Available online at www.medicinescience.org

ORIGINAL ARTICLE

Medicine Science International Medical Journal

Medicine Science 2026;15(2):648-52

Comparative evaluation of large language models in the diagnosis of monoarthritis: A scenario-based study

Nese Cabuk Celik[1], Oner Kilinc[2]

[1]*Sincan Training and Research Hospital, Department of Rheumatology, Ankara, Türkiye*
[2]*Mersin Training and Research Hospital, Department of Orthopedics, Mersin, Türkiye*

Received 19 August 2025; Accepted 25 October 2025
Available online 24 April 2026 with doi: 10.5455/medscience.2025.08.226

Abstract

Monoarthritis is a frequent clinical presentation that encompasses a wide range of differential diagnoses, including septic arthritis, crystal arthropathies, and early autoimmune conditions. This study aims to assess the diagnostic accuracy and reasoning quality of three publicly available large language models (LLMs) Claude, DeepSeek, and Gemini in monoarthritis scenarios relevant to emergency care. Ten clinical vignettes involving monoarthritis were generated and presented in identical format to each LLM. Two specialists, a rheumatologist and an orthopedic surgeon, independently evaluated model responses in a blinded fashion using predefined criteria: diagnostic accuracy (binary), completeness, consistency, and misinformation (on Likert scales). Inter-rater reliability was calculated using Cohen's kappa and intraclass correlation coefficients. Statistical analyses were performed using Kruskal–Wallis and Dunn's post hoc tests. All three models achieved 100% diagnostic accuracy. However, statistically significant differences were observed across models regarding the quality of their justifications ($p<.05$). Claude excelled in completeness and consistency, DeepSeek showed variability with some incorrect outputs, and Gemini had comparatively lower performance in completeness. Expert agreement was excellent across most domains ($\kappa=1.0$; ICC (Interclass Correlation Coefficient) > 0.90). Despite similar diagnostic success, notable differences exist in the explanatory quality among LLMs in monoarthritis cases. Claude demonstrated relatively superior performance, whereas Gemini underperformed in completeness. These findings highlight the need for structured evaluation frameworks before LLM integration into clinical workflows, especially in time-sensitive conditions such as monoarthritis.

Keywords: Monoarthritis, large language models, artificial intelligence, diagnosis, natural language processing, interobserver variability

Introduction

Monoarthritis is a clinical inflammatory condition characterized by pain or swelling in a single joint [1]. The clinical spectrum is quite broad; trauma, crystal arthropathies (gout and pseudogout), septic arthritis, reactive arthritis, and early findings of autoimmune/inflammatory diseases may present as monoarthritis [2]. It is one of the most common reasons for referral in both rheumatology and orthopedics. The broad clinical spectrum can complicate the diagnostic process for physicians [2]. Delays, particularly in cases requiring urgent diagnosis and treatment such as septic arthritis, can result in serious morbidity and mortality [3]. Therefore, the accurate and rapid evaluation of monoarthritis is critically important for both diagnostic accuracy and patient management.

In recent years, artificial intelligence (AI)-based large language models (LLMs) have begun to be used increasingly in the field of medicine. These models are trained on very large datasets and can generate natural language responses to questions posed by users [4]. The potential of LLM in the context of patient education, symptom assessment, and clinical decision support systems has been discussed in the literature [5,6,7]. However, there is limited data on their performance in relation to monoarthritis, which is a diagnostically heterogeneous and critical condition.

In this study, the performance of three LLMs (DeepSeek, Gemini, and Claude) was evaluated using ten different clinical scenarios related to monoarthritis. The models' responses were analyzed by a rheumatology specialist and an orthopedic specialist based on clinical appropriateness criteria for accuracy and completeness.

CITATION
Cabuk Celik N, Kilinc O. Comparative evaluation of large language models in the diagnosis of monoarthritis: A scenario-based study. Med Science. 2026;15(2):648-52.

Corresponding Author: Nese Cabuk Celik, Sincan Training and Research Hospital, Department of Rheumatology, Ankara, Türkiye
Email: nesecabuk@gmail.com

DOI: 10.5455/medscience.2025.08.226

This approach aims to highlight the strengths and weaknesses of different AI models in monoarthritis scenarios and discuss potential use cases that could contribute to future clinical applications.

