Research Article

Statistical Analysis of the Performance of Open-Access Language Models: A Tool in the Clinical Laboratory

Micaela Avellaneda[1], Pamela De Francesco[2], Nilda Fink[3*]

[1]*B.Sc. in Biochemistry Federacion Bioquimica de la Provincia de Buenos Aires*
[2]*Library scientist Fundación Bioquímica Argentina (FBA)*
[3]*PhD, MSc in Biochemistry and Pharmacy. Fundación Bioquímica Argentina (FBA)*

Article Info

[*]***Corresponding Author:***
Nilda Fink
Fundación Bioquímica Argentina (FBA) PROES La Plata, Buenos Aires, Argentina
E-mail: nfink@fba.org.ar

Keywords

Artificial intelligence, language models, comparative evaluation, reproducibility, clinical laboratory

Abstract

Introduction: Artificial intelligence (AI) has been progressively incorporated into the clinical analysis laboratory, optimizing data management, detection of complex patterns, and image-based diagnostic support. Among emerging tools, Large Language Models (LLMs) represent a significant advance by enabling natural language processing for clinical, analytical, and academic tasks. However, their performance is heterogeneous and depends on model architecture, training, and contextual fine-tuning. Objective: To compare the performance of eight openly accessible large language models (LLMs) applied to the biomedical domain by assessing their output quality and inter-rater consistency.

Materials and Methods: A cross-sectional comparative study was carried out. Eight LLMs (ChatGPT-4, DeepSeek V3.1, Gemini, Copilot GPT-5, Claude 3.5 Haiku, Perplexity AI, DeepAI, and Poe) were queried using six standardized questions regarding the analytical evolution of blood glucose measurement. Responses were evaluated by two independent automated systems (Text Cortex AI and Grok Code Fast 1) using a Likert scale from 1 to 10 and considering eight criteria of textual and scientific quality. Friedman and Wilcoxon–Holm tests were used for global and pairwise comparisons, and Pearson and Spearman coefficients were calculated to assess inter-rater agreement.

Results: Significant differences were identified among models ($p < 0.001$). The LLMs Gemini, DeepSeek, and ChatGPT-4 obtained the highest scores (medians > 8.5), whereas DeepAI and Poe showed lower performance. Inter-rater correlation based on median scores showed a markedly strong linear association ($r = 0.91$; $p = 0.0019$) and a robust ordinal agreement ($\rho = 0.83$; $p = 0.011$). The least-squares regression line ($E2 = 1.37 \cdot E1 – 3.25$) indicated a sustained positive relationship between the two systems, with balanced dispersion around the linear model. These findings confirm the stability of the evaluation pattern and support the reproducibility of the method.

Conclusion: The LLMs with greater complexity and contextual fine-tuning demonstrated superior accuracy, coherence, and reproducibility; however, it is suggested that LLMs may serve as complementary tools depending on the clinical or research context. Their use in healthcare requires methodological validation and expert oversight to ensure safety and clinical applicability within established ethical and regulatory frameworks.

Introduction

Artificial intelligence (AI) has become a key pillar in the clinical laboratory due to its ability to analyze large volumes of data, identify complex patterns, and support both clinical and operational decision-making [1,2]. Its applicability extends across all three phases of the diagnostic process. In the pre-analytical phase, recognized as the most vulnerable stage of the diagnostic process, AI and data analytics help to reduce frequent errors (such as those related to patient identification or sample handling), optimize simple traceability, and promote the standardization of procedures [3-5]. In the analytical phase, machine learning approaches have been applied to quality control, outlier detection, and automated validation, thereby improving the reliability and consistency of reported results [2,6]. In the post- analytical phase, integration with EHR systems enables risk stratification and clinical decision support; for instance, predictive models for sepsis can anticipate septic shock with robust performance (AUC ≈ 0.83) prior to clinical onset [7]. Additionally, the use of indirect data-driven methods to estimate reference intervals has expanded. A recent study in a Puerto Rican population applied unsupervised learning and demonstrated age- and sex-specific variations [8]. Outside the strictly analytical setting, AI-based triage systems in emergency departments show potential for improving prioritization and process consistency, however, they require multicenter validation and standardized reporting [9,10]. Finally, large language models (LLMs) provide a natural-language interface that is useful for documentation, synthesis, and diagnostic support. Recent literature and systematic reviews describe their applications, limitations, and safety considerations as well as ongoing clinical trials evaluating their implementation [11-13]. In this context, it is crucial to systematically and critically assess the performance of these tools in tasks that require historical knowledge, factual accuracy, and clarity of exposition. This study aims to compare the performance of eight language models by analyzing their responses to six standardized questions addressing the historical evolution and methodological development of glucose measurement. The evaluation includes an assessment of historical accuracy, the reliability of verifiable sources, and the precision, clarity, and comprehensiveness of the generated content. It also examines originality, factual coherence, argumentative depth, and adherence to academic style based on predefined criteria. Inter-rater agreement will be quantified using correlation coefficients to ensure methodological rigor. In summary, this approach intends to provide a robust estimation of the utility and limitations of large language models as supportive tools in clinical, educational, and research settings.

