


COMMENTARY OPEN ACCESS

Toward the Comprehensive Evaluation of Medical Text Generation by Large Language Models: Programmatic Metrics, Human Assessment, and Large Language Models Judgment

Han Yuan 

Duke-NUS Medical School, National University of Singapore, Singapore

Correspondence: Han Yuan (yuan.han@u.duke.nus.edu)

Received: 15 December 2024 | **Accepted:** 25 December 2024

Funding: The author received no specific funding for this work.

Keywords: generative artificial intelligence | generative pre-trained transformer | large language model evaluation | natural language processing

1 | Large Language Models in Medicine

Large language models (LLMs) such as Generative Pre-trained Transformer 4 (GPT-4), characterized by their extensive parameterization (e.g., exceeding 100 billion parameters) [1], predict the likelihood of subsequent word tokens based on the input context and demonstrate exceptional performance across a broad range of medical specialties, such as radiology [2], nephrology [3], and dermatology [4]. Unlike general tasks where erroneous outputs from LLMs are either readily identifiable or have limited consequences, in the evaluation of medical text, errors are often imperceptible to individuals without specialized medical knowledge and pose significant risks to patient safety [5, 6]. Consequently, a comprehensive evaluation of LLMs in clinical text generation is essential before their real-world release for users such as patients and healthcare professionals [7].

Broadly speaking, medical text generation can be categorized into two types: closed-ended and open-ended generation. Closed-ended generation addresses tasks with predefined answers, as exemplified by Wu et al. [3], who used LLMs to answer multiple-choice questions created by the American Society of Nephrology. By contrast, open-ended generation supports more flexible outputs, which enables the handling of complex and

dynamic tasks. For instance, Wan et al. [8] used LLMs to facilitate conversations in the medical reception area, where interactions encompass the diverse topics of general administration, real-time triaging, and addressing primary care concerns, which are not easily formalized within a closed-ended framework. For closed-ended generation, programmatic metrics serve as the gold standard for evaluation. By contrast, evaluating open-ended generation, as shown in Figure 1, requires a more comprehensive approach that incorporates programmatic metrics, human assessment, and LLMs judgment to ensure a well-rounded analysis.

2 | Programmatic Metrics

Programmatic metrics refer to clearly defined mathematical formulas used to compare the LLMs' generation and ground-truth answers without the involvement of humans. For closed-ended tasks, the LLMs' outputs and the ground-truth labels are structurally predefined. Advanced LLMs, following instruction tuning, can accurately adhere to user prompts and generate corresponding responses. By contrast, LLMs with weaker alignment may require additional processing, such as regular expressions to identify patterns and extract relevant answers [3]. Then, LLMs' answers are compared with the

Abbreviations: BLEU, bilingual evaluation understudy; GPT-4, generative pre-trained transformer 4; LLM, large language model; ROUGE, recall-oriented understudy for gisting evaluation.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2025 The Author(s). *Medicine Advances* published by John Wiley & Sons Ltd on behalf of Tsinghua University Press.

