Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Evaluation Practices for Large Language Models in Healthcare: A Systematic Scoping Review

Ehsan Safaee
Shahed University
Ali Mehrabi
Amirkabir University of Technology
M. Moein Shariatnia
Tehran University of Medical Sciences
Parsa Rouzrokh
Shahid Beheshti University of Medical Sciences
Pouria Rouzrokh
pouria.rouzrokh@yale.edu

Yale University

Article

Keywords:

Posted Date: May 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9576482/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

The rapid integration of Large Language Models (LLMs) into clinical workflows has outpaced the establishment of rigorous frameworks to ensure patient safety and model reliability. We conducted a systematic scoping review of n = 1,054 full-text articles published between 2017 and 2025 to characterize the current landscape of LLM evaluation in healthcare. Using an automated extraction pipeline validated for high-fidelity data retrieval, we mapped study characteristics across the MedHELM taxonomy and 16 distinct evaluation constructs.

Our analysis reveals a profound "safety-utility gap": while technical validity is nearly universal (95.1%), safety-critical dimensions including bias, toxicity, and uncertainty are evaluated in fewer than 10% of clinical decision support tasks. Furthermore, we identify extreme heterogeneity in evaluation metrics, with leading measures such as accuracy accounting for less than 20% of reported evaluations in most constructs. This fragmentation, combined with a lack of "ecological validity" in testing against real-world clinical documentation, prevents meaningful cross-study comparisons and complicates regulatory oversight. These findings demonstrate that current benchmarking is insufficient for clinical deployment. We conclude that a transition toward standardized, safety-first evaluation frameworks is essential to bridge the gap between technical performance and bedside utility.

1. Introduction

The integration of Large Language Models (LLMs) and Vision Language Models (VLMs) into healthcare has transitioned from experimental research to early-stage clinical applications. These models now demonstrate sophisticated capabilities in automated radiology reporting (1), diagnostic decision support(2), and patient-facing communication (3). Yet, this technical velocity has outpaced our ability to verify the safety and reliability of these systems in practice. While LLMs possess extensive latent medical knowledge, their susceptibility to hallucination and the risk of inducing automation bias among clinicians remain significant barriers to safe deployment (4–6).

A fundamental gap exists between achieving state-of-the-art performance on static benchmarks and achieving clinical utility. Current evaluation methodologies are largely fragmented, leading to a lack of comparability across institutional settings (7–9). Most existing evidence relies on *in silico* simulations or exam-style datasets, which fail to replicate the cognitive load and data ambiguity inherent to the clinical environment (9, 10). Furthermore, the "black box" nature of these models limits their generalizability and trustworthiness in diverse settings (11–13).

This gap in readiness stems from a lack of ecological validity in current benchmarking. Standardized medical examinations rarely reflect the complexity of modern diagnostic workflows, which require the synthesis of fragmented data from Electronic Health Records (EHRs), including inconsistent shorthand and temporal dependencies (14–16). Currently, only a small fraction of LLM evaluations utilize actual EHR data, leaving the models untested against the noise of real-world clinical documentation (17).

In this scoping review, we systematically characterize the current landscape of LLM and VLM evaluation in healthcare. We map the breadth of clinical tasks currently being assessed and evaluate the rigor of the benchmarks and metrics employed. By synthesizing these patterns, we identify the structural gaps that must be addressed to move beyond proof-of-concept toward robust, clinically validated AI interventions.

2. Methods

2.1 Search Strategy and Eligibility Criteria

This review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) (18). We included original research articles that used one or more LLMs, VLMs, or other large multimodal models (LMMs) and reported performance metrics or proposed new benchmarks or evaluation setups for healthcare-related tasks. Studies were considered across any healthcare or health-related setting, and only articles written in English were eligible for inclusion. We excluded conference abstracts, reviews and meta-analyses, case reports, non-English publications, and articles without an explicit healthcare focus. These eligibility criteria were applied consistently during title/abstract screening and full-text review.

We systematically queried Scopus, PubMed, and arXiv for relevant studies published between January 2017 and September 2025. We selected the start date to align with the introduction of the Transformer architecture in the academic literature (19). The search strategy combined terms related to (i) artificial intelligence and machine learning (including LLMs and VLMs), (ii) evaluation, benchmarking, or validation, and (iii) healthcare, medicine, or clinical tasks. Details of our search strategy are provided in the supplements.

2.2 Primary Screening

Retrieved records were deduplicated using a custom Python script which relied on matching bibliographic fields. Title and abstract screening was performed using LatteReview framework (20), a Python-based multi-agent framework that leverages LLMs to automate the review process. The screening workflow utilized the Google Gemini API to execute and was structured as follows:

- Initial Tier: Two independent agents (Agent 1: Gemini 2.5 Flash; Agent 2: Gemini 2.5 Flash Lite) screened each title and abstract. Agents assigned an inclusion likelihood score on a five-point ordinal scale (1: Highly unlikely; 5: Highly likely). Records with a cumulative score greater than 6 were automatically included.
- Adjudication Tier: Articles with discordant or borderline cumulative scores (between 3 and 6) were forwarded to a higher-capacity adjudication agent (Agent 3: Gemini 2.5 Pro). These records were included only if Agent 3 assigned a score equal or greater than 3.

Following this tiered screening, full-text versions of all provisionally included studies were retrieved for formal data extraction.

2.3 Data Extraction and Taxonomy Mapping

We engineered an automated extraction pipeline to systematically chart study characteristics from the included full-text articles. Each manuscript was processed by a frontier Large Language Model (Gemini Pro) using a structured prompting strategy designed to map data points directly to a predefined schema. To ensure computational reproducibility and minimize the risk of LLM-generated hallucinations, the pipeline utilized structured output validation via Pydantic v2.12.5 (21) and asynchronous orchestration through LangChain v0.3.x (22). The extraction schema followed two primary healthcare AI frameworks:

1. Task Classification: A granular layer identifying the specific clinical application, categorized using the MedHELM taxonomy (23).
2. Evaluation Constructs: Sixteen distinct evaluation dimensions informed by the Holistic Evaluation of Language Models (HELM) and Vision Holistic Evaluation of Language Models (VHELM) frameworks (24, 25).