Materials and Methods

Study Design

A cross-sectional comparative study with non-parametric statistical analysis was conducted to evaluate the performance of eight LLMs.

Analyzed Models

The models included in the analysis were selected because they represent the main architectures and technological families of freely accessible LLMs available at the time of the study:
M1: ChatGPT-4 (OpenAI)
M2: DeepSeek V3.1
M3: Gemini (Google DeepMind)
M4: Copilot Smart GPT-5 (Microsoft)
M5: Claude 3.5 Haiku (Anthropic)
M6: Perplexity AI
M7: DeepAI
M8: Poe (Quora)
Each model was queried on October 10, 2025, under standardized conditions using a set of six sequential questions designed to assess knowledge of the historical evolution, analytical foundations, and technological applications of blood glucose measurement. These questions addressed, among other aspects, the initial description of the colorimetric method, the introduction of the enzymatic method, key milestones in laboratory automation, and recent innovations in portable devices and integrations with artificial intelligence.
The responses obtained were compiled in text format and evaluated automatically and independently by two AI-based systems: E1: Text Cortex IA and E2: Grok Code Fast 1. Both are freely accessible tools that function as independent evaluators, assigning scores on a Likert scale from 1 to 10 (with 10 being the highest rating) based on eight predefined criteria: historical accuracy, exhaustiveness, narrative clarity, verifiable sources, originality, factual coherence, argumentative depth, and academic style.

Statistical Analysis

Non-parametric methods were applied, as they are suitable for ordinal data and asymmetric distributions. Descriptive measures included the median and interquartile range (IQR). Overall comparisons of model scores were conducted using the Friedman test. Post-hoc analyses were performed with the paired Wilcoxon test, applying Holm's correction to control for multiple-comparison error.
Inter-rater consistency was assessed using Pearson (r) and Spearman (ρ) correlation coefficients. The significance level was set at $\alpha = 0.05$.
All analyses were performed using SPSS Statistics version 20.0.

Results

The comparative analysis demonstrated statistically significant differences among the evaluated models for both scoring systems (E1 and E2), with Friedman test p-values < 0.001. This supports the presence of true variability in the models' performance when addressing standardized blood glucose measurement queries.

According to the scores assigned by evaluator E1, models M3, M2, and M1 constituted the highest-performing group (Table 1). Their medians and interquartile ranges (IQR: Q3–Q1) were 8.9 (0.9), 8.7 (1.0), and 8.6 (1.0), respectively, reflecting a high degree of narrative consistency, argumentative clarity, and historical accuracy (Figure 1). Models M4 and M5 showed intermediate performance, whereas M6, M7, and M8 ranked in the lower tier, with scores below 7 points on the Likert scale. The post-hoc Wilcoxon analysis with Holm correction confirmed significant differences between the top-performing group (M3–M2–M1) and the lower-performing group (M7–M8–M6), validating the robustness of the resulting ranking.

Evaluator E2 identified a similar pattern, although with slight variations in internal ordering (Table 1). In this case, M1 achieved the highest median—9.4 (0.7)—followed by M2 with 8.6 (0.7) and M3 with 8.0 (1.4) (Figure 2). Models M4 and M5 maintained acceptable values, whereas M6, M7, and M8 exhibited significantly lower performance ($p < 0.05$ in pairwise comparisons M1 > M8; M2 > M7; M2 > M8).

Inter-rater correlation calculated from median scores demonstrated a very high linear association ($r = 0.91$; $p = 0.0019$) and a robust ordinal agreement ($\rho = 0.83$; $p = 0.011$). The least-squares regression line ($E2 = 1.37 \cdot E1 - 3.25$) indicated a sustained positive relationship between the two systems, with balanced dispersion around the linear model (Figure 3). These results confirm the stability of the evaluation pattern and support the reproducibility of the method.

