

Retrospective Single-Centre Study

DIAGNOSTIC PERFORMANCE OF MULTIMODAL LARGE LANGUAGE MODELS FOR PROXIMAL FEMORAL FRACTURE DETECTION ON AP HIP RADIOGRAPHS: A RETROSPECTIVE SINGLE-CENTRE STUDY

N.G. Biavardi¹, F. Pezone¹, F. Morlini¹, M. Alessio-Mazzola², G. Placella¹ and V. Salini¹

¹Vita-Salute San Raffaele University, Via Olgettina 60, 20132, Milan

²IRCCS Ospedale San Raffaele, Via Olgettina 60, 20132, Milan

Correspondence to:

Francesco Pezone, MD

Vita-Salute San Raffaele University,

Orthopedics and Traumatology Department,

Via Olgettina 60,

20132, Milan, Italy

e-mail: f.pezone@studenti.unisr.it

ABSTRACT

Multimodal large language models (MLLMs) capable of analyzing both images and text represent a new class of artificial intelligence tools with potential relevance for clinical imaging. Their diagnostic performance on real-world radiographs, particularly for high-stakes conditions such as proximal femoral fractures, has not yet been systematically evaluated. The purpose of this retrospective study is to determine whether contemporary MLLMs can detect proximal femoral fractures on anteroposterior (AP) hip radiographs with diagnostic accuracy non-inferior to that of an experienced musculoskeletal radiologist, and to assess potential efficiency gains in reporting time. In this retrospective, single-center diagnostic accuracy study, which adhered to STARD 2015, 299 AP hip radiographs (200 fracture-positive; 99 fracture-negative) were evaluated. Four commercially available MLLMs (GPT-5, Gemini-2.5, LLaMA-4, Kimi-VLA3B) were tested using a fixed prompt, and their performance was compared with that of a fellowship-trained radiologist. Non-inferiority for AUC (margin -0.05), superiority for F1-score (margin +0.03), and >50% reduction in time-to-report were prespecified. Decision-curve analysis and an exploratory cost model were performed. The fracture-positive subset (n = 200) was used for efficiency and economic analyses. The radiologist achieved an AUC of 0.919. GPT-5 (AUC 0.950; Δ = +0.031), Gemini-2.5 (0.946; Δ = +0.027), and LLaMA-4 (0.941; Δ = +0.022) met the non-inferiority criterion, whereas Kimi-VLA3B did not (0.872; Δ = -0.047). GPT-5 and Gemini-2.5 demonstrated superior F1-scores relative to the radiologist (0.973 and 0.967 vs 0.933), while LLaMA-4 was not significantly different (0.948) and Kimi-VLA3B performed worse (0.912). All models markedly reduced reporting time, with GPT-5 requiring 15.0 ± 2.9 seconds versus 308.0 ± 64.8 seconds for the radiologist (>90% reduction). Decision-curve analysis showed the highest net benefit for GPT-5 across threshold probabilities of 0.05-0.30. In exploratory economic analysis, GPT-5 and Gemini-2.5 produced positive cost estimates within the fracture-positive cohort, while LLaMA-4 was cost-neutral and Kimi-VLA3B unfavourable. Three MLLMs demonstrated non-inferior diagnostic performance compared with an expert radiologist for proximal femoral fracture detection, with GPT-5 and Gemini-2.5 achieving superior composite accuracy and all models

Received: 02 May 2026

Accepted: 05 June 2026

Copyright © by LAB srl 2026. This publication and/or article is for individual use only and may not be further reproduced without written permission from the copyright holder. Unauthorized reproduction may result in financial and other penalties. Disclosure: All authors report no conflicts of interest relevant to this article.

providing substantial efficiency gains. Although limited by its single-center, prevalence-enriched design, this study provides early evidence supporting the potential role of MLLMs as adjunctive tools in emergency radiographic workflows.

KEYWORDS: *hip fracture, orthopedic AI, artificial intelligence, multimodal large language models, diagnostic accuracy*

INTRODUCTION

Proximal femoral fractures are a major cause of morbidity, mortality, and healthcare utilization in older adults. Across Europe, approximately 600,000 hip fractures occur annually, and the global incidence is projected to rise substantially with population ageing (1). These injuries are associated with marked functional decline, loss of independence, and one-year mortality rates of approximately 20-25% (1, 2). They also impose considerable economic burden on healthcare systems, including high acute-care and rehabilitation costs (3). Early diagnosis is therefore clinically important, as delays in definitive management, particularly surgery beyond 24-48 hours, have been associated with worse outcomes, including increased complications and mortality (4, 5).

