Editorial

eISSN 2005-8330
https://doi.org/10.3348/kjr.2024.0843
Korean J Radiol 2024;25(10):865-868

Korean Journal of Radiology
KJR

Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Healthcare (MI-CLEAR-LLM)

Seong Ho Park[1], Chong Hyun Suh[1], Jeong Hyun Lee[2], Charles E. Kahn, Jr[3], Linda Moy[4]

[1]Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Republic of Korea
[2]Department of Radiology and Center for Imaging Science, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea
[3]Department of Radiology and Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, USA
[4]Department of Radiology, New York University Grossman School of Medicine, New York, NY, USA

Keywords: Large language model; Large multimodal model; Chatbot; Generative; Artificial intelligence; Reporting; Guideline; Checklist; Healthcare; Medicine; Radiology

Text-based and multimodal generative foundational models—in this article termed "large language models" (LLMs)—have the capability to process textual data and images due to their multimodal capabilities [1,2]. There has been a remarkable surge in published studies that report on the accuracy of LLMs in medical applications, reflecting LLMs' potential to significantly reshape healthcare [3-5]. These studies represent a new genre of medical research. However, the methodology and presentation of results in these studies are highly variable [6]. Inconsistent and incomplete reporting hampers the ability of the reviewers and readers to evaluate the methodology and results of the studies, as well as to assess the replicability of the findings. Consequently, there is a pressing need for guidelines to improve the quality of research reports that present the accuracy of LLMs in healthcare applications [6,7]. The MInimum reporting items for CLear Evaluation of Accuracy Reports of Large Language Models in healthcare (MI-CLEAR-LLM) checklist aims to provide a set of essential items, rather than an exhaustive list, for the transparent reporting of clinical studies that present the accuracy of LLMs in healthcare applications, thereby promoting clearer evaluation of the study findings.

Received: August 27, 2024 **Accepted:** August 27, 2024
Corresponding author: Seong Ho Park, MD, PhD, Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Republic of Korea
• E-mail: parksh.radiology@gmail.com

Item 1. Identification and Specifications of the LLM Used

LLMs are subject to continuous updates, some of which may not be fully known to the users [3,8]. As a result, it is generally extremely difficult for third parties to directly replicate study results, particularly those obtained with commercial models, due to the evolving nature of these models [8]. Consequently, at the very least, transparent and detailed reporting of the LLM's name, version, manufacturer, and the exact date of querying attempts is critical. It is helpful to note the date through which the LLM was trained and whether it has access to web-based information, known as retrieval-augmented generation, or RAG [9].

Item 2. How Stochasticity Was Handled

Unlike conventional artificial intelligence models that produce consistent outputs for given inputs through deterministic operations, LLMs can generate different

responses even when prompted repeatedly with the exact same input. This phenomenon, known as 'stochasticity,' arises from the randomness in the operation of LLMs, particularly in proprietary models [1,8]. For example, when given the task to complete the sentence "The dog is on the...," the LLM generates probabilities for various words such as "floor," "table," "roof," etc. The final word selection is then made by applying a random factor to these probabilities rather than simply choosing the word with the highest probability. Therefore, although probability plays a significant role, a word with a lower probability might be chosen, and the choice may vary at each time. The degree of randomness in the model's operation can be adjusted. In particular, setting a hyperparameter called the 'temperature' close to zero makes the model almost deterministic, meaning it selects the next word almost entirely according to probability, substantially reducing the effect of stochasticity [1].

Therefore, researchers should clearly describe how stochasticity was managed when reporting study results. This includes specifying the number of querying attempts made and, in cases of repeated querying, explaining how the multiple results generated by these attempts were synthesized for analysis (e.g., using at least one correct answer out of a certain number of querying attempts, averaging the results, majority vote, etc.) and providing the rationale for these choices. Without such clarification, the risk of cherry-picking favorable results after multiple querying attempts cannot be entirely dismissed. Furthermore, if repeated querying was used, the study report should include an analysis of the reliability of the LLM outputs across these attempts. Additionally, it is important to specify the settings of technical parameters, such as the temperature, that modify the level of randomness. Compared to proprietary LLMs, most open-weight models provide more options to adjust stochastic versus deterministic behavior, and the details of any pertinent adjustments, if employed, should be clearly reported [10].