The joint analysis of the assessments provided by both evaluators indicates that models M1, M2, and M3 constitute the group with the highest overall performance. This group is characterized by producing coherent and well-structured responses, demonstrating strong contextual command of scientific information, maintaining high levels of narrative clarity, exhibiting low dispersion (reflected in a narrow interquartile range) and showing a high degree of consistency across independent evaluators.

In contrast, models with simpler architectures (M6–M8) exhibited limitations in factual precision, use of verifiable references, and argumentative development—critical aspects in biomedical applications (Table 3).

Table 1: Comparison of Model Performance: Text Cortex vs. Grok Code Fast 1.

Model	Median [IQR] Text Cortex *	Interpretation (Text Cortex)	Median [IQR] Grok Code Fast 1 **	Interpretation (Grok Code Fast 1)
M1	8.6 [1.0]	Robust and well-balanced performance	9.4 [0.7]	Highest overall performance
M2	8.7 [1.0]	Stable coherence and depth	8.6 [0.7]	High coherence
M3	8.9 [0.9]	Maximum narrative clarity	8.0 [1.4]	Slight variability
M4	8.1 [0.9]	Good intermediate performance	8.1 [0.9]	Good stability
M5	7.3 [1.2]	Moderate variability	6.9 [1.3]	Moderate dispersion
M6	6.8 [1.2]	Lower but acceptable performance	6.9 [0.9]	Intermediate performance
M7	6.6 [0.7]	Limited accuracy	5.6 [1.2]	Low accuracy
M8	6.4 [0.7]	Low and homogeneous performance	4.8 [1.4]	Significantly lower performance
Results	* Text Cortex evaluator: χ^2 (7) = 46.97; $p < 0.001$ → significant differences. Post-hoc (Wilcoxon–Holm): significant differences between the top-performing group (M3–M2–M1) and the lower group (M7–M8–M6).		** Grok Code Fast 1 evaluator: χ^2 (7) = 53.88; $p < 0.001$ → highly significant differences. Post-hoc (Wilcoxon–Holm): M1 > M8 ($p = 0.0413$); M2 > M7 ($p = 0.0210$); M2 > M8 ($p = 0.0108$).	

Values are expressed as median [interquartile range, IQR] on a Likert scale of 1 to 10, where 10 represents the highest possible score (maximum clarity, coherence, and analytical precision).

The Friedman and Wilcoxon–Holm tests were applied to identify global and pairwise differences among models, respectively.

Table 2: Inter-Rater Correlation Between E1 and E2.

Parameter	Value	Interpretation
Pearson's r	0.91	Very high linear correlation (quantitative consistency)
Spearman's ρ	0.83	Robust ordinal correlation (agreement in ranking order)

Table 3: Comparative Summary of Overall Performance by Evaluator.

Evaluator	Overall Median	Overall IQR	Significance (Friedman)
E1	7.7	[6.6–8.8]	$p < 0.001$
E2	7.6	[6.3–8.7]	$p < 0.001$

Figure 1: Comparative Performance Ranking.

