Original Article | Artificial Intelligence

eISSN 2005-8330
https://doi.org/10.3348/kjr.2024.1161
Korean J Radiol 2025;26(4):304-312

Korean Journal of Radiology
KJR

Adherence of Studies on Large Language Models for Medical Applications Published in Leading Medical Journals According to the MI-CLEAR-LLM Checklist

Ji Su Ko[1]*, Hwon Heo[2]*, Chong Hyun Suh[1], Jeho Yi[3], Woo Hyun Shim[1,2]

[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 Medical Science, Asan Medical Institute of Convergence Science and Technology, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Republic of Korea
[3]Asan Medical Library, University of Ulsan College of Medicine, Seoul, Republic of Korea

Objective: To evaluate the adherence of large language model (LLM)-based healthcare research to the Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Healthcare (MI-CLEAR-LLM) checklist, a framework designed to enhance the transparency and reproducibility of studies on the accuracy of LLMs for medical applications.

Materials and Methods: A systematic PubMed search was conducted to identify articles on LLM performance published in high-ranking clinical medicine journals (the top 10% in each of the 59 specialties according to the 2023 Journal Impact Factor) from November 30, 2022, through June 25, 2024. Data on the six MI-CLEAR-LLM checklist items: 1) identification and specification of the LLM used, 2) stochasticity handling, 3) prompt wording and syntax, 4) prompt structuring, 5) prompt testing and optimization, and 6) independence of the test data—were independently extracted by two reviewers, and adherence was calculated for each item.

Results: Of 159 studies, 100% (159/159) reported the name of the LLM, 96.9% (154/159) reported the version, and 91.8% (146/159) reported the manufacturer. However, only 54.1% (86/159) reported the training data cutoff date, 6.3% (10/159) documented access to web-based information, and 50.9% (81/159) provided the date of the query attempts. Clear documentation regarding stochasticity management was provided in 15.1% (24/159) of the studies. Regarding prompt details, 49.1% (78/159) provided exact prompt wording and syntax but only 34.0% (54/159) documented prompt-structuring practices. While 46.5% (74/159) of the studies detailed prompt testing, only 15.7% (25/159) explained the rationale for specific word choices. Test data independence was reported for only 13.2% (21/159) of the studies, and 56.6% (43/76) provided URLs for internet-sourced test data.

Conclusion: Although basic LLM identification details were relatively well reported, other key aspects, including stochasticity, prompts, and test data, were frequently underreported. Enhancing adherence to the MI-CLEAR-LLM checklist will allow LLM research to achieve greater transparency and will foster more credible and reliable future studies.

Keywords: Large language model; Large multimodal model; Chatbot; Generative; Artificial intelligence; Deep learning; Reporting; Guideline; Checklist; Standard; Adherence; Quality

INTRODUCTION

Studies have increasingly reported on the accuracy and utility of large language models (LLMs) or large multimodal models across a wide range of medical contexts, underscoring their potential to transform the healthcare landscape [1-3]. Nevertheless, significant variability in methodologies and reporting practices has been observed

Received: November 12, 2024 **Revised:** December 25, 2024 **Accepted:** January 3, 2025
*These authors contributed equally to this work.
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

Korean Journal of Radiology
KJR

across these studies, leading to challenges in assessing and interpreting their findings [4,5]. As this body of research grows, the lack of standardized reporting has resulted in inconsistent methodological practices, complicating the process of assessing and replicating findings [4-7].

Several reporting checklists and guidelines have been proposed to address these issues. The Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Healthcare (MI-CLEAR-LLM) checklist was recently developed to provide a set of essential items uniquely tailored to the use of LLMs, distinguishing it from those for general artificial intelligence (AI) applications [5]. These items represent the minimum requirements necessary for the transparent reporting of clinical studies that assess the accuracy of LLM use in healthcare applications. The MI-CLEAR-LLM checklist identifies six essential reporting items that address critical areas affecting the reproducibility and interpretability of LLM performance: 1) identification and specifications of the LLM used, 2) how stochasticity was managed, 3) detailed reporting of prompt wording and syntax, 4) how prompts were structured, 5) details of any prompt testing and optimization conducted, and 6) clarification of the independence of the test and training datasets [5]. Although other guidelines are still under development, such as the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD)-LLM (available as a preprint at the time of the current study), the CANGARU project (ChatGPT and Artificial Intelligence Natural Large Language Models for Accountable Reporting and Use), and the Chatbot Assessment Reporting Tool (CHART), they have not yet been formally published [4,8-10].

