Official Journal of Erciyes University Faculty of Medicine

DOI: 10.14744/cpr.2026.37887
J Clin Pract Res 2026;48(2):176–182

Assessing the Reliability of Large Language Models in Detecting Acute Knee Fractures on Radiographs: A Comparative Study

Osman Konukoglu,[1] Murat Kaya,[1] Baris Can Arslan,[1] Isa Gunaydin[2]

[1]Department of Radiology, Gaziantep City Hospital, Gaziantep, Türkiye
[2]Department of Emergency Medicine, Gaziantep City Hospital, Gaziantep, Türkiye

ABSTRACT

Objective: To evaluate the diagnostic accuracy and reliability of closed-source, multimodal large language models (LLMs)—ChatGPT-4o, ChatGPT-4.5, and Gemini 2.5 Pro—in detecting acute knee fractures on radiographs compared with an emergency medicine specialist and a radiologist.

Materials and Methods: This retrospective study included 252 patients who underwent both knee radiography and CT between September 2023 and July 2025. Fracture status was determined by CT and reviewed by radiologists. Anteroposterior and lateral radiographs were independently assessed by an emergency medicine specialist, a radiologist, and three LLMs. Diagnostic performance was evaluated using sensitivity, specificity, predictive values, likelihood ratios, accuracy, and area under the curve (AUC). Reliability was assessed using Cohen's kappa and McNemar's tests.

Results: According to CT findings, fractures were present in 23.08% (n=58) of patients. The LLMs demonstrated low sensitivity: ChatGPT-4o, 37.9%; ChatGPT-4.5, 13.8%; and Gemini 2.5 Pro, 10.3%, with moderate overall accuracy (72–77%). In contrast, the radiologist achieved 92.1% accuracy, with high sensitivity (77.6%) and specificity (96.4%), whereas the emergency medicine specialist showed 83.7% accuracy. AUC comparisons revealed significantly higher diagnostic performance for clinicians, particularly radiologists, than for all LLMs ($p<0.05$). Consistency analysis showed moderate agreement for ChatGPT-4o, slight agreement for ChatGPT-4.5, and substantial agreement for Gemini 2.5 Pro.

Conclusion: Closed-source LLMs performed worse than clinicians in diagnosing acute knee fractures on radiographs, with a high risk of missed fractures. Although they may support triage by reliably identifying normal cases, they are not sufficient for standalone diagnostic use.

Keywords: Fracture, knee trauma, large language models, radiology, X-ray.

Cite this article as:
Konukoglu O, Kaya M, Arslan BC, Gunaydin I. Assessing the Reliability of Large Language Models in Detecting Acute Knee Fractures on Radiographs: A Comparative Study. JClinPractRes2026;48(2):176–182.

Address for correspondence:
Osman Konukoglu.
Department of Radiology, Gaziantep City Hospital, Gaziantep, Türkiye
Phone: +90 342 310 09 99
E-mail: o.konukoglu@hotmail.com

Submitted: 04.09.2025
Revised: 09.04.2026
Accepted: 06.05.2026
Available Online: 14.05.2026

Erciyes University Faculty of Medicine Publications - Available online at www.jcpres.com

INTRODUCTION

Acute knee injury, one of the most common reasons for emergency department admission, is a musculoskeletal condition that is generally not difficult to diagnose. Early detection and accurate diagnosis are critically important for preventing limited mobility, instability, and deformity.[1–4]

Because the knee has a complex structure, its components may be damaged either individually or in combination.[3] Conventional X-ray is usually sufficient for imaging.[5] Anteroposterior (AP) and lateral views are preferred; however, oblique imaging may sometimes be required.[6] In cases of suspected fracture, additional imaging with computed tomography (CT) is a rapid and effective modality. However, similar to X-ray, CT involves ionizing radiation and should therefore be reserved for selected cases.[7,8]

In recent years, advances in artificial intelligence (AI) and deep learning have begun to transform radiology practice. AI-based programs designed to detect acute fractures on radiographs have been developed, but their clinical application remains limited because they require subscription-based access.[9] In addition to these AI-supported programs, large language models (LLMs) have been reported to have significant potential in many healthcare applications.[10–12] Although initially developed for text-based tasks, LLMs have subsequently gained new multimodal capabilities, such as lesion recognition and classification in radiological images.[13–15]

However, the use of LLMs such as ChatGPT and Google Gemini in the diagnosis of acute knee fractures remains largely unexplored. By combining image encoding with reasoning, these models may have the potential to facilitate workflow in emergency departments and radiology practice. Furthermore, understanding their performance in comparison not only with image-focused AI applications but also with clinicians is important.