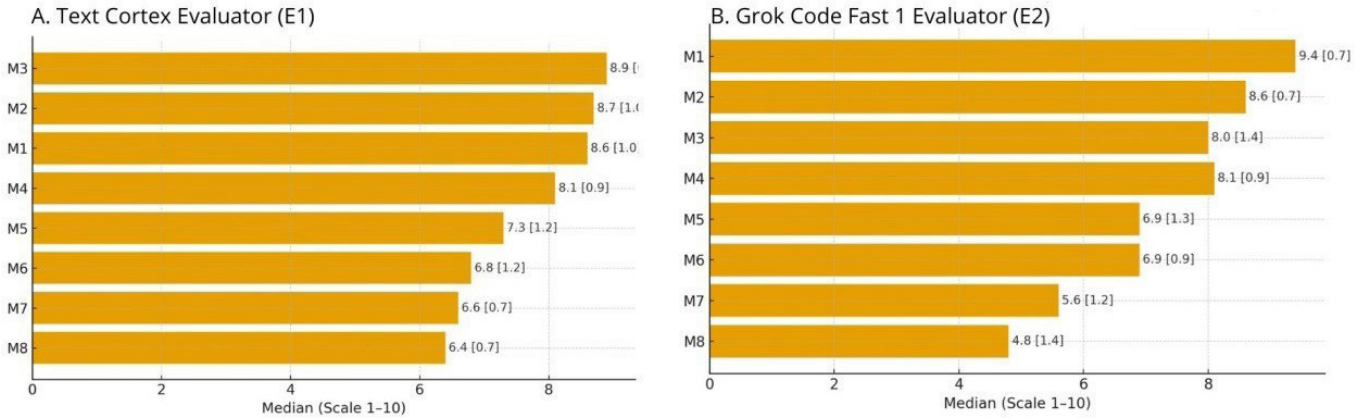
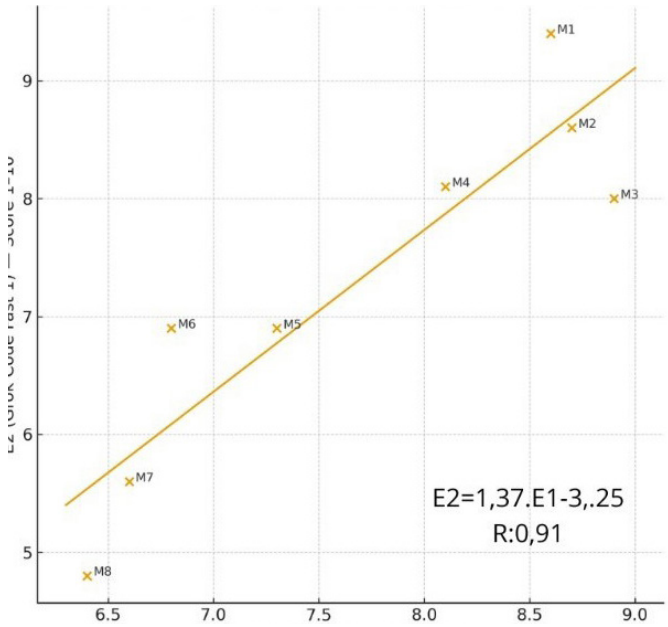


Figure 2: Inter-Rate-Correlation. Text Cortex vs Grok Code Fast 1.



Discussion

The results of this study demonstrate that the evaluated LLMs - particularly models M1, M2, and M3 - exhibit a clear advantage in both analytical capacity and narrative generation compared with general-purpose or minimally trained models. This superiority may be attributed to three key factors: (a) the multimodal scale of training (text and possibly additional modalities), (b) biomedical contextual calibration, and (c) optimization of semantic understanding. Overall, these components enhance stability and accuracy in clinical reasoning tasks, giving specialized models a more robust capacity to process complex information and generate coherent responses within the biomedical domain.

The use of non-parametric tests - such as the Friedman and Wilcoxon–Holm tests - was particularly appropriate for comparing non-normal distributions and minimizing the influence of outliers. Reporting medians and interquartile ranges (IQR) provided a bias-resistant characterization of central tendency, offering an accurate representation under conditions of high variability. Additionally, the incorporation of two independent automated evaluators strengthened the objectivity of the analysis and reduced the risk of subjective bias, demonstrating that automated systems can faithfully reproduce scientific evaluation criteria.

The high inter-rater correlation observed in this study confirms the reproducibility and consistency of the evaluation process. The observed linear concordance ($r = 0.91$; $p = 0.0019$) indicates that both systems largely agree on the magnitude of assigned scores, while the ordinal correlation ($\rho = 0.83$; $p = 0.011$) demostrates strong agreement in the relative ranking of models. These values fall within the range considered strong evidence of validity and stability in reproducibility studies, as reported in methodological literature [14].

The adjusted regression line ($E2 = 1.37 \cdot E1 – 3.25$) exhibited a pronounced positive slope, suggesting that variations in scores assigned by Text Cortex are proportionally reflected in the scores assigned by Grok Code Fast 1. The balanced distribution of points around the fitted line confirms the absence of systematic evaluator-specific biases and indicates that the observed dispersion arises from inherent variability in LLM performance. These results reinforce the methodological rigor of the study, demonstrating that evaluation was consistent, stable, and free of systematic scoring effects.

