Letter to the Editor

eISSN 2005-8330
https://doi.org/10.3348/kjr.2024.0788
Korean J Radiol 2024;25(11):1029-1031

Korean Journal of Radiology
KJR

Insufficient Transparency in Stochasticity Reporting in Large Language Model Studies for Medical Applications in Leading Medical Journals

Chong Hyun Suh[1], Jeho Yi[2], Woo Hyun Shim[1,3], Hwon Heo[3]

[1]Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Republic of Korea
[2]Asan Medical Library, University of Ulsan College of Medicine, Seoul, Republic of Korea
[3]Department of Medical Science, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Republic of Korea

Keywords: Artificial intelligence; Deep learning; Large language model; Large multimodal model; Chatbot; Reporting guideline; Reporting quality; Stochasticity; Healthcare; Medicine; Radiology

Large language models (LLMs) can potentially reshape healthcare [1-3]. Numerous studies continue to report on the performance of LLMs in medical applications. Unlike conventional artificial intelligence models, such as convolutional neural networks, which produce consistent outputs for given inputs through deterministic operations, LLMs can generate varying responses even when prompted repeatedly with the exact same query. This phenomenon, known as 'stochasticity,' results from random elements in the operation of LLMs [4,5].

Stochasticity-related variability in LLM outputs presents critical challenges for both medical practice and scientific research. Maintaining consistent information is vital in medical practice; therefore, understanding the extent and nature of this stochasticity-related variability is crucial for assessing LLMs for medical applications. Failure to adequately address or report stochasticity can hinder the replicability of research findings, as highlighted in a recent editorial [6]. Moreover, the lack of transparency in reporting stochasticity raises concerns that this characteristic of LLMs could be exploited to selectively present favorable results. Despite these limitations, published studies have often overlooked the issue of stochasticity. To address this gap, we conducted a systematic analysis of the published literature, focusing on the reporting practices of stochasticity in research studies evaluating the performance of LLMs in medical applications.

A systematic literature search was conducted to identify research articles evaluating the performance of LLMs in medical applications, as illustrated in Figure 1. This search was performed using PubMed, covering articles published between November 30, 2022 (the release date of ChatGPT by OpenAI) to June 25, 2024. The search query employed was: "(large language model) OR (chatgpt) OR (gpt-3.5) OR (gpt-4) OR (bard) OR (gemini) OR (claude) OR (chatbot)." To manage the large number of results, we focused on those perceived as high-quality publications by selecting studies from journals ranked in the top deciles according to the 2023 Journal Impact Factor. These journals were indexed in the Science Citation Index Expanded and were among the top 10% in each of the 59 subject categories within the Clinical Medicine group as defined by the Journal Citation Reports. The number of querying attempts for each query, methods for handling multiple results, and reliability analysis across repeated queries were extracted from the eligible articles. The proportion of articles that clearly reported stochasticity-related issues was determined. Additionally, a subgroup analysis was conducted by excluding studies that used a temperature setting of zero, as this setting makes the model essentially deterministic and thus minimizes the need to address stochasticity [4]. An experienced medical librarian initially identified the article candidates. All subsequent steps were carried out

Received: August 14, 2024 **Accepted:** August 14, 2024
Corresponding author: Chong Hyun Suh, 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: chonghyunsuh@amc.seoul.kr

independently by two reviewers with expertise in systematic reviews of medical literature. Any disagreements were resolved through consensus.

A total of 159 studies were analyzed (Fig. 1; see Supplementary Table 1 for the full list), of which 147 remained after excluding studies with a temperature of setting of zero. The reporting of stochasticity-related issues is summarized in Table 1. Only 15.1% of the studies (24/159) clearly reported these stochasticity-related issues, while 84.3% (134/159) failed to disclose the number of