While one prior brief study examined adherence to one specific MI-CLEAR-LLM checklist item, stochasticity, no study has comprehensively evaluated adherence to the checklist as a whole [11]. Therefore, we utilized the MI-CLEAR-LLM framework to evaluate current reporting practices for LLM-based medical studies. This study aimed to systematically examine how well these essential items were addressed and to provide an objective assessment of adherence to this framework. By identifying gaps and areas for improvement, our study seeks to highlight aspects of reporting practices that can be improved to support better and more reliable research and applications of LLMs in healthcare.

MATERIALS AND METHODS

Institutional Review Board approval was waived for this retrospective study because it did not involve human participants. This study was performed according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses [12,13] guidelines. The study protocol was registered in PROSPERO (number: CRD42024622870).

Literature Search Strategy and Study Selection

This study adopted the same literature search methodology as a previous study [11] that focused on investigating the reporting practices of stochasticity in original research evaluating the performance of LLMs in medical applications.

A systematic literature search of the PubMed database was conducted to identify research articles on the performance of LLMs in medical applications. The search covered articles published from November 30, 2022 (release date of ChatGPT by OpenAI), through June 25, 2024. The search query used was as follows: "(large language model) OR (chatgpt) OR (gpt-3.5) OR (gpt-4) OR (bard) OR (gemini) OR (claude) OR (chatbot)." By including the term 'large language model,' we expected to capture the majority of studies related to LLMs. Given the large number of potential candidate articles, we decided to focus on high-quality publications; accordingly, only original research from journals ranked within the top decile based on the 2023 Journal Impact Factor and those indexed in the Science Citation Index Expanded within the 'Clinical Medicine' category of the Journal Citation Reports were selected. An experienced medical librarian (J.Y.) conducted the initial search for candidate articles, after which study selection was performed independently by two reviewers (C.H.S. and J.S.K., with 11 and 3 years of experience in conducting systematic reviews, respectively).

Data Extraction and Analysis

Data were independently extracted by two reviewers: J.S.K. (a neuroradiologist and second-year clinical fellow) and H.H. (a research faculty member with expertise in deep learning development). Data on all six key items of the MI-CLEAR-LLM framework were collected and evaluated for each study [5]. The title, abstract, introduction, methods, results, discussion, and supplementary files of all the articles were evaluated for this purpose. In case of disagreement, consensus was reached through consultation with C.H.S., a neuroradiologist with experience in LLM-based research,

 https://doi.org/10.3348/kjr.2024.1161

who supervised the data extraction process. The proportion of articles adhering to each item was calculated, and the following aspects were analyzed:

- Item 1: Identification and specifications of the LLM used, including 1) LLM name, 2) version, 3) manufacturer, 4) cutoff date of training data, 5) whether the LLM had access to web-based information (e.g., retrieval-augmented generation [RAG]), and 6) date of query attempts.
- Item 2: How stochasticity was handled. This analysis focused on reverifying whether stochasticity-related issues were clearly reported. Stochasticity-related reporting details have been omitted because they were specifically addressed in a previous study [11].
- Item 3: Reporting the full text of prompts, including precise spelling, symbols, punctuation, and spaces.
- Item 4: How prompts were employed, including 1) whether each query and corresponding prompt were treated as individual chat sessions or whether multiple queries were processed together in a single session, and 2) whether multiple queries were input simultaneously or sequentially across multiple chat rounds. To categorize LLM access methods, we classified the studies into three types: browser-based public interface access (e.g., ChatGPT), public application programming interface (API) access, and local LLMs or institutional API access.
- Item 5: Prompt testing and optimization details, including 1) steps taken to create the prompts, and 2) rationale behind selecting specific wording over alternatives.
- Item 6: Whether the test dataset was independent, including 1) whether any portion of the test data was used during model training or prompt optimization, and 2) whether data were sourced from the internet and the exact source URLs were provided.

Additionally, the included studies were categorized as radiology-related or in other fields.

Statistical Analysis

Agreement between the results extracted by the two reviewers were analyzed using Cohen's kappa (κ) value to assess interrater reliability. A κ-value >0.8 indicated almost perfect agreement, whereas 0.61–0.80, 0.41–0.60, 0.21–0.40, <0.20 indicated substantial, moderate, fair, and poor agreement, respectively [14]. Adherence percentages were calculated. Adherence rates were compared between radiology-related studies and those in other fields using chi-square and Fisher's exact tests. Statistical significance was set at $P < 0.05$. All statistical analyses were conducted using SPSS software (version 27.0 for Windows; IBM Corp., Armonk, NY, USA).

