Online First
The European Research Journal 2026
Research Article
DOI: https://doi.org/10.18621/eurj.1236
Radiology

Large Language Model Superiority in Guideline-Based Neurointerventional Multiple-Choice Questions: Performance Analysis of Radiologists vs. LLMs Across Experience Levels

Turay Cesur[1], Burcu Orug Köseer[2], Yasin Celal Güneş[3], Eren Çamur[4], Gökhan Karataş[2]

[1]Department of Radiology, Mamak State Hospital, Ankara, Türkiye; [2]Department of Radiology, Trakya University Faculty of Medicine, Edirne, Türkiye; [3]Department of Radiology, Kırıkkale Yuksek Ihtisas Hospital, Kırıkkale, Türkiye; [4]Department of Radiology, Ministry of Health Ankara 29 Mayıs State Hospital, Ankara, Türkiye

Abstract:

Objective: To assess large language models' (LLMs) mastery of contemporary consensus documents in guideline-based neurointerventional decision-making domains and to compare their performance with radiologists of different experience levels.

Methods: In this cross-sectional study, 100 single-best-answer, five-option MCQs were derived from three CIRSE guideline sources. A non-participating general radiologist authored the items (33 carotid stenting, 33 thrombectomy, 34 acute stroke EVT). One neurointerventional radiologist, two general radiologists, and one resident answered independently, blinded to the key. The same set was posed in January 2026 to Claude 4.5 Opus, ChatGPT 5.2 Instant, and Gemini 3 Pro using default settings and a standardized prompt. Accuracy was compared with McNemar's test and Bonferroni correction ($P \leq 0.002$).

Results: The highest accuracy was observed with Claude 4.5 Opus (0.99±0.10); ChatGPT 5.2 Instant (0.96±0.19) and Gemini 3 Pro (0.95±0.21) showed comparable performance. Among human participants, the neurointerventional radiologist achieved the highest accuracy (0.85±0.35), while general radiologists scored 0.67–0.70 and the resident achieved 0.57. Each of the three LLMs significantly outperformed all human groups. Moreover, Claude 4.5 Opus also demonstrated significantly higher performance than the neurointerventional radiologist ($P=0.001$).

Conclusion: LLMs have reached a competitive level with neurointerventional radiologists in recalling and interpreting neurointerventional guideline knowledge. In particular, Claude 4.5 Opus demonstrated strong performance and appears to be a promising candidate for "real-time decision support" applications. Multimodel validation and verification studies using real clinical scenario–based designs are warranted.

Keywords: Large Language Models, Neurointerventional Radiology, Accuracy, Comparative Performance

Submitted: March 13, 2026 **Accepted:** May 22, 2026 **Published Online:** June 16, 2026

How to cite this article: Cesur T, Orug Köseer B, Güneş YC, Çamur E, Karataş G. Large Language Model Superiority in Guideline-Based Neurointerventional Multiple-Choice Questions: Performance Analysis of Radiologists vs. LLMs Across Experience Levels. Eur Res J. 2026. doi: 10.18621/eurj.1236

Corresponding author: Turay Cesur, MD., Phone: +90 312 552 30 00, E-mail: turaycesur93@gmail.com

Cerebrovascular diseases remain among the leading causes of mortality and morbidity worldwide, and in acute ischemic stroke, time is the most critical prognostic determinant [1]. Landmark trials over the past decade have fundamentally reshaped clinical practice by establishing endovascular therapies as standard of care in appropriately selected patients [2–5]. In this rapidly evolving field, the Cardiovascular and Interventional Radiological Society of Europe (CIRSE) and other international organizations regularly publish consensus statements and standards grounded in evidence-based medicine [6–8]. However, the accelerating growth of the literature and frequent guideline updates make it increasingly difficult for clinicians to maintain instant access to the most current recommendations and apply them at the point of care.

