

THE PERFORMANCE OF LARGE LANGUAGE MODELS IN BONE TUMOUR IMAGING: COMPARATIVE ANALYSIS WITH RADIOLOGISTS USING TEXT AND IMAGE-BASED EVALUATION

Eren Çamur^{1✉}, Turay Cesur², Yasin Celal Güneş³,
Semra Duran⁴

Received on October 10, 2025

Presented by D. Damianov, Member of BAS, on November 25, 2025

Abstract

Large language models (LLMs) are emerging as transformative tools in radiology, with potential to enhance diagnostic workflows. However, their performance in bone tumour imaging – a domain requiring both knowledge-based reasoning and visual interpretation – remains unclear. This study compares the diagnostic performance of LLMs with radiologists across text and image-based tasks.

In this cross-sectional study, two LLMs and two radiologists (a junior and a senior) were evaluated using fifty text-based multiple-choice questions (MCQs) and fifty radiographs with clinical vignettes from a public dataset. Participants classified lesions as benign or malignant, identified “don’t-touch” lesions, and provided the most likely diagnosis. Responses were benchmarked against a reference standard using McNemar’s tests.

In MCQs, ChatGPT-5 (92.0%) and Gemini 2.5 Pro (90.0%) achieved accuracies comparable to SR (88.0%) and JR (84.0%) ($p > 0.05$). For benign–malignant classification, LLMs (50.0%, 48.0%) were similar to JR (66.0%) but inferior to SR (94.0%) ($p < 0.05$). In identifying “don’t-touch” lesions, LLMs (46.0%) matched JR (64.0%) yet underperformed compared to SR (92.0%) ($p < 0.05$). For specific diagnosis, LLMs showed low accuracy (38.0%, 30.0%) versus JR (60.0%) and SR (86.0%) ($p < 0.01$).

LLMs may serve as useful adjuncts for clinicians and radiologists in text-based tasks and in distinguishing between benign and malignant bone tumours. However, their diagnostic accuracy remains limited.

Key words: ChatGPT, artificial intelligence, bone tumours, large language models, radiology

Introduction. Large language models (LLMs) have recently emerged as transformative tools in medicine, with applications ranging from natural language processing to clinical decision support [1]. In radiology, their potential extends to knowledge retrieval, structured reporting, and preliminary image interpretation, suggesting that LLMs could complement radiologists by enhancing diagnostic accuracy and workflow efficiency [2].

Bone tumours represent a particularly challenging field due to their wide spectrum of presentations, heterogeneity, and the overlapping radiographic features between benign and malignant lesions [3]. Accurate interpretation of these lesions has direct implications for prevention of unnecessary interventions and avoidance of overtreatment.

Although LLMs have been evaluated across several radiological domains, research focused on musculoskeletal oncologic radiology remains limited [4, 5]. ERDURMUS et al. [5] showed that ChatGPT-4o achieved high accuracy in detecting suspicious lung lesions with low temperature settings. This is highlighting its potential for image-based lung lesion evaluation. However, to our best knowledge, no study has examined LLM performance about bone tumours which is a diagnostically challenging area in radiology.

This study aimed to fill this gap by comparing two current LLMs' performance with radiologists of different experience levels through both text-based and image-based evaluation using X-rays, providing a comprehensive analysis of their potential.

Material and method. This experimental study employed a cross-sectional design to compare the performance of ChatGPT-5), and Gemini 2.5 Pro with that of a junior (JR) and a senior radiologist (SR) on bone tumour imaging.

Since no authentic patient identifiers and data were used and all utilized cases were obtained from publicly open dataset, ethical approval was not applicable for this study. To ensure methodological rigour and reproducibility, the study was conducted in accordance with the Standards for Reporting Diagnostic Accuracy Studies (STARD) guideline [6].

A total of fifty text-based multiple-choice questions (MCQs) and fifty X-rays for image-based questions (IBQs) were used in this study. To ensure the quality and validity of the MCQs, all questions underwent a structured content development in two steps. Radiologist 1 (R1), who has 7 years of general radiology experience, first generated an initial pool of 70 questions based on standard musculoskeletal radiology textbooks, teaching files, and widely accepted diagnostic

criteria for bone tumours. These questions were designed to target four predefined content domains, each serving a specific educational and diagnostic purpose:

(1) Epidemiology questions (12 items) evaluated the participant’s background knowledge of age distribution, anatomical predilection, and tumour prevalence;

(2) Radiographic feature questions (15 items) assessed pattern recognition skills, including matrix identification, periosteal reactions, and hallmark imaging signs;

(3) Benign–malignant differentiation questions (15 items) measured clinical decision-making regarding aggressiveness, transition zones, cortical integrity, and presence of soft-tissue components;

(4) Don’t-touch lesion questions (8 items) evaluated the ability to recognize lesions that should not be biopsied to avoid unnecessary interventions.

