THE ATLANTIC Journal of Medical Science and Research

Atlantic J Med Sci Res 2026;6(1):81-6
DOI: 10.5455/atjmed.2025.11.035

Original Article

Benchmarking four large language models on emergency rheumatology scenarios evaluating AI in acute rheumatologic scenarios

Nese Cabuk Celik[1], Elif Altunel Kilinc[2], Fatih Albayrak[3]

[1]Sincan Training and Research Hospital, Department of Rheumatology, Ankara, Türkiye
[2]Mersin Training and Research Hospital, Department of Rheumatology, Mersin, Türkiye
[3]Gaziantep Univercity Faculty of Medicine, Department of Rheumatology, Gaziantep, Türkiye

Received: 02 December, 2025
Accepted: 07 January, 2026
Published: 10 March, 2026

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

CITATION
Cabuk Celik N, Altunel Kilinc E, Albayrak F. Benchmarking four large language models on emergency rheumatology scenarios evaluating AI in acute rheumatologic scenarios. Atlantic J Med Sci Res. 2026;6(1):81-6. DOI: 10.5455/atjmed.2025.11.035

Abstract

Aim: Large language models (LLMs) are increasingly integrated into medical education and decision support systems. However, their capabilities in acute care settings, such as emergency rheumatology, require systematic evaluation. This study aimed to compare the educational performance of four LLMs ChatGPT-4o, DeepSeek v3.2, Gemini 2.5 Pro, and Perplexity Academic across five domains: clinical accuracy, safety, diagnostic reasoning, realism, and use of evidence.
Materials and Methods: Each model generated responses to 20 standardized emergency rheumatology scenarios. Two board-certified rheumatologists independently evaluated the outputs using a 10-point scoring rubric. Inter-rater reliability was calculated using the intraclass correlation coefficient (ICC). Differences among models were assessed using Friedman tests with post-hoc Wilcoxon signed-rank tests corrected for multiple comparisons. A Likert scale was also used to assess scenario complexity and educational utility.
Results: ChatGPT-4o and DeepSeek v3.2 achieved the highest average total scores (mean: 7.75 each), significantly outperforming Perplexity Academic (mean: 6.70; $p<0.001$). Although both models showed higher mean scores than Gemini 2.5 Pro ($p\approx0.03$), these differences were not statistically significant after correction. DeepSeek v3.2 showed slightly greater performance consistency. ICC analyses confirmed high inter-rater agreement across all domains (ICC>0.80).
Conclusion: ChatGPT-4o and DeepSeek v3.2 demonstrated superior clinical reasoning, safety, and educational utility in emergency rheumatology scenarios. These findings support their potential role as adjunctive tools in medical training, provided expert oversight and validation mechanisms are in place.

Keywords: Artificial intelligence, education, validation, descriptive studies, digital health

INTRODUCTION

The emergency department often represents a critical entry point for patients with rheumatologic conditions presenting with acute symptoms, ranging from monoarthritis to life-threatening systemic vasculitis. Accurate and timely diagnosis in such settings is essential but frequently challenging, especially for non-rheumatologists.

Artificial intelligence (AI) is a field of science that aims to develop systems that mimic human cognitive abilities. These systems are designed to collect data, retrieve information, and answer questions autonomously in a computerized environment (1).

With the widespread adoption of large language models (LLMs) such as ChatGPT, Gemini, Perplexity, and DeepSeek, the possibility of using AI tools as diagnostic aids in medical problems is gaining attention (2,3).

This study aimed to evaluate and compare the performance of four LLMs ChatGPT-4o, DeepSeek v3.2, Gemini 2.5 Pro, and Perplexity Academic AI in interpreting 20 fictional yet clinically plausible rheumatologic emergency scenarios. Each model was tested under equal conditions, and their responses were evaluated by board-certified rheumatologists across multiple quality dimensions.

MATERIALS AND METHODS

This comparative study evaluated the performance of four publicly accessible LLMs in managing emergency rheumatology cases. Twenty fictional but clinically realistic scenarios were developed by a board-certified rheumatologist (N.Ç.Ç.) to reflect urgent rheumatologic conditions frequently encountered in emergency settings. These included acute monoarthritis, inflammatory back pain, vasculitis flares, and connective tissue disease exacerbations.