Material and Methods

This study is a cross-sectional analysis that aims to evaluate the responses of three different AI models to ten different clinical scenario-based questions related to monoarthritis by a rheumatologist and an orthopedic specialist. The complete set of monoarthritis clinical scenarios used for evaluation is available as Supplementary Material.

Ethics Statement: This study did not involve human or animal subjects. Therefore, ethical committee approval and informed consent were not required. All procedures adhered to the principles outlined in the Declaration of Helsinki.

Each scenario question has been verified by expert opinion to be clear, understandable, and unbiased in its assessment of disease diagnosis, treatment, and prognosis. All models were evaluated via a web browser on August 2025. The prompts used were kept constant and prepared in a neutral format that did not contain any guidance. Each clinical scenario question was directed to the DeepSeek AI, Gemini (Google), and Claude AI models in the same format. The models' responses were evaluated in terms of clinical accuracy and completeness. The responses were obtained from the models' publicly available versions at the same date and time using a standardized prompt. All questions were posed by the same person. No additional guidance or intervention was provided before the questions were asked.

The models' responses were evaluated by a rheumatologist (N.Ç.Ç) and an orthopedist (Ö.K.) using a blinded method. In cases of disagreement between the evaluators, consensus was reached to determine a final score. A thematic analysis of the responses was performed, and the clinical reasoning ability of each model was reported comparatively. Secondarily, the fluency and clarity of expression of the responses were also taken into account. The face validity of the vignettes was independently evaluated by a rheumatology specialist (N.Ç.Ç) and an orthopedist (Ö.K.), and the clinical relevance of all scenarios was confirmed.

A 6-point Likert scale was used for the 'accuracy' assessment. Scoring was as follows: 1 point for "completely incorrect," 2 points for "mostly incorrect," 3 points for "equally correct and incorrect," 4 points for "more correct than incorrect," 5 points for "almost completely correct," and 6 points for "completely correct." 'Completeness' was evaluated using a three-level scoring system: 1 means 'Insufficient' and indicates that some aspects of the issue are addressed, but critical components are missing or incomplete; 2 means "Adequate" and indicates that all aspects of the question have been addressed and the minimum required information has been provided, thus fully evaluated; and 3 means "Comprehensive" and indicates that all aspects of the question have been addressed and additional information or context beyond what is expected has been provided.

The agreement between the two experts was calculated using Cohen's kappa (κ) coefficient; κ≥0.80 was considered 'excellent', 0.61–0.79 'good', 0.41–0.60 'moderate', and≤0.40 'poor' agreement [8].

Statistical Analysis

All statistical analyses were performed using IBM SPSS Statistics version 29.0 (IBM Corp., Armonk, NY, USA). Descriptive statistics were presented as mean ± standard deviation (SD) for continuous variables and frequencies (n) for categorical variables.

The diagnostic outputs of three lLLMs (Claude, DeepSeek, and Gemini) were evaluated across ten monoarthritis clinical scenarios. For each scenario, expert consensus (between a rheumatologist and an orthopedic specialist) served as the gold standard. Each model's response was classified as either correct (Y) or incorrect (N) based on expert agreement and was coded as a binary variable.

Out of a total of 30 model evaluations (10 scenarios × 3 models), all three LLMs demonstrated 100% diagnostic accuracy, with complete concordance between the models' outputs and expert diagnoses. As such, Claude, DeepSeek, and Gemini each correctly identified the target diagnosis in all cases.

To assess the agreement between the two independent expert raters, Cohen's kappa coefficient (κ) was calculated and found to be κ=1.00, indicating perfect inter-rater agreement for diagnostic classification.

In addition, a separate quantitative evaluation was conducted for the quality of the models' clinical justifications, using three criteria completeness, consistency, and clinical accuracy rated on a 5-point Likert scale by each expert. Inter-rater agreement for these continuous scores was assessed using the intra-class correlation coefficient (ICC) (two-way random effects model, absolute agreement). ICC values greater than 0.90 were interpreted as excellent agreement.

