The Landscape of LLM Evaluation Metrics in Healthcare: A Rapid Review

Azra ISMAIL[a,1], Manvi S[a], Sumon Kanti DEY[a], Nabile SAFDAR[b], Melissa SOUSA[b] and Selvi RAMALINGAM[b]
[a]*Emory University, Atlanta, GA, USA*
[b]*Emory Healthcare, Atlanta, GA, USA*
ORCiD ID: Azra Ismail https://orcid.org/0000-0002-7570-9474, Manvi S https://orcid.org/0009-0003-1727-9511, Sumon Kanti Dey https://orcid.org/0009-0008-3467-6814, Nabile Safdar https://orcid.org/0000-0002-1131-8052, Melissa Sousa https://orcid.org/0009-0004-8584-6394, Selvi Ramalingam https://orcid.org/0000-0003-4142-7337

Abstract. Large Language Models (LLMs) are entering healthcare rapidly, yet there is little consensus on how to evaluate them. We conducted a rapid review of 38 sources (peer-reviewed papers and grey literature) on LLM evaluation, covering human and LLM-as-a-judge approaches, and extracted their metrics. We found overlaps across efforts, with same or similar metrics varying in definition and usage depending on the context. We organized the metrics into a library grouped by intent, with prototypical metrics and their alternative names, evaluator types, and reported validation evidence for each group. Rather than proposing a single framework, we argue metrics should be selected based on application type and evaluation goals.

Keywords. Evaluation, Large Language Models, Metrics

1. Introduction

As LLM-based applications are increasingly integrated into healthcare workflows, rigorous evaluation is essential to ensuring safe and reliable performance. Yet, little consensus exists on which dimensions of model responses warrant assessment, and using what methods. Existing efforts to address this gap include human evaluation frameworks such as QUEST [1], validated instruments such as PDSQI-9 [2], benchmarks such as MedHELM [3], LLM-based metrics such as DocLens [4], and organizational checklists and reporting guidelines. Each has been developed for a distinct purpose and defines its own metrics according to its intent and application context.

Several reviews have surveyed this literature to identify evaluation trends. Tam et al. reviewed 142 studies of human evaluation, and found gaps in reliability, generalizability, and applicability [1]. Bedi et al. examined 519 studies and found that most assess accuracy on examination-style questions, with limited attention to fairness, bias, and deployment [5]. However, prior reviews rarely compare frameworks or metrics across human, automated, and LLM-as-a-judge approaches. They also miss interrogating where each metric originated, who applies it, and how definitions differ across

[1] Corresponding Author: Azra Ismail, azra.ismail@emory.edu

frameworks. Rather than proposing a new framework, this paper seeks to inform future evaluation by mapping existing efforts and incorporating latest developments in the field.

2. Method

We conducted a rapid review following Cochrane guidance [9], asking what frameworks, benchmarks, and instruments exist for evaluating LLM applications in health, what metrics they define, and how metrics are operationalized. We included papers that propose, validate, or review an evaluation framework, benchmark, instrument, checklist, or reporting guideline, covering human and LLM-as-a-judge evaluation. We included sources in English, published from January 2023 to August 2026. We included preprints and white papers since several widely used frameworks are not yet peer-reviewed, and recorded the peer-review status. We excluded evaluations of a single application that did not propose or adapt a framework, and papers on LLM evaluation outside healthcare.

