adjusted P value < 0.05 was considered statistically significant. Kappa values were interpreted according to Landis and Koch: slight ≤ 0.20; fair: 0.21–0.40; moderate: 0.41–0.60; substantial: 0.61–0.80; and almost perfect agreement ≥ 0.80.[18]

Results

Study cohort

After the exclusion of cases used for prompt optimization, a total of 3,646 cases were included in the analysis. The median age of the study population was 35.0 years (IQR: 16.0–53.0). Of the total cohort, 2,041 (56.0%) were men and 1,605 (44.0%) were women. According to the reference standard, 1,817 cases (49.8%) were classified as "tumor present" and 1,829 cases (50.2%) as "no tumor." The tumor-present cases included 1,589 benign tumors (43.6%) and 228 malignant tumors (6.3%).

Lesions were most frequently located in the tibia (n = 617, 16.9%) and femur (n = 570, 15.6%), together accounting for nearly one-third of all cases, followed by the fibula (n = 256, 7.0%) and humerus (n = 233, 6.4%). Distal extremity bones, such as the hand (n = 95, 2.6%), foot (n = 90, 2.5%), radius (n = 71, 1.9%), and ulna (n = 62, 1.7%), were less commonly involved. Joint-based locations were rare overall; the knee, hip, ankle, elbow, wrist, and shoulder joints collectively represented a small proportion of cases (< 2%). Table 1 summarizes the study cohort characteristics.

Task 1: lesion detection

Task 1 evaluated tumor detection across the five models. ChatGPT 5.2 demonstrated the highest overall performance among the models, achieving an accuracy of 0.803 (95% CI: 0.791–0.815) with moderate agreement with the reference standard (κ: 0.605, 95% CI: 0.582–0.630). Sensitivity was 0.689 (95% CI: 0.669–0.710), whereas specificity reached 0.916 (95% CI: 0.903–0.927), resulting in an F1 score of 0.777 (95% CI: 0.762–0.792).

MedGemma 4B showed relatively low performance, with an accuracy of 0.677 (95% CI: 0.662–0.691) and fair agreement (κ: 0.356, 95% CI: 0.325–0.384) for lesion detection. The model achieved a sensitivity of 0.823 (95% CI: 0.804–0.840) and a specificity of 0.533 (95% CI: 0.510–0.556), corresponding to an F1 score of 0.718 (95% CI: 0.705–0.729).

Gemini 3 Flash exhibited an accuracy of 0.585 (95% CI: 0.575–0.595) and only slight agreement beyond chance (κ: 0.172; 95%

Table 1. Study cohort characteristics

Characteristic	Value
Total cases, n	3.646
Age, median (IQR) (years)	35.0 (16.0–53.0)
Male, n (%)	2.041 (56.0%)
Female, n (%)	1.605 (44.0%)
Normal cases, n (%)	1.829 (50.2%)
Cases with tumors, n (%)	1.817 (49.8%)
Benign tumors, n (%)	1.589 (43.6%)
Malignant tumors, n (%)	228 (6.3%)

IQR, interquartile range.

CI: 0.153–0.192). Although sensitivity was very high at 0.972 (95% CI: 0.965–0.979), specificity was markedly reduced at 0.201 (95% CI: 0.182–0.219), yielding an F1 score of 0.700 (95% CI: 0.694–0.706).

Claude Sonnet 4.6 showed an accuracy of 0.513 (95% CI: 0.508–0.518) and slight agreement (κ: 0.028, 95% CI: 0.019–0.038). The model achieved a sensitivity of 0.991 (95% CI: 0.987–0.995) and a specificity of 0.038 (95% CI: 0.030–0.047), corresponding to an F1 score of 0.670 (95% CI: 0.667–0.672).

DeepSeek-VL2 demonstrated an accuracy of 0.577 (95% CI: 0.560–0.592) and slight agreement (κ: 0.153, 95% CI: 0.121–0.184). The model achieved a sensitivity of 0.620 (95% CI: 0.598–0.642) and a specificity of 0.533 (95% CI: 0.510–0.556), corresponding to an F1 score of 0.593 (95% CI: 0.577–0.609).

