Systematic Review of Reporting guidelines for large language models used in healthcare research

Saif Aldeen Alryalat*[1,2] and Iyad Sultan[3]

[1]University of Illinois Chicago, Chicago, Illinois, USA 2. The University of Jordan, Amman, Jordan 3. King Hussein Cancer Center (KHCC), Amman, Jordan

Keywords: Large Language Models; Reporting Guidelines; CHART; TRIPOD

DOI: 10.59707/hymrUXPX7081

Published on: December 1, 2025

Abstract

This systematic review aims to synthesize existing reporting guidelines for large language models (LLMs) in healthcare research and evaluate their adequacy in addressing gaps in transparency, reproducibility, and clinical applicability. A systematic search was conducted to identify relevant studies on reporting guidelines for LLMs used in healthcare research using the PubMed database. We included 18 studies focused on reporting guidelines for LLMs used in healthcare research. The studies primarily aimed to develop or evaluate reporting frameworks to improve transparency, reproducibility, and methodological rigor in LLM applications. Several studies focused on creating structured reporting checklists for LLM applications in healthcare. The Chatbot Assessment Reporting Tool (CHART) was developed across multiple studies. Similarly, TRIPOD-LLM extended the TRIPOD+AI framework with 19 main items and 50 subitems, emphasizing modular reporting for diverse LLM tasks. Ultimately, while existing reporting guidelines represent an important advancement toward standardizing LLM research, their long-term impact will rely on broad adoption and iterative refinement to meet the evolving challenges of artificial intelligence.

Introduction

The integration of large language models (LLMs) into healthcare research has rapidly expanded, offering transformative potential for tasks ranging from clinical decision support to data analysis. In healthcare, LLMs are increasingly utilized for generating medical reports, interpreting diagnostic results, and even assisting in systematic reviews [1,2]. However, the lack of standardized reporting guidelines for LLM applications in this domain raises concerns about reproducibility, bias, and clinical reliability [3,4]. The absence of robust frameworks to document model architectures, training data, and validation methods undermines the transparency and trustworthiness of LLM-driven research, necessitating the development of tailored reporting standards [3,5]. Current challenges in LLM reporting for healthcare research include inconsistencies in evaluating model performance, inadequate documentation of training datasets, and limited attention to ethical considerations such as bias and fairness [1,5]. For instance, studies reveal that only 5% of LLM evaluations in healthcare use real patient data, while

*Corresponding author: Saif Aldeen Alryalat ¡saifryala@yahoo.com¿

most focus narrowly on accuracy metrics, neglecting critical dimensions like calibration and deployment risks [1,6]. Controversies also persist regarding the optimal methodologies for validating LLM outputs, with some advocating for domain-specific adaptations of existing guidelines like TRIPOD-LLM. In contrast, others call for entirely novel frameworks [3,4]. Additionally, the ”black-box” nature of LLMs complicates interpretability, particularly in high-stakes clinical settings where erroneous outputs could harm patients [7]. This systematic review aims to synthesize existing reporting guidelines for LLMs in healthcare research and evaluate their adequacy in addressing gaps in transparency, reproducibility, and clinical applicability. By critically assessing current practices and proposing evidence-based recommendations, this work seeks to inform policy and standardization efforts, ensuring LLM applications meet rigorous scientific and ethical standards.

Methods

Search Strategy A systematic search was conducted to identify relevant studies on reporting guidelines for large language models (LLMs) used in healthcare research. The PubMed database was queried as follows (”large language model”[Title/Abstract] OR ”LLM”[Title/Abstract] OR ”generative AI”[Title/Abstract] OR ”chatbot”[Title/Abstract] OR ”GPT”[Title/Abstract]) AND (”healthcare”[MeSH Terms] OR ”health care”[Title/Abstract] OR ”medical”[Title/Abstract] OR ”clinical”[Title/Abstract] OR ”medicine”[MeSH Terms]) AND (”reporting guideline”[Title/Abstract] OR ”reporting standard”[Title/Abstract] OR ”writing checklist*”[Title/Abstract]) The search aimed to capture studies addressing reporting standards for LLM applications in healthcare, including clinical advice, evidence synthesis, and medical research. Screening Process We screened titles and abstracts for relevance based on predefined inclusion and exclusion criteria. Full-text screening was then performed to confirm eligibility. Inclusion Criteria: Studies proposing or evaluating reporting guidelines for LLMs in healthcare; Studies addressing generative AI applications in clinical decision-making, medical research, or patient communication; and peer-reviewed articles, protocols, or consensus statements. Exclusion Criteria: Studies not focused on LLMs or generative AI; Non-healthcare applications of LLMs; Articles without explicit reporting guideline development or evaluation. Data Extraction The extracted data elements included the primary objective, secondary objectives, type of reporting guideline addressed (e.g., CHART, TRIPOD-LLM, COREQ+LLM), target audience of the guideline (e.g., researchers, clinicians, journal editors), LLM application area (e.g., clinical advice, diagnosis, research reporting), LLM type (e.g., GPT-4, Claude-3), development stage of the LLM (if specified), study design (e.g., Delphi consensus, systematic review), number of experts involved, criteria for expert selection, data sources used (e.g., literature review, expert panels), reporting guideline components addressed (e.g., model transparency, prompt engineering), evaluation metrics used (e.g., completeness, clarity), and the primary and secondary outcome measures.