We searched Web of Science, Scopus, and Google Scholar with the terms "LLM" or "Large Language Models", "metrics" or "evaluation", and "healthcare" or "clinical", supplemented by a Google search for white papers and references in included papers. The second and third author screened titles and abstracts and read full texts, the rest of the team checked inclusion decisions. We included 38 sources (34 peer-reviewed, three preprints, and one white paper). The full list and methodology are published online: https://azraismail.me/llm-eval-metrics. For each paper, we recorded the framework, its stated intent, every metric defined, rating scale, who performed the rating, the clinical domain, and any reported reliability or validity statistics. For each metric, we recorded its stated definition, related concepts, and the original instrument where cited. We then grouped metrics with overlapping definitions by intent, using the stated definition rather than the paper's heading, and within each group selected as prototypical the metric whose definition most directly named the construct, listing other names as alternatives. We coded reported validation evidence as—no reliability statistics, inter-rater agreement among human raters, reliability and validity testing of the instrument, or agreement between an LLM-as-a-judge and human raters. We coded raters as clinicians or domain experts, patients or other participants, organizational roles, or unspecified. Author disagreements were resolved through discussion.

Table 1. Frameworks developed for LLM evaluation in healthcare, grouped by evaluation approach.

Approach	Framework (with year of publication)	Validation Evidence
Human Review	Med-PaLM rubrics (2023, 2025); QUEST (2024); S.C.O.R.E. (2024); HumanELY (2023); EquityMedQA (2024); PDSQI-9 (2025); CLEVER (2025); AMIE (2025); NOHARM (2025); CREOLA clinical safety framework (2025)	Inter-rater agreement provided at most. Instrument reliability and validity tested for PDSQI-9 only.
Automated (statistical or LLM-based scoring)	Bias evaluation framework (2025); DocLens (2024); clinical decision-making evaluation on MIMIC-IV (2024); MedSafetyBench (2024); Med-HALT (2023)	DocLens reports agreement with medical experts. No reliability statistics reported for others.
Hybrid (human rating plus LLM-based scoring)	MedHELM (2026); CRAFT-MD (2025); LLM-as-a-judge with PDSQI-9 (2025); GREEN (2024)	Agreement between LLM judge and clinician or expert raters
System performance, and reporting guidelines	Translational value framework (2023); Foundation metrics (2024); OPTICA (2024); CHAI Responsible AI Content (2025); TRIPOD-LLM (2025); CHART (2025)	Performance metrics are statistical. No reliability statistics reported for others.

3. Results

Table 1 lists the 26 frameworks, instruments, and benchmarks we identified (hereafter referred to as frameworks). The remaining sources were reviews or studies that applied or tested them. Eleven frameworks relied on human raters, five on automated or LLM-based scoring, six combined human rating with LLM-based scoring, and six addressed system performance or reporting. Only PDSQI-9 underwent full psychometric validation, reporting internal consistency (Cronbach's alpha 0.879) and construct validity through factor analysis and discriminant testing [2], others reported inter-rater agreement at most. Four of the eleven human evaluation frameworks included patients or other non-experts; both versions of Med-PaLM used non-experts alongside physicians, EquityMedQA used consumers alongside health equity experts, and AMIE used trained patient actors to rate communication and empathy. The rest relied on clinicians or domain experts, even for dimensions such as empathy, helpfulness, and trust, or did not specify the rater. The one study that compared experts and non-experts against an LLM judge found that the two groups agreed with it at different rates, pointing to the role of rater expertise [6].

Table 2. Metric library grouped by intent. Each row gives a prototypical metric in italics, followed by alternative names that the same construct goes by in different sources. Full list of references, and library of 116 metrics with definitions and sources are published online: https://azraismail.me/llm-eval-metrics