In the emergency setting, first-line imaging typically consists of anteroposterior (AP) hip radiographs because of their speed, availability, and low cost. However, radiography has recognized limitations. A proportion of proximal femoral fractures, particularly nondisplaced or impacted injuries, may be radiographically occult at initial presentation and therefore missed on first interpretation (2). In patients with persistent clinical suspicion despite negative radiographs, advanced imaging, especially magnetic resonance imaging (MRI), provides substantially higher sensitivity and is widely regarded as the reference problem-solving modality (6). At the same time, rising imaging demand and workforce shortages in radiology may contribute to reporting delays and increase pressure on emergency diagnostic pathways (7).

Artificial intelligence (AI) methods have shown increasing promise in musculoskeletal radiology. Earlier computer-aided detection systems were limited by relatively high false-positive rates and poor clinical integration (8, 9). More recently, deep learning approaches have demonstrated strong performance in fracture detection on radiographs, with a 2022 meta-analysis reporting pooled sensitivities and specificities of 92% and 91%, respectively, comparable to clinician performance (10). Detection architectures such as YOLO have also shown high performance in radiographic fracture identification tasks (11), and AI-based tools may improve reader sensitivity and reduce interpretation time in selected workflows (12, 13). Nevertheless, most currently available models remain highly task-specific, require substantial domain-specific training data, and often function as relatively opaque prediction systems (2).

Multimodal large language models (MLLMs) represent a newer class of foundation models capable of processing both visual and textual information within a single inference framework. Preliminary reports suggest that some MLLMs may perform well on medical image interpretation tasks, including radiographic assessments (14). However, the current evidence base remains limited, particularly for clinically relevant diagnostic scenarios using real radiographs and direct reader comparison. In addition, prior experience with fracture-detection algorithms indicates that strong performance in internal evaluations does not necessarily translate into robust behavior across varied clinical datasets, especially in the presence of altered anatomy or underlying bone pathology (15). Careful benchmarking of these systems in defined radiographic tasks is therefore warranted.

The purpose of this study was to evaluate the diagnostic performance of contemporary MLLMs for detecting proximal femoral fractures on standard AP hip radiographs in a retrospective single-center dataset, using one fellowship-trained musculoskeletal radiologist as the comparator. A secondary objective was to assess reporting latency under the study conditions, given the clinical importance of timely triage and interpretation in emergency care.

MATERIALS AND METHODS

This retrospective single-center diagnostic accuracy study was conducted at a tertiary trauma hospital and is reported in accordance with STARD 2015. Case selection, index-test execution, reference-standard assignment, and statistical analysis planning were completed before formal analysis, and the dataset was locked before outcome evaluation.

Study population and case selection

The study included adults aged 50 to 100 years who presented to the emergency department with suspected proximal femoral fracture and underwent a diagnostically adequate anteroposterior (AP) hip radiograph as part of routine care. Consecutive eligible cases were screened within the study window. Exclusion criteria were irretrievable or non-removable protected health information and nondiagnostic image quality.

A single institutional cohort was assembled for all analyses. All available fracture-positive cases during the study period were included ($n = 200$). Fracture-negative cases were enrolled consecutively and capped at approximately 100 to preserve analytic balance and feasibility ($n = 99$). This produced a prevalence-enriched sample of approximately 67% fracture-positive cases. Because this prevalence does not reflect the routine emergency department case mix, prevalence-dependent metrics were interpreted cautiously. The full cohort ($N = 299$) was used for the primary diagnostic accuracy analyses. Time-to-report and exploratory economic analyses were performed separately as described below.

Reference standard

The reference standard was established using radiologist adjudication and follow-up data. Fractures considered visible on the index AP radiograph were adjudicated by consensus between two fellowship-trained musculoskeletal radiologists. Occult fractures were defined as fractures not clearly identified on the initial radiograph but subsequently confirmed within 90 days by MRI, CT, or intraoperative findings (6). For fracture-negative cases, clinical and imaging follow-up data available within at least 90 days were reviewed to reduce outcome misclassification. All 299 included images had a complete reference-standard assignment before analysis.

Given the inclusion of occult fractures, two related diagnostic constructs were considered during interpretation of the findings: detection of all confirmed fractures and detection of fractures visible on the index radiograph. The primary analysis used all confirmed fractures; occult fractures were also retained in the overall analysis because of their clinical relevance in acute hip pain pathways (6).