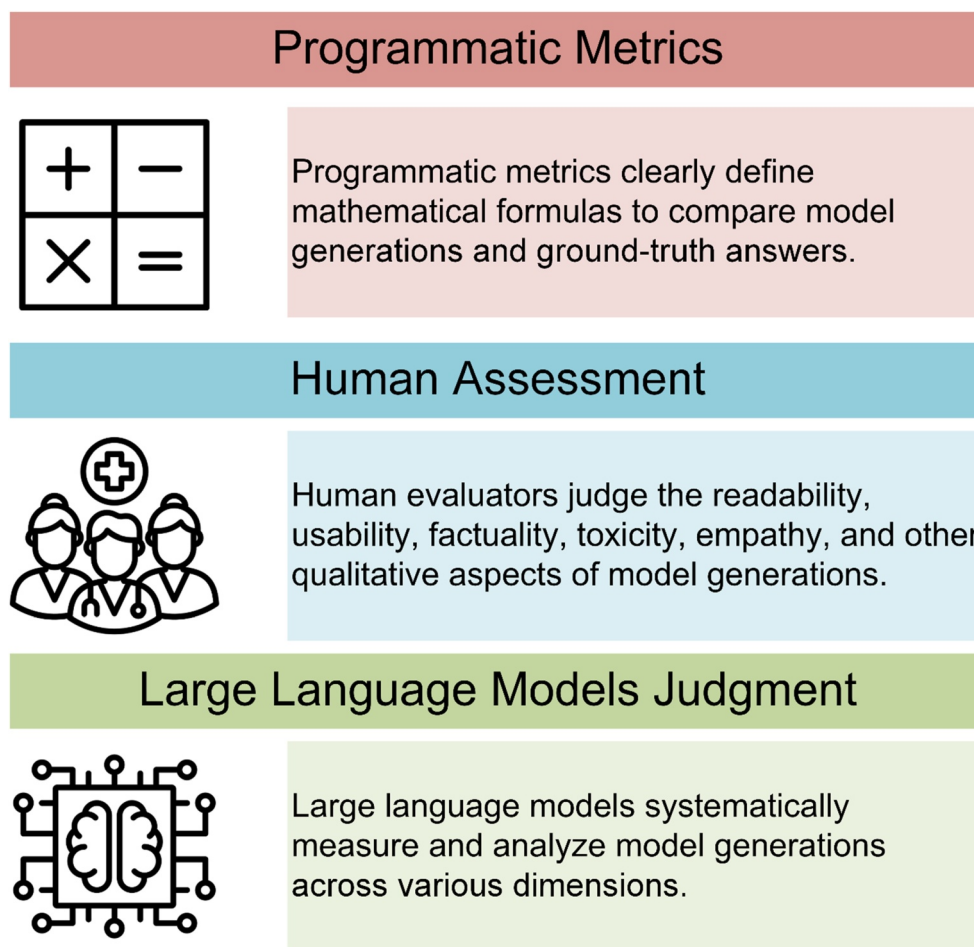


FIGURE 1 | Representative techniques toward the comprehensive evaluation of medical text generation.

ground-truth labels using standard metrics, such as accuracy [3, 9, 10], F_1 -score [11], and area under the receiver operating characteristic curve [11, 12].

For open-ended tasks where LLMs' generation and golden labels are in a free-text format, two commonly used metrics in the general domain are bilingual evaluation understudy (BLEU) [13] and recall-oriented understudy for gisting evaluation (ROUGE) [14]. Both metrics calculate the similarity between LLMs-generated sentences and expert-provided sentences based on contiguous sequences of n tokens, a.k.a. n -grams. BLEU emphasizes the precision of overlapped n -grams with a penalty for overly short outputs, whereas ROUGE offers variants that measure precision, recall, and F_1 -score to provide more flexibility. The two metrics are also widely used in medical text generation [15]. For instance, Sushil et al. [16] quantified LLMs' performance by BLEU and ROUGE in extracting clinically meaningful, complex concepts and relations from oncology reports. In addition to BLEU and ROUGE, programmatic metrics, such as the Word Error Rate [17], Metric for Evaluation of Translation with Explicit Ordering [18], and Bidirectional Encoder Representations from Transformers Score [19], have also been used [3, 20]. For readers seeking detailed computational insights into these programmatic metrics, we recommend consulting the comprehensive survey by Sai, Mohankumar, and Khapra [21].

3 | Human Assessment

Although programmatic metrics provide automatic quantification and are validated to correlate positively with user preferences [19], they fall short of perfectly capturing key aspects such as readability, usability, factuality, toxicity, and empathy [4, 8, 20, 22]. To address these limitations, human assessment is frequently used as an additional safeguard to ensure that the quality of LLMs-generated content meets the stringent criteria required for clinical applicability [23, 24]. For instance, Sandmann et al. [25] systematically analyzed LLMs in the suggestion of initial diagnoses, examination steps, and treatment plans for diverse clinical cases. Two independent physicians evaluated LLMs' outputs based on three criteria: inclusion of relevant options, avoidance of redundancy, and prevention of unjustified statements.

A more comprehensive human assessment was conducted by Singhal et al. [22], which serves as an excellent reference for medical professionals developing in-house assessment frameworks. Specifically, three qualified clinicians manually evaluated LLMs' outputs across five key dimensions: (1) scientific consensus; (2) comprehension, knowledge retrieval, and reasoning capabilities; (3) potential physical or mental-related harm; (4) incorrect or missing content; and (5) bias for medical demographics. Additionally, five laypeople without formal medical training were invited to assess the intent fulfillment,

helpfulness, and actionability of LLMs' generation. Integrating evaluation by clinicians and laypersons, also evidenced in Wan et al. [8], can address the needs of diverse stakeholders. Future researchers are encouraged to conduct human assessment to evaluate both clinical accuracy and user accessibility.