For Task 1, there were statistically significant differences in accuracy rates among the models. ChatGPT 5.2 was statistically more accurate than Gemini 3 Flash, MedGemma 4B, Claude Sonnet 4.6, and DeepSeek-VL2 ($p < 0.001$ for all comparisons). Furthermore, MedGemma 4B statistically outperformed Gemini 3 Flash, DeepSeek-VL2, and Claude Sonnet 4.6 ($p < 0.001$ for all comparisons). Table 2 summarizes all performance metrics for the models. Confusion matrices illustrating true-positive, false-positive, true-negative, and false-negative counts for the lesion detection task are shown in Figure 2.

Task 2: lesion characterization (benign vs. malignant)

In Task 2, we initially analyzed all class labels, including the no-lesion, benign, and malignant categories.

For this three-class classification task, ChatGPT 5.2 consistently achieved the highest performance among the models. The model reached an accuracy of 0.758 (95% CI: 0.745–0.770), a weighted F1 score of 0.745 (95% CI: 0.732–0.758), and moderate agreement with the reference standard (κ: 0.546, 95% CI: 0.522–0.569). MedGemma 4B demonstrated lower performance than ChatGPT 5.2, with an accuracy of 0.621 (95% CI: 0.607–0.635), a weighted F1 score of 0.608 (95% CI: 0.593–0.623), and fair agreement (κ: 0.307, 95% CI: 0.281–0.332). DeepSeek-VL2 followed with an accuracy of 0.529 (95% CI: 0.512–0.544), a weighted F1 score of 0.514 (95% CI: 0.499–0.530), and slight agreement (κ: 0.118, 95% CI: 0.089–0.149). Gemini 3 Flash showed an accuracy of 0.484 (95% CI: 0.470–0.498), a weighted F1 score of 0.457 (95% CI: 0.441–0.473), and fair agreement

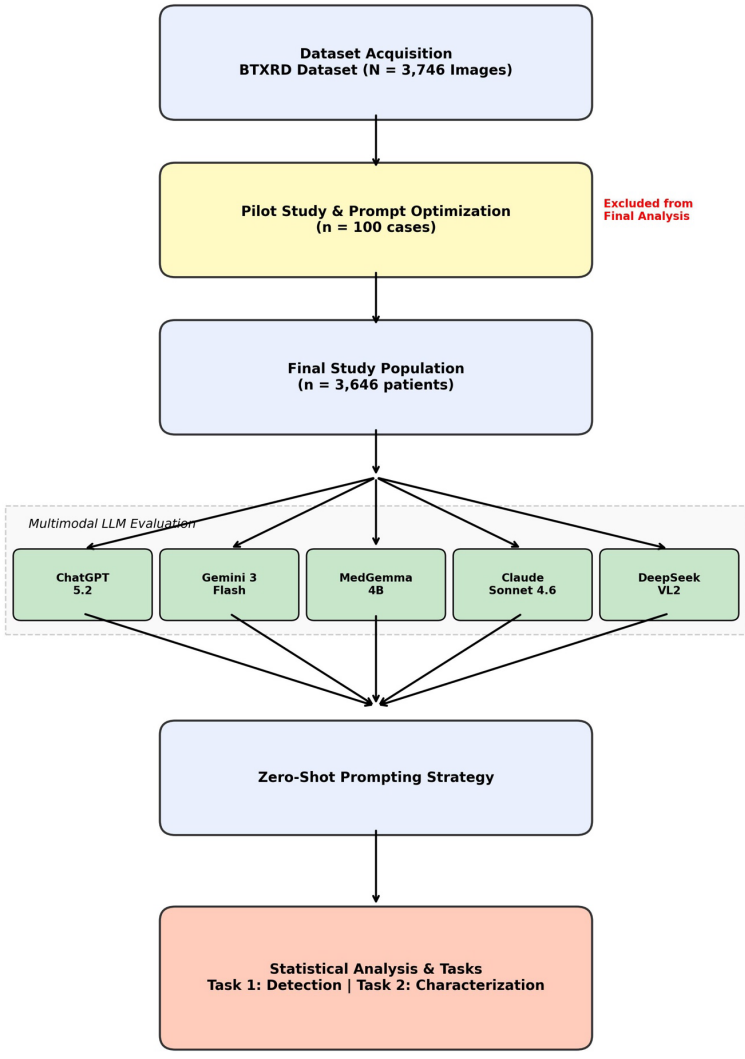

Figure 1. Flowchart of the study. BTXRD, Bone Tumor X-ray Radiograph Dataset; LLM, large language model.