RESULTS

Literature Search

The systematic search initially yielded 13515 articles. A total of 1149 duplicate studies and 10478 articles in journals not ranked within the top decile of each of the 59 categories under the Journal Citation Reports 'Clinical Medicine' group were excluded. The abstracts of the 1888 remaining articles were then screened, of which 1636 were excluded (846 articles that were unrelated to the field of interest, 733 review articles, 42 editorials, 10 surveys, and 5 case series). Full-text reviews of the 252 remaining potentially eligible articles were performed, after which 93 articles were excluded (64 articles unrelated to the field of interest, 19 review articles, 9 editorials, and 1 case series). Finally, 159 original research articles were included in the analysis (Fig. 1). A list of the included articles is provided in Supplementary Table 1.

To ensure transparency, the current study utilized the same dataset as that used in a previous study [11]. However, that study focused exclusively on reporting practices related to stochasticity. By contrast, the current study applied the full MI-CLEAR-LLM checklist, which was not available at the time of the earlier analysis, to provide a comprehensive evaluation of reporting practices across the six key items. This broader approach allowed for a more comprehensive assessment of the reporting quality in LLM-based medical studies.

Characteristics of Included Studies

The subject areas of the studies covered a broad range of medical applications (Table 1). General medicine was the most common field, represented by 54 studies (34.0%, 54/159), followed by radiology and nuclear medicine (11.3%, 18/159) and ophthalmology (8.2%, 13/159). Other subject areas such as oncology, neurosurgery, gastroenterology, urology, neurology, pediatrics, obstetrics and gynecology, emergency medicine, cardiology, orthopedic surgery, and psychiatry were also represented. Eighteen studies were radiology-related and 141 were in other fields.

https://doi.org/10.3348/kjr.2024.1161

Korean Journal of Radiology
KJR

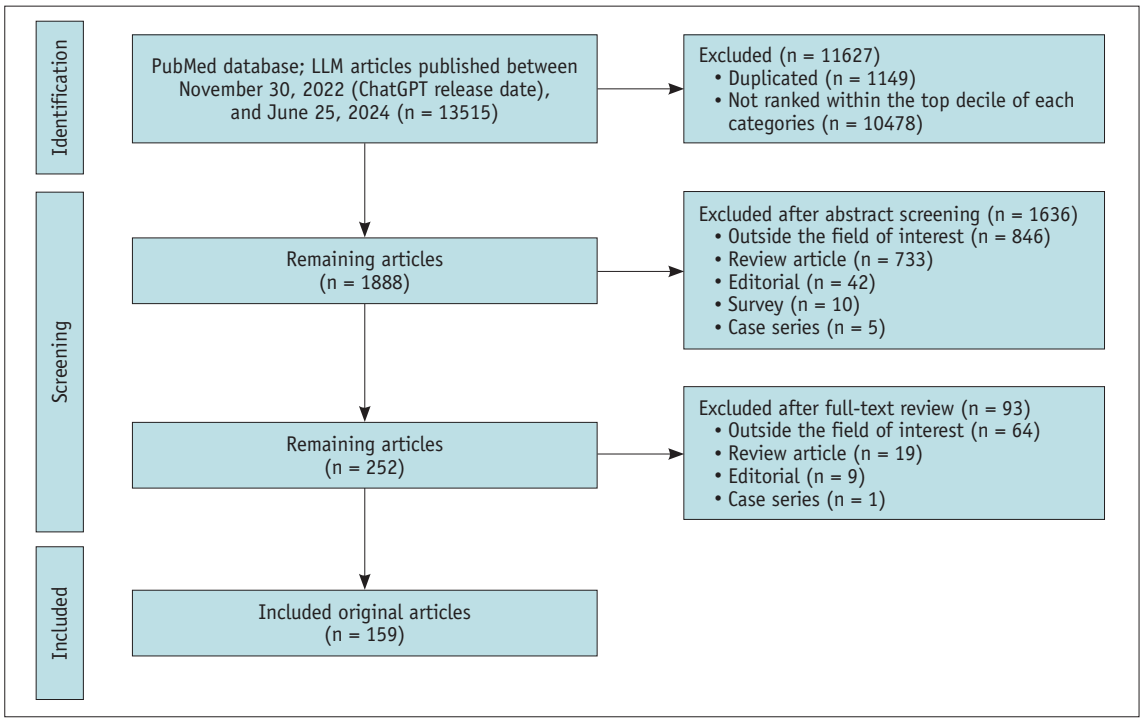

Fig. 1. Flowchart of the literature search process. LLM = large language model

Table 1. Subject fields of included articles (n = 159)

