

## **Title**

A primer on evaluation methods for large language models in healthcare

## **Authors**

Suzannah E McKinney, MBChB<sup>1</sup>

Phuc Vu, MPH<sup>1</sup>

Samuel A Justice, PhD<sup>1</sup>

Christopher Humphries, MBChB<sup>2,3</sup>

Alyssa Pradhan, MBBS<sup>4</sup>

Timothy J Keyes, PhD<sup>5</sup>

Bernardo C Bizzo, MD, PhD<sup>1,6,7</sup>

Keith J Dreyer, DO, PhD<sup>1,6,7</sup>

Sarah F Mercaldo, PhD<sup>1,6,7</sup>

James M Hillis, MBBS, DPhil<sup>1,7,8</sup>

## **Affiliations**

<sup>1</sup>Mass General Brigham AI, Mass General Brigham, Boston, MA, USA

<sup>2</sup>Generative AI Laboratory, School of Informatics, University of Edinburgh, Edinburgh, UK

<sup>3</sup>University of Edinburgh Centre for Cardiovascular Science, Edinburgh, UK

<sup>4</sup>Experimental Medicine Division, Nuffield Department of Medicine, University of Oxford, Oxford, UK

<sup>5</sup>Department of Biomedical Data Science, Stanford University School of Medicine, Stanford University, Stanford, CA, USA

<sup>6</sup>Department of Radiology, Mass General Brigham, Boston, MA, USA

<sup>7</sup>Harvard Medical School, Boston, MA, USA

<sup>8</sup>Department of Neurology, Mass General Brigham, Boston, MA, USA

**Word count: 3824 (including subheadings)**

**Date of revision: September 13**

## **Abstract**

Large language models (LLMs) have a growing range of applications in medicine, and their evaluation is critical for ensuring they provide benefit and not harm. This evaluation can be more challenging than traditional machine learning for many reasons, including probabilistic and open-ended outputs, and behavior that shifts with prompt design and accumulated context. This review covers four key areas of LLM evaluation: principles of study design, statistical methods, capability evaluation and clinical context evaluation. Capability evaluation considers different benchmarks, including multiple-choice, agentic and multi-turn benchmarks, alongside operational metrics like token usage. Clinical context evaluation addresses establishing accuracy of free text outputs, such as human review and LLM-as-a-judge, and clinical trial approaches. Across sections, we describe underlying concepts and potential pitfalls, while emphasizing the importance of aligning evaluation methods with the research question. Together, this article aims to provide a pragmatic basis for designing and executing rigorous evaluations of healthcare LLMs.

## Introduction

The emergence of large language models (LLMs) in medicine has enabled a range of clinical and research uses.<sup>1</sup> They include clinical reasoning with physician-level performance,<sup>2</sup> agents that complete tasks like medical record reviews,<sup>3,4</sup> and chatbots that can converse directly with patients.<sup>5</sup> These applications may involve models built specifically for healthcare or broader general-purpose models that are applied in the healthcare setting,<sup>6</sup> and access to them may be open to the public or restricted to healthcare personnel.

LLMs generate more varied outputs compared to more traditional machine learning (ML) models, such as free text, multi-turn dialogue and agentic task completion. The underlying evaluation task, however, often reduces to familiar items. What is the research question? What is the appropriate reference standard? Is the study adequately powered?

This review aims to provide a pragmatic introduction to the evaluation of LLMs (Figure 1). It begins by considering study design principles and common statistical methods. It then describes capability evaluations including the use of benchmarks and operational metrics. It later addresses clinical context evaluations including how to establish accuracy of free text outputs and clinical trial approaches.

## **Study design principles**

### Study design framework

LLM evaluation is like most medical research in that it requires an objective with associated endpoints that translate to null and alternative hypotheses. These elements are ideally prespecified in a study protocol or statistical analysis plan, which are followed during study conduct. The assessments may aim to demonstrate a minimum performance or compare models. Predetermined thresholds may be set, against which performance passes or fails. Assessments may also provide descriptive analyses about operational metrics like compute usage and cost.

### Reference standard

A reference standard is determined by the research question that an evaluation aims to answer. For task-specific evaluations, it will commonly be a discrete answer such as the presence or absence of a diagnosis in an LLM output. It can also involve a benchmark as described more below. It may be generated with different degrees of rigor including utilizing ICD-10 codes for diagnoses already present in a patient's chart or receiving a consensus opinion from an expert panel.<sup>7</sup> For free-text generation, there may not be a single correct output. The evaluation may therefore rely on structured human or LLM-based assessment across multiple criteria as described more below. It may also compare outputs with an exemplar such as a radiologist-authored report.

