Diagnostic Performance and Confidence Calibration of Large Language Models for Bone Tumor Radiographs

Sanjana Arun B.S.[1], Eujung Park B.A.,[1] Katja Klosterman B.S.[1], Carissa Zhu B.S.[1] Ronak Arun B.S.[2] Palmer Wrigley[3], ,Hamsa Gangaswamiah M.D. [4]

[1]University of Arizona, College of Medicine Phoenix, Phoenix, AZ, United States

[2]University of Michigan, Ann Arbor, MI, United States

[3]Arizona State University,Tempe, AZ, United States

[4]Newport Liberty Medical Center, Jersey City NJ, United States

*Corresponding author(s). E-mail(s): sanjarun@arizona.edu
Contributing authors: eujungpark@arizon.edu ; katjakloste@arizona.edu ;

Abstract

Background/Objectives

Large language models (LLMs) are increasingly applied to medical image interpretation, yet their diagnostic accuracy and reliability in musculoskeletal radiology remain uncertain. This study evaluates the diagnostic performance and confidence calibration of LLMs in detecting and classifying bone tumors on radiographs.

Methods

A dataset of 257 radiographs with confirmed diagnoses was obtained from Radiopaedia, including normal studies and benign and malignant bone tumors. Three LLMs (ChatGPT, X-ray interpreter GPT-4.1, and X-ray interpreter Gemini) evaluated each image using a standardized prompt assessing abnormality detection, tumor detection, classification, and confidence. Outcomes included diagnostic accuracy, false positive abnormality rates, tumor hallucination rates (tumor identification in normal radiographs), and confidence calibration.

Results

Abnormality detection was high across models, with Gemini demonstrating the highest sensitivity (up to 100%). Tumor detection was strongest in lesions with characteristic features, including osteosarcoma and osteochondroma. X-ray interpreter GPT-4.1 achieved the highest primary diagnostic accuracy for osteosarcoma (80%), while ChatGPT 5.3 performed best in benign lesions, with highest diagnostic accuracy in osteochondroma (84.6%) and non-ossifying fibroma (76.9%).

Tumor subtype classification was limited across all models. Performance was poorest for Ewing sarcoma, with 0% primary diagnostic accuracy in ChatGPT 5.3 and X-ray interpreterGPT-4.1 and 10.3% in X-ray interpreter Gemini. Chondrosarcoma classification was also low (4–32%).

False positive abnormality rates were highest in X-ray interpreter GPT-4.1 (40.7%), followed by X-ray interpreter Gemini (25.9%) and ChatGPT 5.3(13.5%). Tumor hallucination occurred only in X-ray interpreter Gemini (12.3%) and was absent in the other models. All models demonstrated confidence miscalibration, with higher confidence in incorrect predictions.

Conclusions

LLMs detect abnormalities but have limited accuracy in tumor subtype classification, particularly for Ewing sarcoma. Performance is strongest in benign lesions and weakest in diagnostically challenging tumors. High false positive rates and overconfidence—especially in GPT-4.1—limit clinical use, supporting a role as adjunctive rather than independent tools.

Keywords: artificial intelligence, radiographs, large language models, tumor classification, bone tumor

Figure 3. Overall Model Performance and Confidence (Corrected)