Uludağ Üniversitesi Tıp Fakültesi Dergisi
Journal of Uludağ University Medical Faculty
52, 1804768, 2026
DOI: https://doi.org/10.32708/uutfd.1804768

ORIGINAL RESEARCH

Diagnostic Performance of Multimodal Large Language Models in Assigning Bone-RADS Categories on CT

Hasan Emin KAYA[1], Abdullah Enes ATAŞ[2]

[1] Bursa Uludağ University, School of Medicine, Department of Radiology, Bursa, Türkiye.
[2] Necmettin Erbakan University, Meram School of Medicine, Department of Radiology, Konya, Türkiye.

ABSTRACT

The aim of the study is to assess the performance of multimodal large language models (MLLMs) in assigning Bone-RADS categories to bone lesions identified on CT images. An MSK radiologist selected one representative slice for 50 bone lesions seen on CT studies and assigned reference Bone-RADS categories using clinical records. Three raters categorized each case: an abdominal radiologist, OpenAI ChatGPT 5, and Google Gemini 2.5 Pro. Accuracy was defined as the correctly labeled Bone-RADS 1 and 4 cases and compared using McNemar test. Agreement with the reference was assessed using weighted Cohen's κ with 95% CIs; pairwise κ differences were tested via bootstrap. Reference categories were Bone-RADS 1, n=23; 2, n=4; 3, n=0; 4, n=23. Accuracy was 84.8% (39/46) for the radiologist, 78.3% (36/46) for Gemini, and 65.2% (30/46) for ChatGPT. The radiologist outperformed ChatGPT (p=0.012); differences between the radiologist vs Gemini (p=0.604) and Gemini vs ChatGPT (p=0.360) were not significant. The radiologist achieved the highest agreement with the reference standard (κ = 0.715, 95% CI: [0.543-0.887]), followed by Gemini (κ = 0.542, 95% CI: [0.313-0.770]) and ChatGPT (κ = 0.292, 95% CI: [0.104-0.479]). Bootstrap comparisons showed that the radiologist's κ was higher than ChatGPT's (95% CI for difference, 0.140-0.675), while radiologist vs Gemini (−0.113-0.434) and Gemini vs ChatGPT (−0.041-0.522) were not significant. In conclusion, general-purpose MLLMs cannot yet replace trained radiologists for Bone-RADS classification, though they may still aid routine clinical practice.

Keywords: MLLMs. LLMs. Bone-RADS. ChatGPT. Gemini.

Multimodal Büyük Dil Modellerinin BT'de Bone-RADS Klasifikasyonundaki Tanısal Performansı

ÖZET

Çalışmamızın amacı multimodal büyük dil modellerinin (MBDM) BT görüntülerinde tespit edilen soliter kemik lezyonlarına Bone-RADS kategorileri atamadaki performanslarını değerlendirmektir. Hastanemizin PACS'ı soliter kemik lezyonu içeren BT tetkikleri için taranmıştır (Ağustos 2024-Ağustos 2025). Her lezyon için bir kas-iskelet radyoloğu tarafından kitleyi en iyi temsil eden bir kesit seçilmiş ve lezyonlara birer referans Bone-RADS skoru atanmıştır. Daha sonra bir abdominal radyolog, ChatGPT 5 ve Gemini 2.5 Pro aynı vakaları kategorilemiştir. Doğruluk, doğru şekilde kategorize edilen Bone-RADS 1 ve 4 vakaları olarak tanımlanmış ve McNemar testi kullanılarak karşılaştırılmıştır. Referansla uyum, ağırlıklı Cohen κ katsayısı kullanılarak değerlendirilmiş ve bootstrap yöntemi ile karşılaştırılmıştır. Referans kategorileri şu şekilde belirlenmiştir: Bone-RADS 1, n=23; 2, n=4; 3, n=0; 4, n=23. Doğruluk, radyolog için %84,8 (39/46), Gemini için %78,3 (36/46) ve ChatGPT için %65,2 (30/46) olarak bulunmuştur. Radyoloğun, ChatGPT'den daha iyi performans gösterdiği (p=0,012); radyolog ile Gemini (p=0,604) ve Gemini ile ChatGPT (p=0,360) arasındaki farkların anlamlı olmadığı görülmüştür. Radyolog, referans standardı ile en yüksek uyumu elde etmiş (κ = 0,715, %95 GA: [0,543-0,887]), bunu Gemini (κ = 0,542, %95 GA: [0,313-0,770]) ve ChatGPT (κ = 0,292, %95 GA: [0,104-0,479]) izlemiştir. Bootstrap ile yapılan karşılaştırmalar, radyoloğun κ değerinin ChatGPT'den daha yüksek olduğunu göstermiş (%95 GA: 0,140-0,675), ancak radyolog ile Gemini (%95 GA: −0,113-0,434) ve Gemini ile ChatGPT (%95 GA: −0,041-0,522) arasındaki fark anlamlı bulunmamıştır. Sonuç olarak genel amaçlı MBDM'ler henüz Bone-RADS kategorizasyonu için eğitimli radyologların yerini tutabilecek durumda görünmemekle beraber bu modellerin günlük pratikte radyologlara yardımcı olabileceği düşünülmektedir.