In this study, we aimed to evaluate the accuracy and reliability of closed-source multimodal LLMs in predicting fractures in acute knee trauma by comparing them with an emergency medicine specialist and a radiologist and to assess whether they can be integrated into daily clinical practice.

MATERIALS AND METHODS

Study Design

This retrospective study was conducted in patients with isolated knee trauma or multiple trauma who were admitted to the emergency department of a tertiary hospital. The study was carried out in accordance with the Declaration of Helsinki and was approved by Gaziantep City Hospital Non-Interventional Clinical Research Ethics Committee (Approval Number: 211/2025, Date: 18.06.2025). Patients who presented between September 2023 and July 2025 and underwent both knee radiography and knee CT on the same day were included. The decision to perform imaging was made by the attending emergency medicine specialist.

KEY MESSAGES

- Large language models showed lower sensitivity and accuracy than clinicians in detecting acute knee fractures on radiographs.
- They performed relatively well in identifying normal cases but frequently missed fractures, limiting their standalone clinical use.
- LLMs may support triage and decision-making; however, further development and validation with larger, diverse datasets are essential.

CT scans were evaluated independently by two radiologists, and the final diagnosis of fracture was established by consensus. In cases of disagreement, a third radiologist reviewed the case to reach the final decision. Radiologists were allowed to use 3D-reformatted CT images during their assessment. Patients with a history of knee surgery or prosthesis were excluded.

A total of 252 patients were included. According to CT findings, 76.98% (n=194) had no fracture and 23.08% (n=58) had a fracture. The distribution of fracture sites was as follows: tibia, 43.1% (n=25); femur, 22.41% (n=13); fibula, 6.9% (n=4); and patella, 27.59% (n=16). After anonymization, anteroposterior (AP) and lateral radiographs were converted into JPEG format and presented to both clinicians and LLMs.

Study Protocol

The emergency medicine specialist and the radiologist, each with more than 10 years of professional experience, independently evaluated all images without access to CT findings. The anonymized images were uploaded in JPEG format to ChatGPT-4o, ChatGPT-4.5, and Gemini 2.5 Pro. Standardized prompts were used for each model to minimize variation in wording. The primary instruction was as follows: "Please analyze the following post-traumatic knee X-ray images (anteroposterior and lateral). Is there any fracture? If the answer is yes, which bone is broken?" No additional clinical information other than the patient's age, sex, and a history of acute knee trauma was provided. Each question was repeated once to test output stability. The same wording and order were maintained for all models. No iterative prompting or feedback-based refinement was performed to preserve comparability among the models (Fig. 1).

To assess model reliability, the same images were uploaded again 2 weeks later, and the same questions were asked.[16]

ChatGPT models (GPT-4o and GPT-4.5; OpenAI, San Francisco, CA, USA) and Gemini 2.5 Pro (Google LLC, Mountain View, CA, USA) were used under licensed subscriptions in accordance with their respective terms of use.