Large language models (LLMs) have recently emerged as tools with potential utility across common cerebrovascular workflows, including diagnostic support, management recommendations, prognostic estimation, and patient communication [9–14]. Prior studies have also shown promising performance in interventional radiology education, including success on the European Board of Interventional Radiology (EBIR) examination and use as a training aid for medical students and interventional radiologists [15, 16]. LLM performance has additionally been evaluated for mechanical thrombectomy decision-making and post-thrombectomy outcome prediction, suggesting potential clinical relevance [17–19]. Nevertheless, their competency has not yet been systematically assessed in highly specialized neurointerventional domains where technical details and guideline-driven decisions are central—such as acute ischemic stroke endovascular therapy, mechanical thrombectomy, and carotid artery stenting. To address this gap, we aimed to evaluate the guideline-level knowledge of contemporary LLMs available in 2026 (Claude 4.5 Opus, Gemini 3 Pro, and ChatGPT 5.2 Instant) across carotid stenting, acute stroke endovascular therapy, and mechanical thrombectomy, and to benchmark their performance against radiologists with different levels of experience. Through this comparison, we sought to clarify the potential role of LLMs in neurointerventional decision-support workflows.

METHODS

Study Design and Reference Sources

This cross-sectional study was designed to compare the extent to which internationally endorsed consensus statements in neurointerventional neuroradiology are understood and interpreted by radiologists versus LLMs. Three core guideline documents published by the CIRSE and collaborating international societies served as the reference standard: the CIRSE Standards of Practice on Carotid Artery Stenting [6], the Multisociety Consensus Quality Improvement Revised Consensus Statement for Endovascular Therapy of Acute Ischemic Stroke [8], and the ESO–Karolinska Stroke Update 2014/2015 consensus statement on mechanical thrombectomy in acute ischemic stroke (supported by ESO, ESMINT, ESNR, and EAN) [7]. Because no real patient data or identifiable patient information were used, institutional review board approval was not required.

Development of the Multiple-Choice Question Set

A total of 100 single-best-answer multiple-choice questions (five options per item, one correct answer) were developed to comprehensively cover the content of the selected guideline documents. The item distribution was predefined according to the scope of each guideline: 33 questions on carotid artery stenting, 33 on mechanical thrombectomy, and 34 on endovascular therapy for acute ischemic stroke. To ensure neutrality and standardize item quality, the full question set was authored by an independent general radiologist (T.C.) with 8 years of clinical experience and European Diploma in Radiology (EDiR) certification, who did not participate in response generation or data analysis. Items were constructed by directly referencing the recommendations and statements within the source documents.

Human Participants and LLM Evaluation Procedure

The question set was administered to two groups: human radiologists and LLMs. The radiologist cohort was structured to represent different levels of expertise and comprised four readers: one neurointerventional radiologist with approximately 5 years of stroke intervention experience, two general radiologists

(Y.C.G. and E.Ç.) each with approximately 8 years of experience and EDiR certification, and one radiology resident currently rotating in interventional radiology. Assessments were conducted in January 2026. Human participants answered the questions independently and were blinded to the answer key and to other participants' responses to ensure independent judgment.

In parallel, the identical question set was administered by the question-writing radiologist (T.C.) to three contemporary LLMs via their official web interfaces using default settings and no parameter modifications: Claude 4.5 Opus (Anthropic), Google Gemini 3 Pro (Google), and ChatGPT 5.2 Instant (OpenAI). Questions were entered one-by-one in separate sessions following a standardized instruction prompt requesting selection of the single best answer according to current guidelines. Model outputs were recorded verbatim.

Statistical Analysis

Statistical analyses were performed using IBM SPSS Statistics (version 26.0; IBM Corp., Armonk, NY). Continuous variables (e.g., accuracy scores) were summarized as mean±standard deviation, as well as median and interquartile range (25th–75th percentiles), as appropriate. Pairwise comparisons of accuracy (correct vs. incorrect) between participants were conducted using McNemar's test, given the paired and categorical nature of the outcome. With 7 participants yielding 21 pairwise comparisons, a Bonferroni-adjusted threshold of $P \leq 0.002$ was considered statistically significant.

RESULTS

Overall Accuracy

The highest accuracy was observed for Claude 4.5 Opus (0.99±0.10), followed by ChatGPT 5.2 Instant (0.96±0.19) and Gemini 3 Pro (0.95±0.21). Performance metrics for all radiologists and LLMs are summarized in Table 1 and Figure 1.

Among human readers, the neurointerventional radiologist achieved the highest accuracy (0.85±0.35). General radiologists scored 0.67±0.47 and 0.70±0.46, respectively, while the resident achieved 0.57±0.49.

Comparative Analyses

Pairwise comparison results are presented in Table 2. No statistically significant differences were observed among the three LLMs (ChatGPT 5.2 Instant vs. Gemini 3 Pro: $P=0.250$; ChatGPT 5.2 Instant vs. Claude 4.5 Opus: $P=1.000$; Gemini 3 Pro vs. Claude 4.5 Opus: $P=0.077$), indicating similarly high performance across models.