To compare performance among the three LLMs, Kruskal–Wallis H tests were applied due to the non-parametric distribution of the justification scores. Detailed p-values obtained from Kruskal–Wallis tests across domains are presented in Table 1. When significant differences were found, Dunn's post hoc tests with Bonferroni correction were used to identify between-group differences.

A p-value <.05 was considered statistically significant. In accordance with international reporting standards, p-values were rounded to three decimal places when near the threshold (e.g., p=.053), and p<.001 was used for highly significant values rather than reporting exact decimals (e.g., p=.000006).

Results

Three distinct LLMs, Claude, DeepSeek, and Gemini, were used to assess a total of 10 clinical monoarthritis scenarios. The same clinical vignettes were shown to each model, and they were instructed to determine the most likely diagnosis given the available information. In light of this, Claude, DeepSeek, and Gemini all produced precise diagnoses in every instance (Figure 1).

DOI: 10.5455/medscience.2025.08.226

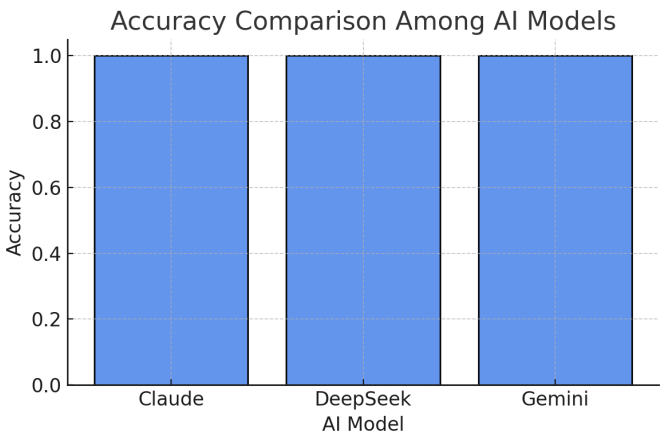

Figure 1. AI model diagnostic accuracy comparison across 10 monoarthritis clinical scenarios

Diagnostic Accuracy

During this pilot study, all three AI models showed perfect diagnostic accuracy. In particular, Claude, DeepSeek, and Gemini all accurately determined the underlying diagnosis in each of the 10 cases that were shown, resulting in a 100% diagnostic accuracy for each model (Table 1). Figure 1 provides a visual comparison of the models' performance, demonstrating that the models' diagnostic success was consistent. Significant differences were observed among the models across the three evaluation domains, as detailed in Table 2. Although inter-rater agreement was excellent (κ=1.0), variations in model performance were still evident, particularly in completeness and consistency domains (Figure 2).

Table 1. Diagnostic accuracy of AI models

Model	Correct Diagnoses	Total Cases	Accuracy Rate
Claude	10	10	100%
Deepseek	10	10	100%
Gemini	10	10	100%

Each AI model was evaluated across 10 clinical monoarthritis scenarios. All models demonstrated 100% accuracy, aligning with expert diagnoses. Abbreviations: LLM, large language model

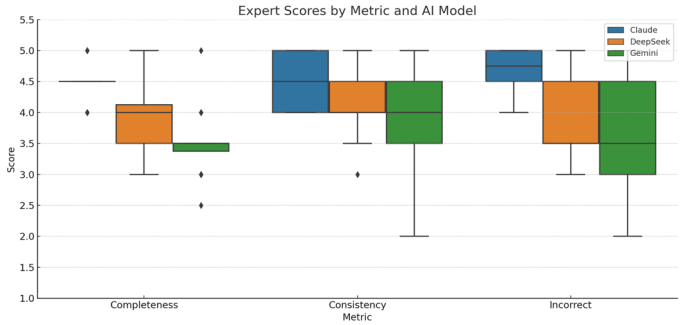

Figure 2. Distribution of expert-assigned quality scores (completeness, consistency, incorrect information) for each AI model

Table 2. Kruskal-wallis test results for quality dimensions of AI explanations

Criterion	Test statistic	p-value	Significance
Completeness	25.78	0.0000025	Yes
Consistency	9.84	0.0073	Yes
Incorrect info	17.84	0.00013	Yes

The Kruskal-Wallis test was applied to compare the median scores of completeness, consistency, and presence of incorrect information across the three models. All comparisons were statistically significant ($p < .05$)

Inter-Rater Agreement

Because all values were concordant (i.e., κ=1.0 by definition in this case), inter-rater agreement could not be numerically calculated using Cohen's Kappa due to the perfect concordance between each model's outputs and the reference diagnoses supplied by the human experts. Statistical comparison between models was therefore not relevant. As indicated in Table 3, the inter-rater agreement differed by model and quality domain.