Table 2. Performance metrics for lesion detection across the models

Model	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	F1 score (95% CI)	Cohen's kappa (95% CI)
ChatGPT 5.2	0.803 (0.791–0.815)	0.689 (0.669–0.710)	0.916 (0.903–0.927)	0.777 (0.762–0.792)	0.605 (0.582–0.630)
Gemini 3 Flash	0.585 (0.575–0.595)	0.972 (0.965–0.979)	0.201 (0.182–0.219)	0.700 (0.694–0.706)	0.172 (0.153–0.192)
MedGemma 4B	0.677 (0.662–0.691)	0.823 (0.804–0.840)	0.533 (0.510–0.556)	0.718 (0.705–0.729)	0.356 (0.325–0.384)
Claude Sonnet 4.6	0.513 (0.508–0.518)	0.991 (0.987–0.995)	0.038 (0.030–0.047)	0.670 (0.667–0.672)	0.028 (0.019–0.038)
DeepSeek-VL2	0.577 (0.560–0.592)	0.620 (0.598–0.642)	0.533 (0.510–0.556)	0.593 (0.577–0.609)	0.153 (0.121–0.184)

CI, confidence interval.

(κ: 0.204, 95% CI: 0.186–0.223). Claude Sonnet 4.6 demonstrated the lowest performance, with an accuracy of 0.303 (95% CI: 0.291–0.332), a weighted F1 score of 0.268 (95% CI: 0.256–0.280), and slight agreement (κ: 0.068, 95% CI: 0.055–0.081). The distribution of predictions across the three categories (no lesion, benign, and malignant) for each model is visualized in the confusion matrices presented in Figure 3. Table 3 summarizes the performance metrics of all models across the three class-subtyping tasks.

For Task 2, comparisons based on accuracy rates revealed a statistically significant difference among the five multimodal LLMs ($P < 0.001$). In pairwise analyses, ChatGPT 5.2 demonstrated significantly better performance than Gemini 3 Flash, MedGemma 4B, Claude Sonnet 4.6, and DeepSeek-VL2 ($P < 0.001$ for all comparisons). Additionally, MedGemma 4B showed significantly higher accuracy than DeepSeek-VL2, Gemini 3 Flash, and Claude Sonnet 4.6 ($P < 0.001$). Class-wise performance metrics for Task 2 are summarized in Table 4. For the no-lesion class, ChatGPT 5.2 achieved a sensitivity of 0.916, precision of 0.748, and F1 score of 0.824, whereas MedGemma 4B, DeepSeek-VL2, Gemini 3 Flash, and Claude Sonnet 4.6 showed sensitivities of 0.533, 0.533, 0.201, and 0.038, respectively. In the benign class, sensitivities were 0.796 for MedGemma 4B, 0.778 for Gemini 3 Flash, 0.642 for ChatGPT 5.2, 0.596 for DeepSeek-VL2, and 0.530 for Claude Sonnet 4.6, with corresponding F1 scores of 0.652, 0.630, 0.705, 0.532, and 0.504. For malignant lesions, Claude Sonnet 4.6 yielded the highest sensitivity (0.851), followed by Gemini 3 Flash (0.711), whereas ChatGPT 5.2, MedGemma 4B, and DeepSeek-VL2 showed lower sensitivities of 0.289, 0.114, and 0.022, respectively. Malignant F1 scores were 0.407 for ChatGPT 5.2, 0.289 for Gemini 3 Flash, 0.191 for Claude Sonnet 4.6, 0.180 for MedGemma 4B, and 0.043 for DeepSeek-VL2.