### Training and testing datasets

While traditional ML models commonly use highly curated training datasets, LLMs are often trained on large, undisclosed corpora. Layered applications of LLMs may then require additional development and training including for prompt generation or agentic architecture design. When it comes to selecting a testing dataset, data contamination and generalizability are two key

concerns. Data contamination can manifest if testing data were part of training corpora; for example, it may not be possible to use published case reports to test diagnostic performance.<sup>8</sup> Benchmarks, which are discussed more below, aim to avoid this contamination although may themselves become part of training corpora in due course. A related item is the temporality of data elements, whereby an evaluation rubric should only provide the data available at the relevant point in the clinical pathway and not inadvertently include downstream diagnostic or treatment steps. Generalizability involves ensuring that LLMs and their layered applications work across all intended settings. Especially if the layered applications were developed in a single hospital or healthcare system, it is critical to evaluate the performance at other sites. Similarly, the demographic representativeness of a test dataset warrants explicit consideration; prespecified subgroup analyses should typically be performed for age group, sex, race, ethnicity, and site to detect performance disparities that aggregate metrics may obscure. Given clinical practice involves mild and equivocal presentations rather than only textbook presentations, representation of real-world spectrum of disease is also important; it may be facilitated by consecutive case selection when using actual clinical cases.

#### Model stability over time

Training data shift can additionally occur when models are updated with data from user activity. As mitigation, evaluations can specify a locked model version rather than a continuously updated model.<sup>9</sup> Furthermore, given LLMs can retain and be influenced by information retained within a single session, evaluations can consider using stateless single-turn application programming interface (API) calls so that earlier cases cannot influence later outputs.<sup>10</sup> Repeated testing can also be used to check for response evolution.

#### Run-to-run variability

LLM evaluation also faces a challenge from stochastic output: the same input can produce different outputs, complicating comparison within and across models. Mitigation strategies include conducting repeated runs for each case and pairing analyses on the same cases when comparing models. Results should report sampling settings and variance across runs.

Additionally, evaluations can utilize advanced features available via API access, including sampling or temperature settings to minimize output randomness.<sup>11</sup> A fixed random seed, if supported, may also make any residual randomness reproducible across runs.

### Using LLMs for the evaluation

Beyond LLMs being the subject of evaluations, they may be used to perform evaluations, referred to here as a methodologic LLM. For example, an LLM could assist with case selection by reviewing records against inclusion criteria as they have already been used in clinical trials.<sup>12,13</sup> An LLM used to make evaluation decisions, known as “LLM-as-a-judge”, is discussed more below. To ensure fairness and consistency, multiple methodologic LLMs can perform the same task and manual review can be performed on a sample of outputs.<sup>14</sup> Prompts provided to a methodologic LLM should be unambiguous and reported in a publication.

### Reporting checklists

Reporting guidelines have become a requirement for many journals to ensure manuscripts include critical elements. They improve comparability, reproducibility and appraisal, and can serve as a design-stage checklist. TRIPOD-LLM provides a checklist of items in a modular format that accommodates different LLM study designs and tasks, emphasizing transparency and task-specific performance reporting.<sup>9</sup> Discipline-specific instruments include FLAIR for radiology.<sup>15</sup>

### Sample size and powering

The number of cases needed in a study depends on the research question and degree of variation expected in the results. In practice, case availability, time and funding can also impose constraints. Studies comparing systems should include enough cases to detect a meaningful or clinically important difference; this difference should be specified a priori. Before the study begins, a sample-size calculation may estimate the number of cases required. This calculation may utilize the statistical power needed to detect a specified difference or a confidence interval of an acceptable width.

### Confidence, significance and meaning

Statistical methods should be appropriate for the research question and the type of data collected. While results are often reported using p-values, a threshold of  $p < 0.05$  does not establish that a difference is real, clinically meaningful, or large; a very small difference can be statistically significant in a large sample. It is therefore often desirable to report confidence intervals that help readers understand both the possible size of an effect and the uncertainty around it. A 95% confidence interval shows the range of values that are reasonably consistent with the observed data; in repeated studies conducted in the same way, approximately 95% of the confidence intervals would be expected to contain the true value. There are also alternative statistical paradigms to the standard frequentist methods that this review focuses on.<sup>16</sup> Beyond individual evaluations, tiered evidence frameworks can situate findings in a broader hierarchy of evidence: from retrospective benchmarking through silent prospective deployment to randomized comparison against routine care.<sup>17,18</sup> For teams planning an evaluation program, these frameworks can provide a staged roadmap with criteria for advancement.