Item 3. Full Text of Prompts With Exact Wording and Syntax Used

Even slight modifications in prompts—such as the change of a single word or word order—can lead to significantly different results from LLMs, a phenomenon known as 'prompt brittleness' [8,11,12]. For instance, while the phrases "Calculate the LI-RADS category" and "Determine the LI-RADS category" may seem semantically identical, altering just one word can drastically change the model's output [12]. Given the sensitivity of LLM outputs to prompt variations, complete transparency in reporting the exact wording and syntax of prompts is essential. Providing the exact full text of the prompts used in a study is crucial for replicability and for a clear understanding of the results. This includes precise spellings, symbols, punctuation, spaces, and any other relevant details.

Item 4. A Detailed Explanation of How the Prompts Were Specifically Employed

Not only does the specific wording of prompts affect the LLM output, but also how these prompts are employed also plays a critical role. When testing an LLM with multiple queries, it is crucial to provide a detailed explanation of how the prompts were structured and utilized. Specifically, it should be clarified whether each query and its corresponding prompts were treated as individual chat sessions or if multiple queries were processed together within a single chat session. In the latter scenario, it is important to specify whether the multiple queries were input all at once or sequentially across multiple chat rounds, with new queries being treated as continuations of previous interactions or other prior queries within the session. These distinctions are significant because LLM responses are influenced by prior interactions within the same chat session, which may impact the model's output.

Item 5. Whether Prompt Testing and Optimization Were Used and, If so, Their Details

Given the sensitivity of LLM outputs to prompt variations, researchers often employ various strategies to optimize prompts in order to achieve the desired LLM performance—a process commonly referred to as 'prompt engineering' [12-14]. A detailed elaboration of the optimization process, if used, should include the steps taken to create the prompts, the rationale behind selecting specific wording over alternatives (e.g., by referencing clinical practice guidelines, standardized terminology, or through consultation with subject matter experts), and whether any prompt testing was conducted. If data were used during the prompt testing and optimization process, it is essential to clarify whether the data were entirely independent from the data used to evaluate LLM performance, to ensure a fair assessment

https://doi.org/10.3348/kjr.2024.0843

Korean Journal of Radiology
KJR

Table 1. Minimum items for the transparent reporting of clinical studies that present the performance of LLMs

Item number	Checklist item	Details	Yes/No/NA
1	Identification and specifications of the LLM used	• Name • Version • Manufacturer • Cutoff date for the data used to train the LLM • Whether the LLM has access to web-based information, known as RAG • Date of querying attempts	
2	How stochasticity was handled	• Number of querying attempts made • How the multiple results generated by multiple attempts were synthesized for analysis, and the rationale behind it • Analysis of the reliability of the LLM outputs across multiple attempts • Settings of technical parameters, such as the temperature, that modify the level of randomness	
3	Full text of prompts with exact wording and syntax used	• Precise spellings, symbols, punctuation, spaces, and any other relevant details	
4	Detailed explanation of how the prompts were specifically employed	• Whether each query and its corresponding prompts were treated as individual chat sessions or if multiple queries were processed together in a single session • Whether the multiple queries were input all at once or sequentially across multiple chat rounds	
5	Whether prompt testing and optimization were used and, if so, their details	• Steps taken to create the prompts • Rationale behind selecting specific wording over alternatives	
6	Whether the test dataset was independent	• Whether any portion of the test data was used in the model training or prompt testing and optimization • If sourced from the internet, the exact URLs where they can be found	

LLM = large language model, NA = not applicable, RAG = retrieval-augmented generation

without data leakage.

Item 6. Whether the Test Dataset Was Independent

Transparency is essential regarding the independence of the dataset used to evaluate the performance of LLMs. This includes clarifying whether any portion of the test data was used or referenced during prompt optimization, as mentioned above, and whether the test data might have been included in the model's training process. Since LLMs are often developed using extensive scraping of internet content, there is a risk that test data may have inadvertently been part of the training dataset, potentially leading to data leakage [8,14]. If the test data were sourced from the internet, such as publicly available sets of test questions, the exact URLs where they can be found must be clearly identified.