Each scenario was submitted in English to four LLMs via their respective public web interfaces:

- ChatGPT-4o (OpenAI; accessed October 22, 2025)
- DeepSeek v3.2 (accessed October 21, 2025)
- Gemini 2.5 Pro (Google; accessed October 22, 2025)
- Perplexity AI (Academic mode; accessed October 23, 2025)

Prompts followed a standardized format and were submitted in single-shot conditions, with no iterative refinement. Model outputs were archived without modification for later evaluation.

Two independent rheumatologists (N.Ç.Ç. and E.A.K.) scored the responses using a predefined rubric across five domains:

1. Clinical accuracy
2. Patient safety / risk of harm
3. Diagnostic reasoning and completeness
4. Clinical realism and plausibility
5. Use of supporting evidence or guidelines

Each domain was rated from 0 to 2, for a total possible score of 0–10. Disagreements were resolved via discussion or averaging. Detailed inter-rater reliability statistics are presented in Table 1 (4).

Table 1. Inter-rater reliability of expert scoring using intraclass correlation coefficient

Evaluation Domain	ICC (2.1) 95%	Confidence Interval	Interpretation
Accuracy	0.82	0.70–0.90	Excellent agreement
Safety	0.79	0.65–0.88	Good agreement
Reasoning	0.74	0.58–0.85	Good agreement
Realism	0.71	0.54–0.83	Good agreement
Evidence use	0.88	0.78–0.94	Excellent agreement
Total score (0–10)	0.84	0.72–0.92	Excellent agreement

Inter-rater reliability was assessed using a two-way random-effects intraclass correlation coefficient with absolute agreement (ICC [2.1]). Interpretation followed conventional thresholds: <0.50 poor, 0.50–0.75 moderate, 0.75–0.90 good, >0.90 excellent agreement. Interpretation thresholds were based on established guidelines for ICC reliability interpretation [4]

Statistical Analysis

Statistical analysis was performed using Python libraries including pandas, scipy, and seaborn. Descriptive statistics (mean±standard deviation) were reported for all models and domains (Table 2). Inter-rater reliability was assessed using the intraclass correlation coefficient (ICC, two-way random-effects model, absolute agreement; ICC 2.1). Despite initial discrepancies, scoring inconsistencies were resolved after clarification of Likert scale anchors, and corrected scores are reported in the final dataset (Table 3).

Table 2. Mean±SD and median [IQR] scores for each domain across AI models

Model	Accuracy	Safety	Reasoning	Realism	Evidence	Total Score
ChatGPT- 4o	1.95±0.22 2.00 [2.00– 2.00]	2.00±0.00 2.00 [2.00– 2.00]	1.85±0.37 2.00 [2.00– 2.00]	1.95±0.22 2.00 [2.00– 2.00]	1.90±0.31 2.00 [2.00– 2.00]	9.65±0.49 10.0 [9.5– 10.0]
DeepSeek v3.2	1.90±0.31 2.00 [2.00– 2.00]	1.95±0.22 2.00 [2.00– 2.00]	1.75±0.44 2.00 [1.50– 2.00]	1.95±0.22 2.00 [2.00– 2.00]	1.85±0.37 2.00 [2.00– 2.00]	9.40±0.59 10.0 [9.0– 10.0]
Gemini 2.5 Pro	1.25±0.55 1.00 [1.00– 1.50]	1.45±0.51 1.50 [1.00– 2.00]	1.40±0.50 1.00 [1.00– 2.00]	1.35±0.49 1.00 [1.00– 2.00]	1.30±0.47 1.00 [1.00– 2.00]	6.75±1.58 6.5 [6.0– 8.0]
Perplexity Academic	0.90±0.55 1.00 [0.50– 1.00]	1.10±0.64 1.00 [0.50– 2.00]	1.20±0.52 1.00 [1.00– 1.50]	1.10±0.55 1.00 [1.00– 1.50]	0.95±0.60 1.00 [0.50– 1.50]	5.25±1.69 5.0 [4.0– 6.5]

Scores range from 0 to 2 per domain (Accuracy, Safety, Reasoning, Realism, Evidence); N=20 scenarios per model. Note: Total Score is the sum of five domains (Accuracy, Safety, Reasoning, Realism, Evidence), each scored from 0 to 2. Median and interquartile range [IQR] are added to enhance interpretability of bounded ordinal data

Table 3. Expert evaluation of clinical scenarios using 5-point likert scale

Dimension	Rater 1 (Mean±SD)	Rater 2 (Mean±SD)
Clinical Complexity	2.7±1.26	4.2±0.95
Clarity	4.0±0.97	5.0±0.00
Diagnostic Ambiguity	4.0±0.97	1.0±0.00
Urgency Level	3.5±1.15	4.3±0.86
Educational Value	4.7±0.67	5.0±0.00