Subject field	n (%)
General medicine	54 (34.0)
Radiology and nuclear medicine	18 (11.3)
Ophthalmology	13 (8.2)
Oncology and radiation oncology	9 (5.7)
Neurosurgery	9 (5.7)
Gastroenterology	8 (5.0)
Urology	6 (3.8)
Neurology	5 (3.1)
Pediatrics	5 (3.1)
Obstetrics and gynecology	4 (2.5)
Emergency medicine	4 (2.5)
Cardiology	4 (2.5)
Orthopedic surgery	4 (2.5)
Psychiatry	4 (2.5)
Other fields	12 (7.5)

Adherence to the MI-CLEAR-LLM Checklist

Among the 159 studies analyzed, the number of MI-CLEAR-LLM checklist components that were satisfied by each study varied widely. Specifically, 1 study satisfied only 2 components, whereas the highest number of components satisfied was 11, which was achieved by 2 studies. The distribution of checklist adherence was as follows: 2 components (1 study), 3 components (6 studies), 4 components (9 studies), 5 components (39 studies), 6 components (34 studies), 7 components (30 studies), 8 components (24 studies), 9 components (9 studies), 10 components (5 studies), and 11 components (2 studies). A comparison between radiology-related studies (n = 18) and studies in other fields (n = 141) revealed significant differences for item 1-4 (cutoff date for data used to train the LLM, $P = 0.017$) and item 3-1 (full text of prompts with exact wording and syntax used, $P = 0.01$). No statistically significant differences were observed for the other checklist items. The interrater reliability between the two reviewers, as measured by Cohen's κ, was 0.69, indicating substantial agreement. Details regarding the extent to which the checklist items were addressed are provided in Table 2 and illustrated in Figure 2.

Item 1: Identification and Specification of the LLM Used

All 159 studies (100%, 159/159) reported the name of the LLM used to ensure clarity regarding which model was used. The version of LLM used was mentioned in 154 studies (96.9%, 154/159). Additionally, 146 studies (91.8%, 146/159) provided the name of the manufacturer, which further enhanced transparency. However, the cutoff date for the training data, which is an important factor in understanding the knowledge scope and limitations of LLMs, was reported in only 86 studies (54.1%, 86/159). Furthermore, only 10 studies (6.3%, 10/159) mentioned

https://doi.org/10.3348/kjr.2024.1161

Table 2. Adherence to the MI-CLEAR-LLM checklist

Item #	Checklist item	All	Radiology studies	Non-radiology studies
Item 1. Identification and specification of the LLM used				
#1-1	Name	159 (100)	18 (100)	141 (100)
#1-2	Version	154 (96.9)	18 (100)	136 (96.5)
#1-3	Manufacturer	146 (91.8)	18 (100)	128 (90.8)
#1-4	Cutoff date for the data used to train the LLM	86 (54.1)	5 (27.8)	81 (57.4)
#1-5	Whether the LLM has access to web-based information (RAG)	10 (6.3)	3 (16.7)	7 (5.0)
#1-6	Date of querying attempts	81 (50.9)	11 (61.1)	70 (49.6)
Item 2. How stochasticity was handled				
#2-1	Clear documentation of stochasticity-related factors*	24 (15.1)	4 (22.2)	20 (14.2)
Item 3. Full text of prompts with exact wording and syntax used				
#3-1	Precise spellings, symbols, punctuation, spaces, and any other relevant details	78 (49.1)	14 (77.8)	64 (45.4)
Item 4. Detailed explanation of how the prompts were specifically employed				
#4-1	Whether each query and its corresponding prompts were treated as individual chat sessions or if multiple queries were processed together in a single session	54 (34.0)	5 (27.8)	49 (34.8)
#4-2	Whether the multiple queries were input all at once or sequentially across multiple chat rounds	55 (34.6)	8 (44.4)	47 (33.3)
Item 5. Whether prompt testing and optimization were used and, if so, their details				
#5-1	Steps taken to create the prompts	74 (46.5)	8 (44.4)	66 (46.8)
#5-2	Rationale behind selecting specific wording over alternatives	25 (15.7)	3 (16.7)	22 (15.6)
Item 6. Whether the test data were independent				
#6-1	Whether any portion of the test data was used in the model training or prompt testing and optimization	21 (13.2)	1 (5.6)	20 (14.2)
#6-2	If sourced from the internet, the exact URLs where they can be found†	43 (56.6)	6 (75.0)	37 (54.4)

Data are number of studies with percentages in parentheses.
*The results for item 2 were cited from Suh et al. [11].
†The percentages were calculated based on the number of studies that sourced test data from the internet, with 76, 8, and 68 being the total number of such studies in all, radiology, and non-radiology categories, respectively.
MI-CLEAR-LLM = Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Healthcare, RAG = retrieval-augmented generation

whether the LLM had access to web-based information such as RAG. The dates of the query attempts were reported in 81 studies (50.9%, 81/159).