As expected, the neurointerventional radiologist significantly outperformed the resident ($P<0.001$) and one general radiologist ($P=0.002$). The difference between the neurointerventional radiologist and the second general radiologist did not meet the Bonferroni-adjusted significance threshold ($P=0.009$; adjusted $P \leq 0.002$). No significant difference was found between the two general radiologists ($P=0.508$).

Notably, all three LLMs significantly outperformed every human reader group, including the neurointerventional radiologist. All LLMs showed clear superiority over the resident and general

TABLE 1. Accuracy of Participating Radiologists and Large Language Models on the 100Study-Specific Fictional Guideline-Based Questions

Participant	Accuracy	Standard deviation	Median	Percentiles (25P-75P)
Interventional radiologist	0.85	0.35	1	1-1
General radiologist 1	0.67	0.47	1	0-1
General radiologist 2	0.7	0.46	1	0-1
Assistant radiologist	0.57	0.49	1	0-1
ChatGPT 5.2 Instant	0.96	0.19	1	1-1
Gemini 3 Pro	0.95	0.21	1	1-1
Claude 4.5 Opus	0.99	0.10	1	1-1

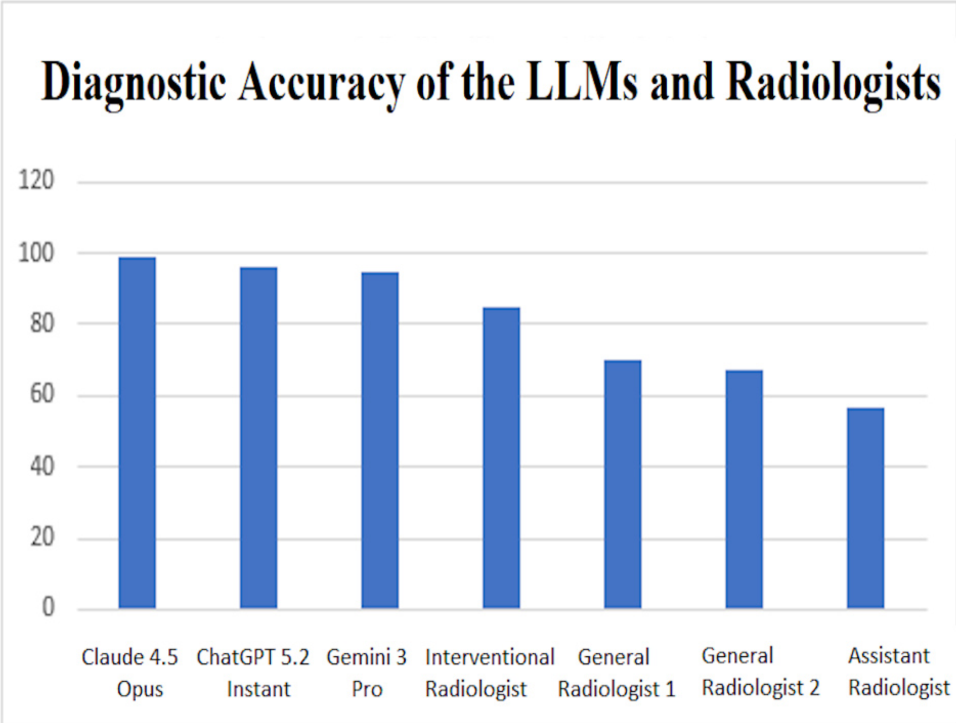

FIGURE 1. Performance of radiologists and large language models on the 100 study-specific multiple-choice questions developed from neurointerventional radiology guideline documents.

radiologists (all P<0.001). Compared with the neurointerventional radiologist, Claude 4.5 Opus achieved significantly higher accuracy (P=0.001), whereas the higher accuracy of ChatGPT 5.2 Instant (P=0.019) and Gemini 3 Pro (P=0.031) did not reach Bonferroni-adjusted significance (P≤0.002).

DISCUSSION

This study provides objective evidence that rapidly advancing LLMs can achieve exceptionally high performance on guideline-based knowledge tasks within a highly specialized, technically detailed, and protocol-driven neurointerventional domain. The principal finding was that all evaluated LLMs significantly outperformed both general radiologists and the radiology resident. Importantly, Claude 4.5 Opus not only exceeded the performance of non-specialist readers but also outperformed the neurointerventional radiologist, whose daily clinical practice centers on these procedures (P=0.001), highlighting the models' strong capability for guideline retrieval and interpretation in a structured question format.