Image acquisition and pre-processing

Radiographs were exported from the institutional picture archiving and communication system (PACS) as 2048 $\times$ 2048 pixel 8-bit JPEG images in sRGB colour space. Window and level settings were standardised at export, and all protected health information was removed before analysis. Each radiograph was assigned a unique anonymised identifier. Clinical metadata were stored separately and were not provided to the models.

Index tests: multimodal large language models

Four commercially available multimodal large language models (MLLMs) were evaluated: GPT-5 Chat (OpenAI), Gemini-2.5 Pro (Google), LLaMA-4 Maverick (Meta), and Kimi-VL-A3B (MoonshotAI). All models were accessed through the OpenRouter API, and model/provider information was logged at the time of inference. Each model received the same single standardised prompt, without iterative prompting, follow-up questions, or manual correction:

"Identify any proximal femur fracture. Return JSON: {fracture_present: yes/no, subtype: neck/intertrochanteric/pertrochanteric, side: left/right, rationale: ≤ 30 words}."

Only the first complete response was accepted. No retries were performed. Inference metadata, including timestamps and latency, was recorded systematically.

Because the prompt constrained the models to produce binary fracture classifications, all principal diagnostic analyses were based on case-level classification performance rather than on calibrated probability outputs. ROC/AUC analyses were retained as supplementary discrimination analyses and interpreted with caution in light of the binary-output format.

Comparator reader

The comparator was one fellowship-trained musculoskeletal radiologist. The radiologist interpreted the same AP radiographs independently under study conditions. No model outputs were available to the radiologist during review. The comparison should therefore be interpreted as a model-versus-single-expert-reader analysis rather than a general estimate of radiologist-level performance.

Study outcomes

The primary outcomes were case-level diagnostic performance and reporting latency under study conditions. Diagnostic performance was assessed principally using sensitivity and specificity, with additional reporting of accuracy, positive predictive value, negative predictive value, and F1-score. Confusion-matrix-based metrics were emphasized because all index tests produced binary outputs. AUC was reported as a supplementary discrimination metric.

Secondary and exploratory outcomes included decision-curve analysis across threshold probabilities from 0.05 to 0.30 and an exploratory economic model estimating the potential impact of changes in false-negative classifications on bed-day utilisation and cost.

Time-to-report measurement

For the MLLMs, time-to-report was defined as the interval from API request submission to receipt of the first complete JSON output. For the radiologist, time-to-report was defined as the interval required to review the radiograph and produce the study interpretation under the study conditions. Because these workflows are inherently different, timing results were interpreted as a comparison between model inference latency and reader interpretation time under controlled conditions, rather than as a direct simulation of a real-world reporting workflow.

Timing measurements were summarised for the fracture-positive subset, including occult fractures, because the exploratory downstream analyses focused on missed-fracture consequences.

Statistical analysis

Categorical outcomes are reported as counts and proportions; continuous variables are reported as means with standard deviations unless otherwise specified. For the main diagnostic comparison, confusion matrices were generated for each model and for the radiologist. Sensitivity, specificity, accuracy, positive predictive value, negative predictive value, and F1-score were calculated with corresponding confidence intervals. Paired differences between each model and the radiologist were estimated using bootstrap resampling with 5,000 iterations and a fixed random seed.

ROC curves and AUC values were also estimated as supplementary analyses to facilitate comparison with prior diagnostic-AI literature (10,15). Confidence intervals for AUC differences between each model and the radiologist were derived using paired bootstrap resampling; DeLong-based sensitivity analyses were performed secondarily. Because the index tests produced binary outputs, ROC/AUC findings were interpreted as supportive rather than definitive measures of discrimination.

Time-to-report comparisons were treated as paired repeated measurements across the same case set. Mean differences between each model and the radiologist were assessed at the case level using paired testing. Overall between-reader timing comparisons were interpreted descriptively, with emphasis placed on the magnitude and consistency of paired differences rather than on omnibus group testing alone.

Decision-curve analysis was performed across threshold probabilities from 0.05 to 0.30. Net benefit was calculated using standard methodology. Because the study cohort was prevalence-enriched, these analyses were considered exploratory and interpreted with caution.

The exploratory economic analysis was restricted to the fracture-positive subset. Differences in false-negative classifications between each model and the radiologist were translated into bed-days avoided using the predefined assumption that each missed fracture corresponded to 0.6 inpatient bed-days, with a cost of €300 per bed-day (3). Fixed per-case infrastructure and API costs were included. These estimates were intended only as exploratory scenario analyses and not as formal health-economic evaluations.