Anahtar Kelimeler: MBDM. BDM. Bone-RADS. ChatGPT. Gemini.

Date Received: 16.October.2025
Date Accepted: 18.December.2025

Dr. Hasan Emin KAYA
Bursa Uludağ University, School of Medicine, Department of Radiology, Görükle Campus, 16059, Bursa, Türkiye
hasaneminkaya@gmail.com

AUTHORS' ORCID INFORMATION
Hasan Emin KAYA: 0000-0002-7411-4102
Abdullah Enes ATAŞ: 0000-0001-6623-3024

With the increasing availability of imaging examinations and reduced costs, radiologists are encountering a growing number of incidental bone lesions in routine clinical practice[1]. Distinguishing bone lesions that can be safely regarded as "don't touch" from those requiring further work-up or tissue diagnosis is clinically important. Because incidental bone lesions are often interpreted by non-musculoskeletal (MSK) radiologists, structured systems that assist in their diagnostic management

may be particularly valuable. To address this need, several reporting and data systems (RADS) for bone lesions have been developed to enhance inter-reader agreement, improve diagnostic performance, and facilitate communication across clinical disciplines. These include the Radiological Evaluation Score for Bone Tumors (REST)[2], the American College of Radiology (ACR) Bone-RADS[3], and the Osseous Tumor Reporting and Data System (OT-RADS)[4]. The most recent addition is the Society of Skeletal Radiology (SSR) Bone-RADS, designed for use in classifying bone lesions encountered on CT or MRI examinations[5]. The SSR Bone-RADS categories are as follows: Bone-RADS-1, likely benign, no further work-up; Bone-RADS-2, incompletely assessed, requiring an alternative imaging modality; Bone-RADS-3, indeterminate, warranting follow-up imaging; and Bone-RADS-4, suspicious for malignancy or requiring treatment, with referral to orthopedic oncology and consideration of biopsy. Several studies have evaluated the reproducibility and effectiveness of SSR Bone-RADS for both CT and MRI[6,7].

The role of large language models (LLMs) in radiology reporting has been investigated extensively. Prior work has assessed their performance for various tasks such as generating impressions for radiology reports[8], simplifying interventional radiology reports for patients[9], and assigning RADS categories[10,11]. Multimodal LLMs (MLLMs) with image interpretation capabilities are also being evaluated for their potential applications in radiology. One area of assessment is their performance in assigning RADS categories to radiologic images. Recent studies have explored the use of MLLMs in assigning BI-RADS categories for breast ultrasound, and TI-RADS categories for thyroid ultrasound examinations[12,13]. The aim of the current study is to evaluate the performance of two widely used MLLMs in assigning SSR Bone-RADS categories to bone lesions identified on CT images. If demonstrated to be effective for this task, MLLMs may serve as a valuable adjunct in routine radiology practice, particularly for radiologists without subspecialty training in musculoskeletal imaging.

Material and Method

Following approval from our institutional ethics committee (Decision number: 2025/865/16-16), we queried our department's picture archiving and communication system for CT examinations demonstrating solitary bone lesions between August 2024 and August 2025. For each lesion, a single representative image was selected by a musculoskeletal (MSK) radiologist with 5 years of experience. The optimal plane and slice were chosen to best depict key characteristics (e.g., cortical disruption, pathologic fracture, matrix). An SSR Bone-RADS category was assigned to each image by the same MSK radiologist, who had access to the patients' electronic medical records; these assignments served as the reference standard. Subsequently, three readers performed Bone-RADS categorization on the images: an abdominal radiologist with 2 years of experience, OpenAI's ChatGPT (ChatGPT 5), and Google's Gemini (Gemini 2.5 Pro). For the MLLMs, a zero-shot prompt derived from the flowchart in the SSR Bone-RADS white paper (5) was used (see Supplementary File). Models were provided with the image and relevant clinical information. All queries were conducted through the models' native user interfaces, initiating a new chat session for each case.