Results

This systematic review included 18 studies focused on reporting guidelines for large language models (LLMs) used in healthcare research after screening a total of 42 studies (Figure 1). Detailed Extraction table for the included studies is provided as a supplementary material, with the references mentioned in the results cited in supplementary material table.

The studies primarily aimed to develop or evaluate reporting frameworks to improve transparency, reproducibility, and method-

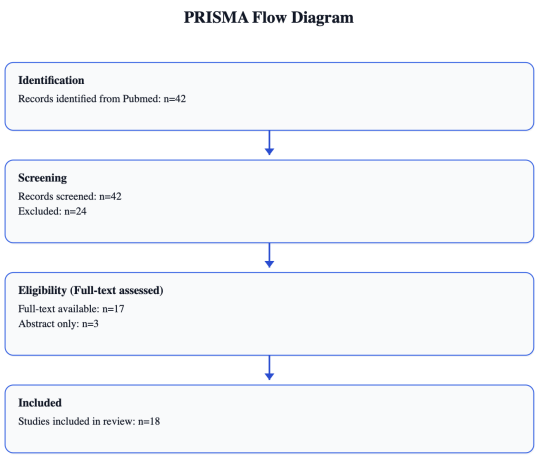

Figure 1: W
orkflow from the initial search to the final included studies

ological rigor in LLM applications. The majority employed consensus-based methodologies, including Delphi processes (Bright Huo 2025 [JAMA Network Open], Bright Huo 2025 [Artificial Intelligence in Medicine]), expert panel reviews (Jack Gallifant 2025 [Nature Medicine]), and scoping reviews (Xufei Luo 2025 [JMIR Research Protocols]). Study populations varied, with stakeholder involvement ranging from 26 Delphi participants (Jack Gallifant 2025 [Nature Medicine]) to 531 in a modified Delphi process (Bright Huo 2025 [BMC Medicine]). Target users included clinicians, researchers, journal editors, policymakers, and AI developers. Development of Reporting Guidelines Several studies focused on creating structured reporting checklists for LLM applications in healthcare. The Chatbot Assessment Reporting Tool (CHART) was developed across multiple studies (Bright Huo 2025 [BMC Medicine], Bright Huo 2025 [JAMA Network Open], Bright Huo 2025 [Artificial Intelligence in Medicine]) to standardize reporting in chatbot health advice (CHA) studies. CHART includes 12 main items and 39 subitems covering model identification, prompt engineering, query strategy, and performance evaluation. Similarly, TRIPOD-LLM (Jack Gallifant 2025 [Nature Medicine]) extended the TRIPOD+AI framework with 19 main items and 50 subitems, emphasizing modular reporting for diverse LLM tasks. Reporting Completeness and Variability A recurring finding was the inconsistent reporting of key LLM-related details in existing studies. Bright Huo 2025 [JAMA Network Open] noted that studies often omitted critical information such as model versions, prompt design, and query strategies. Similarly, Jeff Choi 2025 highlighted gaps in transparency for AI clinical decision support models, particularly in high-stakes applications. Zhong et al. 2024 found that only 43.9% of radiological journals had explicit LLM use policies, and none fully addressed all six assessed reporting items (e.g., model name, verification methods). Methodological Rigor in LLM Evaluation Studies assessing LLM performance revealed variability in evaluation frameworks. Tim Woelfle 2024 benchmarked LLMs (Claude-3-Opus, GPT-4) against human consensus for evidence appraisal tools (PRISMA, AMSTAR, PRECIS-2), reporting classification accuracies ranging from 83.81% (CONSORT-AI) to 92.11% (CONSORT). However, limitations included potential data leakage and reliance on text-only models. Zeming Li 2025 introduced RAPID, an automated checklist generator achieving high accuracy (92.11%) but with variability across LLMs. Stakeholder Engagement and Consensus Building Consensus-driven approaches were central to guideline development. The CHART studies involved 531 Delphi participants and 48 expert panelists (Bright Huo 2025 [BMC Medicine]), while TRIPOD-LLM engaged 26 Delphi participants (Jack Gallifant 2025 [Nature Medicine]). Feedback highlighted the need for multidisciplinary collaboration, particularly between clinicians and AI researchers (Bright Huo 2025 [JAMA Network Open]). However, Leonard Fehring 2025 noted challenges in achieving balanced representation, as their Delphi study excluded patient representatives due to limited AI expertise. Barriers to Guideline Adoption Anticipated barriers included limited awareness among target audi-