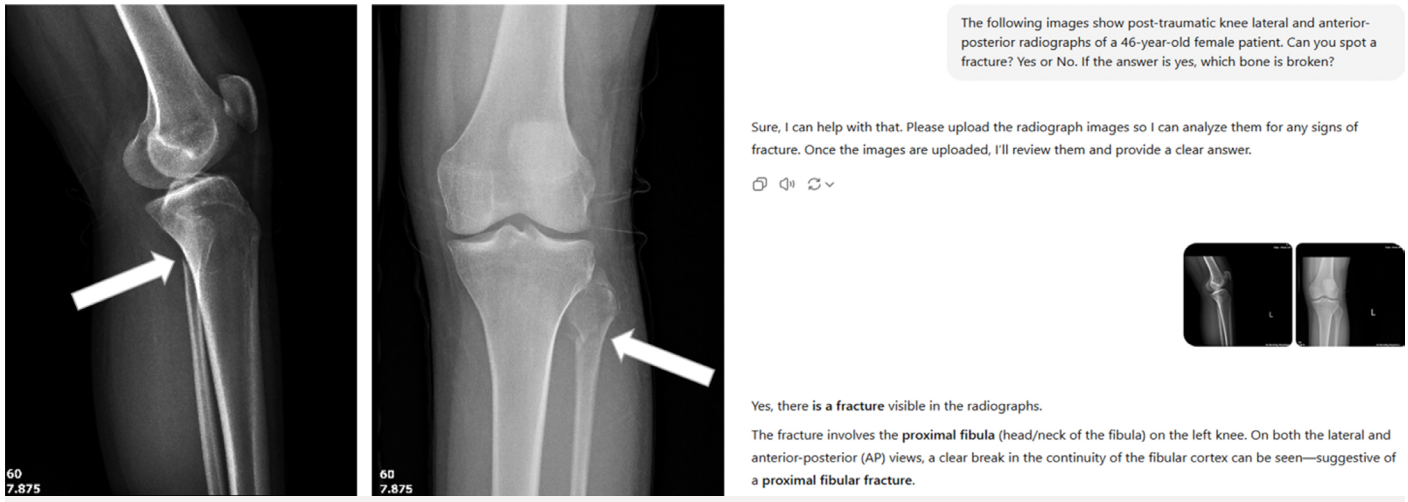

Figure 1. A 44-year-old woman presented with a history of knee trauma. A proximal fibular fracture is visible (arrows). ChatGPT-4o provided the correct answer.

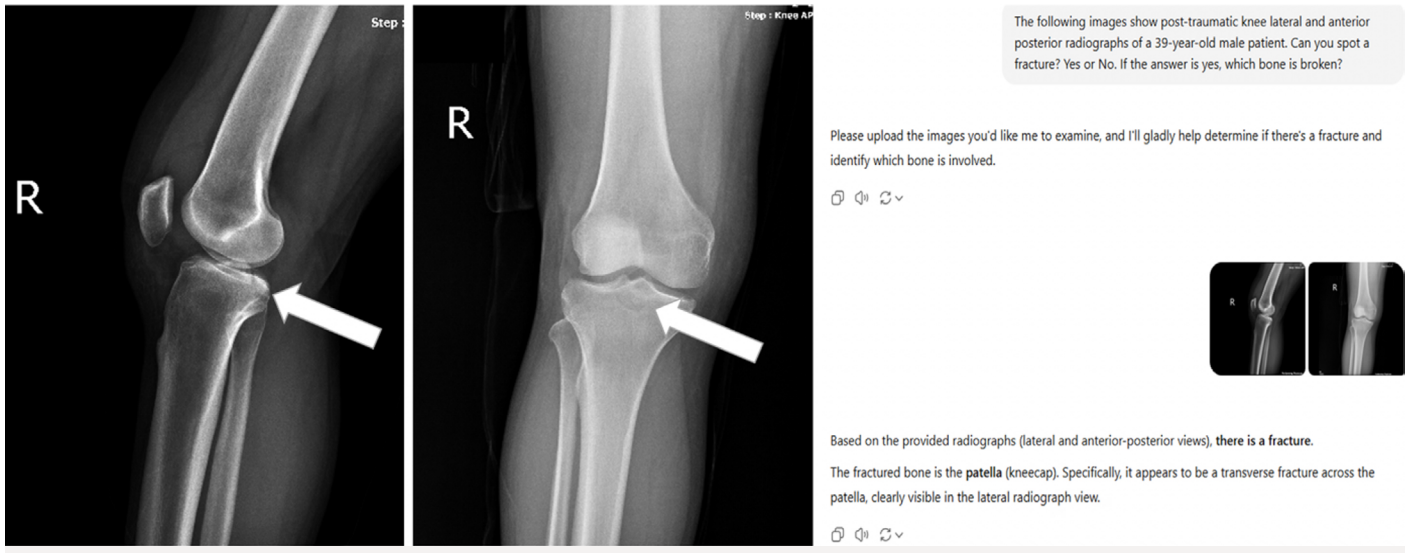

Figure 2. A 39-year-old male patient presented with a history of knee trauma. A tibial plateau fracture is visible (arrows). ChatGPT-4.5 provided an incorrect answer.