Quality control and bias mitigation

Quality-control procedures included automated checks for malformed or logically inconsistent model outputs, source verification for flagged cases, and dataset locking before formal analysis. All models were evaluated using the same fixed prompt and the same image set, without user interaction, to preserve comparability. Consecutive sampling within the available strata and follow-up review of fracture-negative cases were used to reduce selection and verification bias where feasible. Nevertheless, the single-centre design, enriched prevalence, and use of a single expert comparator were recognised sources of limited generalisability.

Software and reproducibility

All analyses were performed using R version 4.3.0 (packages including pROC, boot, and dcurves) and Python 3.11 (including numpy, scipy, and statsmodels), with random processes initialised using seed 12345. De-identified data, API logs, and analysis scripts will be made available upon publication or upon reasonable request, subject to institutional and privacy constraints.

RESULTS

A total of 299 AP hip radiographs were included in the diagnostic accuracy analyses, comprising 200 fracture-positive and 99 fracture-negative images. No included image was incomplete, and all 299 cases had complete reference-standard assignment. The fracture-positive subset of 200 cases, including 20 occult fractures, was used for time-to-report assessment and exploratory economic analysis.

Diagnostic performance

The comparator radiologist achieved an AUC of 0.919 (95% CI, 0.882-0.956). Among the MLLMs, GPT-5 achieved the highest AUC at 0.950 (95% CI, 0.920-0.976), followed by Gemini-2.5 at 0.946, LLaMA-4 at 0.941, and Kimi-VL-A3B at 0.872. Relative to the radiologist, the estimated AUC differences were +0.031 for GPT-5 (95% CI, -0.017 to 0.077), +0.027 for Gemini-2.5 (95% CI, -0.017 to 0.074), +0.022 for LLaMA-4 (95% CI, -0.018 to 0.064), and -0.047 for Kimi-VL-A3B (95% CI, -0.101 to 0.007). Because the index tests produced binary outputs, these ROC-based results should be interpreted as supplementary discrimination analyses rather than the sole basis for comparative performance assessment. Case-level diagnostic performance metrics are summarized in Table I.

Table I. Case-level diagnostic performance of the radiologist and multimodal large language models for proximal femoral fracture detection.

Reader	Confusion counts				matrix					Classification metrics			ROC summary		
Reader / model	TP	FP	TN	FN	Sensitivity	Specificity	Accuracy	PPV	NPV	F1-score	AUC (95% CI)	ΔAUC vs radiologist (95% CI)			
Radiologist	180	6	93	20	0.900	0.939	0.913	0.968	0.823	0.933	0.919 (0.882-0.956)	Reference			
GPT-5	196	7	92	4	0.980	0.929	0.963	0.966	0.958	0.973	0.950 (0.920-0.976)	+0.031 (-0.017 to 0.077)			
Gemini-2.5	192	5	94	8	0.960	0.949	0.957	0.975	0.922	0.967	0.946	+0.027 (-0.017 to 0.074)			
LLaMA-4	184	4	95	16	0.920	0.960	0.933	0.979	0.856	0.948	0.941	+0.022 (-0.018 to 0.064)			
Kimi-VL-A3B	176	10	89	24	0.880	0.899	0.886	0.946	0.788	0.912	0.872	-0.047 (-0.101 to 0.007)			

Abbreviations: **TP**: true positive; **FP**: false positive; **TN**: true negative; **FN**: false negative; **PPV**: positive predictive value; **NPV**: negative predictive value; **AUC**: area under the receiver operating characteristic curve.

Confusion-matrix-based performance showed a similar ranking. The radiologist’s F1-score was 0.933, based on 180 true positives, 6 false positives, and 20 false negatives. GPT-5 achieved the highest F1-score at 0.973, followed by Gemini-2.5 at 0.967, LLaMA-4 at 0.948, and Kimi-VL-A3B at 0.912. Compared with the radiologist, the estimated F1-score differences were +0.040 for GPT-5 (95% CI, +0.013 to +0.068), +0.035 for Gemini-2.5 (95% CI, +0.008 to +0.064), +0.016 for LLaMA-4 (95% CI, -0.012 to +0.044), and -0.021 for Kimi-VL-A3B (95% CI, -0.054 to +0.010). Overall, GPT-5 and Gemini-2.5 showed the strongest case-level classification performance in this dataset, LLaMA-4 was broadly similar to the radiologist, and Kimi-VL-A3B performed less well. Differences in F1-score relative to the radiologist are shown in Fig. 1.