Item 2: How Stochasticity Was Handled

Only 15.1% of the studies (24/159) provided clear documentation of stochasticity-related factors, whereas 84.3% (134/159) did not report the number of query attempts. Furthermore, only 12.7% (20/158, excluding one study that specified a single attempt) included a reliability analysis of the results of repeated queries [11].

Item 3: Full Text of Prompts with Exact Wording and Syntax Used

Of the 159 studies, 78 (49.1%) provided detailed information on at least one element related to the exact wording and syntax of the prompts employed. These elements typically included the following.

- Precise spelling: 78 studies (49.1%) ensured that the spelling of each term in the prompts was consistent across multiple query attempts.
- Symbols used: 28 studies (17.6%) described the use of special characters or symbols within prompts. These included curly braces for the JavaScript Object Notation (JSON) structure and square brackets as placeholders.
- Punctuation: 30 studies (18.9%) included the strategic use of quotation marks, commas, and colons to structure prompt templates.
- Spaces: 78 studies (49.1%) focused on the use of spaces, including line breaks.
- Other relevant syntax: 30 studies (18.9%) mentioned additional syntactic elements such as capitalization, indentation, or formatting cues (e.g., the use of [ANSWER] placeholders).

https://doi.org/10.3348/kjr.2024.1161

Korean Journal of Radiology
KJR

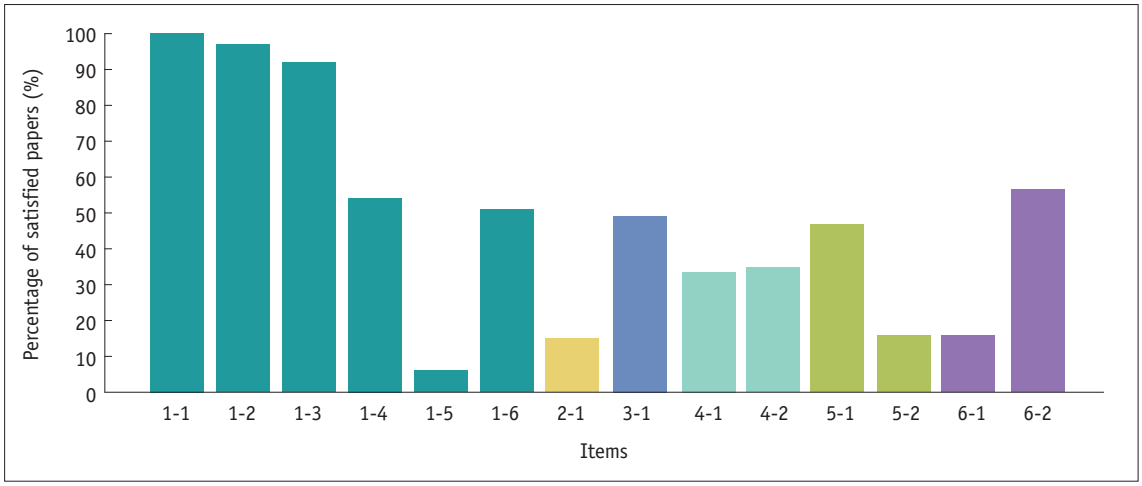

Fig. 2. Proportions of articles that satisfied each item of the MI-CLEAR-LLM checklist (the results for item 2 were cited from Suh et al. [11]). 1-1 = LLM name, 1-2 = version, 1-3 = manufacturer, 1-4 = cutoff date for the data used to train the LLM, 1-5 = whether the LLM had access to web-based information (e.g., RAG), 1-6 = date of querying attempts, 2-1 = how stochasticity was handled, 3-1 = full text of prompts with exact wording and syntax used, 4-1 = whether each query and its corresponding prompts were treated as individual chat sessions or if multiple queries were processed together in a single session, 4-2 = whether the multiple queries were input simultaneously or sequentially across multiple chat rounds, 5-1 = steps taken to create the prompts, 5-2 = rationale behind selecting specific wording over alternatives, 6-1 = whether any portion of the test data was used in the model training or prompt testing and optimization, and 6-2 = if sourced from the internet, the exact URLs where they can be found. The percentage for 6-2 was calculated based on 76 studies that sourced test data from the internet. Otherwise, the denominator was 159 studies. MI-CLEAR-LLM = Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Healthcare, RAG = retrieval-augmented generation

However, a large proportion of the studies (50.9%, 81/159) did not provide sufficient details regarding the exact syntax or structure of the prompts utilized, including information on special characters or spacing.