We adopted the MedHELM taxonomy as our primary framework for classifying the object of each study (23). Each article was mapped to one of five mutually exclusive categories: *Clinical Decision Support, Clinical Note Generation, Patient Communication and Education, Medical Research Assistance, Administration and Workflow*.

Regarding evaluation constructs, we mapped each study against a multidimensional taxonomy derived from HELM, defining 16 distinct constructs (Table 1) spanning technical performance and clinical safety. For each construct, the pipeline executed a four-stage extraction sequence:

Table 1
Definitions of evaluation constructs used for data extraction in this scoping review

Evaluation Constructs	Definition (used for data extraction)
Validity	The degree to which a model's output matches a reference or gold-standard answer, assessed through explicit correctness-oriented evaluations and independent of reasoning, explanation, or safety considerations.
Knowledge	The ability of a model to recall factual medical information not present in the input, evaluated via explicit factual recall or factfulness assessments rather than reasoning, summarization quality, or task performance.
Reasoning	The ability of a model to perform explicit multi-step logical or clinical inference, evaluated using dedicated reasoning benchmarks or tests that assess inference processes beyond final answer correctness.
Comprehension	The ability of a model to correctly understand and interpret the provided input context (textual or multimodal) and respond appropriately, without requiring multi-step inference or assessing response completeness.
Comprehensiveness	The extent to which a model's response fully covers all required facts, sub-questions, or criteria needed to address a task, as evaluated through reference coverage, checklist-based, or completeness-focused methods.
Readability	The degree to which model outputs are easy to read and linguistically accessible to the intended audience, assessed via readability formulas or human judgments of clarity and fluency.
Coherence	The extent to which generated text is logically and structurally well-organized, maintaining consistent discourse flow and connectivity across sentences and paragraphs.
Consistency	The ability of a model to produce non-contradictory and stable outputs across paraphrased prompts, repeated queries, or multi-turn interactions, while remaining aligned with provided evidence rather than ground-truth correctness alone.
Robustness	The ability of a model to maintain performance under explicitly introduced invariant perturbations (e.g., noise, typos, corruption, or controlled distribution shifts) designed to test robustness.
Bias	The extent to which a model avoids systematic performance disparities or stereotypical associations across demographic or spurious attributes, evaluated using explicit bias or fairness assessment methods.
Multilingualism	The ability of a model to understand, generate, or perform comparably across multiple languages, demonstrated through explicit cross-lingual or multilingual evaluation rather than monolingual testing.
Toxicity	The tendency of a model to generate offensive, abusive, or harmful language, assessed independently of refusal behavior or broader safety mechanisms.
Safety	The ability of a model to refuse, redirect, or safely respond to prompts that could cause real-world harm, evaluated through explicit safety, red-teaming, or policy-compliance assessments.

Evaluation Constructs	Definition (used for data extraction)
Uncertainty	The extent to which a model represents and communicates uncertainty about its outputs, and how well those confidence signals are calibrated to actual correctness or risk.
Explainability	The use of explicit interpretability or explanation methods to make a model's decision-making process understandable to humans, beyond producing outputs or applying prompt engineering alone.
Resource Utilization	The efficiency with which a model uses computational, financial, or environmental resources, evaluated through explicit measurements such as latency, memory usage, energy consumption, or inference cost.

1. Open-Ended Methodological Identification: Extraction of the specific methods or benchmarks employed.
2. Closed-Vocabulary Verification: A choice-based query to map the previous response to established benchmarks identified in current literature.
3. Open-Ended Metric Extraction: Identification of the specific metrics used to report performance.
4. Closed-Vocabulary Metric Categorization: A choice-based query to map the previous response to established metrics identified in current literature.

Quantitative analysis relied on choice-based outputs only after cross-verification against open-ended responses. Discrepancies between the free-text synthesis and categorical selections were flagged for manual adjudication. A construct was coded as "present" only if the study reported formal empirical measurements or a rigorous, evidence-based qualitative assessment. Superficial mentions or theoretical discussions without corresponding data were marked as *not reported*.

Despite these safeguards, preliminary runs revealed that the models struggled with specific semantic nuances, for example in distinguishing between comprehension and comprehensiveness. These terms were frequently reported interchangeably, which threatened the precision of our taxonomy. To resolve this, we conducted a targeted validation phase on a pilot sample of the studies. We manually reviewed the full-text articles to identify where the automated pipeline diverged from expert clinical interpretation. This calibration allowed us to refine our prompting logic and ensure that the final extracted data accurately reflected the distinct clinical and technical dimensions of each evaluation.

2.6 Sampling Strategy

To validate the automated pipeline, we determined a representative sample size using Cochran's formula. For a population of N = 1,054, we selected n = 42 articles to achieve a 95% confidence level (Z = 1.96) with a 15% margin of error (c = 0.15), balancing statistical rigor against the feasibility of manual full-text review.

Because preliminary tests suggested that document complexity rather than content was the primary driver of extraction errors, we utilized stratified random sampling based on non-semantic structural

metadata. By stratifying the sample by article layout, publication year, and document length, we ensured the validation set captured varying PDF parsing difficulties without introducing thematic bias. Two independent reviewers (E.S. and A.M.) manually extracted data from this subset and compared their findings against the LLM-generated outputs to quantify agreement and identify systematic errors. The resulting accuracy metrics for each construct are detailed in Table 2. During the final validation phase, we observed that the models exhibited insufficient discriminative performance when distinguishing between certain conceptually proximal dimensions. Due to this persistent semantic ambiguity and the risk of conflating distinct clinical attributes, three constructs—Consistency, Comprehensiveness, and Comprehension—were excluded from the final full-scale data extraction to ensure the high specificity and reliability of the reported results.