4 | LLMs Judgment

Human assessment, implemented by experienced medical practitioners, remains the gold standard for evaluation. However, it is time-consuming, lacks scalability in high-volume contexts, and requires multiple evaluators to reduce variance and ensure consistency. Programmatic metrics, while scalable, primarily focus on surface-level evaluation and often fail to assess deeper qualities such as contextual relevance, logical coherence, or factual accuracy. Recent advancements in LLMs have enhanced their language comprehension and knowledge synthesis capabilities to a near-human level [9], which has led to their use as an intermediary evaluation method between human assessment and programmatic metrics [26]. Fast et al. [27] validated the average 97% F_1 -score between GPT-4 judgment and human assessment across 20 diagnostic scenarios spanning 13 specialties. Notably, this study revealed no significant bias in GPT-4's self-assessment and highlighted the potential of using LLMs judgment as a robust and objective method for assessing both their own performance and that of other LLMs.

To summarize, in this commentary, three evaluation approaches were discussed for assessing LLMs' generation in healthcare and their applications were illustrated in exemplificative cases. No single approach can address all challenges; however, the combination of these three methods, exemplified by Qiu et al. [9], provides a pipeline toward the comprehensive evaluation of medical text generation.

Author Contributions

Han Yuan: conceptualization, formal analysis, investigation, software, validation, visualization, writing—original draft, writing—review and editing.

Acknowledgments

The author has nothing to report.

Ethics Statement

This study is exempted from review by the ethics committee as it does not involve human participants, animal subjects, or the collection of sensitive data.

Consent

The author has nothing to report.

Conflicts of Interest

The author declares no conflicts of interest.

Data Availability Statement

The author has nothing to report.

References

1. Y. Chang, X. Wang, J. Wang, et al., "A Survey on Evaluation of Large Language Models," *ACM Transactions on Intelligent Systems and Technology* 15, no. 3 (2024): 1–45, <https://doi.org/10.1145/3641289>.
2. D. Van Veen, C. Van Uden, L. Blankemeier, et al., "Adapted Large Language Models Can Outperform Medical Experts in Clinical Text Summarization," *Nature Medicine* 30, no. 4 (2024): 1134–1142, <https://doi.org/10.1038/s41591-024-02855-5>.
3. S. Wu, M. Koo, L. Blum, et al., "Benchmarking Open-Source Large Language Models, GPT-4 and Claude 2 on Multiple-Choice Questions in Nephrology," *NEJM AI* 1, no. 2 (2024), <https://doi.org/10.1056/aidbp2300092>.
4. J. Zhou, X. He, L. Sun, et al., "Pre-Trained Multimodal Large Language Model Enhances Dermatological Diagnosis Using SkinGPT-4," *Nature Communications* 15, no. 1 (2024): 5649, <https://doi.org/10.1038/s41467-024-50043-3>.
5. H. Yuan, "Natural Language Processing for Chest X-Ray Reports in the Transformer Era: BERT-Like Encoders for Comprehension and GPT-Like Decoders for Generation," *iRADIOLOGY* 3, no. 1 (2025): 1–8, <https://doi.org/10.1002/ird3.115>.
6. H. Yuan, "Agentic Large Language Models for Healthcare: Current Progress and Future Opportunities," *Medicine Advances* 3, no. 1 (2025), <https://doi.org/10.1002/med4.70000>.
7. H. Yuan, "Toward Real-World Deployment of Machine Learning for Health Care: External Validation, Continual Monitoring, and Randomized Clinical Trials," *Health Care Science* 3, no. 5 (2024): 360–364, <https://doi.org/10.1002/hcs2.114>.
8. P. Wan, Z. Huang, W. Tang, et al., "Outpatient Reception via Collaboration Between Nurses and a Large Language Model: A Randomized Controlled Trial," *Nature Medicine* 30, no. 10 (2024): 2878–2885, <https://doi.org/10.1038/s41591-024-03148-7>.
9. P. Qiu, C. Wu, X. Zhang, et al., "Towards Building Multilingual Language Model for Medicine," *Nature Communications* 15, no. 1 (2024): 8384, <https://doi.org/10.1038/s41467-024-52417-z>.
10. D. M. Levine, R. Tuwani, B. Kompa, et al., "The Diagnostic and Triage Accuracy of the GPT-3 Artificial Intelligence Model: An Observational Study," *Lancet Digital Health* 6, no. 8 (2024): e555–e561, [https://doi.org/10.1016/S2589-7500\(24\)00097-9](https://doi.org/10.1016/S2589-7500(24)00097-9).
11. B. K. Beaulieu-Jones, M. F. Villamar, P. Scordis, et al., "Predicting Seizure Recurrence After an Initial Seizure-Like Episode From Routine Clinical Notes Using Large Language Models: A Retrospective Cohort Study," *Lancet Digital Health* 5, no. 12 (2024): e882–e894, [https://doi.org/10.1016/S2589-7500\(23\)00179-6](https://doi.org/10.1016/S2589-7500(23)00179-6).
12. L. Y. Jiang, X. C. Liu, N. P. Nejatian, et al., "Health System-Scale Language Models Are All-Purpose Prediction Engines," *Nature* 619, no. 7969 (2023): 357–362, <https://doi.org/10.1038/s41586-023-06160-y>.
13. K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," in *Proceedings of the Annual Meeting on Association for Computational Linguistics* (Philadelphia, PA, 2001), 311–318, <https://doi.org/10.3115/1073083.1073135>.
14. C. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Proceedings of the Workshop on Text Summarization Branches Out* (2004), <https://aclanthology.org/W04-1013/>.
15. A. Soroush, B. S. Glicksberg, E. Zimlichman, et al., "Large Language Models Are Poor Medical Coders: Benchmarking of Medical Code Querying," *NEJM AI* 1, no. 5 (2024), <https://doi.org/10.1056/aidbp2300040>.
16. M. Sushil, V. E. Kennedy, D. Mandair, B. Y. Miao, T. Zack, and A. J. Butte, "CORAL: Expert-Curated Oncology Reports to Advance Language Model Inference," *NEJM AI* 1, no. 4 (2024), <https://doi.org/10.1056/aidbp2300110>.