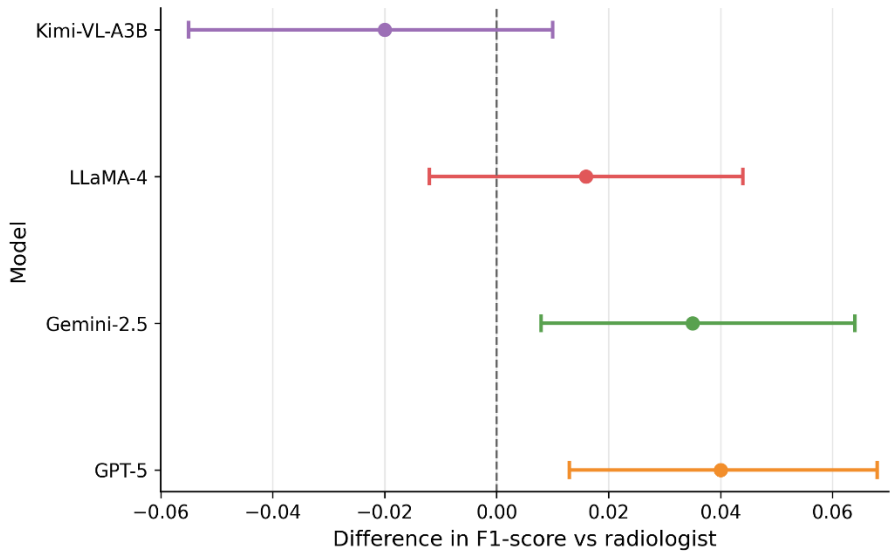


Fig. 1. Difference in F1-score between each multimodal large language model and the radiologist, with 95% confidence intervals.

Time-to-report

Within the fracture-positive subset, the radiologist's mean interpretation time under the study conditions was 308.0 ± 64.8 seconds. Model output latency was substantially shorter for all four MLLMs: 15.0 ± 2.9 seconds for GPT-5, 17.1 ± 2.9 seconds for Gemini-2.5, 27.1 ± 4.3 seconds for LLaMA-4, and 40.2 ± 6.0 seconds for Kimi-VL-A3B. These corresponded to approximate reductions of 95.1%, 94.5%, 91.2%, and 86.9%, respectively, relative to the radiologist's study reading time. Because these measures compare model inference latency with reader interpretation time under controlled study conditions, they should not be interpreted as direct equivalents of real-world clinical reporting workflow. Time-to-report under the study conditions is shown in Fig. 2.

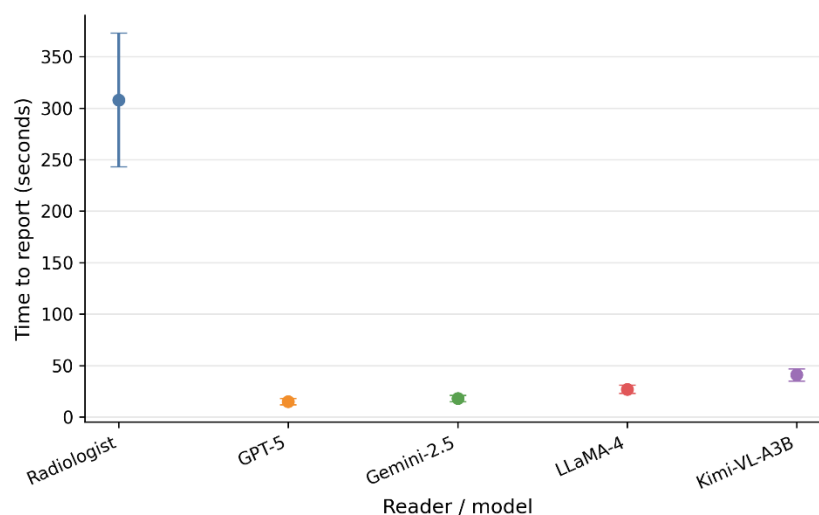


Fig. 2. Time to report under study conditions for the radiologist and multimodal large language models, shown as mean $\pm$ standard deviation.