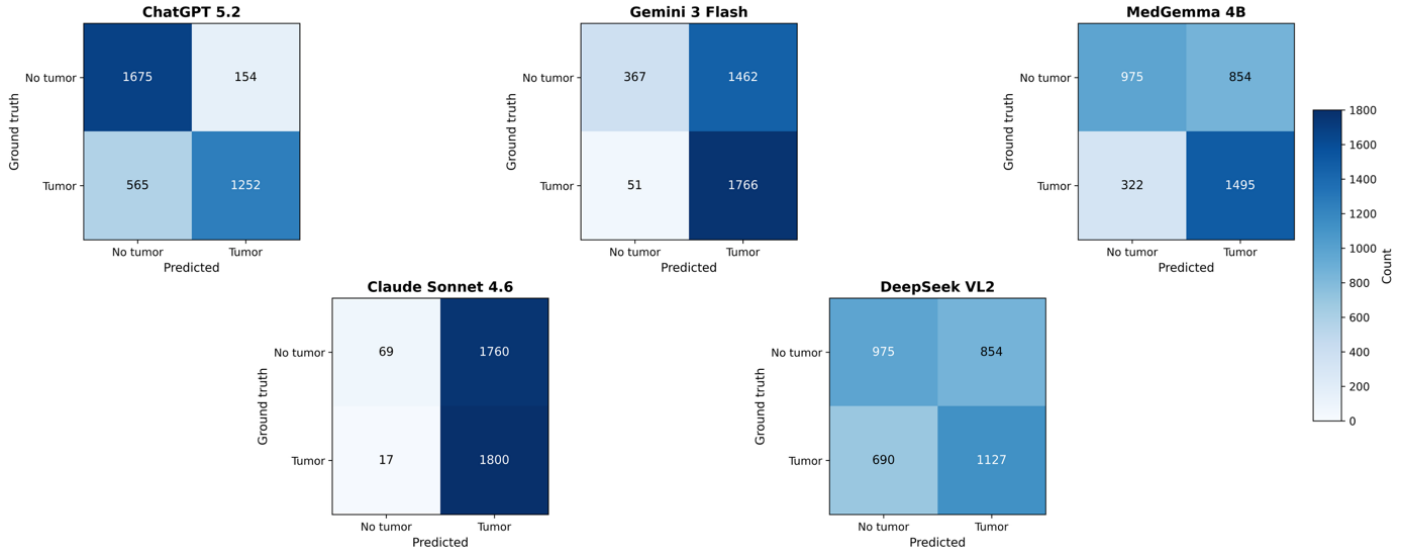

Figure 2. Confusion matrices for Task 1 (lesion detection). The panels display the performance of LLMs in detecting the presence of bone lesions. LLM, large language model.