Table 2
Error rates after final round of prompt optimization for extracted results

Evaluation Construct	After Optimization (± 15%)
Uncertainty	4.76%
Coherence	0.00%
Toxicity	0.00%
Safety	4.76%
Readability	2.38%
Validity	0.00%
Robustness	4.76%
Bias	7.14%
Resource Utilization	7.14%
Multilingualism	9.52%
Explainability	14.29%
Task	7.14%
Knowledge	2.38%
Comprehensiveness	26.19%
Reasoning	9.52%
Comprehension	38.10%
Consistency	42.86%

3. Results

3.1 Study Selection and Characteristics

The systematic search yielded 21,895 citations. After removing 4,172 duplicates, 17,723 unique records underwent automated title and abstract screening. This process identified 2,023 studies for data extraction, of which 1,054 full-text articles were successfully retrieved and analyzed (Fig. 1).

Clinical applications were predominantly centered on Clinical Decision Support (CDS) (n = 574; 54.5%), followed by non-clinical applications (n = 184; 17.5%), patient communication and education (n = 119; 11.3%), and clinical note generation (n = 107; 10.2%). Research assistance (n = 65; 6.2%) and administrative workflows (n = 5; 0.5%) represented the smallest shares of the literature (Fig. 2). Regarding model architecture, the vast majority of studies evaluated text-only LLMs (n = 932; 88.4%), while VLMs accounted for 11.6% (n = 122) of the identified research.

3.2 Automated vs. Human Evaluation

Evaluation strategies followed two major trajectories: automated computational and human-centric assessments. Automated methods were the primary vehicle for assessing technical performance, with classification tasks relying on accuracy, F1-score, and sensitivity/specificity or generative tasks frequently quantifying via linguistic similarity metrics, including BLEU, ROUGE, and METEOR. A growing subset of recent studies implemented "LLM-as-a-judge" frameworks to automate the scoring of factuality and clinical relevance. Notably, constructs such as validity (94%), uncertainty (89%), and multilingualism (94%) were almost exclusively evaluated through automated methods (Fig. 4).

Conversely, human evaluation remained the most common for qualitative constructs requiring domain expertise. These assessments utilized validated instruments such as DISCERN and the JAMA benchmark criteria or Likert-scale ratings for subjective attributes like readability and helpfulness. Human-rated methods were predominantly used to evaluate medical knowledge (60%), safety (71%), and toxicity (60%).

3.3 The Landscape of Evaluated Constructs

The distribution of evaluation constructs revealed a significant imbalance between technical verification and clinical validation (Fig. 3). While validity was nearly universal across all task categories (95.1%–100%), other critical clinical constructs remained under-investigated.

- Cognitive and Linguistic: Evaluation of reasoning was present in fewer than half of the studies, peaking in clinical note generation (57.9%). Readability and coherence assessments were highly task-specific, appearing frequently in patient-facing applications (75.6%) but rarely in research or non-clinical tasks (10%).
- Trust and Reliability: Dimensions essential for clinical trust, such as explainability (35.5%) and uncertainty quantification (27.0%, were infrequently reported, even in high-stakes CDS tasks.

- The Safety and Bias Gap: Despite the clinical risks associated with LLM deployment, safety was evaluated in fewer than 30% of studies across all categories. In the context of CDS, safety evaluations dropped to 9.6%. Furthermore, systemic risks including algorithmic bias (8.4% in CDS) and toxicity (1.6% in CDS) were largely neglected in the current literature. Robustness showed moderate coverage (41.1% in CDS), yet the lack of standardized stress-testing against clinical noise remains a visible trend.

3.4 Metric Heterogeneity

Our analysis reveals profound heterogeneity in the metrics employed to evaluate LLMs, with no single dominant standard emerging across any of the nine identified constructs. While a core set of general-purpose metrics—primarily Accuracy, F1-score, and Manual Evaluation—appears across most dimensions, they rarely account for more than 30% of the reporting instances within a given category. This fragmentation is most pronounced in the Validity construct, where, despite Accuracy (21.4%) and F1 (14.8%) being the most frequent, a "long tail" of 45 additional distinct metrics accounts for a massive share of the literature (n = 575 reporting instances; Fig. 5).

The distribution of metrics suggests a reliance on traditional machine learning benchmarks that may not capture the nuances of clinical utility. For example, in Knowledge and Reasoning tasks, Accuracy remains the primary metric (27.2% and 29.6%, respectively), yet these constructs also exhibit significant diversity, with 24 to 32 "Other" metrics identified in each. This suggests that while researchers default to familiar classification metrics, they frequently supplement them with idiosyncratic measures to address the specificities of medical logic.

In safety-critical domains, the evaluation strategy shifts markedly toward human-centric oversight. Manual Evaluation is the leading methodology for Safety (30.2%) and Toxicity (22.2%), reflecting the current technical limitations of automated benchmarks in capturing complex clinical risks. However, even within these categories, consensus remains elusive; for instance, Safety evaluation is split across various measures including "Safety" labels (13.4%), Accuracy (12.3%), and Hallucination Rates (10.1%).

4. Discussion

The rapid integration of LLMs and VLMs into clinical workflows has created an urgent need for standardized evaluation frameworks. Our systematic characterization of 1,054 articles reveals a profound gap between technical performance and clinical readiness. While models frequently achieve high scores on static, exam-style benchmarks, the lack of standardized metrics and the neglect of safety-critical evaluations suggest that current evidence is insufficient to support widespread clinical deployment.

4.1 Clinical Validation Gap

Our data highlights a concerning imbalance between technical verification and clinical validation, particularly regarding the "Safety and Bias Gap". Safety evaluations drop to a mere 9.6% in high-stakes CSD tasks, suggesting a "technical-first" bias where the ease of running automated benchmarks outweighs the rigorous, human-centric validation required for patient safety. This oversight extends to systemic risks such as algorithmic bias and toxicity, which remain largely neglected in the current literature despite their potential for clinical harm. Additionally, the infrequent evaluation of uncertainty quantification and explainability undermines the transparency of these models, directly hindering the development of clinician trust and the mitigation of automation bias.