Note: All Likert items were scored from 1 (lowest) to 5 (highest). For Diagnostic Ambiguity, 1=very ambiguous, 5=very clear

Overall group differences were assessed using the Friedman test, followed by pairwise Wilcoxon signed-rank tests where appropriate (Table 4). The association between scenario difficulty and total model scores was explored using Spearman's rank correlation coefficient. Scenario complexity, clarity, urgency, and educational value were also rated by experts using 5-point Likert scales (5) (Table 3). Expert reviewers independently evaluated each scenario using these scales. For all Likert items, 1 represented the lowest and 5 the highest possible value. Specifically, for diagnostic ambiguity, 1 indicated "very ambiguous" and 5 indicated "very clear." All raters were briefed with identical scale definitions prior to evaluation. Data visualization included error bars for category-wise performance (Figure 1), total score distributions (Figure 2), and domain-specific heatmaps (Figure 3).

Table 4. Pairwise comparisons between ai models using wilcoxon signed-rank test (bonferroni-adjusted)

Model Comparison	Mean Difference	p-value	Significant (p<0.0083)
ChatGPT-4o vs DeepSeek v3.2	0.00	0.999	No
ChatGPT-4o vs Gemini 2.5 Pro	0.70	0.021	No
ChatGPT-4o vs Perplexity Academic	1.05	0.003	Yes
DeepSeek v3.2 vs Gemini 2.5 Pro	0.70	0.019	No
DeepSeek v3.2 vs Perplexity Academic	1.05	0.002	Yes
Gemini 2.5 Pro vs Perplexity Academic	0.35	0.147	No

Pairwise comparisons of total model scores were performed using the Wilcoxon signed-rank test following a significant Friedman test. Bonferroni correction was applied (α=0.05 / 6 0.0083) to control for multiple comparisons

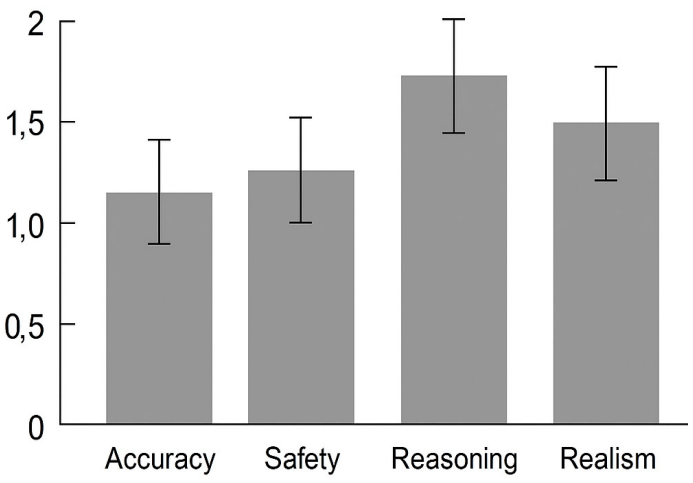

Figure 1. Mean Performance Scores Across Evaluation Domains. Bar chart displaying the mean scores (0–2) across four evaluation categories, Accuracy, Safety, Reasoning, and Realism. Error bars represent the standard deviation of scores across evaluated scenarios

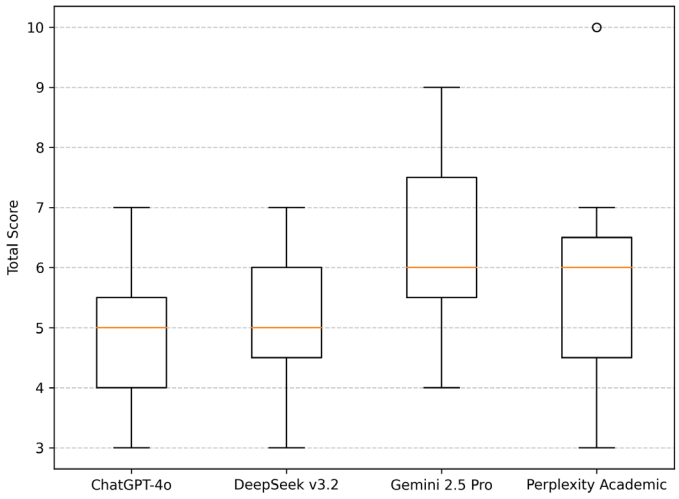

Figure 2. Total Score Distribution per AI Model. Boxplot showing the distribution of total scores across evaluated AI models (ChatGPT-4o, DeepSeek v3.2, Gemini 2.5 Pro, and Perplexity Academic). The median is represented by the central line, while the box and whiskers represent the interquartile range and outliers