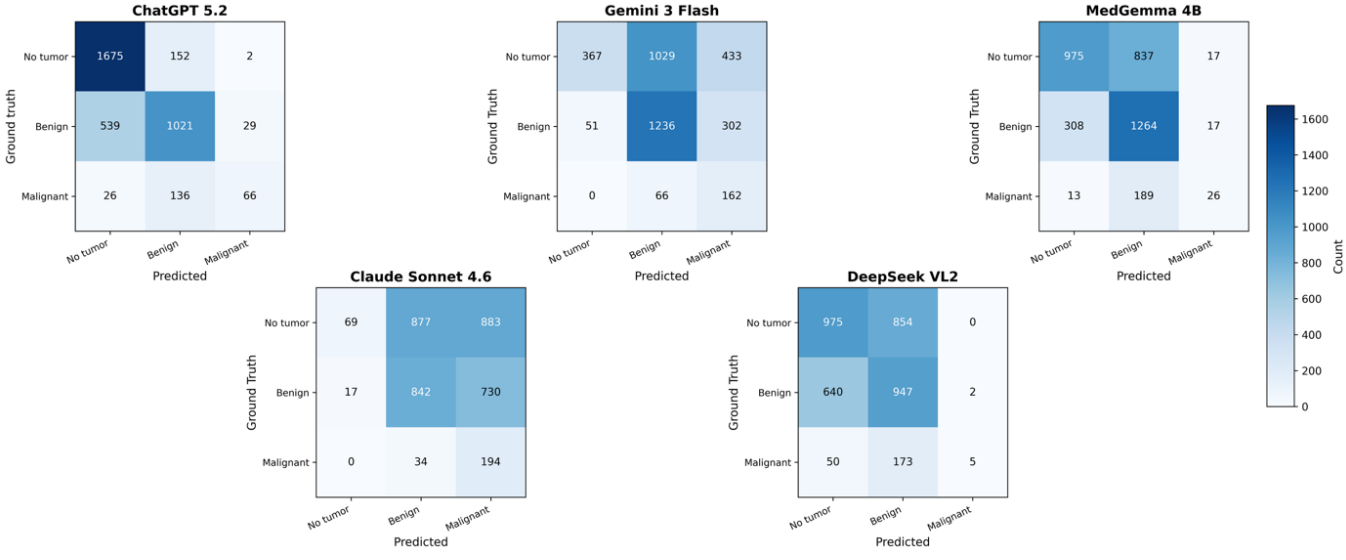

Figure 3. Confusion matrices for the three-class classification task. The matrices illustrate the classification performance for distinguishing between "no lesion," "benign," and "malignant" categories across the five evaluated multimodal LLMs. LLM, large language model.

Table 3. Three-class classification performance (no tumor, benign, and malignant)

Model	Accuracy (95% CI)	Weighted F1 score (95% CI)	Cohen's kappa (95% CI)
ChatGPT 5.2	0.758 (0.745–0.770)	0.745 (0.732–0.758)	0.546 (0.522–0.569)
Gemini 3 Flash	0.484 (0.470–0.498)	0.457 (0.441–0.473)	0.204 (0.186–0.223)
MedGemma 4B	0.621 (0.607–0.635)	0.608 (0.593–0.623)	0.307 (0.281–0.332)
Claude Sonnet 4.6	0.303 (0.291–0.332)	0.268 (0.256–0.280)	0.068 (0.055–0.081)
DeepSeek-VL2	0.529 (0.512–0.544)	0.514 (0.499–0.530)	0.118 (0.089–0.149)

CI, confidence interval.

Table 4. Sub-class performance metrics for lesion characterization

Model	Class	Precision	Sensitivity	F1 score	Number of cases
MedGemma 4B	No lesion	0.752	0.533	0.624	1.829
	Benign	0.552	0.796	0.652	1.589
	Malignant	0.433	0.114	0.180	228
Gemini 3 Flash	No lesion	0.878	0.201	0.327	1.829
	Benign	0.530	0.778	0.630	1.589
	Malignant	0.181	0.711	0.289	228
ChatGPT 5.2	No lesion	0.748	0.916	0.824	1.829
	Benign	0.780	0.642	0.705	1.589
	Malignant	0.680	0.289	0.407	228
Claude Sonnet 4.6	No lesion	0.802	0.038	0.072	1.829
	Benign	0.480	0.530	0.504	1.589
	Malignant	0.107	0.851	0.191	228
DeepSeek-VL2	No lesion	0.586	0.533	0.558	1.829
	Benign	0.480	0.596	0.532	1.589
	Malignant	0.714	0.022	0.043	228

Discussion

In this head-to-head exploratory study of five multimodal LLMs, we observed that although all models demonstrated the capacity to generate diagnostic outputs and ChatGPT 5.2 provided the most robust baseline performance, the discordance between models highlights the volatility of zero-shot inference in complex diagnostic tasks. Although we initially hypothesized that medical pretraining might confer an advantage, ChatGPT 5.2 outperformed the medical-specific MedGemma 4B, whereas the remaining general-purpose models did not demonstrate superior performance. Notably, MedGemma 4B exceeded several other general-purpose models, suggesting that both model scale and domain-specific training contribute to performance and require further optimization.