Prior work has examined LLM performance on interventional radiology–focused questions and board-style assessments; however, most studies have emphasized broad interventional knowledge rather than the granular, recommendation-level detail contained in specific CIRSE and multisociety consensus documents [6–8, 15, 20]. For example, Barat *et al*. evaluated the appropriateness of ChatGPT responses to a custom set of 20 interventional radiology questions and reported fully appropriate answers for only 40% of items [20]. In an editorial letter, Güneş *et al*. reported that ChatGPT-3.5/4 achieved 66% (10/15) accuracy on 15 example EBIR questions hosted on the CIRSE website, outperforming several other general-purpose systems. Against this backdrop, the near-ceiling accuracies observed in our study—99% for Claude 4.5 Opus, 96% for ChatGPT 5.2 Instant, and 95% for Gemini 3 Pro—suggest substantial recent gains in consistency for guideline-based recall and selection in single-best-answer formats. Moreover, by benchmarking across

TABLE 2. Pairwise Comparison of Accuracy Between Large Language Models and Radiologists Using P-Values Derived From Mcnemar's Test.

R	Interventional radiologist	General radiologist 1	General radiologist 2	Assistant radiologist	ChatGPT 5.2 Instant	Gemini 3 Pro	Claude 4.5 Opus
Interventional radiologist	-	**0.002**	0.009	**<0.001**	0.019	0.031	**0.001**
General radiologist 1	**0.002**	-	0.508	0.013	**<0.001**	**<0.001**	**<0.001**
General radiologist 2	0.009	0.508	-	0.011	**<0.001**	**<0.001**	**<0.001**
Assistant radiologist	**<0.001**	0.013	0.011	-	**<0.001**	**<0.001**	**<0.001**
ChatGPT 5.2 Instant	0.019	**<0.001**	**<0.001**	**<0.001**	-	0.250	1
Gemini 3 Pro	0.031	**<0.001**	**<0.001**	**<0.001**	0.250	-	0.077
Claude 4.5 Opus	**0.001**	**<0.001**	**<0.001**	**<0.001**	1	0.219	-

Because 21 pairwise subgroup comparisons were performed using McNemar's test, a Bonferroni-adjusted p-value of **≤ 0.002** was considered statistically significant. Bolded values indicate statistically significant differences in accuracy.

multiple human experience levels, our data also indirectly speak to item difficulty: the neurointerventional radiologist's 85% accuracy indicates that the question set was non-trivial and required domain expertise, whereas the 57–70% range for the resident and general radiologists supports the highly specialized nature of the guideline content.

These findings should not be interpreted as evidence that LLMs can replace physicians. Clinical decision-making in neurointervention extends beyond selecting guideline-consistent statements; it integrates patient-specific variables, imaging-based pattern recognition, procedural nuance, and real-time management of complications under uncertainty. Nevertheless, the observed performance—particularly for Claude 4.5 Opus—supports the plausibility of LLMs functioning as "real-time decision support" tools in the angiography suite or outpatient setting, where rapid access to guideline-level recommendations may be valuable. In practice, an LLM could serve as a high-speed reference layer when clinicians encounter uncommon complications, borderline indications, or rapidly evolving recommendations, potentially acting as a "digital colleague" for guideline retrieval—provided that outputs are transparently presented, critically appraised, and embedded within appropriate governance and safety frameworks.

Strengths and Limitations

Our study has several strengths. First, it addresses a timely and clinically relevant question by evaluating contemporary large language models in a highly specialized, guideline-driven neurointerventional domain. Second, the multiple-choice questions were systematically derived from internationally recognized CIRSE and multisociety consensus documents, providing a clear and reproducible reference standard. Third, the study compared LLM performance with radiologists at different levels of experience, including a neurointerventional radiologist, general radiologists, and a radiology resident, thereby enabling a clinically meaningful benchmark across expertise levels. Finally, the use of identical questions, standardized prompting, blinded human assessment, and pairwise statistical comparisons strengthened the methodological consistency of the evaluation. Several limitations also merit consideration. First, evaluation was restricted to text-only multiple-choice items, which do not capture the multimodal complexity of real clinical workflows that require integration of imaging, laboratory trends, and evolving neurological examinations. Second, the study design focused on accuracy in a single-best-answer format and did not assess calibration, explanation quality, uncertainty expression, or the propensity for plausible-sounding but incorrect rationales. Third, model performance was measured under default platform settings at a single time point, and LLM behavior may vary across versions, updates, or prompting strategies. Future work should therefore prioritize multimodal, scenario-based validation studies incorporating imaging and clinical context, assess reliability and uncertainty handling, and evaluate prospective utility within supervised clinical decision-support pipelines.