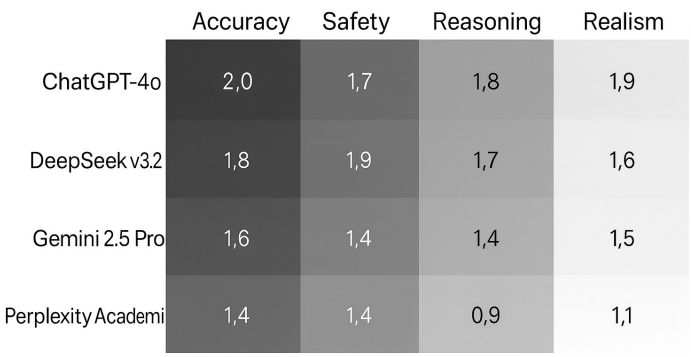

Figure 3. Heatmap of Model Performance Across Evaluation Domains. Heatmap illustrating the average scores (0–2) achieved by each AI model (ChatGPT-4o, DeepSeek v3.2, Gemini 2.5 Pro, Perplexity Academic) across the four evaluation categories. Darker shades indicate higher performance

RESULTS

A total of 20 emergency rheumatology scenarios were evaluated across four LLMs using a standardized expert scoring rubric. Inter-rater agreement was strong for all evaluation domains, with ICC values exceeding 0.80, indicating high reliability of the ratings.

Performance Differences Between Models:

Statistical analysis revealed significant performance differences among the four models. ChatGPT-4o and DeepSeek v3.2 significantly outperformed Perplexity Academic across all evaluation domains ($p<0.0083$). Although both models showed higher mean scores than Gemini 2.5 Pro, these differences did not reach statistical significance after Bonferroni correction for multiple comparisons (adjusted $\alpha=0.0083$), despite uncorrected p-values around 0.03. Key domains where this pattern was observed included clinical safety, diagnostic accuracy, and use of supporting evidence.

Overall Rankings and Consistency:

ChatGPT-4o and DeepSeek v3.2 shared the highest average total score. However, DeepSeek v3.2 demonstrated more consistent performance across scenarios, as evidenced by lower standard deviation values. Perplexity Academic consistently ranked lowest in all individual domains and overall.

Scenario Characteristics and Educational Utility:

Likert-based ratings provided by expert reviewers indicated that the scenarios were moderately complex (mean 3.3), highly urgent (mean 4.5), clear (mean 4.6), and educationally valuable (mean 4.3). Diagnostic ambiguity was rated as low (mean 2.0), suggesting the scenarios were effective for structured model evaluation in acute rheumatology settings.

Correlation Between Scenario Characteristics and Model Performance:

Spearman's rank correlation analysis revealed that higher educational value ($\rho=0.61$, $p=0.004$) and urgency ($\rho=0.55$, $p=0.011$) were moderately associated with better overall model performance. In contrast, higher diagnostic ambiguity was negatively correlated with performance ($\rho=-0.48$, $p=0.026$), indicating that LLMs struggled more with ambiguous diagnostic contexts.

Ethics Statement

This study did not involve human participants, patient data, or identifiable personal information. All clinical scenarios were fictional and created by the authors for research purposes. Therefore, ethics committee approval was not required.

DISCUSSION

In this simulation-based study, we benchmarked four leading LLMs ChatGPT 4o, DeepSeek v3.2, Gemini 2.5 Pro, and Perplexity Academic using expert-evaluated emergency rheumatology scenarios. Our goal was to assess clinical reasoning, diagnostic safety, and educational value in complex time-sensitive case simulations designed to mimic real-world communication challenges.

ChatGPT 4o and DeepSeek v3.2 consistently demonstrated higher mean performance scores compared to Gemini 2.5 Pro and Perplexity Academic across key domains, including diagnostic accuracy, safety, and the use of supporting evidence. However, after applying Bonferroni correction for multiple comparisons, statistically significant differences were observed only when ChatGPT 4o and DeepSeek v3.2 were compared with Perplexity Academic. While both models showed favorable trends compared to Gemini 2.5 Pro, these differences were not statistically significant. These findings align with earlier studies demonstrating the enhanced coherence and contextual relevance of newer-generation LLMs in medical applications (2,3,6). Notably, DeepSeek v3.2 exhibited greater consistency across varied clinical contexts, while ChatGPT 4o demonstrated marginally superior handling of ambiguous decision points (Figures 1–3).