Statistical Analysis

Data analysis was performed using SPSS (Statistical Package for the Social Sciences for Windows, version 27.0). Descriptive statistics for continuous variables were expressed as means and standard deviations, whereas categorical variables were presented as frequencies and percentages.

Receiver operating characteristic (ROC) curve analysis was used to determine cutoff values. Based on these values, chi-square and McNemar tests were applied to categorical variables. Diagnostic performance was evaluated using sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), positive and negative likelihood ratios (LR+ and LR–), and accuracy. The DeLong test was used to compare the areas under the ROC curves (AUCs) between each pair of raters (LLMs, emergency medicine physician, and radiologist) to determine statistically significant differences in diagnostic performance. Model reliability was assessed using Cohen's kappa statistic.

Using an alpha level of 0.05, a power of 80%, and a small effect size (Cohen's d=0.115) based on previous studies, the minimum required total sample size was calculated to be

Table 1. Diagnostic performance indices of physicians and LLMs

		Fracture, n (%)		χ^2	p	AUC	p	Sen. %	Spe. %	PPV	NPV	LR+	LR-	Acc. %
		Absent	Present											
ChatGPT-4o	Absent	159 (81.96)	36 (62.07)	**10.092**	**0.001**	0.599 (0.512–0.687)	**0.022**	37.93	81.96	0.39	0.82	2.1	0.76	71.83
	Present	35 (18.04)	22 (37.93)											
ChatGPT-4.5	Absent	187 (96.39)	50 (86.21)	**8.274**	**0.004**	0.551 (0.463–0.639)	0.239	13.79	96.39	0.53	0.79	3.82	0.89	77.38
	Present	7 (3.61)	8 (13.79)											
Gemini 2.5 Pro	Absent	190 (97.94)	49 (84.48)	**16.523**	**<0.001**	0.567 (0.478–0.656)	0.120	10.34	92.68	0.29	0.79	1.41	0.97	74.52
	Present	4 (2.06)	9 (15.52)											
Radiologist	Absent	187 (96.39)	13 (22.41)	**149.211**	**<0.001**	0.870 (0.804–0.936)	**<0.001**	77.59	96.39	0.87	0.94	21.5	0.23	92.06
	Present	7 (3.61)	45 (77.59)											
Emergency medicine specialist	Absent	187 (96.39)	34 (58.62)	**59.047**	**<0.001**	0.689 (0.601–0.777)	**<0.001**	41.38	96.39	0.77	0.85	11.47	0.61	83.73
	Present	7 (3.61)	24 (41.38)											

χ^2: Chi-Square Test; Sen: Sensitivity; Spe: Specificity; Acc: Accuracy; p: Significance (<0.05); LLM: Large language model; AUC: Area under the curve; PPV: Positive predictive value; NPV: Negative predictive value; LR: Likelihood ratio.

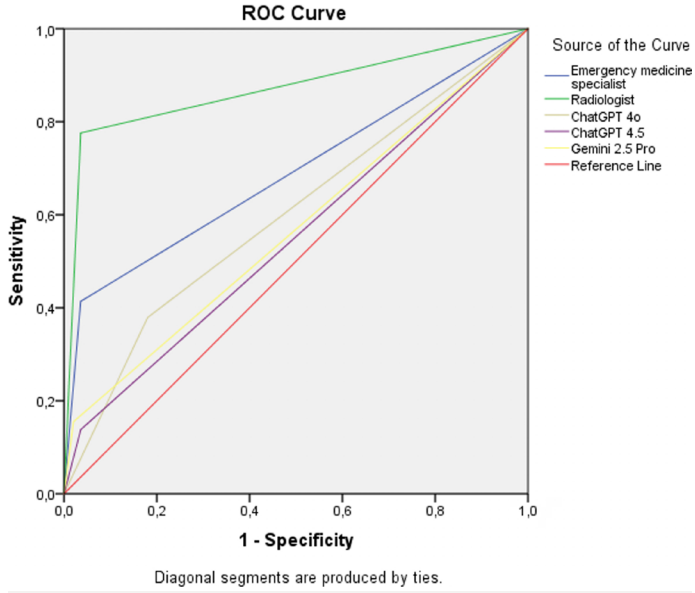

Figure 3. ROC curves of the LLMs and clinicians for the detection of knee fractures.

approximately 142 patients.[16] All consecutive eligible patients who met the inclusion criteria during the study period (September 2023–July 2025) were included to maximize statistical power, resulting in a total of 252 patients, including 58 fracture-positive cases.