CONCLUSION

In conclusion, as of 2026, contemporary LLMs demonstrate guideline-level mastery in neurointerventional neuroradiology that is competitive with—and in the case of Claude 4.5 Opus, surpassing—human expert performance on structured guideline-based questions. This technological step-change has direct implications for medical education and for the design of safe, auditable, clinician-in-the-loop decision-support systems in neurointerventional practice.

Ethics Approval and Consent to Participate

In this cross-sectional study, we designed to compare the extent to which internationally endorsed consensus statements in neurointerventional neuroradiology are understood and interpreted by radiologists versus large language models (LLMs). Because no real patient data or identifiable patient information were used, institutional review board approval was not required.

Data Availability

All data generated or analyzed during this study are included in this published article. The data that support the findings of this study are available on request from the corresponding author, upon reasonable request.

Authors' Contribution

Study Conception: TC, GK; Study Design: TC, YCG; Supervision: GK; Funding: N/A; Materials: TC; Data Collection and/or Processing: TC, BOK, EÇ; Statistical Analysis and/or Data Interpretation: TC; Literature Review: TC; Manuscript Preparation: TC; and Critical Review: YCG, TC.

Conflict of Interest

The author(s) disclosed no conflict of interest during the preparation or publication of this manuscript.

Financing

The author(s) disclosed that they did not receive any grant during the conduction or writing of this study.

Acknowledgments

The authors have no acknowledgments to declare.

Generative Artificial Intelligence Statement

During the preparation of this work the authors used ChatGPT (Open AI) in order to improve language and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication. The all content of the study was produced by the author(s) in accordance with scientific research methods and academic ethical principles.

Editor's Note

All statements made in this article are solely those of the authors and do not represent the views of their affiliates or the publisher, editors, or reviewers. Any claims made by any product or manufacturer that may be evaluated in this article are not guaranteed or endorsed by the publisher.

REFERENCES

1. Krishnamurthi RV, Ikeda T, Feigin VL. Global, Regional and Country-Specific Burden of Ischaemic Stroke, Intracerebral Haemorrhage and Subarachnoid Haemorrhage: A Systematic Analysis of the Global Burden of Disease Study 2017. Neuroepidemiology. 2020;54(2):171-179. doi: 10.1159/000506396.
2. Campbell BC, Mitchell PJ, Kleinig TJ, et al.; EXTEND-IA Investigators. Endovascular therapy for ischemic stroke with perfusion-imaging selection. N Engl J Med. 2015;372(11):1009-1018. doi: 10.1056/NEJMoa1414792.