## Statistical methods

Many traditional statistical methods are used for LLM evaluations given LLM outputs are often reduced to categorical or numeric forms.

### Categorical outputs

Classification tasks involve categorical outputs and are split into several types. Binary classification involves a model picking one of two options, like the presence or absence of a condition. Multiclass and ordinal classification involve a model picking one of three or more options, such as in assigning a disease severity grade. Multilabel classification involves a model potentially picking more than one option from a list, such as an LLM generating diagnoses on a problem list. Beyond returning the respective labels for classification tasks, an LLM may provide an underlying probability or confidence score for those labels. It is commonly a numeric value between 0 and 1, and allows assessment of model discrimination for how well it ranks cases. If the score reflects a true probability that intends to match frequency (i.e., 0.8 corresponds to an 80% occurrence), calibration can also be assessed.

For a binary classification task, the standard metrics are sensitivity, specificity, positive and negative predictive value, and accuracy.<sup>19,20</sup> Accuracy alone can be misleading when one class is rare.<sup>21</sup> If only 5% of notes document the condition, a model that always answers “no” reaches 95% accuracy while missing every true case. Balanced accuracy,<sup>22</sup> which averages sensitivity and specificity, and the F1 score,<sup>23</sup> which rewards true positive outputs without too many false positive outputs, are useful in this setting.

For multiclass tasks, raw agreement does not correct for chance and can mislead when classes are imbalanced. Cohen’s kappa measures agreement while correcting for chance.<sup>24</sup> For ordinal categories, where disagreement between categories equates to a larger error, weighted kappa

is useful.<sup>25</sup> Per-class performance can be summarized by macro-averaging, which weights every category equally, or micro-averaging, which weights every case equally; the two diverge under class imbalance, so the choice should be reported.

For multilabel classification tasks, a key question is how much outputs overlap a reference set. The Dice coefficient<sup>26</sup> and the Jaccard index<sup>27</sup> (intersection over union) both measure overlap and differ in normalization. The Tversky index<sup>28</sup> weights incorrect additions and missed items differently, which helps when one error is more costly such as omitting a diagnosis rather than adding one.

Discrimination may be conceptualized through thinking that each score reflects a threshold that yields its own sensitivity and specificity. Varying these thresholds and plotting sensitivity against the false-positive rate ( $1 - \text{specificity}$ ) creates a receiver operating characteristic (ROC) curve.<sup>29</sup> The area under a ROC curve (AUC) shows the probability that the model gives a randomly chosen positive case a higher score than a randomly chosen negative one: 0.5 is chance, 1 is perfect. If there is a negative class imbalance and the impact of false positive cases needs to be considered, the area under the precision-recall curve (AUPRC) can provide an informative metric by plotting precision (positive predictive value) against recall (true positive rate).<sup>30</sup>

Calibration matters when a probability is shown to a clinician or feeds a downstream decision rule. The Brier score gives a single summary number, though it blends calibration with discrimination rather than isolating either.<sup>31</sup> Decision curve analysis calculates the net benefit of using the model to guide clinical decisions across a range of threshold probabilities, and compares this benefit to treating all patients or none. It is most useful when a model's output feeds directly into a consequential clinical action.<sup>32</sup>

## Numeric outputs

Correlation, agreement, and error each answer a different question about numeric outputs and should be reported separately. Correlation asks whether predicted and reference values vary together: Pearson for a linear relationship, Spearman for a monotonic one.<sup>33,34</sup> Correlation does not incorporate agreement, which asks whether the values coincide on the same scale.

Agreement can instead be assessed with the intraclass correlation coefficient and the concordance correlation coefficient.<sup>35,36</sup> Error quantifies the typical deviation. The root mean square error (RMSE) and mean absolute error report error in the units of the output, with RMSE penalizing larger misses more heavily.

## Capability evaluation

Capability evaluation asks whether a model possesses the broad clinical knowledge, reasoning, and task-completion abilities required for potential use in healthcare.

### Benchmarks

The term ‘benchmark’ refers to a fixed dataset of tasks with predefined correct answers and an associated scoring procedure. The scoring procedure draws on the same underlying methods used to judge any model output against a reference, including statistical, language-based, human, and LLM-as-a-judge approaches as described in other sections. Critically, it applies them through a fixed, predetermined protocol. Benchmarks facilitate easy comparison of models.