Although ChatGPT 5.2 achieved moderate agreement (κ: 0.605) and 80.3% accuracy in lesion detection, its zero-shot capabilities remain inferior to benchmarks established by supervised deep learning algorithms. Across prior studies, diagnostic performance varied by both model architecture and imaging modality. High accuracy and area under the curve values exceeding 0.79 were reported for X-ray–based deep learning approaches, particularly when combined with clinical or demographic features. Furthermore, computed tomography or MRI-based models achieved similar discrimination with or without added patient-level information.[11,19-23]

It is also noteworthy that while prior applications of LLMs in orthopedic oncology have predominantly focused on textual report analysis or data extraction, this study represents one of the first evaluations of their direct visual capabilities on radiographs in this field.[24,25] However, despite the nascent stage of this technology, the observed performance deficits suggest considerable limitations in current architectures. Consequently, these findings serve as a cautionary signal for radiologists, clinicians, and patients, indicating that general-purpose multimodal LLMs are not currently reliable for diagnostic decision-making and should be approached with skepticism until rigorous validation is achieved.

An analysis of model performance reveals distinct diagnostic profiles characterized by an imbalance between sensitivity and specificity, potentially limiting their clinical integration. Two divergent interpretative patterns were observed. First, Gemini 3 Flash and Claude Sonnet 4.6 demonstrated a high-sensitivity, low-specificity profile in lesion detection, resulting in a high rate of false positives. This tendency toward

over-identification in Gemini models is consistent with reported performance in other clinical domains, such as emergency triage and cervical cytology.[26,27] In contrast, MedGemma 4B systematically misclassified malignant lesions as benign during the characterization task. Although this model maintained high specificity, its low sensitivity for malignancy poses a substantial diagnostic risk, as malignant lesions are frequently categorized as nonaggressive. This discrepancy between models—ranging from excessive false positives to the failure to identify malignant features—underscores that current zero-shot multimodal architectures lack the calibrated decision thresholds necessary for autonomous oncologic risk stratification.

Despite having lower performance than dedicated task-specific AI models, multimodal LLMs may offer complementary value in radiology in the future. Unlike conventional deep learning systems that require local deployment and integration into imaging infrastructure, multimodal LLMs are broadly accessible through web-based interfaces, allowing for use across institutions with varying technical resources, including resource-limited settings.[28] In addition, recent research suggests that the most promising role for LLMs might not be as standalone diagnostic classifiers, but as components of agentic systems in which the LLM functions as a reasoning engine coordinating specialized tools, such as task-specific CNN pipelines.[29-31] For instance, an agentic approach in orthopedic oncology could structure the diagnostic process through a series of specialized, interoperable models. In the first stage, a CNN-based model performs lesion detection and segmentation. In the second stage, a separate model accesses relevant clinical data—including prior malignancy history and laboratory findings—and integrates these with the imaging outputs to generate a differential diagnosis and preliminary assessment. In the final stage, a third model synthesizes all available information into a structured radiology report and presents it to the radiologist for review and validation.

Our findings indicate that this accessibility currently comes with unacceptable diagnostic risks. The inability of the highest-performing models to reliably detect malignancy or exclude pathology renders them unsuitable for diagnostic decision-making. Consequently, these tools should currently be viewed as experimental rather than as diagnostic alternatives. Radiologists, clinicians, and patients must be cautioned that the fluent textual output of these models does not equate to radiologic competence; thus, human-in-the-loop verification remains non-negotiable.

Our study has limitations inherent to its exploratory design. First, the use of a zero-shot prompting strategy—without providing reference exemplars (few-shot learning) or chain-of-thought reasoning—may have underestimated the models' true potential. Second, the reliance on single-view radiographs contrasts with clinical practice, which typically incorporates orthogonal views and cross-sectional imaging. Third, the nature of commercial closed-source models precludes an analysis of the specific visual features driving their decisions due to their proprietary, non-interpretable architectures. Furthermore, model outputs may vary over time with undocumented updates.[32] Fourth, the direct comparison between massive, proprietary frontier models and the significantly smaller MedGemma 4B and DeepSeek-VL2 models is intrinsically unbalanced. The observed performance differences are likely driven by the vast discrepancy in parameter counts rather than strictly by the difference between general and medical-specific training. Additionally, the MedGemma 27B model was not included due to the lack of multimodal capability at the time of study design, which limits our ability to assess performance scaling within the same model family.