Item 4: Detailed Explanation of How the Prompts Were Specifically Employed

Of the 159 studies, only 54 (34.0%, 54/159) offered explicit descriptions of their query-handling approach. These studies used clear terminology such as "individual sessions," "separate conversations," or "continuous dialogue" to describe their LLM interaction structure, indicating whether queries were processed individually or in conjunction with other prompts. Among the studies, 84 (52.8%) specified their LLM access methods: 55 employed public interface access (e.g., ChatGPT), 29 utilized public API access, and 17 used local LLMs or institutional API access.

In addition, 55 studies (34.6%, 55/159) documented their query input methodologies such as simultaneous input (batch processing) or sequential input across multiple chat rounds. In the context of our analysis, "multiple chat rounds" refers to back-and-forth interactions occurring either within a single session or across different dates. When queries were introduced sequentially, they were typically independent of each other, unlike chain prompting, in which subsequent queries build upon previous responses. For example, a researcher may first inquire about disease symptoms, followed by an independent question about treatment options, with each query standing alone rather than building upon previous responses.

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

Among the 159 studies examined, 74 (46.5%, 74/159) provided insights into the steps taken to create and refine their prompts. These studies often discussed methodologies such as "prompt engineering," "iterative development," or "prompt design processes," detailing how prompts were tested and optimized for improved results.

In contrast, only 25 studies (15.7%, 25/159) addressed the rationale behind selecting a specific wording over alternatives. These studies explained why particular words, phrases, or syntax were often chosen through comparative analyses of different prompt versions.

Item 6: Whether the Test Data Were Independent of the Model's Training Data

Only 21 studies (13.2%, 21/159) clearly reported whether any portion of the test data had been used during model training or prompt optimization, ensuring that the

 https://doi.org/10.3348/kjr.2024.1161

evaluation results were not biased by prior exposure to the test data. Additionally, 43 (56.6%) of the 76 studies that sourced test data from the internet included the exact URLs from which the data were obtained.

DISCUSSION

This study assessed reporting practices in clinical research utilizing LLMs, focusing on key elements such as model identification, prompt usage, query handling, and dataset independence. As highlighted in a previous review, reproducibility issues in AI research often stem from incomplete documentation and variations in model performance across different runs, even when using the same code, owing to factors such as random initialization and specific training parameters [15]. Our analysis revealed considerable variation in reporting practices across studies, with some aspects being consistently reported, while others were characterized by notable gaps in documentation, reflecting the broader challenges in AI reproducibility [16].

All the studies analyzed (159/159) clearly identified the LLM used by name, and a high percentage (96.9%) reported the version number, allowing readers to accurately trace the specific models used. In addition, the manufacturer was identified in 91.8% of the studies, reflecting a strong trend in reporting basic technical details. However, only 54.1% of the studies included the training data cutoff date, which is a critical element for understanding the relevance of the results obtained using a model. Mitchell et al. [17] emphasized that such temporal boundaries are essential components of model documentation and significantly influence the interpretation of the results. Furthermore, only 6.3% of the studies reported whether the LLM had access to web-based information (i.e., RAG), and 50.9% mentioned the dates on which the model was queried. These specifications are crucial for ensuring the reproducibility and validation of machine learning research results [18].

Only 15.1% of the studies provided clear documentation of stochasticity-related factors, as previously reported [11]. This finding indicates a significant deficiency in the documentation of stochasticity-related factors in studies assessing LLM performance in medical applications. This underscores the urgent need to improve transparency and detailed reporting of stochasticity, ideally through stricter adherence to established reporting checklists.

Only 49.1% of the studies provided detailed information regarding the specific wording and syntax of the prompts used. While traditional journal space limitations may constrain the inclusion of comprehensive prompt details in the main text, Belz et al. [19] demonstrated that supplementary materials provide an ideal platform for documenting critical information in natural language processing research. The importance of complete prompt documentation aligns with the reproducibility standards proposed by Heil et al. [18], which emphasizes the need for comprehensive methodological documentation in machine learning studies.