Intent	Prototypical metric (with sample alternatives)	Who scores it
Technical performance *Does the output match gold standard?*	*Accuracy against gold labels* (F1, sensitivity, specificity) *Similarity* (BLEU, ROUGE, BERTScore)	Automated; LLM judge for some metrics
Safety and harm *Could the output hurt someone?*	*Harm* (safety, harm severity, omission) *Bias* (objectivity, fairness, equity harms) *Hallucination* (fabrication, confabulation) *Guideline adherence* (consensus, agreement) *Escalation and fallback* (low-confidence trigger, escalation success)	Human raters; automated and LLM judge for some benchmarks
Quality of information *Is the content correct, complete, and relevant?*	*Accuracy* (correctness, factuality) *Comprehensiveness* (completeness, coverage) *Relevance* (appropriateness) *Consistency* (reproducibility)	Human raters; LLM judge
Understanding/reasoning *Was the query understood correctly?*	*Reasoning* (logical processing, explainability) *Understanding* (comprehension) *History-taking* (information gathering)	Human raters; LLM judge
Expression & communication *Is the output readable and appropriate in tone?*	*Clarity* (readability, fluency) *Conciseness* (succinctness) *Empathy* (emotional support, communication skill)	Human raters; Automated for readability
User experience & trust *Does the user gain confidence and value?*	*Trust* (assurance, reliability) *Usefulness* (helpfulness, actionability) *Satisfaction*	Human raters; LLM judge in some cases
System and deployment *Is the system feasible and fit for the organization?*	*Robustness* (generalization) *Computation* (memory, FLOPs, token limits) *Reliability* (uptime, latency, error rate) *Translational value* (utility, adoption)	Automated; organizational review

Across the 38 sources, we identified 116 metrics whose vocabulary overlapped in two ways. First, different names often described the same construct. Hallucination, for example, was a type of harm in HumanELY, fabrication in QUEST, an error class in CREOLA, and part of the "accurate" attribute in PDSQI-9 [1,2]. QUEST alone listed 66 related concepts for its 17 dimensions [1]. Because only some papers stated where their metrics came from (e.g., PDSQI-9 adapted the Physician Documentation Quality Instrument [2]), we could not trace this overlap to a common origin. Second, the same

name often carried different definitions. Accuracy meant factual correctness in QUEST, correctness plus currency in HumanELY, and freedom from hallucination in PDSQI-9 [1,2]. Safety meant adherence to guidelines in S.C.O.R.E., protection of information and systems in the foundation metrics, and compliance with AMA Principles of Medical Ethics in MedSafetyBench, while QUEST treated harm as a separate dimension [1]. Even raters using the same instrument could disagree on a critical dimension such as safety [7]. Table 2 organizes metrics by intent, grouping metrics that appeared in more than one framework under a "prototypical" metric with their alternative names.

Seven papers proposed or tested an LLM-as-a-judge method, and all measured the judge against clinician or expert raters. The strongest validation came from Croxford et al. [8]: LLM judges scored real-world EHR summaries on the nine PDSQI-9 attributes, and a reasoning model reached an ICC of 0.818 with a median score difference of zero from physicians. MedHELM's multi-model jury agreed with clinicians (ICC 0.47) more closely than clinicians agreed with each other (0.43) [3]. LLM judges were applied only to metrics that could be assessed from the text itself: completeness, attribution, conciseness, relevance, correctness against a reference, and error counts. Trust, satisfaction, and empathy were rated by humans in every framework that included them. Human judgment thus remained the reference standard across evaluations. Additionally, the strongest agreement was obtained when the judge was scored with the only instrument that had tested reliability and validity [8].

4. Discussion

No single evaluation approach in our review performed well in isolation; each carried distinct limitations. Human raters are well positioned to judge critical dimensions such as safety, trust, and appropriateness, yet their judgments can diverge, and even clinicians may disagree sharply on a dimension as central as safety [7]. Automated and statistical metrics, by contrast, offer reproducibility but remain insensitive to context and clinical consequence. LLM judges occupy a middle ground: they can approximate human ratings on text-based metrics and can be verified against expert ratings or annotated reference text, but they fall short on more subjective dimensions [6]. Organizational checklists, for their part, assess whether an institution is prepared to deploy a model rather than whether the model itself performs well. The pertinent question, therefore, is not which approach is best, but which combination of approaches suits a given application. A patient-facing chatbot, for instance, requires evaluation of safety, expression, and user experience, dimensions that only human raters can adequately judge and that ideally warrant the involvement of patients themselves. A clinical note summarizer presents a different case: where gold standard reference notes are available, its output may be scored by an LLM judge [8], with clinician review reserved for identifying potential safety concerns. Our review and accompanying metric library are intended to support such application-specific selection of metrics in future evaluations.