17. A. Ali and S. Renals, "Word Error Rate Estimation for Speech Recognition: e-WER," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (Melbourne, Australia, 2018), 20–24, <https://doi.org/10.18653/v1/p18-2004>.
18. G. Murray, S. Renals, J. Carletta, and J. D. Moore, "Word Error Rate Estimation for Speech Recognition: e-WER," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (2005), <https://aclanthology.org/P18-2004/>.
19. T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation With BERT," in *Proceedings of the International Conference on Learning Representations* (2020), <https://openreview.net/forum?id=SkeHuCVFDr>.
20. C. Pais, J. Liu, R. Voigt, V. Gupta, E. Wade, and M. Bayati, "Large Language Models for Preventing Medication Direction Errors in Online Pharmacies," *Nature Medicine* 30, no. 6 (2024): 1574–1582, <https://doi.org/10.1038/s41591-024-02933-8>.
21. A. B. Sai, A. K. Mohankumar, and M. M. Khapra, "A Survey of Evaluation Metrics Used for Natural Language Generation Systems," *arXiv* (2020), <https://doi.org/10.48550/arXiv.2008.12009>.
22. K. Singhal, S. Azizi, T. Tu, et al., "Large Language Models Encode Clinical Knowledge," *Nature* 620, no. 7972 (2023): 172–180, <https://doi.org/10.1038/s41586-023-06291-2>.
23. H. Yuan, L. Kang, Y. Li, and Z. Fan, "Human-in-the-Loop Machine Learning for Healthcare: Current Progress and Future Opportunities in Electronic Health Records," *Medicine Advances* 2, no. 3 (2024): 318–322, <https://doi.org/10.1002/med4.70>.
24. H. Yuan, K. Yu, F. Xie, M. Liu, and S. Sun, "Automated Machine Learning With Interpretation: A Systematic Review of Methodologies and Applications in Healthcare," *Medicine Advances* 2, no. 3 (2024): 205–237, <https://doi.org/10.1002/med4.75>.
25. S. Sandmann, S. Riepenhausen, L. Plagwitz, and J. Varghese, "Systematic Analysis of ChatGPT, Google Search and Llama 2 for Clinical Decision Support Tasks," *Nature Communications* 15, no. 1 (2024): 2050, <https://doi.org/10.1038/s41467-024-46411-8>.
26. L. Zheng, W. Chiang, Y. Sheng, et al., "Judging LLM-as-a-Judge With MT-Bench and Chatbot Arena," in *Proceedings of the Advances in Neural Information Processing Systems* (2023), 46595–46623, <https://openreview.net/forum?id=uccHPGDlao>.
27. D. Fast, L. C. Adams, F. Busch, et al., "Autonomous Medical Evaluation for Guideline Adherence of Large Language Models," *NPJ Digital Medicine* 7, no. 1 (2024): 358, <https://doi.org/10.1038/s41746-024-01356-6>.