Exploratory decision-curve analysis

Across threshold probabilities from 0.05 to 0.30, all readers demonstrated positive net benefit relative to a treat-none strategy. GPT-5 showed the highest net benefit across this range. At threshold probabilities of 0.10, 0.20, and 0.30, GPT-5 yielded net benefits of 0.648, 0.640, and 0.645, compared with 0.631, 0.591, and 0.583 for the radiologist. Gemini-2.5 and LLaMA-4 showed intermediate performance, generally similar to or slightly above the radiologist, whereas Kimi-VL-A3B showed the lowest net benefit. Because the study cohort was prevalence-enriched, these findings should be interpreted as exploratory rather than definitive estimates of clinical utility. Exploratory decision-curve analysis is shown in Fig. 3.

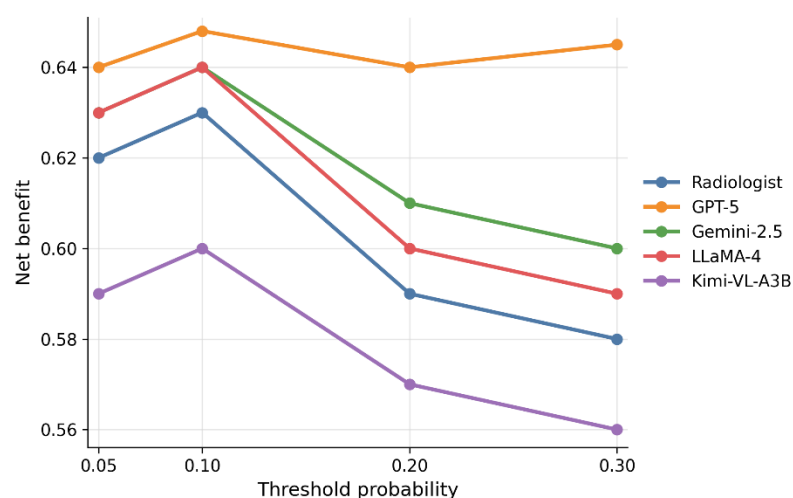


Fig. 3. Exploratory decision-curve analysis showing net benefit across threshold probabilities for the radiologist and multimodal large language models.

Exploratory economic analysis

Within the 200 fracture-positive cases, the radiologist missed 20 fractures. GPT-5 missed 4 fractures ($\Delta FN = -16$), Gemini-2.5 missed 8 ($\Delta FN = -12$), LLaMA-4 missed 16 ($\Delta FN = -4$), and Kimi-VL-A3B missed 24 ($\Delta FN = +4$). Bootstrap confidence intervals for ΔFN were -24 to -7 for GPT-5, -20 to -3 for Gemini-2.5, -13 to +5 for LLaMA-4, and -5 to +13 for Kimi-VL-A3B. Using the predefined exploratory assumptions of 0.6 bed-days avoided per missed fracture prevented and €300 per bed-day, together with fixed infrastructure and API costs, the estimated net cost impact over 200 fracture-positive cases was €2,128 for GPT-5, €1,412 for Gemini-2.5, approximately cost-neutral for LLaMA-4 (-€22), and unfavorable for Kimi-VL-A3B (-€1,460). These results should be interpreted strictly as exploratory scenario estimates rather than as a formal health-economic evaluation.

DISCUSSION

In this retrospective single-centre study, several contemporary multimodal large language models demonstrated promising performance for proximal femoral fracture detection on AP hip radiographs when compared with one fellowship-trained musculoskeletal radiologist. Among the evaluated systems, GPT-5 and Gemini-2.5 showed the strongest overall performance in this dataset, with LLaMA-4 yielding similar but somewhat less consistent results, whereas Kimi-VL-A3B performed less well. These findings support the feasibility of multimodal foundation models for a focused musculoskeletal radiographic task, but they should be interpreted cautiously given the study design and the absence of external or multi-reader validation.

The clinical importance of this task is well established. Hip fractures are associated with substantial morbidity, mortality, and healthcare costs, and timely diagnosis remains critical because delays in operative management are associated with worse outcomes (1-5). In this context, the possibility that MLLMs may assist with rapid identification of proximal femoral fracture on first-line radiographs is of genuine interest. At the same time, diagnostic performance in a retrospective enriched dataset should not be conflated with real-world clinical effectiveness. The present findings therefore suggest technical promise rather than readiness for autonomous deployment.

Compared with the radiologist, the highest-performing models achieved similar or better case-level classification metrics within this dataset. However, that observation should be framed carefully. First, the study included only one expert comparator, which precludes any inference about average radiologist performance or radiologist-level capability in a broader sense. Second, the evaluated task was deliberately narrow: all readers assessed a single AP radiograph in isolation, without the wider clinical context, additional radiographic projections, prior studies, or multidisciplinary workflow information that commonly inform real emergency imaging interpretation. Accordingly, the results should be understood as a constrained reader-comparison experiment rather than a surrogate for comprehensive clinical reporting.