At the second step, R1 and a second radiologist (R2) who has also 7 years of general radiology experience (not involved in testing) reviewed all MCQs for clarity, clinical relevance, and ambiguity, and discrepancies were resolved by in person discussion and consensus. Based on this review, the final 50 MCQs were selected.

For the image-based evaluation, X-rays were selected from the publicly available bone tumour dataset [7]. The dataset included a balanced mix of benign, malignant, and “don’t-touch” lesions across various anatomical sites. For each case, R1 prepared a brief vignette detailing patient age and presenting complaint, accompanied by X-rays in JPEG format. Vignettes were provided to the models as text during the chat session. LLMs and radiologists answered three identical image-based questions (IBQs): (1) Is the lesion benign or malignant? (2) What is the most likely diagnosis? (3) Is it a “don’t-touch” lesion?

The workflow of the study is provided in Fig. 1.

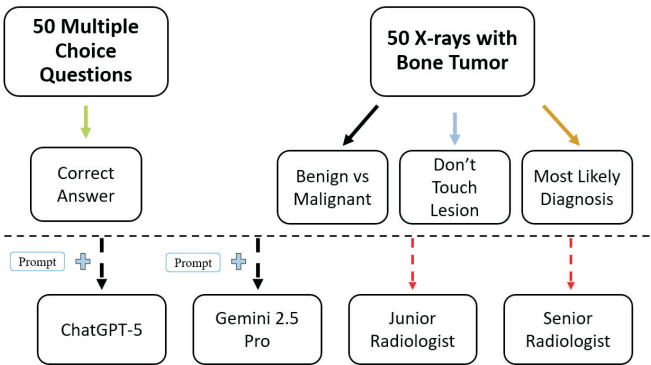


Fig. 1. The workflow of the study

To minimize variability in LLM performance, standardized following prompts were used. No fine-tuning, additional training, or external resources were applied, and default hyperparameters were used for all platforms.

The study included two radiologists: JR with 7 years of experience, and SR with 20 years of experience, including 10 in musculoskeletal radiology.

Radiologists answered the identical MCQs and IBQs that are given to LLMs. They independently evaluated the questions under blinded conditions using electronic forms with their computers without internet access or external resources.

Reference answers for all questions were determined by R1 after reviewing the dataset documentation. The responses of LLMs and radiologists were subsequently compared with the reference standard established by R1, and each response was scored as correct (1) or incorrect (0).

Statistical analyses assessed performance differences across groups. Data normality was tested using the Kolmogorov–Smirnov test. Descriptive statistics (minimum, maximum, median, interquartile range, percentages) summarized results. As data were non-normally distributed, non-parametric methods were applied. Paired group comparisons used McNemar’s test, while chi-square tested categorical differences. A p -value ≤ 0.05 indicated statistical significance.

Results. ChatGPT-5 and Gemini 2.5 Pro demonstrated high MCQ performance, achieving accuracies of 92.0% and 90.0%, respectively, with no significant difference between them ($p > 0.05$). SR achieved 88.0% and JR 84.0%, showing no significant difference from the LLMs ($p > 0.05$) (Table 1).

T a b l e 1
Performance comparison between LLMs and radiologists in MCQs

Participants	Accuracy	p -values (McNemar test)			
		ChatGPT-5	Gemini 2.5 Pro	JR	SR
ChatGPT-5	% 92.0	–	0.412	0.236	0.318
Gemini 2.5 Pro	% 90.0	0.412	–	0.528	0.472
JR	% 84.0	0.236	0.528	–	0.317
SR	% 88.0	0.318	0.472	0.317	–

JR: Junior radiologist, SR: Senior radiologist

For benign–malignant classification, LLMs showed moderate accuracy (ChatGPT-5 50.0%, Gemini 48.0%), comparable to JR (66.0%) ($p > 0.05$) but lower than SR (94.0%). No significant difference existed between the two LLMs ($p > 0.05$). In identifying “don’t touch” lesions, both LLMs correctly classified 46.0%, similar to JR (64.0%) ($p > 0.05$) but significantly below SR (92.0%) ($p < 0.01$). No significant difference was noted between LLMs ($p > 0.05$) (Table 2).