Statistical analyses

Accuracy was defined as the proportion of cases with reference labels Bone-RADS 1 or 4 that were correctly classified. Cases labeled Bone-RADS 2 or 3 were excluded because their classification may vary depending on the reader's experience and judgment, as well as the patient's attitude and socioeconomic background[6]. Pairwise comparisons between the radiologist and each model were performed using two-sided McNemar's exact tests with Bonferroni correction for multiple comparisons. To evaluate the inter-rater agreement between each rater and the reference standard, we calculated the weighted Cohen's kappa. This metric was chosen as appropriate for ordinal data like the Bone-RADS classification, penalizing larger disagreements more heavily than smaller ones. Kappa statistics was interpreted as follows: poor, < 0.20; fair, 0.20–0.40; moderate. 0.40–0.60; good, 0.60–0.80; and excellent, ≥ 0.80[14]. To compare the performance between raters, we employed a non-parametric bootstrapping procedure. We generated bootstrap replicates to estimate the bias-corrected and accelerated 95% confidence intervals for the pairwise differences between the kappa coefficients. A statistically significant difference in performance was determined if the 95% confidence interval for the difference did not contain zero. Statistical analyses were performed using R (version 4.5.1, R Foundation for Statistical Computing, Vienna, Austria) and SPSS (version 29.0, IBM Corp., Armonk, NY, USA).

Results

Among 50 patients, 20 were female; the mean age was 47.4 years (range, 20–82). Lesion characteristics are summarized in Table I. Reference Bone-RADS classifications made by the MSK radiologist were Bone-RADS 1, n=23; Bone-RADS 2, n=4; Bone-RADS 3, n=0; and Bone-RADS 4, n=23.

Table I. Lesion characteristics.

Diagnosis	n
Metastasis	7
Non-ossifying fibroma/fibrous cortical defect	7
Subchondral cyst/geode	6
Giant cell tumor	3
Simple bone cyst	3
Osteoid osteoma	3
Bone island	2
Enchondroma	2
Lipoma	2
Aneurysmal bone cyst	2
Multiple myeloma	2
Chondromyxoid fibroma	1
Lymphoma	1
Osteochondral lesion	1
Chondrosarcoma	1
Chordoma	1
Atypical cartilaginous tumor	1
Fibrous dysplasia	1
Paget's disease	1
Hemangioma	1
Osteomyelitis	1
Bone infarction	1

The accuracy was 39/46 (84.78%), 36/46 (78.26%), and 30/46 (65.22%) for the radiologist, Gemini, and ChatGPT, respectively. Pairwise McNemar tests showed the radiologist significantly outperformed ChatGPT ($p = 0.012$), whereas differences between the radiologist and Gemini ($p = 0.604$) and between Gemini and ChatGPT ($p = 0.360$) were not significant.

The radiologist achieved the highest agreement with the reference standard ($\kappa = 0.715$, 95% CI: [0.543-0.887]), followed by Gemini ($\kappa = 0.542$, 95% CI: [0.313-0.770]) and ChatGPT ($\kappa = 0.292$, 95% CI: [0.104-0.479]) (Table II) (Figure 1). To determine if these performance differences were statistically significant, we conducted a bootstrap comparison with 2000 replicates. The analysis revealed that the radiologist's performance was statistically significantly superior to ChatGPT's (95% CI for the difference: [0.140-0.675]). However, the observed differences between the radiologist and Gemini (95% CI: [−0.113-0.434]) and between Gemini and ChatGPT (95% CI: [−0.041-0.522]) were not statistically significant (Table III).