Our findings are consistent with recent evidence showing that LLM performance depends significantly on model architecture, training volume, and the degree of fine-tuning to the biomedical domain. For instance, one review study reported that LLMs can answer free-text medical queries with promising levels of aptitude but still show heterogeneous results [15]. In another study, which evaluated an LLM on the United States Medical Licensing Examination a performance exceeding 85% accuracy was reported, highlighting the model’s capacity for contextualized clinical reasoning [16]. Similarly, recent methodological reviews indicate that factual accuracy and narrative consistency in advanced models depend not only on training corpus size but also on hallucination control and reinforcement with clinically verified data via Reinforcement Learning with Human Feedback [17]. These factors may explain the observed superiority models M1 and M3 over less specialized models such as M6 and M8, which lack systematic biomedical fine-tuning.

The integration of LLMs with structured knowledge bases (e.g., UMLS, SNOMED CT, PubMedQA) has been shown to improve factual coherence and reduce interpretive errors, reinforcing the contextual validity of their responses [15]. This observation aligns with the results of the present study: the models with greater semantic integration and medical contextualization - M1 and M3 - achieved the highest clarity and accuracy scores.

From a statistical standpoint, the choice of non-parametric methods is well supported in the AI evaluation literature, given the frequent non-normality of score distributions between models [18]. Likewise, the high inter-rater correlation obtained is consistent with values reported in previous studies as markers of reproducibility in automated evaluation contexts [14].

From a broader perspective, our results reinforce the premise that the quality of AI in healthcare depends not only on algorithmic scale but also on ethical, regulatory, and human supervision frameworks. In this regard, the findings align with conclusions established in other works [15,19]. Consequently, the integration of AI into clinical laboratory workflows requires prospective validation, documentation traceability, and expert review before implementation in critical medical decisions.

From an applied perspective, the results confirm that LLMs can play a complementary role in the generation, analysis, and interpretation of biomedical knowledge, provided their performance is validated through objective, reproducible, and comparative criteria. In particular, model M1 emerges as the most consistent in precision and conceptual clarity, suggesting potential applications in clinical laboratory support, scientific writing, and medical education.

Nevertheless, their use should be conceived as a form of cognitive assistance intended to complement, not to replace professional human judgment and critical review. In high-stakes settings, the presence of biases, inaccuracies, or errors may carry serious consequences.

Among the main limitations of this study, it should be noted that the analyzed responses represent a single temporal snapshot of performance; therefore, potential variations due to model updates or architectural modifications were not evaluated. Future research should incorporate longitudinal analyses to examine performance evolution across new versions or training adjustments. Additionally, we recommend including further metrics of semantic coherence, expert cross-validation, and thematic sensitivity testing in specific biomedical areas such as oncology, genetics, or molecular diagnostics, in order to strengthen the evaluative framework

and consolidate clinical applicability.

Conclusions

The results of this study indicate that M1 achieved the highest overall performance, followed by M2 and M3. However, it is suggested that LLMs may operate in a complementary manner, as one model may outperform another in specific criteria, providing differentiated strengths depending on the evaluation conducted. The correlation analysis revealed a high level of consistency between the two automated evaluators ($r = 0.91$; $\rho = 0.83$), reinforcing the methodological robustness of the study. The fitted regression line, $E2 = 1.37 \cdot E1 - 3.25$, showed a stable positive relationship between the systems, with no evidence of differential bias. Taken together, these findings confirm that the inter-model comparisons were carried out under reliable and reproducible evaluation conditions.

These results suggest that LLMs represent promising tools for analytical and documentary support in the clinical laboratory context, provided their use is accompanied by appropriate ethical, regulatory, and methodological safeguards.

Ethical approval

This study did not involve human or animal subjects and therefore did not require ethical approval, in accordance with the Declaration of Helsinki (different versions).

Informed consent

Not applicable.

Conflict of interest

None declared.

Funding

This study did not receive financial support from public or private agencies.

Use of AI

This manuscript was drafted with the assistance of ChatGPT GPT-5 (OpenAI, 2025) for language refinement. The authors verified the accuracy and coherence of all content.

Author Contributions (CRediT)

Role	Author(s)
Conceptualization	Nilda Fink, Micaela Avellaneda, Pamela De Francesco
Methodology	Nilda Fink, Micaela Avellaneda, Pamela De Francesco
Formal analysis	Nilda Fink, Micaela Avellaneda, Pamela De Francesco
Investigation	Nilda Fink, Micaela Avellaneda, Pamela De Francesco
Validation	Nilda Fink
Writing – original draft	Micaela Avellaneda, Pamela De Francesco
Writing – review & editing	Nilda Fink, Micaela Avellaneda, Pamela De Francesco
Supervision	Nilda Fink

Data Availability and Transparency Statement

All data supporting the findings of this study are available upon reasonable request to the corresponding author. Anonymized datasets, scoring matrices, and figure source files are preserved for verification and reproducibility. The authors affirm full transparency in data handling, statistical analysis, and manuscript preparation, in accordance with IFCC and COPE guidelines.