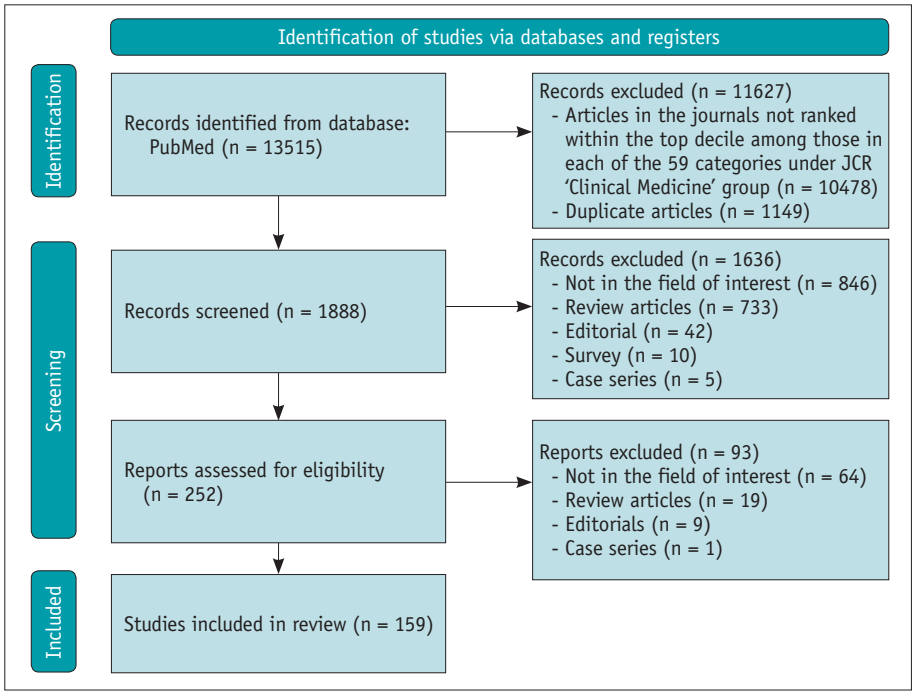

Fig. 1. PRISMA flow diagram for systematic literature analysis. JCR = Journal Citation Reports

Table 1. Reporting of stochasticity-related issues in the published papers

Reporting characteristic	All studies (n = 159)	Studies, excluding temperature = 0 (n = 147)
1. Reporting of querying attempts for obtaining results used for performance analysis		
Clear reporting: specification of querying attempts number and, if applicable, methods to handle multiple results	24 (15.1)	22 (15.0)
• Single attempt	1 (0.6)	1 (0.7)
• Multiple repeat attempts with specified methods for selecting/creating results for analysis	23 (14.5)	21 (14.3)
- Individual attempts (e.g., first attempt) analyzed exclusively	17	15
- Averaging the results from multiple attempts	3	3
- Counted as correct if any attempt gave correct result	2	2
- Majority vote among multiple attempts	1	1
Unclear reporting	135 (84.9)	125 (85.0)
• Lack of disclosure on the number of querying attempts	134 (84.3)	125 (85.0)
• Multiple repeat attempts disclosed without specifying the methods for selecting/creating results for analysis	1 (0.6)	0 (0)
2. Reporting of reliability analysis across results from repeat querying attempts (for 158 and 146 studies, excluding one study that explicitly reported the use of single attempt)		
Present	20 (12.7)	17 (11.6)
Absent	138 (87.3)	129 (88.4)

Data are shown as the number or number of studies (%)

https://doi.org/10.3348/kjr.2024.0788

query attempts. Additionally, only 12.7% of the studies (20/158, excluding one study that explicitly reported using a single attempt) included a reliability analysis of the results from repeated querying attempts. These results were consistent across the 147 studies.

The literature analysis was limited by the use of PubMed alone. Nevertheless, the findings revealed an unequivocal substantial deficiency in the reporting of stochasticity-related issues in studies on the performance of LLMs in medical applications. As our analysis focused on studies published in leading medical journals, the reporting quality of studies from lower-tier journals might even be more deficient. A compelling need exists to enhance transparency and thoroughness in reporting stochasticity, particularly through the reporting guidelines [6].

Supplement

The Supplement is available with this article at https://doi.org/10.3348/kjr.2024.0788.

Conflicts of Interest

Chong Hyun Suh, an Assistant to the Editor of the *Korean Journal of Radiology*, was not involved in the editorial evaluation or decision to publish this article. The remaining author has declared no conflicts of interest.

Author Contributions

Conceptualization: Chong Hyun Suh. Funding acquisition: Woo Hyun Shim. Investigation: all authors. Methodology: all authors. Supervision: Chong Hyun Suh. Writing—original draft: Chong Hyun Suh. Writing—review & editing: Jeho Yi, Woo Hyun Shim, Hwon Heo.

ORCID IDs

Chong Hyun Suh
https://orcid.org/0000-0002-4737-0530
Jeho Yi
https://orcid.org/0009-0002-2322-7454
Woo Hyun Shim
https://orcid.org/0000-0002-7251-2916
Hwon Heo
https://orcid.org/0000-0002-6103-4680

Funding Statement

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (HR20C0026) and a grant from the Asan Institute for Life Sciences, Asan Medical Center, Seoul, Republic of Korea (2024IP0060-1). The funders had no specific roles in this study.

REFERENCES

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med* 2023;29:1930-1940
2. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med* 2023;6:120
3. 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
4. Bhayana R. Chatbots and large language models in radiology: a practical primer for clinical and research applications. *Radiology* 2024;310:e232756
5. Kaddour J, Harris J, Mozes M, Bradley H, Raileanu R, McHardy R. Challenges and applications of large language models. arXiv [Preprint]. 2023 [accessed on August 10, 2024]. Available at: https://doi.org/10.48550/arXiv.2307.10169
6. 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