Table II. Weighted Cohen's kappa for inter-rater agreement with the reference standard. (CI = confidence interval)

Rater	Weighted kappa (95% CI)
Radiologist	0.715 (0.543, 0.887)
Gemini	0.542 (0.313, 0.770)
ChatGPT	0.292 (0.104, 0.479)

Table III. Bootstrap comparison of weighted kappa coefficients. (CI = confidence interval)

Comparison	Difference in kappa (Δκ)	Bootstrap 95% CI for difference
Radiologist vs. Gemini	0.173	[−0.113, 0.434]
Radiologist vs. Gemini	0.250	[−0.041, 0.522]
Radiologist vs. ChatGPT	0.423	[0.140, 0.675]

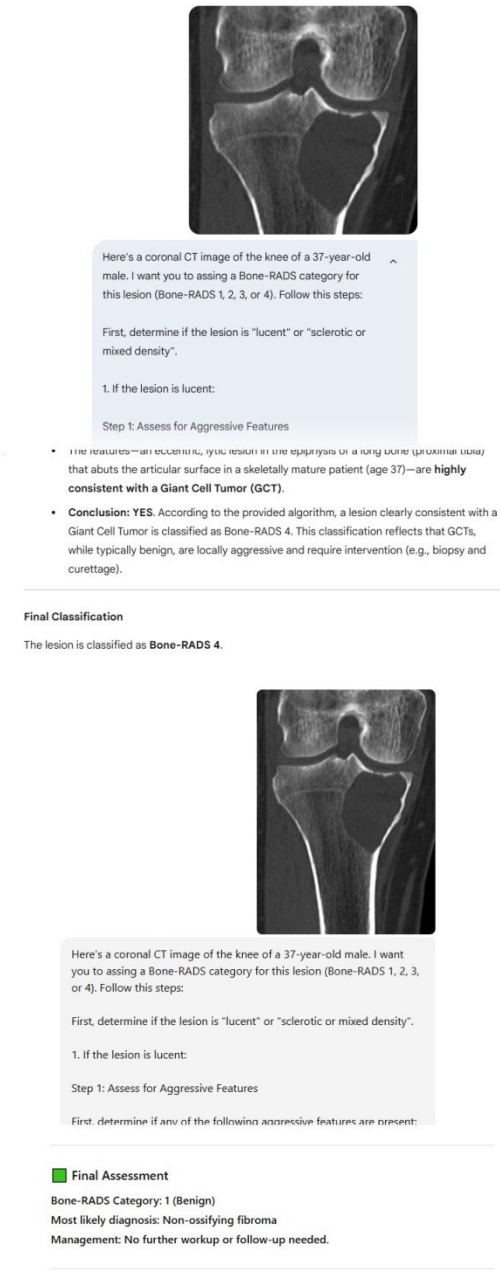

Figure 1.
Coronal CT image of the knee in a 37-year-old man demonstrating a proximal tibial lesion consistent with a giant cell tumor. Both the radiologist and Gemini (a) correctly characterized the lesion and categorized it as Bone-RADS 4, whereas ChatGPT (b) misclassified it as a non-ossifying fibroma and categorized the lesion as Bone-RADS 1.

Discussion and Conclusion

In this study, the non-MSK radiologist showed good agreement with the MSK radiologist for Bone-RADS categorization of solitary bone lesions on CT, whereas agreement was fair for Gemini and poor for ChatGPT. Several RADS frameworks have been developed to standardize reporting and management across specialties (e.g., BI-RADS, LI-RADS, and PI-RADS)[15]. Similar tools exist for bone tumors. The REST system scores radiographic lesions across eight features, with scores <3 suggesting benignity and ≥3 suggesting malignancy[2]. The ACR Bone-RADS assesses margin, periosteal reaction, endosteal erosion, pathologic fracture, extraosseous soft-tissue mass, and history of primary cancer; total score determines the category, with ≥7 indicating a lesion highly suspicious for malignancy[3]. OT-RADS, conceptually analogous to BI-RADS, applies to MRI and includes six categories: 0 (incomplete), I (negative), II (definitely benign), III (probably benign), IV (suspicious/ indeterminate), V (highly suggestive of malignancy), and VI (known/recurrent malignancy)[4]. The most recent system, the SSR Bone-RADS, can be applied to CT or MR[5]. A key distinction is its management-oriented approach: Bone-RADS 4 denotes the need for referral or treatment and does not necessarily imply malignancy. For example, a simple bone cyst with cortical involvement and risk of pathologic fracture would be classified as Bone-RADS 4 to prompt orthopedic evaluation. The algorithm for CT first stratifies lesions by density (lucent vs sclerotic/mixed) using separate flowcharts. The presence of aggressive features (cortical involvement, soft-tissue extension, or pathologic fracture) assigns Bone-RADS 4. In their absence, additional factors such as history of a malignancy with osseous metastatic propensity, lesion density and matrix, or typical appearances (e.g., fibrous dysplasia, non-ossifying fibroma, osteoid osteoma, giant cell tumor) guide categorization. Validation studies suggest high sensitivity and good interreader reliability for lucent lesions but lower specificity, potentially reflecting dependence on clinical features such as pain and oncologic history[16]. Another CT-based validation found the system useful for triaging lesions that may require treatment, with moderate interreader agreement[6].