Our map also reveals where AI evaluation efforts in healthcare require further strengthening. Validated instruments remain rare, PDSQI-9 was the only instrument in our review with tested reliability and validity. LLM judges show considerable promise, but their use requires rigorous calibration against validated human ratings. Patient perspectives were absent from most frameworks, even for applications that directly concern the patient's own experience. Bias, moreover, was assessed either through rubrics or through statistical audits, but no framework we identified undertook both.

Finally, multi-turn interaction and retrieval-augmented pipelines were evaluated by only a few frameworks, representing an important opportunity for future work.

Our review has several limitations. As a rapid review conducted without a registered protocol, it trades exhaustive coverage for timeliness [9]. Our search terms may have missed frameworks that do not use "evaluation" or "metrics" in the title or abstract, and although citation searching mitigates this risk, relevant work may remain unidentified. Several included frameworks were also preprints that had not yet undergone peer review. Our account of who performs evaluations reflects only what the papers reported. Finally, because grouping metrics by intent and coding validation evidence required interpretive judgment, another team might organize the metrics differently. To address this, we will maintain the online library as a living resource that we will continue to update, and readers can submit frameworks or metrics to add through the website.

5. Conclusion

LLM evaluation frameworks in healthcare overlap in the metrics they name and diverge in how they define them. Grouping metrics by intent across human, automated, and LLM-as-a-judge approaches makes both visible and shows where the field stands: one validated instrument, LLM judges applied to text-based metrics and validated against clinicians, and patient perspectives present in a minority of frameworks. Rather than adopting a single framework, evaluation should be tailored to the application and goals at hand and could be more patient-centered overall.

Acknowledgments. This work was supported by a grant from the Emory Healthcare Innovation Challenge.

References

[1] Tam TYC, Sivarajkumar S, Kapoor S, et al. A framework for human evaluation of large language models in healthcare derived from literature review. NPJ Digit Med. 2024;7:258, doi:10.1038/s41746-024-01258-7

[2] Croxford E, Gao Y, Pellegrino N, et al. Development and validation of the provider documentation summarization quality instrument for large language models. J Am Med Inform Assoc. 2025;32(6):1050-60. doi:10.1093/jamia/ocaf068

[3] Bedi S, Cui H, Fuentes M, et al. Holistic evaluation of large language models for medical tasks with MedHELM. Nat Med. 2026;32(3):943-51., doi:10.1038/s41591-025-04151-2

[4] Xie Y, Zhang S, Cheng H, Liu P, Gero Z, Wong C, et al. Doclens: Multi-aspect fine-grained medical text evaluation. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024. p. 649–79. doi:10.18653/v1/2024.acl-long.39

[5] Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA. 2025;333(4):319-28, doi:10.1001/jama.2024.21700

[6] Szymanski A, Ziems N, Eicher-Miller HA, et al. Limitations of the LLM-as-a-judge approach for evaluating LLM outputs in expert knowledge tasks. In: Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25); 2025 Mar 24-27; Cagliari, Italy. New York (NY): ACM; 2025. p. 952-66. doi:10.1145/3708359.3712091

[7] Deva R, S M, Zhou J, et al. A mixed-methods evaluation of LLM-based chatbots for menopause. Stud Health Technol Inform. 2025;327:1175-9, doi:10.3233/SHTI250575

[8] Croxford E, Gao Y, First E, et al. Evaluating clinical AI summaries with large language models as judges. NPJ Digit Med. 2025;8:640. doi:10.1038/s41746-025-02005-2

[9] Garritty C, Gartlehner G, Nussbaumer-Streit B, et al. Cochrane Rapid Reviews Methods Group offers evidence-informed guidance to conduct rapid reviews. J Clin Epidemiol. 2021;130:13-22, doi: 10.1016/j.jclinepi.2020.10.007