For most likely diagnosis, ChatGPT-5 and Gemini 2.5 Pro achieved 38.0% and 30.0%, both significantly lower than JR (60.0%) and SR (86.0%) ($p < 0.01$). SR outperformed JR, while LLMs did not differ significantly from each other ($p > 0.05$) (Table 3).

T a b l e 2

Benign–Malignant differentiation and “Don’t Touch” lesion detection performance of LLMs and radiologists

		<i>p</i> -values (McNemar test)			
Participants	Accuracy	ChatGPT-5	Gemini 2.5 Pro	JR	SR
Benign vs. Malignant					
ChatGPT-5	%50.0	–	0.472	0.318	0.023
Gemini 2.5 Pro	%48.0	0.472	–	0.276	0.019
JR	%66.0	0.318	0.276	–	0.041
SR	%94.0	0.023	0.019	0.041	–
“Don’t Touch” Lesion					
ChatGPT-5	%46.0	–	0.563	0.244	0.021
Gemini 2.5 Pro	%46.0	0.563	–	0.217	0.018
JR	%64.0	0.244	0.217	–	0.033
SR	%92.0	0.021	0.018	0.033	–

JR: Junior radiologist, SR: Senior radiologist

T a b l e 3

Accuracy and the performance differences of LLMs and radiologists in determining the most likely diagnosis

		<i>p</i> -values (McNemar test)			
Participants	Accuracy	ChatGPT-5	Gemini 2.5	JR	SR
ChatGPT-5	%38.0	–	0.472	< 0.01	< 0.01
Gemini 2.5 Pro	%30.0	0.472	–	< 0.01	< 0.01
JR	%60.0	< 0.01	< 0.01	–	0.028
SR	%86.0	< 0.01	< 0.01	0.028	–

JR: Junior radiologist, SR: Senior radiologist

Discussion. One of the most striking findings of this study is that both LLMs performed at a level comparable to, and slightly higher than radiologists in MCQs. This aligns with prior studies that LLMs excel at codified or rule-based tasks [8,9]. Li et al. [8] examined ChatGPT-3.5 and 4 against German evidence-based guidelines for soft-tissue sarcoma using 80 clinical questions, showing that ChatGPT-4 outperformed its predecessor but still generated potentially harmful recommendations, such as inappropriate use of radiotherapy instead of wide surgical resection. Similarly, WANG et al. [9] demonstrated that embedding NCCN bone cancer guidelines into GPT-4, GPT-3.5, and PaLM-2 markedly improved treatment decision accuracy compared with unaugmented models. YANG et al. [10] evaluated ChatGPT on 1366 bone tumour imaging reports and found very high sensitivity (95%) but poor specificity (58%), leading to frequent false posi-

tives. After minor fine-tuning, accuracy improved to 87% with sensitivity rising to 99%, though 56 benign cases were still misclassified as malignant. Similarly, our findings highlight that LLMs have sufficient text-based knowledge about bone tumours and are strong when operating within structured frameworks or guideline-scaffolded contexts. This superior performance of ChatGPT-5 and Gemini 2.5 Pro in MCQs in this study reflects this trend and suggests that LLMs can serve as reliable supports for knowledge-driven tasks, provided guideline alignment or expert oversight is ensured.

Another important finding of this study is that, in classifying lesions as benign or malignant and recognizing “don’t-touch” entities, LLMs performed comparably to a junior radiologist but were outperformed by the senior radiologist. Our results uniquely show that LLMs can achieve sensitivity adequate for triage yet lack the nuanced judgment needed to avoid misclassifying benign or low-risk cases, underscoring the potential risks of unsupervised clinical use. To our knowledge, no prior study has evaluated LLMs across both text- and image-based tasks within a unified framework. For determining the most likely diagnosis, LLMs demonstrated poor performance, markedly inferior to both radiologists, reflecting limited higher-order reasoning for lesion-specific interpretation. The clear performance gradient between junior and senior radiologists further highlights the value of cumulative clinical experience and exposure to diverse case series. This study thus provides a novel benchmark of current LLM capabilities in radiographic diagnosis.

The broader AI literature in MSK oncology has been focused on deep learning and convolutional neural networks tailored for radiographic or MRI detection, which have demonstrated high accuracy in classifying bone tumours and assessing treatment response. Unlike these vision-specialized models, LLMs are generalists, not optimized for image pattern recognition [11,12]. Our findings highlight that although LLMs – core elements of modern AI – excel in text-based reasoning, they remain inferior to vision-specific AI models designed for direct lesion detection on radiographs. This gap stems from their architecture, which prioritizes general text-driven reasoning over image analysis. While recent multimodal developments have introduced visual encoders to merge perception with reasoning, image-based interpretation remains limited. With targeted fine-tuning and domain-specific training, multimodal LLMs could markedly improve their visual diagnostic capabilities beyond current text-centred performance.