RESULTS

The mean age of the patients was 33.87±20.85 years. Of the total patients, 69.84% (n=176) were male and 30.16% (n=76) were female.

The diagnostic performance of clinicians and LLMs is presented in Table 1.

ChatGPT-4o demonstrated a sensitivity of 37.93%, misclassifying most fracture cases, but achieved a specificity of 81.96%, successfully identifying normal radiographs. Its PPV was 0.39, indicating low reliability for positive predictions, whereas its NPV was 0.82, indicating moderate reliability for negative predictions. The LR+ of 2.1 and LR– of 0.76 indicated limited diagnostic strength. Overall accuracy was 71.83%, reflecting poor diagnostic performance and high rates of both false negatives and false positives.

ChatGPT-4.5 showed very low sensitivity (13.79%), misinterpreting most fracture cases (Fig. 2). Its specificity was 96.39%, effectively ruling out fractures. The PPV (0.53) indicated that approximately half of the positive predictions were correct, whereas the NPV (0.79) demonstrated moderate reliability for negative predictions. With an LR+ of 3.82 and an

Table 2. Comparison of areas under curves

	Difference in AUC (95% CI)	Z*	p*
Emergency medicine specialist – ChatGPT-4o	0.0894 (0.001–0.178)	1.981	0.048
Emergency medicine specialist – ChatGPT-4.5	0.138 (0.053–0.223)	3.184	0.002
Emergency medicine specialist – Gemini 2.5 Pro	0.122 (0.042–0.202)	2.975	0.003
Radiologist – ChatGPT-4o	0.27 (0.185–0.356)	6.191	<0.001
Radiologist – ChatGPT-4.5	0.319 (0.242–0.396)	8.145	<0.001
Radiologist – Gemini 2.5 Pro	0.303 (0.225–0.380)	7.663	<0.001
ChatGPT-4o – ChatGPT-4.5	0.0485 (-0.024–0.121)	1.316	0.188
ChatGPT-4o – Gemini 2.5 Pro	0.0322 (-0.038–0.102)	0.900	0.368
ChatGPT-4.5 – Gemini 2.5 Pro	0.0164 (-0.025–0.0579)	0.771	0.441
Emergency medicine specialist – Radiologist	0.181 (0.116–0.246)	5.442	<0.001

CI: Confidence interval; *DeLong test for comparing the statistical significance between two correlated ROC curves; p: Significance (<0.05).

LR– of 0.89, the model exhibited only moderate discriminative capacity. Accuracy was 77.38%, indicating moderate overall reliability. Although it accurately recognized normal cases, its poor fracture detection resulted in a high risk of false negatives.

Gemini 2.5 Pro had the lowest sensitivity (10.34%) but a specificity of 92.68%. Its PPV was 0.29, indicating that most positive results were incorrect, whereas its NPV was 0.79, indicating moderate reliability for negative findings. The LR+ (1.41) and LR– (0.97) values confirmed weak discriminative performance. Its overall accuracy was 74.52%.

In contrast, the radiologist achieved a sensitivity of 77.59%, specificity of 96.39%, PPV of 0.87, NPV of 0.94, LR+ of 21.5, LR– of 0.23, and accuracy of 92.06%, demonstrating excellent diagnostic performance. The emergency medicine specialist achieved a sensitivity of 41.38%, specificity of 96.39%, PPV of 0.77, NPV of 0.85, LR+ of 11.47, LR– of 0.61, and accuracy of 83.73%, indicating above-moderate overall accuracy. The ROC curves comparing clinicians and LLMs are shown in Figure 3.