Some studies provided sufficient information on how prompts were employed, such as whether they were processed as individual chat sessions or as part of a continuous session (34.0%) or whether multiple queries were input simultaneously or sequentially (34.6%). As identified in a previous review, these technical details significantly influence how AI systems interact with inputs and consequently affect study outcomes [16]. Underreporting of session structures and query input methods creates significant challenges for the reproducibility of results. Andaur Navarro et al. [20] found systematic issues in reporting standards across machine learning-based prediction studies, highlighting the need for more rigorous documentation of methodological details, including protocols for interaction with AI systems.

While 46.5% of the studies detailed the steps taken to create and refine prompts, only 15.7% explained the rationale behind specific word choices. This finding aligns with broader concerns about the reproducibility of AI research [15]. Our analysis indicates that researchers should provide comprehensive documentation of their prompt testing processes as supplementary material, following the documentation standards proposed by Mitchell et al. [17].

Only 13.2% of the studies explicitly reported whether the test data used were independent of the training data, and 56.6% provided URLs for internet-derived test data. This lack of transparency regarding test data independence significantly complicates the evaluation of LLM performance [21]. Paullada et al. [22] further emphasized that clear documentation of data sources and their independence are crucial for establishing the reliability of the reported results. The absence of clear documentation regarding the potential overlap between the training and test data makes it difficult to assess whether the model truly demonstrates generalization capabilities. This concern echoes the findings of Andaur Navarro et al. [20] regarding systematic issues in reporting the standards and potential biases in machine

https://doi.org/10.3348/kjr.2024.1161

Korean Journal of Radiology
KJR

learning studies. A recent critical assessment of 23 state-of-the-art LLM benchmarks highlighted key limitations, including biases, difficulties in measuring genuine reasoning, and risks of contamination between the training and evaluation processes [23,24]. Such issues emphasize the need for standardized methodologies and dynamic evaluation frameworks that can better capture the complex behavior of LLMs while mitigating potential biases.

This study aimed to analyze adherence to the MI-CLEAR-LLM checklist, focusing specifically on methodological elements unique to LLM research. Although LLM studies potentially encompass a broader range of evaluation dimensions, the MI-CLEAR-LLM checklist emphasizes the essential items necessary for assessing the clinical accuracy and reliability of LLM-based evaluations. A recent systematic review [25] highlighted that most LLM evaluations focus on assessing knowledge using sources such as the United States Medical Licensing Examination (USMLE) test questions, with limited attention paid to real clinical data and issues such as fairness, bias, toxicity, and deployment considerations. Although it is possible to conduct a more extensive evaluation of LLM research, the MI-CLEAR-LLM checklist concentrates on specific methodological reporting elements to ensure reproducibility and transparency. Thus, although this study's focus on LLM-specific reporting elements could be considered a limitation, it was beyond the scope of our study, which aimed to address the unique features of LLM methodologies in healthcare applications.

In summary, this analysis revealed significant gaps in how key aspects of LLM-related research are currently reported, particularly in areas such as stochasticity handling, prompt usage, and test data independence. While the technical specifications of LLMs, such as the model name and version, are generally well documented, other methodological details affecting LLM performance are often overlooked. This hinders the reproducibility of studies and limits our understanding of LLMs. By improving the transparency of LLM research, the broader scientific community can foster a deeper understanding of these powerful models and ensure that future studies are conducted with greater credibility and reliability.

Supplement

The Supplement is available with this article at https://doi.org/10.3348/kjr.2024.1161.

Availability of Data and Material

The datasets generated or analyzed during the study are included in this published article and its supplement.

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, Woo Hyun Shim. Data curation: Ji Su Ko, Hwon Heo, Jeho Yi. Formal analysis: Ji Su Ko, Hwon Heo, Chong Hyun Suh. Funding acquisition: Chong Hyun Suh. Investigation: Ji Su Ko, Hwon Heo, Chong Hyun Suh, Woo Hyun Shim. Methodology: Hwon Heo, Chong Hyun Suh. Project administration: Chong Hyun Suh. Resources: Ji Su Ko, Hwon Heo. Software: Hwon Heo, Woo Hyun Shim. Supervision: Chong Hyun Suh. Visualization: Ji Su Ko. Writing—original draft: Ji Su Ko, Hwon Heo. Writing—review & editing: Chong Hyun Suh.