This study has several limitations. First, the dataset was derived from a public source and may not reflect the full heterogeneity of bone tumours in clinical practice. Second, the analysis was limited to plain radiographs; while X-rays are primary diagnostic tools, incorporating CT and MRI could better represent real-world evaluation. Future studies should include multimodal imaging and larger, more diverse datasets. Third, only one standardized prompt per question was used, without assessing alternative phrasing effects. Since prompt design can affect LLM performance, further research should explore optimized prompting.

Finally, as LLMs evolve continuously, our results reflect their performance at a specific time point; newer versions may yield improved accuracy and clinical utility.

Conclusion. In conclusion, LLMs may serve as useful supportive tools for clinicians and radiologists in text-based tasks and in distinguishing between benign and malignant bone tumours. However, their diagnostic accuracy remains limited. These findings suggest that LLMs are better suited as expert-assisted support tools to enhance clinical workflow rather than as primary diagnostic instruments in bone tumour imaging.

REFERENCES

- [1] THIRUNAVUKARASU A. J., D. S. J. TING, K. ELANGOVAR et al. (2023) Large language models in medicine, *Nat. Med.*, **29**(8), 1930–1940.
- [2] MEDDEB A., S. LÜKEN, F. BUSCH et al. (2024) Large language model ability to translate CT and MRI free-text radiology reports into multiple languages, *Radiology*, **313**(3), e241736.
- [3] GONZÁLEZ-HUETE A., A. SALGADO-PARENTE, C. SUEVOS-BALLESTEROS et al. (2023) Radiographic evaluation of bone tumors, *Radiographics*, **43**(11), <https://doi.org/10.1148/rg.230048>.
- [4] ÇAMUR E., T. CESUR, Y. C. GUNES (2024) A comparative study: Can large language models beat radiologists on PI-RADSv2.1-related questions?, *J. Med. Biol. Eng.*, **44**(8), 821–830.
- [5] ERDURMUS O. Y., E. UYANIK, A. GURUF (2025) Diagnostic evaluation of computed tomography images in patients with suspected lung lesions using GPT-4o: A comparison of temperature settings and diagnostic accuracy, *C. R. Acad. Bulg. Sci.*, **78**(9), 1360–1370.
- [6] BOSSUYT P. M., J. B. REITSMA, D. E. BRUNS et al. (2015) STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies, *Radiology*, **277**(3), 826–832.
- [7] Bone tumor dataset. Available at: <https://www.kaggle.com/datasets/antimoni/bone-tumor>.
- [8] LI C. P., J. JAKOB, F. MENGE et al. (2024) Comparing ChatGPT-3.5 and ChatGPT-4’s alignments with the German evidence-based S3 guideline for adult soft tissue sarcoma, *iScience*, **27**(12), 111493.
- [9] WANG Y., X. WU, L. CARLSON, D. ONIANI (2024) Generative AI enhanced with NCCN clinical practice guidelines for clinical decision support: A case study on bone cancer, *J. Clin. Oncol.*, **42**(16_suppl), e13623–e13623.
- [10] YANG F., D. YAN, Z. WANG (2024) 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.*, **44**, 100525.
- [11] KANEMARU N., K. YASAKA, N. FUJITA et al. (2024) The fine-tuned large language model for extracting the progressive bone metastasis from unstructured radiology reports, *J. Imaging Inform. Med.*, **38**(2), 865–872.

- [12] YUAN D., H. ZHANG, L. TONG et al. (2025) Artificial intelligence for bone: Theory, methods, and applications, Adv. Intell. Discov., <https://doi.org/10.1002/aidi.202500070>.

¹*Department of Radiology, Ankara 29 Mayıs State Hospital, Ankara, Türkiye*
e-mail: eren.camur@outlook.com, <https://orcid.org/0000-0002-8774-5800>

²*Department of Radiology, Ankara Mamak State Hospital, Ankara, Türkiye*
e-mail: turaycesur93@gmail.com, <https://orcid.org/0000-0002-2726-8045>

³*Department of Radiology, Kırıkkale Yüksek İhtisas Hastanesi, Kırıkkale, Türkiye*
e-mail: gunesyasincelal@gmail.com, <https://orcid.org/0000-0001-7631-854X>

⁴*Department of Radiology, Ankara Bilkent City Hospital, Ankara, Türkiye*
e-mail: semraduran91@gmail.com, <https://orcid.org/0000-0003-0863-2443>