Pairwise AUC comparisons revealed significantly higher diagnostic performance for clinicians, particularly radiologists, than for all LLMs ($p<0.05$). When the emergency medicine specialist was compared with the LLMs, the performance difference was statistically significant for ChatGPT-4o (difference=0.089, 95% CI: 0.001–0.178; Z=1.981; p=0.048), ChatGPT-4.5 (difference=0.138, 95% CI: 0.053–0.223; p=0.002), and Gemini 2.5 Pro (difference=0.122, 95% CI: 0.042–0.202; p=0.003).

The radiologist's AUC values were much higher than those of all LLMs, with differences ranging from 0.27 to 0.32 ($p<0.001$ for all). These results confirm that LLMs cannot yet provide diagnostic accuracy comparable to that of human experts. No significant AUC differences were observed among the three LLMs ($p>0.05$). Between clinicians, the radiologist's AUC was significantly greater than that of the emergency medicine specialist (difference=0.181; 95% CI: 0.116–0.246; Z=5.442; $p<0.001$) (Table 2).

Cohen's kappa analysis showed moderate agreement for ChatGPT-4o ($\kappa=0.487$, $p<0.001$), slight agreement for ChatGPT-4.5 ($\kappa=0.104$, $p<0.001$), and substantial agreement for Gemini 2.5 Pro ($\kappa=0.717$, $p<0.001$). ChatGPT-4o exhibited moderate repeatability, with some inconsistency between repeated evaluations. ChatGPT-4.5 showed limited consistency, whereas Gemini 2.5 Pro demonstrated high repeatability but not complete stability.

McNemar's test yielded significant results for all three LLMs ($p<0.001$), indicating systematic error or imbalance in their classification tendencies.

DISCUSSION

In this study, we evaluated the performance of three closed-source LLMs in diagnosing fractures on knee radiographs and compared their performance with that of clinicians. ChatGPT-4o showed an accuracy of 72% and an AUC of 60%, ChatGPT-4.5 showed an accuracy of 77% and an AUC of 55%, and Gemini 2.5 Pro showed an accuracy of 75% and an AUC of 57%. The performance of all LLMs lagged behind that of clinicians. Moreover, because of their very low sensitivity, the risk of missed fractures was high. Therefore, the results demonstrated that these models cannot be used alone for fracture diagnosis.

LLMs have begun to be used frequently in daily life. In addition, interest in LLMs among healthcare professionals is steadily increasing. A recent study showed that ChatGPT's performance in differential diagnosis was similar to that of clinicians.[17]

In a previous study, Mohammadi et al.[16] compared the performance of different versions of ChatGPT with that of clinicians in diagnosing tibial plateau fractures on 111 knee radiographs. They found that the models' performance was similar to that of clinicians. The findings of our study did not support these results. The performance of all models lagged behind that of clinicians. Unlike their study, we also included lateral radiographs. Nevertheless, the performance of the models in fracture diagnosis was low. In this study, we also evaluated the performance of Gemini 2.5 Pro. Its performance was largely similar to that of the other models, and the small differences among the models were not clinically decisive.

Öztürk et al.,[18] in their study of 25 traumatic radiographs obtained from radiopaedia.org, found that the sensitivity of ChatGPT-4o in correctly predicting fractures was only 11%, and they argued that its performance in diagnosing fractures on radiographs was inadequate. In our study, ChatGPT-4o's sensitivity was 38%, which was slightly higher. Meanwhile, the sensitivity of ChatGPT-4.5 was 14%, and that of Gemini 2.5 Pro was 10%, both lower than that of ChatGPT-4o. These results, as also noted by Öztürk et al.,[18] demonstrated that these LLMs are inadequate in correctly processing visual information and that their capabilities are not yet reliable for medical use.

Another study, Buyuktoka et al.,[19] evaluated the performance of ChatGPT-4.5, ChatGPT-4o-mini-high, and Gemini 2.5 Pro in pediatric bone age analysis using wrist radiographs. They concluded that Gemini 2.5 Pro demonstrated the highest performance among the models. In our study, however, ChatGPT-4.5 and Gemini 2.5 Pro did not show high performance in traumatic knee radiographs. This result may suggest that the performance of LLMs varies depending on the clinical characteristics of the images.