A further issue concerns the inclusion of occult fractures. These cases are clinically important because missed occult hip fractures can delay definitive care and worsen outcomes (6). However, they also complicate interpretation of model-versus-reader performance on single AP radiographs. A fracture may be present in the patient but not clearly visible on the index radiograph, meaning that the task of detecting all confirmed fractures is not identical to the task of identifying fractures demonstrably visible on that image. For that reason, the present findings should not be interpreted as evidence that the models consistently detect radiographically occult injuries. Rather, they indicate that in a dataset containing both visible and occult fractures, some models retained favorable overall classification performance. Future studies should report visible-fracture and occult-fracture analyses separately.

The timing findings also require careful interpretation. All four MLLMs generated outputs much faster than the radiologist's study reading time. This observation is unsurprising given the difference between automated API inference and human image interpretation. Although shorter output latency may be relevant for triage-oriented workflows, these measurements do not capture the full complexity of clinical reporting, including worklist prioritization, review of additional projections or prior imaging, communication with clinical teams, and report finalization. The results, therefore, support only the narrower conclusion that model inference latency is substantially shorter than expert reader interpretation time under the controlled conditions of this study.

The exploratory decision-curve and economic analyses should likewise be interpreted conservatively. Decision-curve analysis suggested higher net benefit for the strongest-performing models across the tested threshold range, and the exploratory cost scenarios favored GPT-5 and Gemini-2.5. However, both analyses are sensitive to assumptions and case mix. The present cohort was prevalence-enriched, which limits the clinical transportability of threshold-based utility estimates, and the economic model relied on simplified assumptions regarding avoided bed-days and costs while not fully incorporating the downstream consequences of false-positive classifications. These analyses are therefore best viewed as hypothesis-generating rather than as evidence of real-world economic value.

Our findings are broadly consistent with the growing literature showing that AI systems can perform well on targeted fracture-detection tasks (10, 11, 13). They also extend the discussion from conventional task-specific deep learning systems to more general multimodal foundation models (14). This distinction is relevant because MLLMs may offer practical advantages in flexibility and interaction, but it also introduces new concerns regarding reproducibility, model drift, proprietary updating, and limited transparency of internal decision processes. In contrast to fixed locally deployed algorithms, commercially hosted MLLMs may change over time, making rigorous version control and independent replication especially important for medical applications (14, 15).

This study has several limitations. First, it was retrospective and single-centre in design, which limits generalisability. Second, the dataset was intentionally enriched with fracture-positive cases, and therefore does not reflect routine emergency department prevalence. Third, the comparison involved only one fellowship-trained musculoskeletal radiologist, without assessment of inter-reader variability. Fourth, the use of binary model outputs constrained the statistical framework and makes ROC/AUC less informative than it would be for calibrated probabilistic outputs. Fifth, radiographs were evaluated as exported JPEG images rather than in a full PACS reading environment, and only AP views were provided. Sixth, the reference standard combined consensus review of visible fractures with follow-up confirmation of occult fractures, which, although clinically pragmatic, is not equivalent to a single uniform reference procedure. Finally, no external validation cohort was included, and performance may differ in other institutions, imaging protocols, hardware environments, and patient populations.

Notwithstanding these limitations, the study has some strengths. It addresses a clinically relevant emergency imaging task, evaluates multiple contemporary multimodal systems using the same standardized prompt and image set, includes an expert human comparator, and examines not only classification performance but also timing and exploratory downstream implications. These features make the work a useful pilot assessment of whether general-purpose multimodal foundation models merit further study in musculoskeletal radiology.

Future work should focus on prospective multi-reader studies, external validation across centres, and more clinically realistic workflows incorporating additional radiographic views, clinical metadata, and calibrated probability outputs. Separate reporting of radiographically visible versus occult fractures will be particularly important. Any serious evaluation for clinical adoption should also include formal workflow studies and more rigorous health-economic modelling.

CONCLUSIONS

In this retrospective single-centre pilot study, several multimodal large language models demonstrated promising performance for detecting proximal femoral fractures on AP hip radiographs when compared with one fellowship-trained musculoskeletal radiologist under the study conditions. GPT-5 and Gemini-2.5 showed the strongest overall results in this dataset, with substantially shorter output latency than expert reader interpretation time. However, these findings should be interpreted cautiously, given the enriched case mix, the use of binary model outputs, the mixed reference standard, and the absence of external and multi-reader validation. Larger prospective studies with clinically representative populations and more rigorous workflow-based evaluation are required before any clinical implementation can be considered.