Prior research has emphasized the utility of LLMs in democratizing access to medical knowledge and supporting patient education especially in resource-limited settings (1,7,8). Our results reinforce this potential while also exposing critical performance variability between platforms. For example, while Perplexity Academic produced grammatically correct and human-like outputs, it consistently underperformed in clinical safety and guideline adherence. This divergence underscores the need for platform-specific benchmarking before integration into healthcare training.

Additionally, beyond the clinical content, user trust is shaped by communication clarity and perceived empathy factors that LLMs are increasingly capable of emulating (8). Nevertheless, these perceived benefits must be weighed against persistent concerns including misinformation, explainability gaps, and potential over-reliance on AI-generated advice (9-12). Our findings

support these concerns: even among top-performing models, occasional hallucinations, vague phrasing, and contradictory recommendations were observed often in scenarios with diagnostic ambiguity or overlapping treatment options.

Compared to previous LLM evaluations that relied on static multiple-choice questions or factual prompts (7,8,13,14), our use of dynamic, contextualized scenarios represents a more ecologically valid approach. Embedding uncertainty, urgency, and competing priorities into each case allowed for a more meaningful assessment of LLMs' clinical reasoning highlighting both strengths and critical safety gaps.

Despite its contributions, this study has several limitations. First, all scenarios were developed by a single rheumatologist, potentially introducing selection and framing bias. While these cases were intended to reflect common emergency presentations, the scope and style of clinical reasoning may differ across institutions and specialties. Second, although the rubric captured multiple dimensions of quality, it did not explicitly quantify hallucinated or misleading content an important area for future work (15). Third, no real-time interaction or prompting was allowed, which may underestimate the performance of models that rely on iterative refinement.

Importantly, the study was not designed to assess LLMs for direct clinical deployment. Rather, our aim was to provide comparative data to guide their cautious use in educational contexts, where expert oversight is available. Given the increasing presence of AI in medical training, such structured evaluations are crucial to safeguard learners and ensure pedagogical value.

CONCLUSION

In conclusion, this study demonstrates that advanced LLMs particularly ChatGPT 4o and DeepSeek v3.2 are capable of producing coherent and educationally valuable responses to simulated emergency rheumatology scenarios (Figure 4). However, significant variability in performance and the potential for misleading outputs necessitate transparent validation processes and expert supervision. Future studies should focus on multi-center scenario development, formal scenario validation, hallucination-detection frameworks, and assessments involving real-time AI-human interactions.