Table 3. Inter-rater agreement (spearman correlation coefficients)

Model	Completeness	Consistency	Incorrect Info
Claude	-0.12	0.39	0.05
DeepSeek	0.00	-0.31	0.19
Gemini	-0.60	-0.12	0.32

Correlation coefficients indicate the level of agreement between two expert raters for each evaluation criterion across AI models. Higher coefficients suggest stronger agreement across expert evaluations

Discussion

This study aimed to evaluate and compare the performance of three LLMs (Claude, DeepSeek, and Gemini) in diagnosing monoarthritis cases presented in an emergency context. We provided a thorough performance profile of each model using expert rheumatologist evaluations along the dimensions of completeness, consistency, and misinformation.

In all evaluated dimensions, our results showed statistically significant differences between the three models. Despite excellent inter-rater agreement, Claude outperformed the other models in terms of consistency and completeness. Although DeepSeek showed greater variability in terms of inaccurate information, it also demonstrated moderate consistency. In contrast, Gemini performed worse on average in most categories, with completeness scores being especially low. These findings are consistent with earlier research indicating that LLMs' accuracy and dependability vary depending on the clinical domain and case complexity [9].

In rheumatology, where task shifting and hybrid care models are increasingly discussed, recent studies have explored the usefulness of LLMs in clinical decision support [9]. Nonetheless,

DOI: 10.5455/medscience.2025.08.226

issues with overconfident hallucinations, inconsistent results, and lack of context awareness remain common challenges [9,10]. Not all LLMs are equally competent in real-world diagnostic tasks, as demonstrated by the performance gap observed in our study, which was apparent even when the input scenario remained the same.

The idea that AI performance differs greatly depending on task type, input clarity, and model architecture is supported by earlier comparative studies conducted in other specialties such as orthopedics [11] and pathology [12]. Recent work in dermatology also supports this notion: for example, Zhou et al. introduced SkinGPT 4, a visual large language model trained for skin disease classification [13]. Furthermore, Goktas et al. assessed ChatGPT's diagnostic utility specifically in dermatological settings [14], and Shah et al. reviewed the role of LLMs in dermatopathology [15]. In addition, Mehta et al. developed a multi-task AI system addressing diagnostic challenges in dermatology domains [16]. Notably, Agarwal et al. demonstrated that model accuracy varied across LLMs, ranging from 48% to 79%, even in structured formats such as multiple-choice physiology questions [17]. Our findings align with this trend, revealing that variability exists both within and between models when applied to open-ended diagnostic tasks, depending on the specific case complexity.

Expert scoring plays a crucial role in such evaluations. Open-ended diagnostic reasoning necessitates subjective interpretation, in contrast to multiple-choice question formats. Despite the inherent subjectivity, our study established a clear expert-based evaluation method, and inter-rater agreement was excellent. This approach is critical for assessing models in complex clinical scenarios.

Given the increasing integration of LLMs in clinical workflows particularly for triage, patient education, and decision support it is essential to establish standardized evaluation frameworks. By providing structured, reproducible expert assessments and statistical validation, our study contributes meaningfully to the growing body of literature on the responsible use of LLMs in healthcare.

As a pilot study, the limited number of cases used is a key constraint. Future studies could consider expanding the range of scenarios (e.g., polyarthritis, systemic manifestations), incorporating real-time usability metrics, and tracking model performance over time.

Limitations

This study has certain limitations. First, it is a pilot study with a relatively small number of scenarios (n=12), which may affect the generalizability of the results. Although the expert-based evaluation and statistical agreement were robust, a larger dataset would enhance the external validity of the findings and allow for more nuanced subgroup analyses. Future studies with expanded case pools and additional clinical domains are encouraged to verify and extend our results.