4.2 Ecological Validity and the EHR Gap

Our results indicate a saturation of evaluations in CDS, yet these studies frequently rely on curated, exam-style datasets rather than the "noisy" reality of Electronic Health Records (EHRs). As noted in our introduction, the synthesis of fragmented, temporal, and often contradictory clinical documentation is a prerequisite for clinical utility. However, the paucity of evaluations utilizing real-world EHR data limits the generalizability of reported performance. When models are tested against clean datasets, their propensity for hallucination and sensitivity to data ambiguity—key themes in recent literature—are often masked. Moving forward, evaluation frameworks must prioritize "ecological validity," stress-testing models against the unstructured and inconsistent data formats that define modern medical practice.

4.3 Addressing the Safety and Bias Deficit

Perhaps the most concerning trend is the relative neglect of safety, bias, and toxicity assessments, particularly in CDS tasks where safety evaluation was present in fewer than 10% of studies. The rapid technical velocity of LLM development has outpaced the development of protective frameworks. In clinical settings, algorithmic bias can exacerbate existing health disparities, while a lack of uncertainty quantification can lead to automation bias, where clinicians over-rely on model suggestions despite potential inaccuracies. For LLMs to transition from academic interest to clinical tools, evaluation constructs must shift from "can the model answer correctly?" to "how does the model fail, and who is harmed when it does?"

4.4 The Crisis of Metric Heterogeneity

A central finding of this review is the extreme fragmentation of evaluation metrics across all clinical and technical constructs, with no single dominant standard emerging in any of constructs. The "long tail" of idiosyncratic metrics most notably in the validity construct, where 45 distinct measures were identified indicates that the field has yet to coalesce around a unified benchmarking standard. This heterogeneity presents a significant barrier to the maturation of medical AI because it renders the evidence base incommensurable; without a consensus on metrics, it is logically and statistically impossible to conduct accurate cross-study comparisons. Consequently, current efforts to perform meta-analyses on LLM performance are fundamentally limited, as pooling data from studies using divergent definitions of accuracy or hallucination rates lacks clinical meaning and logical coherence. Furthermore, the continued

reliance on general-purpose linguistic metrics like BLEU or ROUGE often fails to capture the high-stakes precision required for clinical documentation or diagnostic reasoning.

4.4 Limitations

This scoping review is subject to several limitations. First, our search was restricted to English-language publications, which may overlook significant LLM developments in other linguistic and regional healthcare contexts. Second, the rapid evolution of the field means that new models and benchmarks may have emerged since our final search in September 2025. Finally, while our automated extraction pipeline utilized Pydantic-validated structures to minimize hallucinations, the inherent "black box" nature of LLMs means that subtle semantic nuances in study descriptions may have been categorized with varying degrees of precision, despite our rigorous manual calibration phase.

5. Conclusion and Future Directions

To move beyond the current state of fragmented experimentation, the biomedical informatics community must converge on a multidimensional evaluation standard. We propose the development of a "Clinical LLM Report Card" that mandates the reporting of safety, bias, uncertainty and explainability metrics alongside traditional validity scores. Additionally, as VLMs begin to occupy a larger share of the research landscape, these frameworks must be adapted to account for multimodal data integration. Only through standardized, ecologically valid, and safety-first evaluation can we ensure that LLMs serve as reliable partners in clinical practice.

Declarations

Data Availability:

The datasets generated and/or analysed during the current study are available in the *Evaluation Practices for Large Language Models in Healthcare: Article Dataset* repository, and can be accessed via this link https://www.doi.org/10.5281/zenodo.19915394/.

Code Availability:

The underlying code [and training/validation datasets] for this study is available in *llm-evaluation-scoping-review-paper* and can be accessed via this link https://www.github.com/alimium/llm-evaluation-scoping-review-paper.git/.

Acknowledgements:

This study received no funding.

Author Contributions:

All authors contributed to the conceptualization and overall design of the study. P.R. (senior author) initiated the core research idea and supervised the project. E.S. led the study execution, including coordination, data extraction, manual validation, and primary manuscript drafting. A.M. developed the automated extraction pipeline, performed programming, and conducted the data analysis. M.M.S. contributed to data interpretation and assisted in manuscript writing. P.R. provided critical revisions, methodological guidance, and oversight throughout the study. All authors reviewed and approved the final manuscript.

Competing Interests:

All authors declare no financial or non-financial competing interests.

References

1. Gupta, P. *et al.* Rapid review: Growing usage of Multimodal Large Language Models in healthcare. *J. Biomed. Inform.* **169**, 104875; 10.1016/j.jbi.2025.104875 (2025).
2. Nazi, Z. A. & Peng, W. Large language models in healthcare and medical domain: A review. *Informatics***11,** 57; 10.3390/informatics11030057 (2024).
3. Lin, C. & Kuo, C. F. Roles and potential of Large language models in healthcare: A comprehensive review. *Biomed. J.***48,** 100868; 10.1016/j.bj.2025.100868 (2025).
4. Ullah, E., Parwani, A., Baig, M. M. & Singh, R. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology - a recent scoping review. *Diagn. Pathol.***19,** 43; 10.1186/s13000-024-01464-7 (2024).
5. Goh, E. et al. Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Netw. Open***7,** e2440969; 10.1001/jamanetworkopen.2024.40969 (2024).
6. Jung, K. H. Large language models in medicine: Clinical applications, technical challenges, and ethical considerations. *Healthc. Inform. Res.***31,** 114–124 (2025).
7. Jacob, C. et al. AI for IMPACTS framework for evaluating the long-term real-world impacts of AI-powered clinician tools: Systematic review and narrative synthesis. *J. Med. Internet Res.***27,** e67485; 10.2196/67485 (2025).
8. Mäkitie, A. A. et al. Artificial intelligence in head and neck cancer: A systematic review of systematic reviews. *Adv. Ther.* **40,** 3360–3380 (2023).
9. Tsopra, R. et al. A framework for validating AI in precision medicine: considerations from the European ITFoC consortium. *BMC Med. Inform. Decis. Mak.***21,** 274; 10.1186/s12911-021-01634-3 (2021).
10. Yin, J., Ngiam, K. Y. & Teo, H. H. Role of artificial intelligence applications in real-life clinical practice: Systematic review. *J. Med. Internet Res.***23,** e25759; 10.2196/25759 (2021).
11. Ennab, M. & Mcheick, H. Enhancing interpretability and accuracy of AI models in healthcare: a comprehensive review on challenges and future directions. *Front. Robot. AI***11,** 1444763;

10.3389/frobt.2024.1444763 (2024).