ences (Bright Huo 2025 [Artificial Intelligence in Medicine]) and rapid technological evolution necessitating frequent updates (Jack Gallifant 2025 [Nature Medicine]). Zhong et al. 2024 identified inconsistent journal policies as a hurdle, while Zhen Ling Teo 2023 emphasized heterogeneity in AI study designs as a limitation for standardization. Dissemination Strategies Proposed dissemination methods included interactive platforms (e.g., TRIPOD-LLM's web-based tool; Jack Gallifant 2025 [Nature Medicine]), EQUATOR Network listings (Bright Huo 2025 [BMC Medicine]), and integration into journal submission systems (Zhong et al. 2024). Pilot testing of CHART received positive feedback on usability, though some users requested clarifications on checklist items (Bright Huo 2025 [Artificial Intelligence in Medicine]). Quantitative Findings - Classification accuracy: RAPID achieved 92.11% for CONSORT and 83.81% for CONSORT-AI (Zeming Li 2025). - Policy adoption: Only 43.9% of radiological journals had LLM use policies (Zhong et al. 2024). - Stakeholder involvement: CHART development engaged 531 Delphi participants and 48 panelists (Bright Huo 2025 [BMC Medicine]).

Discussion

The systematic review included 18 studies focused on reporting guidelines for large language models (LLMs) in healthcare research. The majority of these studies aimed to develop or adapt reporting checklists to improve transparency and methodological rigor in evaluating LLM applications. Key themes included: • Development of Custom Reporting Guidelines: Several studies introduced new reporting frameworks, such as the Chatbot Assessment Reporting Tool (CHART) for chatbot health advice studies, TRIPOD-LLM for predictive modeling, and COREQ+LLM for qualitative research. These guidelines emphasized structured reporting of model details, prompt engineering, query strategies, and performance evaluation. • Consensus-Based Approaches: Many guidelines were developed through modified Delphi processes, expert panel discussions, and pilot testing, involving diverse stakeholders (e.g., clinicians, AI researchers, ethicists, and policymakers). • Identified Reporting Gaps: Common deficiencies in existing LLM studies included inadequate descriptions of model versions, prompt development, bias assessment, and real-world clinical validation. • Target Applications: Most guidelines addressed LLM use in clinical decision support, evidence synthesis, patient communication, and qualitative research, with limited focus on administrative or operational healthcare tasks. The findings align with broader concerns in LLM healthcare research, as highlighted in recent reviews. For instance, Bedi et al found that only 5% of LLM evaluations used real patient data, with most studies focusing on medical knowledge assessments (e.g., exam questions) rather than clinical deployment [1]. Similarly, Gallifant et al noted inconsistencies in LLM-generated mental health advice, underscoring the need for standardized reporting of model limitations and ethical risks [3]. The emphasis on transparency in prompt engineering and model identification in the reviewed guidelines addresses gaps identified [7], where intervention reporting in autism research lacked methodological detail. Additionally, another study highlighted the underreporting of LLM evaluation metrics in nursing, reinforcing the need for structured checklists like CHART or TRIPOD-LLM [4]. However, the reviewed guidelines diverged from broader critiques in some areas. A call for ethical frameworks to govern LLM use has been issued, but only a subset of the included studies explicitly addressed bias, fairness, or regulatory compliance in their checklists [8]. This suggests a need for future guidelines to integrate ethical reporting standards more robustly. This review has several limitations, including: Heterogeneity of Guidelines: The included studies proposed diverse checklists (e.g., CHART, TRIPOD-LLM) with varying scopes, making direct comparisons challenging. Lack of Val-