3. Goyal M, Demchuk AM, Menon BK, et al.; ESCAPE Trial Investigators. Randomized assessment of rapid endovascular treatment of ischemic stroke. N Engl J Med. 2015;372(11):1019-1030. doi: 10.1056/NEJMoa1414905.
4. Jovin TG, Chamorro A, Cobo E, et al.; REVASCAT Trial Investigators. Thrombectomy within 8 hours after symptom onset in ischemic stroke. N Engl J Med. 2015;372(24):2296-2306. doi: 10.1056/NEJMoa1503780.
5. Saver JL, Goyal M, Bonafe A, et al.; SWIFT PRIME Investigators. Stent-retriever thrombectomy after intravenous t-PA vs. t-PA alone in stroke. N Engl J Med. 2015;372(24):2285-2295. doi: 10.1056/NEJMoa1415061.
6. Spiliopoulos S, Blanc R, Gandini R, Müller-Hülsbeck S, Reith W, Moschovaki-Zeiger O. CIRSE Standards of Practice on Carotid Artery Stenting. Cardiovasc Intervent Radiol. 2024;47(6):705-716. doi: 10.1007/s00270-024-03707-y.
7. Wahlgren N, Moreira T, Michel P, et al.; ESO-KSU, ESO, ESMINT, ESNR and EAN. Mechanical thrombectomy in acute ischemic stroke: Consensus statement by ESO-Karolinska Stroke Update 2014/2015, supported by ESO, ESMINT, ESNR and EAN. Int J Stroke. 2016;11(1):134-147. doi: 10.1177/1747493015609778.
8. From the American Association of Neurological Surgeons (AANS), American Society of Neuroradiology (ASNR), Cardiovascular and Interventional Radiology Society of Europe (CIRSE), Canadian Interventional Radiology Association (CIRA), Congress of Neurological Surgeons (CNS), European Society of Minimally Invasive Neurological Therapy (ESMINT), European Society of Neuroradiology (ESNR), European Stroke Organization (ESO), Society for Cardiovascular Angiography and Interventions (SCAI), Society of Interventional Radiology (SIR), Society of NeuroInterventional Surgery (SNIS), and World Stroke Organization (WSO); Sacks D, Baxter B, Campbell BCV, et al. Multisociety Consensus Quality Improvement Revised Consensus Statement for Endovascular Therapy of Acute Ischemic Stroke. Int J Stroke. 2018;13(6):612-632. doi: 10.1177/1747493018778713.
9. Chlorogiannis DD, Apostolos A, Chlorogiannis A, et al. The Role of ChatGPT in the Advancement of Diagnosis, Management, and Prognosis of Cardiovascular and Cerebrovascular Disease. Healthcare (Basel). 2023;11(21):2906. doi: 10.3390/healthcare11212906.
10. Altunisik E. Artificial Intelligence and Cerebrovascular Diseases: ChatGPT Model. Cerebrovasc Dis. 2024;53(3):354-358. doi: 10.1159/000533967.
11. Kuzan BN, Meşe İ, Yaşar S, Kuzan TY. A retrospective evaluation of the potential of ChatGPT in the accurate diagnosis of acute stroke. Diagn Interv Radiol. 2025;31(3):187-195. doi: 10.4274/dir.2024.242892.
12. Chowdary VR. Management of Stroke: An Exploration with ChatGPT. Neurol India. 2025;73(5):1092-1094. doi: 10.4103/neurol-india.Neurol-India-D-24-00919.
13. Wang X, Ye S, Feng J, Feng K, Yang H, Li H. Performance of ChatGPT on prehospital acute ischemic stroke and large vessel occlusion (LVO) stroke screening. Digit Health. 2024;10:20552076241297127. doi: 10.1177/20552076241297127.
14. Koyun M, Taskent I. Evaluation of Advanced Artificial Intelligence Algorithms' Diagnostic Efficacy in Acute Ischemic Stroke: A Comparative Analysis of ChatGPT-4o and Claude 3.5 Sonnet Models. J Clin Med. 2025;14(2):571. doi: 10.3390/jcm14020571.
15. Güneş YC, Cesur T. Diagnostic Accuracy of Large Language Models in the European Board of Interventional Radiology Examination (EBIR) Sample Questions. Cardiovasc Intervent Radiol. 2024;47(6):836-837. doi: 10.1007/s00270-024-03674-4.
16. Ebel S, Ehrengut C, Denecke T, Gößmann H, Beeskow AB. GPT-4o's competency in answering the simulated written European Board of Interventional Radiology exam compared to a medical student and experts in Germany and its ability to generate exam items on interventional radiology: a descriptive study. J Educ Eval Health Prof. 2024;21:21. doi: 10.3352/jeehp.2024.21.21.
17. Chen TC, Couldwell MW, Singer J, et al. Assessing the clinical reasoning of ChatGPT for mechanical thrombectomy in patients with stroke. J Neurointerv Surg. 2024;16(3):253-260. doi: 10.1136/jnis-2023-021163.
18. Pedro T, Sousa JM, Fonseca L, et al. Exploring the use of ChatGPT in predicting anterior circulation stroke functional outcomes after mechanical thrombectomy: a pilot study. J Neurointerv Surg. 2025;17(3):261-265. doi: 10.1136/jnis-2024-021556.
19. Kottlors J, Hahnfeldt R, Görtz L, et al. Large Language Models-Supported Thrombectomy Decision-Making in Acute Ischemic Stroke Based on Radiology Reports: Feasibility Qualitative Study. J Med Internet Res. 2025;27:e48328. doi: 10.2196/48328.
20. Barat M, Soyer P, Dohan A. Appropriateness of Recommendations Provided by ChatGPT to Interventional Radiologists. Can Assoc Radiol J. 2023;74(4):758-763. doi: 10.1177/08465371231170133.