12. Payrovnaziri, S. N. et al. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *J. Am. Med. Inform. Assoc.***27,** 1173–1185 (2020).
13. Albahri, A. S. et al. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Inf. Fusion***96,** 156–191 (2023).
14. Nori, H. et al. Capabilities of GPT-4 on medical challenge problems. *arXiv* 2303.13375 (2023).
15. Raji, I. D., Daneshjou, R. & Alsentzer, E. It's time to bench the medical exam benchmark. *NEJM AI***2,** AIe2401235; 10.1056/AIe2401235 (2025).
16. Pal, A., Umapathi, L. K. & Sankarasubbu, M. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. *Proc. Conf. Health, Inference, Learn.*, in *Proc. Mach. Learn. Res.***174,** 248–260 (2022).
17. Bedi, S. et al. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA* **333,** 319–328 (2025).
18. Tricco, A. C. et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann. Intern. Med.* **169,** 467–473 (2018).
19. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.***30,** (2017).
20. Rouzrokh, P., Khosravi, B., Rouzrokh, P. & Shariatnia, M. LatteReview: A multi-agent framework for systematic review automation using large language models. *arXiv* 2501.05468 (2025).
21. Colvin, S. et al. Pydantic validation. 2.12.5. *Pydantic*https://github.com/pydantic/pydantic (n.d.).
22. Chase, H. et al. LangChain. 1.1.3. *LangChain*https://github.com/langchain-ai/langchain (2022).
23. Bedi, S. et al. Holistic evaluation of large language models for medical tasks with MedHELM. *Nat. Med.***32,** 943–951 (2026).
24. Liang, P. et al. Holistic evaluation of language models. *arXiv* 2211.09110 (2023).
25. Lee, T. et al. VHELM: a holistic evaluation of vision language models. *Adv. Neural Inf. Process. Syst.***37,** 140632–140666 (2024).

Figures

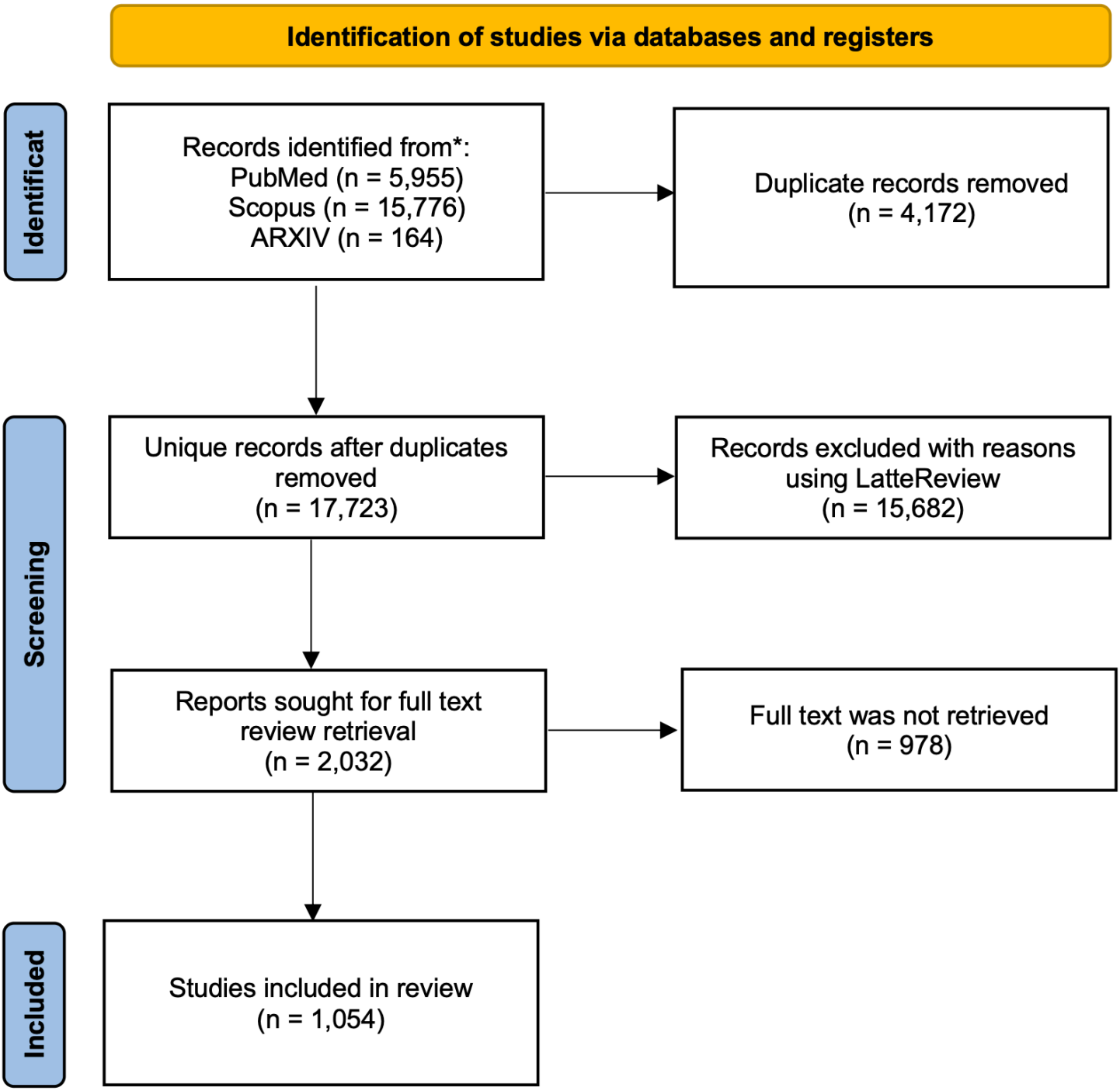


Figure 1

Study selection flow diagram for the scoping review.PRISMA-style overview of the identification, screening, and inclusion process, showing records retrieved from PubMed, Scopus, and arXiv, duplicate removal, automated exclusion using the LatteReview system, full-text retrieval, and the final set of studies included in the review.