idation: Most guidelines were newly developed and lacked empirical validation of their impact on reporting quality. Limited Coverage of Non-Clinical Tasks: Administrative applications (e.g., billing, documentation) were underrepresented, despite their growing use in healthcare [1]. Geographic Bias: Stakeholder consensus processes predominantly involved experts from high-income countries, potentially limiting global applicability. The implications of this study underscore the need for both practical and research-oriented measures to enhance the quality and reproducibility of large language model (LLM)–based research. Journals and funding agencies should mandate adherence to LLM-specific reporting frameworks, such as CHART, to ensure consistency and transparency in published work. Clinicians and researchers are encouraged to document model versions, prompts, and bias mitigation strategies to maintain clinical relevance and traceability [1,8]. Future research should include validation studies to assess whether guideline adoption improves reporting completeness, as previously demonstrated with PRISMA-driven research [7]. Additionally, the scope of existing guidelines should be expanded to include modules addressing multimodal LLMs, real-world deployment, and ethical auditing, as these aspects remain insufficiently explored in current literature [3,4]. Engaging diverse stakeholders—including patients, policymakers, and representatives from low-resource settings—in the development process is also essential to promote inclusivity and equitable implementation [8]. Ultimately, while existing reporting guidelines represent an important advancement toward standardizing LLM research, their long-term impact will rely on broad adoption and iterative refinement to meet the evolving challenges of artificial intelligence.

Conflict of Interest

The authors declare that they have no competing interests.

Acknowledgements

This article was developed through the Artificial Intelligence Powered Research Automation (AIPRA) platform, which utilizes several large language models integrated, including proprietary models. It was used in several parts of the article. It was used under strict human supervision and control. Additionally, authors carefully reviewed and polished the content generated by the model.

Financial Support

There was no funding.

References

[1] ↑ Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, Fries JA, Wornow M, Swaminathan A, Lehmann LS, Hong HJ, Kashyap M, Chaurasia AR, Shah NR, Singh K, Tazbaz T, Milstein A, Pfeffer MA, Shah NH. Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. JAMA. 2025

[2] ↑ Zhang L, Zhao Q, Zhang D, Song M, Zhang Y, Wang X. Application of large language models in healthcare: A bibliometric analysis. Digital health. 2025

[3] ↑ Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, Demner-Fushman D, Dligach D, Daneshjou R, Fernandes C, Hansen LH, Landman A, Lehmann L, McCoy LG, Miller T, Moreno A, Munch N, Restrepo D, Savova G, Umeton R, Gichoya JW, Collins GS, Moons KGM, Celi LA, Bitterman DS. The TRIPOD-LLM reporting guideline for studies using large language models. Nature medicine. 2025

[4] ↑ Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, Demner-Fushman D, Dligach

D, Daneshjou R, Fernandes C, Hansen LH, Landman A, Lehmann L, McCoy LG, Miller T, Moreno A, Munch N, Restrepo D, Savova G, Umeton R, Gichoya JW, Collins GS, Moons KGM, Celi LA, Bitterman DS. The TRIPOD-LLM Statement: A Targeted Guideline For Reporting Large Language Models Use. medRxiv : the preprint server for health sciences. 2024

[5] ↑ Fareed M, Fatima M, Uddin J, Ahmed A, Sattar MA. A systematic review of ethical considerations of large language models in healthcare and medicine. Frontiers in digital health. 2025. doi: 10.3389/fdgth.2025.1653631 DOI: 10

[6] ↑ Iqbal U, Tanweer A, Rahmanti AR, Greenfield D, Lee LT, Li YJ. Impact of large language model (ChatGPT) in healthcare: an umbrella review and evidence synthesis. Journal of biomedical science. 2025

[7] ↑ Guo Z, Lai A, Thygesen JH, Farrington J, Keen T, Li K. Large Language Models for Mental Health Applications: Systematic Review. JMIR mental health. 2024

[8] ↑ Hobensack M, von Gerich H, Vyas P, Withall J, Peltonen LM, Block LJ, Davies S, Chan R, Van Bulck L, Cho H, Paquin R, Mitchell J, Topaz M, Song J. A rapid review on current and potential uses of large language models in nursing. International journal of nursing studies. 2024