Future research should prioritize prompt engineering strategies to mitigate the observed biases and investigate whether fine-tuning methods can realign these generalist models with the sensitivity requirements of orthopedic oncology. In addition, future studies should evaluate larger multimodal variants within the same model family; higher-parameter models, such as the MedGemma 27B version, may demonstrate improved performance and provide further insight into scaling effects. Finally, prospective reader studies evaluating radiologist performance with and without LLM assistance are needed to determine clinical utility while accounting for ongoing model updates and version changes. Despite these limitations, our study provides key insights into the use of multimodal LLMs in orthopedic oncology and highlights the current limitations of these models within clinical decision algorithms, such as the interpretation of direct radiographs in routine practice.

This study establishes a baseline for the zero-shot capabilities of multimodal LLMs in orthopedic oncology. Our findings demonstrate heterogeneous performance across models: ChatGPT 5.2 achieved the highest overall results, whereas the domain-specific MedGemma 4B outperformed several other general-purpose models. However, despite these results, all evaluated models exhibited critical performance deficits compared with established supervised deep-learning benchmarks. The distinct operating profiles observed—ranging from high sensitivity and low specificity to the dangerous overcalling of benign lesions and low malignancy detection—indicate that current multimodal LLMs lack the calibrated decision boundaries required for safe oncologic decision support. Consequently, although these models offer potential for future workflow integration and increased accessibility, they are not currently suitable for independent diagnostic decision-making. To ensure oncologic safety before clinical adoption, future development should pivot from utilizing LLMs as standalone classifiers toward integrating them into agentic workflows.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Miller TT. Bone tumors and tumorlike conditions: analysis with conventional radiography. *Radiology*. 2008;246(3):662-674. [Crossref]
2. Lalam R, Bloem JL, Noebauer-Huhmann IM, et al. ESSR consensus document for detection, characterization, and referral pathway for tumors and tumorlike lesions of bone. *Semin Musculoskelet Radiol*. 2017;21(5):630-647. [Crossref]
3. Mehta K, McBee MP, Mihal DC, England EB. Radiographic analysis of bone tumors: a systematic approach. *Semin Roentgenol*. 2017;52(4):194-208. [Crossref]
4. Chang CY, Garner HW, Ahlawat S, et al. Society of Skeletal Radiology- white paper. Guidelines for the diagnostic management of incidental solitary bone lesions on CT and MRI in adults: bone reporting and data system (Bone-RADS). *Skeletal Radiol*. 2022;51(9):1743-1764. [Crossref]
5. Park C, Azhideh A, Pooyan A, et al. Diagnostic performance and inter-reader reliability of bone reporting and data system (Bone-RADS) on computed tomography. *Skeletal Radiol*. 2025;54:209-217. [Crossref]
6. Akinci D'Antonoli T, Stanzione A, Bluethgen C, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagn Interv Radiol*. 2024;30(2):80-90. [Crossref]