MLLMs, capable of processing images in addition to text, are an active area of investigation in radiology[17]. Prior work has evaluated performance on RSNA 2023 Case of the Day Questions[18], New England Journal of Medicine Image Challenge Cases[19] or Japanese Diagnostic Radiology Board Examinations[20]. A 2024 report indicated limited accuracy for pediatric imaging[21]. However, in a 2025 study, it was reported that MLLMs showed improvements in answering image-based quiz questions over the course of one year[22]. Evidence specific to RADS assignment remains sparse. For example, ChatGPT 4 showed mixed accuracy for TI-RADS on ultrasound, high for low-risk nodules (93.6%) but lower for high-risk nodules (42.1%)[13]. Among the five MLLMs tested for BI-RADS categorization on breast ultrasound images, Claude 3.5 Sonnet achieved the highest accuracy (59%)[12]. To the best of our knowledge, no prior study has evaluated the performance of MLLMs in assigning Bone-RADS categories on CT images. Therefore, direct comparison of model performance in this specific task is not feasible. Nonetheless, the overall accuracy observed in our study appears comparable to that reported in previous studies involving other RADS classification systems.

This study has several limitations. First, the small sample size and retrospective design may limit the generalizability of our findings. Second, using a single representative slice per lesion may not reflect the full performance of either human readers or MLLMs; however, we carefully selected slices that most clearly demonstrated the characteristic features of each lesion. Third, the evaluated models were not fine-tuned for radiologic interpretation; while specialized models exist, they are not universally accessible. Fourth, Bone-RADS 2 and 3 lesions were under-represented. Notably, prior studies focused on establishing reliable reference standards for Bone-RADS 1 and 4, underscoring the system's emphasis on confidently distinguishing clearly benign lesions from those requiring further evaluation or intervention[6].

In conclusion, ChatGPT and Gemini, as general-purpose MLLMs, demonstrated inferior agreement with the MSK radiologist compared with a non-MSK radiologist for Bone-RADS categorization of solitary bone lesions encountered on CT when limited to single-slice inputs. While MLLMs are advancing rapidly, they are not yet substitutes for trained radiologists in this task. Despite their current limitations, these models may provide valuable assistance in routine clinical practice. Larger, prospective studies, ideally using full image stacks and fine-tuned models, are needed to assess their performance and clinical utility more definitively.

Researcher Contribution Statement:
Idea and design: H.E.K, A.E.A.; Data collection and processing: H.E.K, A.E.A.; Analysis and interpretation of data: H.E.K, A.E.A.; Writing of significant parts of the article: H.E.K, A.E.A.
Support and Acknowledgement Statement:
ChatGPT 5 was used to improve the grammar and readability of the manuscript.
Conflict of Interest Statement:
The authors of the article have no conflict-of-interest declarations.
Ethics Committee Approval Information:
Approving Committee: Bursa Uludağ University School of Medicine Health Research Ethics Board
Approval Date: 24.09.2025
Decision No: 2025/864/16-16