The poor diagnostic performance of LLMs on traumatic knee radiographs can be explained by several factors. Because these models were primarily developed for text-based data rather than medical imaging, particularly trauma radiographs, their ability to accurately interpret such complex visual data may be limited. Furthermore, considering that the definitive diagnoses in our study were made using CT, another important consideration is potential selection bias. Because our cohort included only patients who underwent both knee radiography and CT, the sample likely represented more complex or ambiguous trauma cases in which initial radiographs were insufficient. Consequently, the dataset may have been skewed toward diagnostically challenging images, possibly underestimating the actual performance of LLMs in an emergency department population. Another factor that may have affected diagnostic fidelity is the conversion of DICOM images into JPEG format for compatibility with LLMs. This process could have introduced compression artifacts and reduced grayscale depth, which are essential for detecting subtle contrast cues and fracture lines. Although the images were converted at the highest possible resolution, even slight loss of pixel detail could have influenced the LLMs' visual interpretation.

Additionally, the McNemar test performed for all LLMs yielded significant results ($p<0.001$). This demonstrated that these models had an imbalance or systematic error in their classification abilities. Therefore, they currently lack sufficient capacity for fracture diagnosis and require further development.

Although not measured, the average response time for each model was up to 20 seconds after image upload, which may be considered rapid for triage settings. However, total processing time also depends on upload latency and user interaction; therefore, a comprehensive evaluation of time efficiency requires a separate analysis.

Our study had several limitations. First, it was conducted solely on radiographs. Even in cases requiring further evaluation, additional tests and physical examination findings were not provided. Therefore, some clinical parameters that might have affected performance were absent. Second, we did not measure the response times of the models. Although the response times were relatively short, we did not evaluate their potential contribution to saving time in emergency clinical practice when considering the image upload, questioning, and response evaluation processes. Third, in our study, using CT as the gold standard for fracture diagnosis may have caused case selection bias and affected the performance of LLMs.

CONCLUSION

In conclusion, our study demonstrated that the current capabilities of both ChatGPT versions and Gemini in diagnosing fractures on traumatic knee radiographs remain inferior to those of experienced clinicians. Therefore, although these LLMs are not sufficient for standalone use in fracture diagnosis, they may be used as supportive tools in the clinical decision-making process. In particular, their relatively better performance in interpreting normal radiographs highlights their potential utility in triage. Future studies with larger patient populations should incorporate diverse, high-quality datasets within the clinical context.

Ethics Committee Approval: Ethics committee approval was obtained from Gaziantep City Hospital Non-Interventional Clinical Research Ethics Committee (Approval Number: 211/2025, Date: 18.06.2025).

Informed Consent: Written informed consent was not required due to the retrospective nature of this study.

Conflict of Interest: The authors have no conflicts of interest to declare.

Funding: The authors declared that this study received no financial support.

Use of AI for Writing Assistance: No use of AI-assisted technologies was declared by the authors.

Author Contributions: Concept – OK; Design – OK, MK, BCA; Supervision – OK, IG; Resource – OK, MK, BCA, IG; Materials – OK, MK, BCA, IG; Data Collection and/or Processing - BCA, IG; Analysis and/or Interpretation - OK; Literature Review – OK, MK, BCA, IG; Writing – OK, MK, BCA, IG; Critical Review – OK, MK, BCA, IG.

Peer-review: Externally peer-reviewed.