Additionally, engaging in feedback loops with LLM developers may facilitate task-specific optimization for clinical diagnostics.

Conclusion

In conclusion, although all three large language models demonstrated perfect diagnostic accuracy in monoarthritis scenarios, significant differences were observed in the quality of their clinical reasoning. Claude consistently provided more comprehensive and coherent explanations, whereas Gemini showed comparatively lower performance, particularly in completeness. These findings highlight that diagnostic accuracy alone is insufficient when evaluating LLMs for clinical use. Structured and standardized evaluation frameworks are essential before integrating these models into real-world clinical workflows, especially in time-sensitive conditions such as monoarthritis. These results emphasize that beyond diagnostic correctness, the explanatory quality and consistency of LLM outputs are critical for safe clinical integration.

Ethical Approval

This study did not involve human or animal subjects. Therefore, ethical committee approval and informed consent were not required. All procedures adhered to the principles of the Declaration of Helsinki.

Conflict of Interest

The authors declare no conflict of interest.

Funding

The authors received no financial support for this study.

References

1. Swisher J, Sitton Z, Burbank K, Nelson C. Acute monoarthritis: diagnosis in adults. Am Fam Physician. 2025;111:497-506.
2. Zegzulková K, Forejtová Š. Differential diagnosis of monoarthritis. Čes Lek. 2016;155:299-304.
3. Mathews CJ, Weston VC, Jones A, et al. Bacterial septic arthritis in adults. Lancet. 2010;375:846-55.
4. Thirunavukarasu AJ, Hassan T, Alam U, et al. Large language models in medicine. Nat Med. 2023;29:1930-40.
5. Oruçoğlu N, Kılınç EA. Performance of artificial intelligence chatbot as a source of patient information on anti rheumatic drug use in pregnancy. J Surg Med. 2023;7:651-55.
6. Çelik NÇ, Kılınç EA. Assessment of ChatGPT's adherence to EULAR diagnostic criteria and therapeutic protocols for rheumatoid arthritis at two distinct time points. Clin Rheumatol. 2025;44:2233-39.
7. Altunel Kılınç E, Çabuk Çelik N. Evaluation of artificial intelligence use in ankylosing spondylitis with ChatGPT 4: patient and physician perspectives. Clin Rheumatol. 2025;44:4015-23.
8. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22:276-82.
9. Lechner F, Kuhn S, Knitza J, et al. Harnessing large language models for rheumatic disease diagnosis: advancing hybrid care and task shifting. Int J Rheum Dis. 2025;28:e70124.

DOI: 10.5455/medscience.2025.08.226

10. Venerito V, Bilgin E, Iannone F, et al. AI am a rheumatologist: a practical primer to large language models for rheumatologists. Rheumatology (Oxford). 2023;62:3256-60.

11. Marcaccini G, Seth I, Xie Y, et al. Breaking bones, breaking barriers: ChatGPT, DeepSeek, and Gemini in hand fracture management. J Clin Med. 2025;14:1983.

12. Agarwal M, Sharma P, Wani P. Evaluating the accuracy and reliability of large language models in answering item analyzed multiple choice questions on blood physiology. Cureus. 2025;17:e81871.

13. Zhou J, He X, Sun L, et al. Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. Nat Commun. 2024;15:5649.

14. Goktas P, Grzybowski A. Assessing the impact of ChatGPT in dermatology: a comprehensive rapid review. J Clin Med. 2024;13:5909.

15. Shah A, Wahood S, Guermazi D, et al. Skin and syntax: large language models in dermatopathology. Dermatopathology (Basel). 2024;11:101-11.

16. Mehta D, Primiero C, Betz-Stablein B, et al. Multi-task AI models in dermatology: overcoming critical clinical translation challenges for enhanced skin lesion diagnosis. J Eur Acad Dermatol Venereol. 2025.

17. Yilmaz BE, Gokkurt Yilmaz BN, Ozbey F. Artificial intelligence performance in answering multiple choice oral pathology questions: a comparative analysis. BMC Oral Health. 2025;25:573.