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56. doi: 10.1038/s41591-018-0300-7.
2. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med. 2019;25(1):24-29. doi: 10.1038/s41591-018-0316-z.
3. Van Moll C, Egberts T, Wagner C, Zwaan L, Ten Berg M. The Nature, Causes, and Clinical Impact of Errors in the Clinical Laboratory Testing Process Leading to Diagnostic Error: A Voluntary Incident Report Analysis. J Patient Saf. 2023;19(8):573-579. doi: 10.1097/PTS.0000000000001166.
4. Nordin N, Ab Rahim SN, Wan Omar WFA, Zulkarnain S, Sinha S, Kumar S, Haque M. Preanalytical Errors in Clinical Laboratory Testing at a Glance: Source and Control Measures. Cureus. 2024;16(3):e57243. doi: 10.7759/cureus.57243.
5. Simundic AM, Lippi G. Preanalytical phase--a continuous challenge for laboratory professionals. Biochem Med (Zagreb). 2012;22(2):145-149. doi: 10.11613/bm.2012.017.
6. Xie H, Jia Y, Liu S. Integration of artificial intelligence in clinical laboratory medicine: Advancements and challenges. Interdisciplinary Medicine 2024;2. doi: 10.1002/inmd.20230056.
7. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. Sci Transl Med. 2015;7(299):299ra122. doi: 10.1126/scitranslmed.aab3719.
8. Velev J, LeBien J, Roche-Lima A. Unsupervised machine learning method for indirect estimation of reference intervals for chronic kidney disease in the Puerto Rican population. Sci Rep. 2023;13:17198. Doi: 10.1038/s41598-023-43830-3.

9. Tyler S, Olis M, Aust N, Patel L, Simon L, Triantafyllidis C, et al. Use of Artificial Intelligence in Triage in Hospital Emergency Departments: A Scoping Review. Cureus. 2024;16(5):e59906. doi: 10.7759/cureus.59906.
10. Porto BM. Improving triage performance in emergency departments using machine learning and natural language processing: a systematic review. BMC Emerg Med. 2024;24(1):219. doi: 10.1186/s12873-024-01135-2.
11. Meng X, Yan X, Zhang K, Liu D, Cui X, Yang Y, et al. The application of large language models in medicine: A scoping review. iScience. 2024;27(5):109713. doi: 10.1016/j.isci.2024.109713.
12. Omar M, Nadkarni GN, Klang E, Glicksberg BS. Large language models in medicine: A review of current clinical trials across healthcare applications. PLOS Digit Health. 2024;3(11):e0000662. doi: 10.1371/journal.pdig.0000662.
13. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 Technical Report. arXiv. 2023. doi: 10.48550/arXiv.2303.08774.
14. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern Med. 2023;183(6):589-596. doi: 10.1001/jamainternmed.2023.1838.
15. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29:1930-1940. doi: 10.1038/s41591-023-02448-8.
16. Ho CN, Tian T, Ayers AT, Aaron RE, Phillips V, Wolf RM, Mathioudakis N, Dai T, Klonoff DC. Qualitative metrics from the biomedical literature for evaluating large language models in clinical decision-making: a narrative review. BMC Med Inform Decis Mak. 2024;24(1):357. doi: 10.1186/s12911-024-02757-z.
17. Lee J, Park S, Shin J, Cho B. Analyzing evaluation methods for large language models in the medical field: a scoping review. BMC Med Inform Decis Mak. 2024;24:366. doi: 10.1186/s12911-024-02709-7.
18. Guo Y, Ovadje A, Al-Garadi MA, Sarker A. Evaluating large language models for health-related text classification tasks with public social media data. J Am Med Inform Assn. 2024;31:2181-2189. doi: 10.1093/jamia/ocae210.
19. Tang YD, Dong ED, Gao W. LLMs in medicine: The need for advanced evaluation systems for disruptive technologies. Innovation (Camb). 2024;5(3):100622. doi: 10.1016/j.xinn.2024.100622.