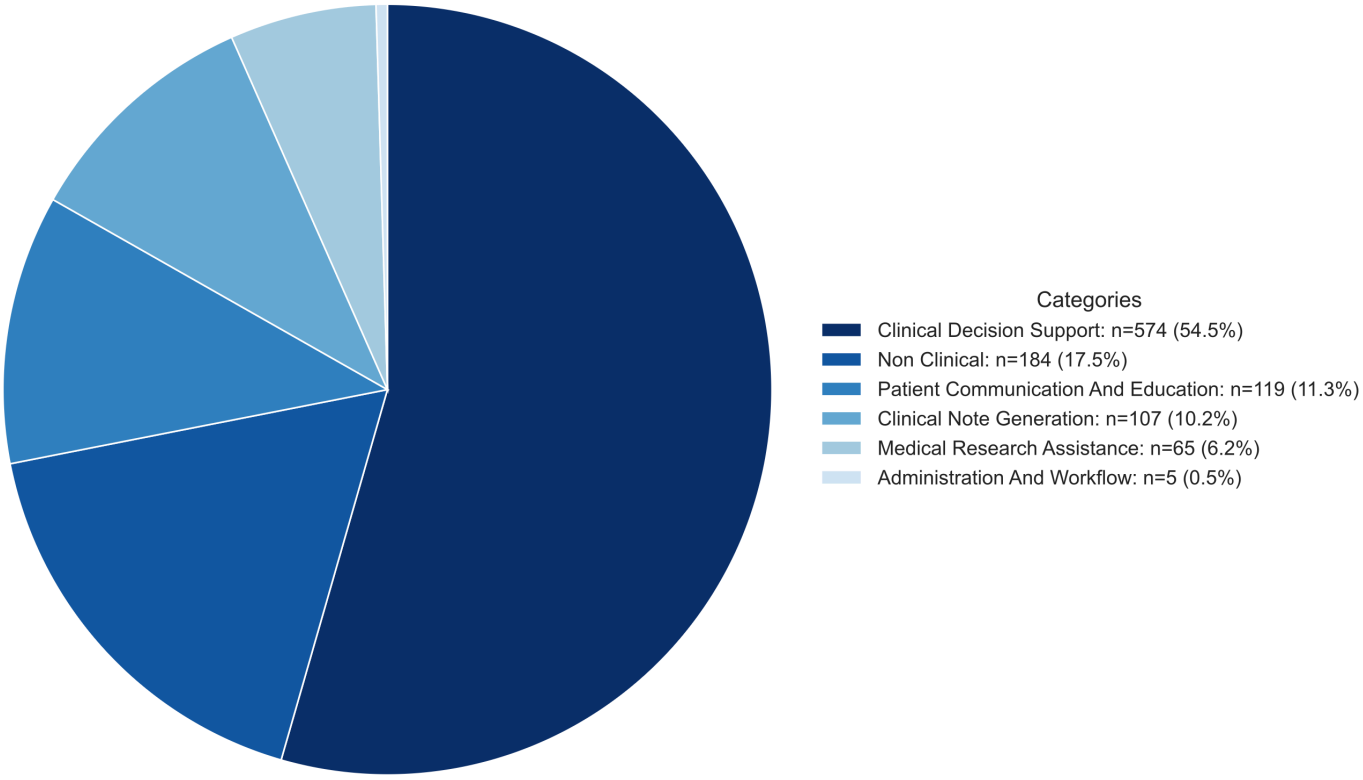


Figure 2

Distribution of MedHELM tasks across reviewed articles. The chart depicts the proportional representation of six primary categories of artificial intelligence applications across N=1,054 entries. Clinical Decision Support constitutes the largest share of the distribution, accounting for 54.5% (n=574) of the total. This is followed by non-clinical applications at 17.5% (n=184) and Patient Communication and Education at 11.3% (n=119). Clinical Note Generation and Medical Research Assistance represent 10.2% (n=107) and 6.2% (n=65) of the cohort, respectively. Administration and Workflow tools represent the smallest functional group, comprising 0.5% (n=5) of the sampled applications.