7. Bhayana R. Chatbots and large language models in radiology: a practical primer for clinical and research applications. *Radiology*. 2024;310(1):e232756. [Crossref]
8. Salbas A, Buyuktoka RE. Performance of large language models in recognizing brain MRI sequences: a comparative analysis of ChatGPT-4o, Claude 4 Opus, and Gemini 2.5 Pro. *Diagnostics (Basel)*. 2025;15(15):1919. [Crossref]
9. von Schacky CE, Wilhelm NJ, Schäfer VS, et al. Multitask deep learning for segmentation and classification of primary bone tumors on radiographs. *Radiology.* 2021;301(2):398-406. [Crossref]
10. Engin O, Durmaz Engin C, Çilengir AH, Dirim Mete B. Detection and classification of supraspinatus pathologies on shoulder magnetic resonance images using a code-free deep learning application. *Asia Pac J Sports Med Arthrosc Rehabil Technol*. 2025;42:1-7. [Crossref]
11. Li J, Li S, Li X, et al. Primary bone tumor detection and classification in full-field bone radiographs via YOLO deep learning model. *Eur Radiol.* 2023;33(6):4237-4248. [Crossref]
12. OpenAI Team. Introducing GPT-5. Accessed January 2025. [Crossref]
13. Comanici G, Bieber E, Schaekermann M, et al. Gemini 2.5: pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv.* 2025;arXiv:2507.06261. [Crossref]
14. Salbas A, Baysan EK. Assessment of large language models in musculoskeletal radiological anatomy: a comparative study with radiologists. *Jt Dis Relat Surg*. 2026;37(1):190-199. [Crossref]
15. Zamora T, Salas P, Zuñiga S, Botello E, Andia ME. Generative artificial intelligence, large language models and ChatGPT in musculoskeletal oncology: current applications and future potential. *J Clin Orthop Trauma*. 2025;69:103161. [Crossref]
16. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat Med*. 2025;31(1):60-69. [Crossref]
17. Yao S, Huang Y, Wang X, et al. A radiograph dataset for the classification, localization, and segmentation of primary bone tumors. *Sci Data*. 2025;12(1):88. [Crossref]
18. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174. [Crossref]
19. von Schacky CE, Wilhelm NJ, Schäfer VS, et al. Development and evaluation of machine learning models based on X-ray radiomics for the classification and differentiation of malignant and benign bone tumors. *Eur Radiol.* 2022;32(9):6247-6257. [Crossref]
20. Liu R, Pan D, Xu Y, et al. A deep learning-machine learning fusion approach for the classification of benign, malignant, and intermediate bone tumors. *Eur Radiol*. 2022;32(2):1371-1383. [Crossref]
21. He Y, Pan I, Bao B, et al. Deep learning-based classification of primary bone tumors on radiographs: a preliminary study. *EBioMedicine*. 2020;62:103121. [Crossref]
22. Eweje FR, Bao B, Wu J, et al. Deep learning for classification of bone lesions on routine MRI. *EBioMedicine*. 2021;68:103402. [Crossref]
23. Yildiz Potter I, Yeritsyan D, Mahar S, et al. Automated bone tumor segmentation and classification as benign or malignant using computed tomographic imaging. *J Digit Imaging*. 2023;36(3):869-878. [Crossref]
24. Kanemaru N, Yasaka K, Fujita N, Kanzawa J, Abe O. The fine-tuned large language model for extracting the progressive bone metastasis from unstructured radiology reports. *J Imaging Inform Med*. 2025;38(2):865-872. [Crossref]
25. Yang F, Yan D, Wang Z. Large-scale assessment of ChatGPT's performance in benign and malignant bone tumors imaging report diagnosis and its potential for clinical applications. *J Bone Oncol*. 2024;44:100525. [Crossref]
26. Lee S, Jung S, Park JH, Cho H, Moon S, Ahn S. Performance of ChatGPT, Gemini and DeepSeek for non-critical triage support using real-world conversations in emergency department. *BMC Emerg Med*. 2025;25(1):176. [Crossref]
27. Laohawetwanit T, Apornvirat S, Asaturova A, Li H, Lami K, Bychkov A. Evaluation of general-purpose large language models as diagnostic support tools in cervical cytology. *Pathol Res Pract*. 2025;274:156159. [Crossref]
28. Stogiannos N, Cuocolo R, Akinci D'Antonoli T, et al. Recognising errors in AI implementation in radiology: a narrative review. *Eur J Radiol*. 2025;191:112311. [Crossref]
29. Tzanis E, Adams LC, Akinci D'Antonoli T, et al. Agentic systems in radiology: principles, opportunities, privacy risks, regulation, and sustainability concerns. *Diagn Interv Imaging*. 2026;107(1):7-16. [Crossref]
30. Tzanis E, Klontzas ME. ReclAIm: a multi-agent framework for degradation-aware performance tuning of medical imaging AI. *arXiv*. 2025. [Crossref]
31. Tzanis E, Klontzas ME. mAIstro: an open-source multi-agent system for automated end-to-end development of radiomics and deep learning models for medical imaging. *Eur J Radiol Artif Intell*. 2025;4:100044. [Crossref]
32. Gu Y, Fu J, Liu X, et al. The illusion of readiness: stress testing large frontier models on multimodal medical benchmarks. *arXiv*. 2025;arXiv:2509.18234. [Crossref]