REFERENCES

1. Oei EH, Nikken JJ, Ginai AZ, Krestin GP, Verhaar JA, van Vugt AB, et al. Acute knee trauma: value of a short dedicated extremity MR imaging examination for prediction of subsequent treatment. Radiology 2005;234(1):125-33. [CrossRef]
2. Mustonen AO, Koskinen SK, Kiuru MJ. Acute knee trauma: analysis of multidetector computed tomography findings and comparison with conventional radiography. Acta Radiol 2005;46(8):866-74. [CrossRef]
3. Avci M, Kozaci N. Comparison of X-Ray Imaging and Computed Tomography Scan in the Evaluation of Knee Trauma. Medicina (Kaunas) 2019;55(10):623. [CrossRef]
4. Teh J, Kambouroglou G, Newton J. Investigation of acute knee injury. BMJ 2012;344:e3167. [CrossRef]
5. Pinto A, Berritto D, Russo A, Riccitiello F, Caruso M, Belfiore MP, et al. Traumatic fractures in adults: missed diagnosis on plain radiographs in the Emergency Department. Acta Biomed 2018;89(1-S):111-23.
6. Venkatasamy A, Ehlinger M, Bierry G. Acute traumatic knee radiographs: beware of lesions of little expression but of great significance. Diagn Interv Imaging 2014;95(6):551-60. [CrossRef]
7. Chen Y, Zhang K, Qiang M, Li H, Dai H. Comparison of plain radiography and CT in postoperative evaluation of ankle fractures. Clin Radiol 2015;70(8):e74-82. [CrossRef]
8. Caracchini G, Pietragalla M, De Renzis A, Galluzzo M, Carbone M, Zappia M, et al. Talar fractures: radiological and CT evaluation and classification systems. Acta Biomed 2018;89(1-S):151-65.
9. Bousson V, Attané G, Benoist N, Perronne L, Diallo A, Hadid-Beurrier L, et al. Artificial Intelligence for Detecting Acute Fractures in Patients Admitted to an Emergency Department: Real-Life Performance of Three Commercial Algorithms. Acad Radiol 2023;30(10):2118-39. [CrossRef]
10. Akinci D'Antonoli T, Stanzione A, Bluethgen C, Vernuccio F, Ugga L, Klontzas ME, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. Diagn Interv Radiol 2024;30(2):80-90. [CrossRef]
11. Kim K, Cho K, Jang R, Kyung S, Lee S, Ham S, et al. Updated Primer on Generative Artificial Intelligence and Large Language Models in Medical Imaging for Medical Professionals. Korean J Radiol 2024;25(3):224-42. [CrossRef]
12. Bhayana R. Chatbots and Large Language Models in Radiology: A Practical Primer for Clinical and Research Applications. Radiology 2024;310(1):e232756. [CrossRef]
13. Polis B, Zawadzka-Fabijan A, Fabijan R, Kosińska R, Nowosławska E, Fabijan A. Comparative Evaluation of Large Language and Multimodal Models in Detecting Spinal Stabilization Systems on X-Ray Images. J Clin Med 2025;14(10):3282. [CrossRef]
14. Horiuchi D, Tatekawa H, Oura T, Shimono T, Walston S, Takita H, et al. Comparison of the diagnostic accuracy among GPT-4 based ChatGPT, GPT-4V based ChatGPT, and radiologists in musculoskeletal radiology. medRxiv. 2023 December 09. doi: 10.1101/2023.12.07.23299707 [Epub ahead-of-print]. [CrossRef]
15. Ozenbas C, Engin D, Altinok T, Akcay E, Aktas U, Tabanli A. ChatGPT-4o's Performance in Brain Tumor Diagnosis and MRI Findings: A Comparative Analysis with Radiologists. Acad Radiol 2025;32(6):3608-17. Erratum in: Acad Radiol 2025;32(11):6955. [CrossRef]
16. Mohammadi M, Parviz S, Parvaz P, Pirmoradi MM, Afzalimoghaddam M, Mirfazaelian H. Diagnostic performance of ChatGPT in tibial plateau fracture in knee X-ray. Emerg Radiol 2025;32(1):59-64. [CrossRef]
17. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic Accuracy of Differential-Diagnosis Lists Generated by Generative Pretrained Transformer 3 Chatbot for Clinical Vignettes with Common Chief Complaints: A Pilot Study. Int J Environ Res Public Health 2023;20(4):3378. [CrossRef]
18. Öztürk A, Günay S, Ateş S, Yiğit Yavuz Yigit Y. Can Gpt-4o Accurately Diagnose Trauma X-Rays? A Comparative Study with Expert Evaluations. J Emerg Med 2025;73:71-9. [CrossRef]
19. Büyüktoka RE, Salbas A. Multimodal Large Language Models for Pediatric Bone-Age Assessment: A Comparative Accuracy Analysis. Acad Radiol 2025;32(11):6905-12. [CrossRef]