CONCLUSION

Paying close attention to these items list in Table 1 will facilitate an adequate evaluation of clinical studies on LLM performance in healthcare applications. As these are key minimum items for transparent reporting, researchers should also make an effort to refer to any other relevant reporting guidelines when applicable for additional requirements [6,7,15]. Ensuring clarity and thoroughness in reporting will promote clearer evaluation of studies involving LLMs and help advance research replicability. Going forward, the *Korean Journal of Radiology* will ask authors and/or reviewers to use this checklist when assessing LLM papers.

Conflicts of Interest

Seong Ho Park: Editor-in-Chief without involvement in the editorial evaluation or decision to publish this article; honoraria from Bayer and Korean Society of Radiology; support for travel from Korean Society of Radiology.
Chong Hyun Suh: Assistant to the Editor without

 https://doi.org/10.3348/kjr.2024.0843

involvement in the editorial evaluation or decision to publish this article.
Charles E. Kahn, Jr: Salary support from the Radiological Society of North America (RSNA) paid to employer.
Linda Moy: Salary and travel support from Radiological Society of North America (RSNA) paid to employer for service as Editor of Radiology; grants from Siemens, Gordon and Betty Moore Foundation, Mary Kay Foundation, and Google; consulting fees from Lunit Insight, ICAD, Guerbet, and Medscape; payment or honoraria from ICAD, Lunit, and Guerbet; support for travel from British Society of Breast Radiology, European Society of Breast Imaging, and Korean Society of Radiology; participation on board of ACR DSMB; Society of Breast Imaging board; stock/stock options in Lunit.
The remaining author has declared no conflicts of interest.

Author Contributions

Writing—original draft: Seong Ho Park. Writing—review & editing: Chong Hyun Suh, Jeong Hyun Lee, Charles E. Kahn, Jr, Linda Moy.

ORCID IDs

Seong Ho Park
https://orcid.org/0000-0002-1257-8315
Chong Hyun Suh
https://orcid.org/0000-0002-4737-0530
Jeong Hyun Lee
https://orcid.org/0000-0002-7125-8899
Charles E. Kahn, Jr
https://orcid.org/0000-0002-6654-7434
Linda Moy
https://orcid.org/0000-0001-9564-9360

Funding Statement

None

REFERENCES

1. Bhayana R. Chatbots and large language models in radiology: a practical primer for clinical and research applications. *Radiology* 2024;310:e232756
2. Jung KH. Uncover this tech term: foundation model. *Korean J Radiol* 2023;24:1038-1041
3. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med* 2023;29:1930-1940
4. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med* 2023;6:120
5. Li R, Kumar A, Chen JH. How chatbots and large language model artificial intelligence systems will reshape modern medicine: fountain of creativity or pandora's box? *JAMA Intern Med* 2023;183:596-597
6. CHART Collaborative. Protocol for the development of the chatbot assessment reporting tool (CHART) for clinical advice. *BMJ Open* 2024;14:e081155
7. Park SH, Suh CH. Reporting guidelines for artificial intelligence studies in healthcare (for both conventional and large language models): what's new in 2024. *Korean J Radiol* 2024;25:687-690
8. Kaddour J, Harris J, Mozes M, Bradley H, Raileanu R, McHardy R. Challenges and applications of large language models [accessed on August 27, 2024]. Available at: https://doi.org/10.48550/arXiv.2307.10169
9. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks [accessed on August 26, 2024]. Available at: https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf
10. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: state-of-the-art natural language processing [accessed on August 26, 2024]. Available at: https://doi.org/10.18653/v1/2020.emnlp-demos.6
11. Kim W. Seeing the unseen: advancing generative AI research in radiology. *Radiology* 2024;311:e240935
12. Lee JH, Shin J. How to optimize prompting for large language models in clinical research. *Korean J Radiol* 2024;25:869-873
13. Gu K, Lee JH, Shin J, Hwang JA, Min JH, Jeong WK, et al. Using GPT-4 for LI-RADS feature extraction and categorization with multilingual free-text reports. *Liver Int* 2024;44:1578-1587
14. Sahoo SS, Plasek JM, Xu H, Uzuner Ö, Cohen T, Yetisgen M, et al. Large language models for biomedicine: foundations, opportunities, challenges, and best practices. *J Am Med Inform Assoc* 2024;31:2114-2124
15. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM statement: a targeted guideline for reporting large language models use. medRxiv [Preprint]. 2024 [accessed on August 26, 2024]. Available at: https://doi.org/10.1101/2024.07.24.24310930