ORCID IDs

Ji Su Ko
https://orcid.org/0000-0001-6589-2431
Hwon Heo
https://orcid.org/0000-0002-6103-4680
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

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 (2024IP0060-1) from the Asan Institute for Life Sciences, Asan Medical Center, Seoul, Korea.

REFERENCES

1. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States medical licensing examination (USMLE)? The implications of

large language models for medical education and knowledge assessment. *JMIR Med Educ* 2023;9:e45312
2. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023;2:e0000198
3. Suh PS, Shim WH, Suh CH, Heo H, Park CR, Eom HJ, et al. Comparing diagnostic accuracy of radiologists versus GPT-4V and Gemini Pro Vision using image inputs from diagnosis please cases. *Radiology* 2024;312:e240273
4. CHART Collaborative. Protocol for the development of the Chatbot assessment reporting tool (CHART) for clinical advice. *BMJ Open* 2024;14:e081155
5. Park SH, Suh CH, Lee JH, Kahn CE, Moy L. Minimum reporting items for clear evaluation of accuracy reports of large language models in healthcare (MI-CLEAR-LLM). *Korean J Radiol* 2024;25:865-868
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
7. Huo B, Cacciamani GE, Collins GS, McKechnie T, Lee Y, Guyatt G. Reporting standards for the use of large language model-linked chatbots for health advice. *Nat Med* 2023;29:2988
8. Tejani AS, Yi P, Bluethgen C, D'Antonoli TA, Huisman M. Best practices in large language model reporting [accessed on November 5, 2024]. Available at: https://pubs.rsna.org/page/ai/blog/2024/9/ryai_editorsblog093024?doi=10.1148%2Fryai&publicationCode=ai
9. 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 November 10, 2024]. Available at: https://doi.org/10.1101/2024.07.24.24310930
10. Cacciamani GE, Collins GS, Gill IS. ChatGPT: standard reporting guidelines for responsible use. *Nature* 2023;618:238
11. Suh CH, Yi J, Shim WH, Heo H. Insufficient transparency in stochasticity reporting in large language model studies for medical applications in leading medical journals. *Korean J Radiol* 2024;25:1029-1031
12. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71
13. Park HY, Suh CH, Woo S, Kim PH, Kim KW. Quality reporting of systematic review and meta-analysis according to PRISMA 2020 guidelines: results from recently published papers in the Korean Journal of Radiology. *Korean J Radiol* 2022;23:355-369
14. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-174
15. Hutson M. Artificial intelligence faces reproducibility crisis. *Science* 2018;359:725-726
16. Gundersen OE, Kjensmo S. State of the art: reproducibility in artificial intelligence [accessed on November 10, 2024]. Available at: https://doi.org/10.1609/aaai.v32i1.11503
17. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model cards for model reporting [accessed on November 10, 2024]. Available at: https://doi.org/10.1145/3287560.3287596
18. Heil BJ, Hoffman MM, Markowetz F, Lee SI, Greene CS, Hicks SC. Reproducibility standards for machine learning in the life sciences. *Nat Methods* 2021;18:1132-1135
19. Belz A, Agarwal S, Shimorina A, Reiter E. A systematic review of reproducibility research in natural language processing. arXiv [Preprint]. 2021 [accessed on November 10, 2024]. Available at: https://doi.org/10.48550/arXiv.2103.07929
20. Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Systematic review finds "spin" practices and poor reporting standards in studies on machine learning-based prediction models. *J Clin Epidemiol* 2023;158:99-110
21. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in ML-based science. arXiv [Preprint]. 2022 [accessed on November 10, 2024]. Available at: https://doi.org/10.48550/arXiv.2207.07048
22. Paullada A, Raji ID, Bender EM, Denton E, Hanna A. Data and its (dis)contents: a survey of dataset development and use in machine learning research. *Patterns (N Y)* 2021;2:100336
23. McIntosh TR, Susnjak T, Arachchilage N, Liu T, Watters P, Halgamuge MN. Inadequacies of large language model benchmarks in the era of generative artificial intelligence. arXiv [Preprint]. 2024 [accessed on December 15, 2024]. Available at: https://doi.org/10.48550/arXiv.2402.09880
24. Jegorova M, Kaul C, Mayor C, O'Neil AQ, Weir A, Murray-Smith R, et al. Survey: leakage and privacy at inference time. *IEEE Trans Pattern Anal Mach Intell* 2023;45:9090-9108
25. Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, et al. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA* 2024 Oct 15 [Epub]. http://doi.org/10.1001/jama.2024.21700