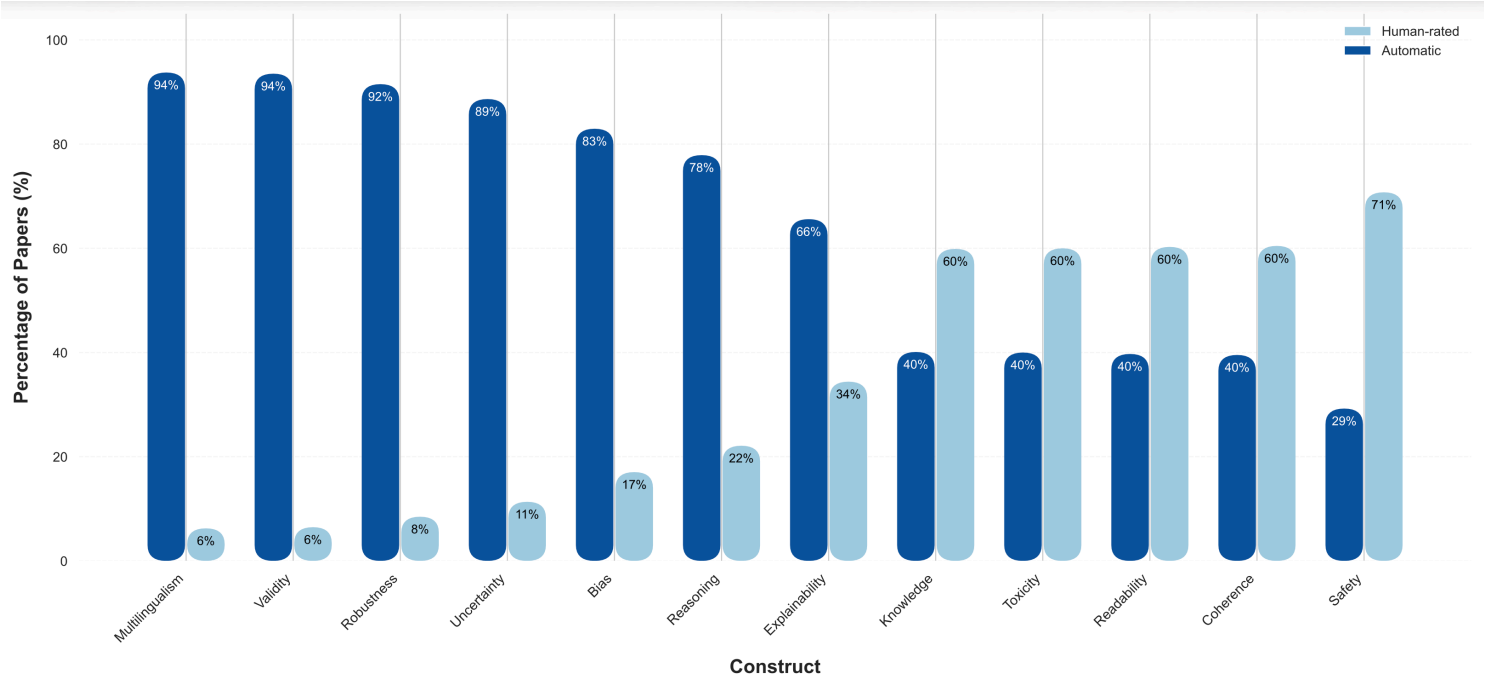


Figure 3

Comparison of automatic versus human-rated benchmarking across evaluation constructs. This grouped bar chart illustrates the methodological preference for evaluating specific AI performance dimensions, categorized by automatic (dark blue) and human-rated assessments (light blue). Technical and linguistic constructs are predominantly evaluated via automatic methods, with **Multilingualism** (94%), **Validity**(94%), and **Robustness** (92%) showing the highest reliance on automated benchmarking. Conversely, **Safety** is the only construct primarily evaluated through human oversight, with human-rated assessments accounting for 71% of papers compared to 29% for automatic metrics. Qualitative dimensions such as **Knowledge**, **Toxicity**, **Readability**, and **Coherence**exhibit a balanced distribution, each evaluated by human raters in 60% of cases and by automatic tools in 40%. These data suggest a significant reliance on automated scales for structural and technical validation, while human judgment remains the gold standard for assessing safety and nuances in content quality.