REFERENCES

1. Werner M, Macke C, Gogol M, Krettek C, Liodakis E. Differences in hip fracture care in Europe: a systematic review of recent annual reports of hip fracture registries. *European Journal of Trauma and Emergency Surgery*. 2021;48(3):1625-1638. doi:<https://doi.org/10.1007/s00068-021-01797-8>
2. González G, Galant J, Salinas JM, et al. Classification and segmentation of hip fractures in x-rays: highlighting fracture regions for interpretable diagnosis. *Insights into Imaging*. 2025;16(1). doi:<https://doi.org/10.1186/s13244-025-01958-y>
3. Blaschke BL, Parikh HR, Vang SX, Cunningham BP. Time-Driven Activity-Based Costing: A Better Way to Understand the Cost of Caring for Hip Fractures. *Geriatric Orthopaedic Surgery & Rehabilitation*. 2020;11:215145932095820. doi:<https://doi.org/10.1177/2151459320958202>
4. Welford P, Jones CS, Davies G, et al. The association between surgical fixation of hip fractures within 24 hours and mortality. *The Bone & Joint Journal*. 2021;103-B(7):1176-1186. doi:<https://doi.org/10.1302/0301-620x.103b7.bjj-2020-2582.r1>
5. Klestil T, Röder C, Stotter C, et al. Impact of Timing of Surgery in Elderly Hip Fracture patients: a Systematic Review and meta-analysis. *Scientific Reports*. 2018;8(1). doi:<https://doi.org/10.1038/s41598-018-32098-7>
6. Bartolotta RJ, Ha AS, Bateni CP, et al. ACR Appropriateness Criteria® Acute Hip Pain: 2024 Update. *Journal of the American College of Radiology*. 2025;22(5):S3-S13. doi:<https://doi.org/10.1016/j.jacr.2025.02.039>

7. Meng F, Zhang L, Liu S, Zhang H. The Growing Problem of Radiologist Shortage: China's Perspective. *Korean Journal of Radiology*. 2023;24(11):1046-1046. doi:<https://doi.org/10.3348/kjr.2023.0839>
8. Jing AB, Garg N, Zhang J, Brown JJ. AI solutions to the radiology workforce shortage. *npj Health Systems*. 2025;2(1):20. doi:<https://doi.org/10.1038/s44401025000236>
9. Elmore JG, Lee CI. Artificial Intelligence in Medical Imaging—Learning From Past Mistakes in Mammography. *JAMA Health Forum*. 2022;3(2):e215207. doi:<https://doi.org/10.1001/jamahealthforum.2021.5207>
10. Kuo RYL, Harrison C, Curran TA, et al. Artificial Intelligence in Fracture Detection: A Systematic Review and Meta-Analysis. *Radiology*. 2022;304(1). doi:<https://doi.org/10.1148/radiol.211785>
10. Ahmed A, Ali Shariq Imran, Manaf A, Kastrati Z, Sher Muhammad Daudpota. Enhancing wrist abnormality detection with YOLO: Analysis of state-of-the-art single-stage detection models. *Biomedical signal processing and control*. 2024;93:106144-106144. doi:<https://doi.org/10.1016/j.bspc.2024.106144>
11. Liu Y, Liu W, Chen H, et al. Artificial intelligence versus radiologist in the accuracy of fracture detection based on computed tomography images: a multi-dimensional, multi-region analysis. *Quantitative Imaging in Medicine and Surgery*. 2023;13(10):6424-6433. doi:<https://doi.org/10.21037/qims-23-428>
12. Fu T, Viswanathan V, Attia A, et al. Assessing the Potential of a Deep Learning Tool to Improve Fracture Detection by Radiologists and Emergency Physicians on Extremity Radiographs. *Academic Radiology*. 2023;31(5). doi:<https://doi.org/10.1016/j.acra.2023.10.042>
13. Jin Q, Chen F, Zhou Y, et al. Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine. *npj Digital Medicine*. 2024;7(1):1-6. doi:<https://doi.org/10.1038/s41746-024-01185-7>
14. Oakden-Rayner L, Gale W, Bonham TA, et al. Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department: a diagnostic accuracy study. *The Lancet Digital Health*. 2022;4(5):e351-e358. doi:[https://doi.org/10.1016/s2589-7500\(22\)00004-8](https://doi.org/10.1016/s2589-7500(22)00004-8)