References

1. Blackburn CW, Richardson SM, Devita RR, et al. What Is the Prevalence of Clinically Important Findings Among Incidentally Found Osseous Lesions? *Clin Orthop Relat Res*. 2023;481(10):1993-2002. doi:10.1097/CORR.0000000000002630
2. Salunke AA, Nandy K, Puj K, et al. A proposed "Radiological Evaluation Score for Bone Tumors" (REST): An objective system for assessment of a radiograph in patients with suspected bone tumor. *Musculoskelet Surg*. 2022;106(4):371-382. doi:10.1007/S12306-021-00711-0
3. Caracciolo JT, Ali S, Chang CY, et al. Bone Tumor Risk Stratification and Management System: A Consensus Guideline from the ACR Bone Reporting and Data System Committee. *Journal of the American College of Radiology*. 2023;20(10):1044-1058. doi:10.1016/j.jacr.2023.07.017
4. Chhabra A, Gupta A, Thakur U, et al. Osseous Tumor Reporting and Data System-Multireader Validation Study. *J Comput Assist Tomogr*. 2021;45(4):571-585. doi:10.1097/RCT.0000000000001184
5. 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. doi:10.1007/S00256-022-04022-8
6. Xing Y, Ding D, Dai S, et al. Bone reporting and data system on CT (Bone-RADS-CT): a validation study by four readers on 328 cases from three local and two public databases. *Insights Imaging*. 2025;16(1):174. doi:10.1186/S13244-025-02057-8
7. Xing Y, Hu Y, Liu X, et al. Bone Reporting and Data System on MRI (Bone-RADS-MRI): a validation study by four readers on 275 cases from three local and two public databases. *Insights Imaging*. 2025;16(1). doi:10.1186/S13244-025-02040-3
8. Doshi R, Amin KS, Khosla P, Bajaj S, Chheang S, Forman HP. Quantitative Evaluation of Large Language Models to Streamline Radiology Report Impressions: A Multimodal Retrospective Analysis. *Radiology*. 2024;310(3). doi: 10.1148/radiol.231593.
9. Can E, Uller W, Vogt K, et al. Large Language Models for Simplified Interventional Radiology Reports: A Comparative Analysis. *Acad Radiol*. Published online September 30, 2024. doi:10.1016/J.ACRA.2024.09.041
10. Bhayana R, Jajodia A, Chawla T, et al. Accuracy of Large Language Model–based Automatic Calculation of Ovarian-Adnexal Reporting and Data System MRI Scores from Pelvic MRI Reports. *Radiology*. 2025;315(1). doi: 10.1148/radiol.241554.
11. Arnold PG, Russe MF, Bamberg F, et al. Performance of large language models for CAD-RADS 2.0 classification derived from cardiac CT reports. *J Cardiovasc Comput Tomogr*. 2025;0(0). doi:10.1016/j.jcct.2025.03.007
12. Güneş YC, Cesur T, Çamur E, Günbey Karabekmez L. Evaluating text and visual diagnostic capabilities of large language models on questions related to the Breast Imaging Reporting and Data System Atlas 5th edition. *Diagn Interv Radiol*. 2025;31(2):111-129. doi:10.4274/DIR.2024.242876
13. Cabezas E, Toro-Tobon D, Johnson T, et al. ChatGPT-4's Accuracy in Estimating Thyroid Nodule Features and Cancer Risk From Ultrasound Images. *Endocrine Practice*. 2025;31(6):716-723. doi:10.1016/j.eprac.2025.03.008
14. Gisev N, Bell JS, Chen TF. Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*. 2013;9(3):330-338. doi:10.1016/j.sapharm.2012.04.004
15. An JY, Unsdorfer KML, Weinreb JC. BI-RADS, C-RADS, CAD-RADS, LI-RADS, Lung-RADS, NI-RADS, O-RADS, PI-RADS, TI-RADS: Reporting and Data Systems. *Radiographics*. 2019;39(5):1435. doi:10.1148/RG.2019190087
16. 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(2):209-217. doi:10.1007/S00256-024-04721-4
17. Shen Y, Xu Y, Ma J, et al. Multi-modal large language models in radiology: principles, applications, and potential. *Abdominal Radiology*. 2025;50(6):2745-2757. doi:10.1007/S00261-024-04708-8/FIGURES/2
18. Mukherjee P, Hou B, Suri A, et al. Evaluation of GPT large language model performance on RSNA 2023 case of the day questions. *Radiology*. 2024;313(1). doi:10.1148/RADIOL.240609
19. Suh PS, Shim WH, Suh CH, et al. Comparing Large Language Model and Human Reader Accuracy with New England Journal of Medicine Image Challenge Case Image. *Radiology*. 2024;313(3). doi:10.1148/RADIOL.241668
20. Nakaura T, Yoshida N, Kobayashi N, et al. Performance of Multimodal Large Language Models in Japanese Diagnostic Radiology Board Examinations (2021-2023). *Acad Radiol*. 2025;32(5):2394-2401. doi:10.1016/j.acra.2024.10.035
21. Reith TP, D'Alessandro DM, D'Alessandro MP. Capability of multimodal large language models to interpret pediatric radiological images. *Pediatr Radiol*. 2024;54(10):1729-1737. doi:10.1007/S00247-024-06025-0
22. Hou B, Mukherjee P, Batheja V, Wang KC, Summers RM, Lu Z. One Year On: Assessing Progress of Multimodal Large Language Model Performance on RSNA 2024 Case of the Day Questions. *Radiology*. 2025;316(2). doi:10.1148/RADIOL.250617