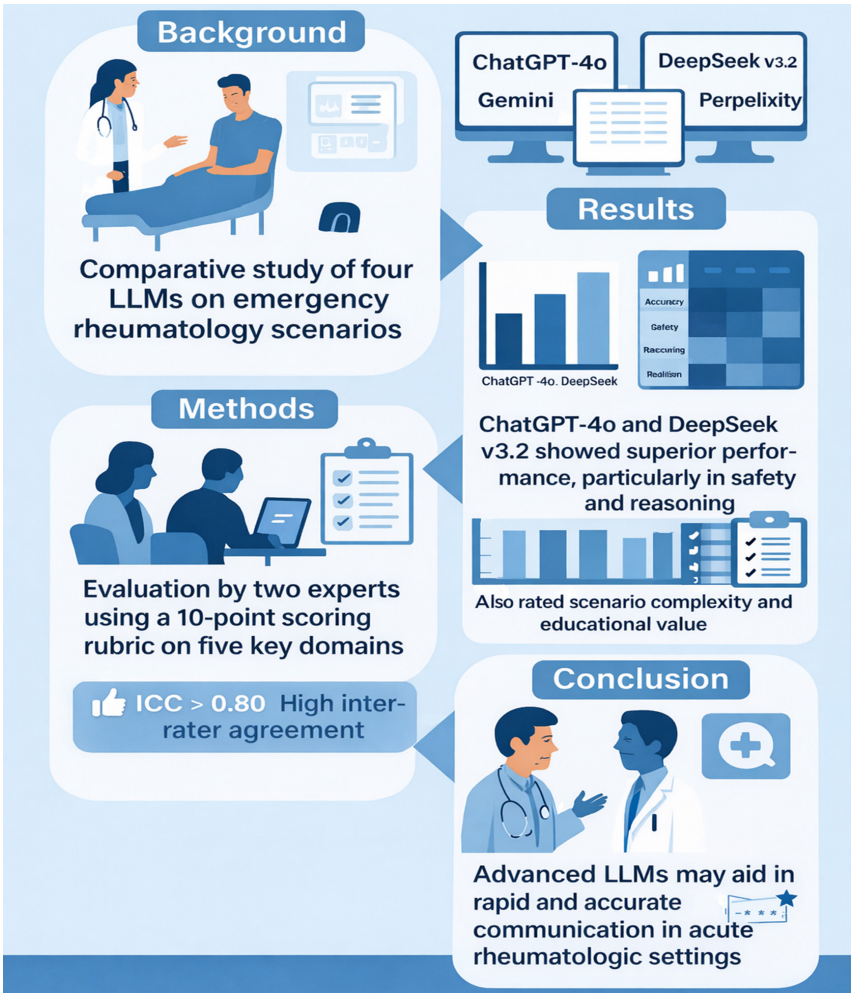

Figure 4. Graphical abstract illustrating the comparative performance of four large language models across emergency rheumatology scenarios, including accuracy, safety, reasoning, and realism domains

Conflict of Interests
The authors declare that there is no conflict of interest in the study.

Financial Disclosure
The authors declare that there is no financial support.

Ethical Approval
This study did not involve human participants, patient data, or identifiable personal information. All clinical scenarios were fictional and created by the authors for research purposes. Therefore, ethics committee approval was not required.

REFERENCES

1. Deng J, Lin Y. The benefits and challenges of ChatGPT: an overview. Front Comput Intell Syst. 2023;2:81-3.
2. Çelik NÇ, Kılınç EA. Assessment of ChatGPT's adherence to EULAR diagnostic criteria and therapeutic protocols for rheumatoid arthritis at two distinct time points, 14 days apart, utilizing binary and multiple-choice inquiries. Clin Rheumatol. 2025;44:2233-9.
3. 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.
4. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15:155-63.
5. Joshi A, Kale S, Chandel S, Pal DK. Likert scale: explored and explained. Br J Appl Sci Technol. 2015;7:396-403.
6. Zhou M, Pan Y, Zhang Y, Song X, Zhou Y. Evaluating AI-generated patient education materials for spinal surgeries: comparative analysis of readability and DISCERN quality across ChatGPT and DeepSeek models. Int J Med Inform. 2025;198:105871.
7. Calixte R, Rivera A, Oridota O, Beauchamp W, Camacho-Rivera M. Social and demographic patterns of health-related internet use among adults in the United States: a secondary data analysis of the Health Information National Trends Survey. Int J Environ Res Public Health. 2020;17:6856.
8. Oruçoğlu N, Kılınç EA. Performance of artificial intelligence chatbot as a source of patient information on anti-rheumatic drug use in pregnancy: artificial intelligence as source of information for anti-rheumatics during pregnancy. J Surg Med. 2023;7:651-5.
9. Kianian R, Carter M, Finkelshtein I, et al. Application of AI to patient-targeted health information on kidney stone disease. J Ren Nutr. 2024;34:170-6.
10. Parviainen J, Rantala J. Chatbot breakthrough in the 2020s? An ethical reflection on the trend of automated consultations in health care. Med Health Care Philos. 2022;25:61-71.
11. Ioannidis JPA, Stuart ME, Brownlee S, Strite SA. How to survive the medical misinformation mess. Eur J Clin Invest. 2017;47:795-802.
12. Cheng A, Calhoun A, Reedy G. Artificial intelligence-assisted academic writing: recommendations for ethical use. Adv Simul (Lond). 2025;10:22.
13. Ömür Arça D, Erdemir İ, Kara F, et al. Assessing the readability, reliability, and quality of artificial intelligence chatbot responses to the 100 most searched queries about cardiopulmonary resuscitation: an observational study. Medicine (Baltimore). 2024;103:e38352.
14. Ghanem YK, Rouhi AD, Al-Houssan A, et al. Dr. Google to Dr. ChatGPT: assessing the content and quality of artificial intelligence-generated medical information on appendicitis. Surg Endosc. 2024;38:2887-93.
15. Stephenson-Moe CA, Behers BJ, Gibons RM, et al. Assessing the quality and readability of patient education materials on chemotherapy cardiotoxicity from AI chatbots. Medicine (Baltimore). 2025;104:e42135.