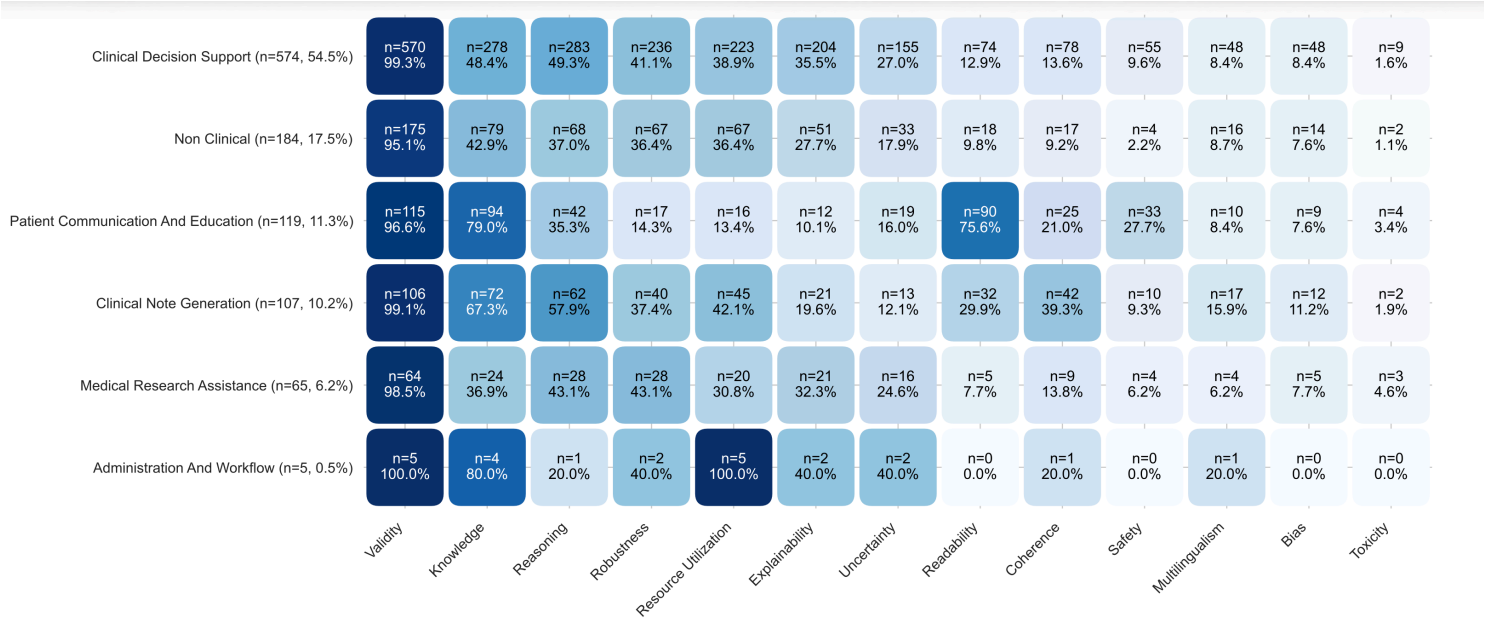


Figure 4

Heatmap of evaluation constructs analyzed across healthcare AI categories. This figure illustrates the frequency and proportion of specific evaluation constructs. Each cell indicates the number of studies (n) and the relative percentage of the subcategory evaluating a given construct. **Validity** serves as the near-universal construct across all domains, with the highest prevalence in Administration and Workflow (100%, n=5) and Clinical Decision Support (99.3%, n=570). **Knowledge** and **Reasoning**show significant focus in clinical tasks, appearing in 48.4% and 49.3% of Clinical Decision Support studies, respectively. Notably, **Readability** is a prominent focus for Patient Communication and Education (75.6%, n=90), while ethical dimensions such as **Bias** and **Toxicity** are consistently the least evaluated across all subcategories, with Toxicity frequently falling below 2%.

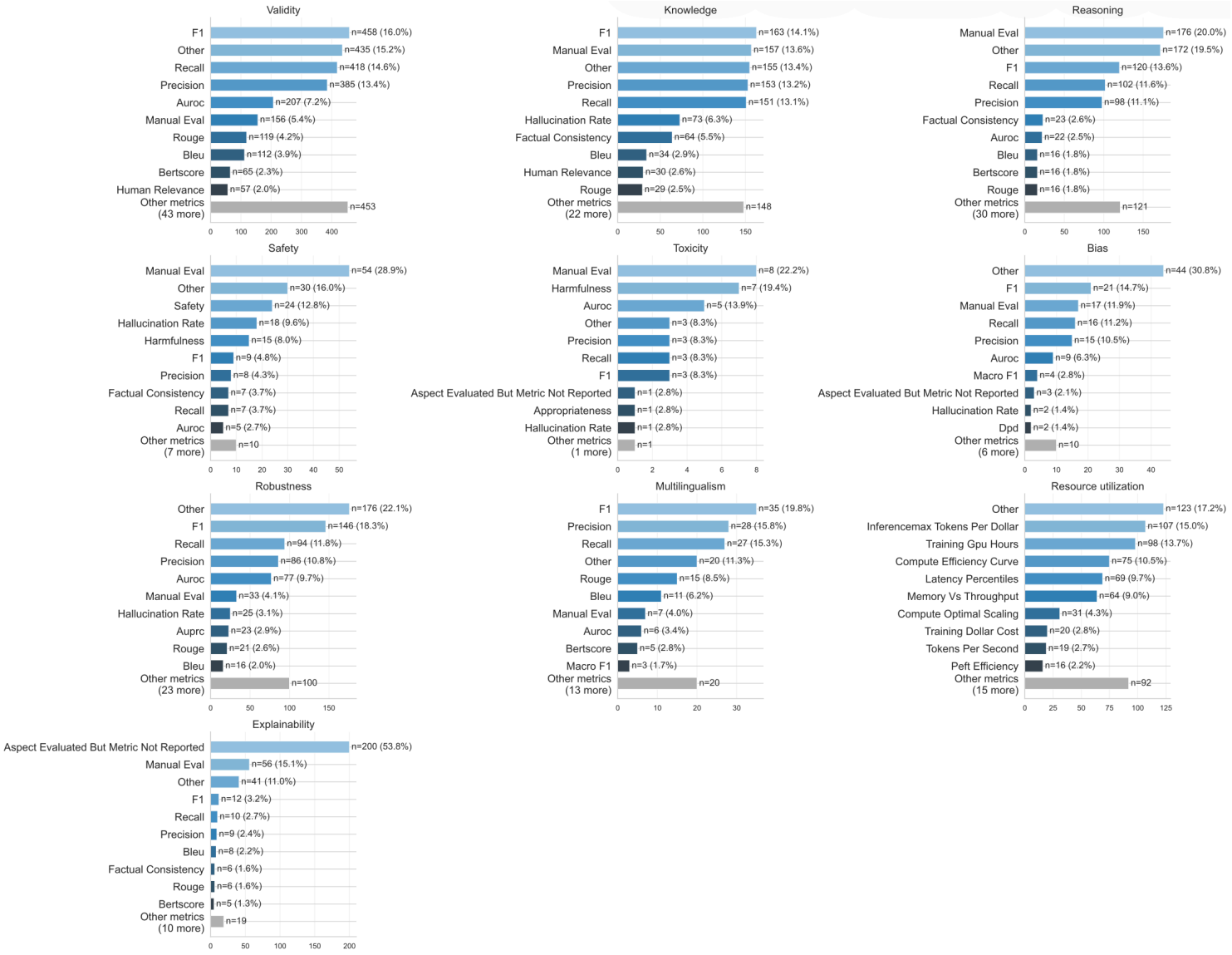


Figure 5

Distribution of evaluation metrics across performance and safety constructs. This panel provides a detailed breakdown of the metrics employed to quantify different evaluation construct of LLMs. Validity and Knowledge are primarily assessed using standard classification and retrieval metrics such as F1-score, Precision, and Recall. Conversely, human-centric constructs like Reasoning, Safety, and Toxicity rely heavily on Manual Evaluation, which constitutes 20.0%, 28.9%, and 22.2% of their respective metric distributions. Notably, Resource Utilization is characterized by specialized hardware and cost-efficiency metrics, including Inference Tokens Per Dollar (15.0%) and Training GPU Hours (13.7%). Emerging qualitative metrics such as Hallucination Rate and Factual Consistency appear across multiple constructs but remain a secondary focus compared to traditional overlap-based metrics like ROUGE and BLEU. The category Explainability shows a significant portion (53.8%) of assessments where a construct was evaluated but a specific metric was not reported or relied solely on manual review.