#### *Multiple-choice and knowledge benchmarks*

The initial benchmarks were knowledge-oriented and included MedQA<sup>37</sup>, a set of US Medical Licensing Examination-style multiple-choice questions. Such benchmarks can be helpful as an early signal for model performance but are increasingly considered insufficient for testing clinical readiness, capturing factual recall more than real-world utility.<sup>38,39</sup> They also often focus on a single timepoint rather than a multistep, iterative clinical workflow process.

#### *Multi-task and agentic benchmarks*

Multi-task frameworks have been established as a more robust approach to knowledge benchmarking. They assemble many clinical tasks and domains into a single standardized and reproducible approach. MedHELM assesses models against a clinician-validated taxonomy of 121 real clinical tasks spanning decision support, note generation, patient communication, research and administration, rather than against exam-style knowledge questions.<sup>40</sup> The value of these frameworks is breadth: they characterize performance across different tasks and

domains. Limitations include that aggregate scores can obscure task-level variation, constituent tasks may not resemble the entire clinical workflow, and public benchmarks can contribute to training-data contamination.

AI agents involve systems that can autonomously plan and execute a sequence of actions, including using tools or external data sources. They therefore require evaluation frameworks that assess the quality of individual outputs but also an agent's ability to navigate complex environments. The current benchmarks include Medical Intelligence for Reasoning and Action (MIRA), MedAgentBench, Physician Bench and HealthAgentBench.<sup>3,41–43</sup>

### *Multi-turn benchmarks*

A multi-turn conversation involves a model producing a series of exchanges with the user, as could be imagined with a chatbot. General frameworks commonly use an LLM-as-a-judge that scores the responses to a fixed set of benchmark conversations against several criteria, for example factual accuracy and communication quality.<sup>4,44–47</sup> To mimic the stochastic nature of real conversations, other benchmark designs use patient actors to generate the dialogue instead of a fixed script or LLM agents prompted with a predefined case vignette.<sup>48</sup> These simulated users typically require their own further validation.

### *Reasoning benchmarks*

Typical methods of scoring LLM outputs assume that an output is correct if it matches the reference answer, irrespective of the reasoning used to get there. A model that reasons to the correct diagnosis is not equivalent to one that anchors on a keyword and returns the most common answer; an assessment based on only the final label cannot distinguish them.

Reasoning assessment evaluates the process rather than an endpoint. This distinction matters

because the black box nature of LLMs can limit their reliability and trustworthiness. Models can be prompted to divide tasks into smaller steps to demonstrate their step-by-step reasoning, known as Chain-of-Thought (CoT) reasoning.<sup>49,50</sup> Multiple benchmarks exist where clinical reasoning is present in the referenced answers, including MedThink-Bench,<sup>51</sup> MedR-Bench,<sup>52</sup> Open-XDDx<sup>53</sup> and ER-Reason.<sup>54</sup> These benchmarks vary in whether the reference reasoning is expert-authored or machine-generated. Separately, script concordance testing investigates how added data adjust outputs, aiming to reflect the often uncertain and dynamic nature of clinical decision making.<sup>55</sup>

While generated CoT reasoning can provide insight into how a model reaches its answer, several studies have shown divergence between the stated reasoning and the process driving a model's output. One study found that truncating or inserting errors into the reasoning chain often leaves the final answer unchanged, indicating that the intervening steps were not load-bearing.<sup>56</sup> Another study introduced input features aiming to bias a model's answer: models frequently produced CoT explanations that justified an incorrect answer and did not mention the biasing features in those explanations.<sup>57</sup> Parametric methods also allow CoT assessments; they involve erasing information in a model's weights about individual reasoning steps and testing whether predictions consequently change.<sup>58</sup>

These methods commonly depend on model-stated CoTs and a future direction for reasoning assessments utilizes separate insights. The J-space by Anthropic considers a set of internal representations identified by an interpretability method called the Jacobian lens.<sup>59</sup> It surfaces intermediate assessments a model has formed but not verbalized. Further methods include probing classification<sup>60</sup> and activation patching.<sup>61</sup>

## Operational metrics

Operational metrics like cost and latency are critical to consider alongside accuracy, especially as there can be a tradeoff between them.<sup>62</sup> Furthermore, energy footprint and sustainability have drawn increasing attention as generative AI is deployed at scale.<sup>63</sup>

Tokens are defined as the basic piece of text or data that an AI model receives as an input, such as a word, part of a word or punctuation.<sup>64</sup> For example, “Unreal!” may contain three tokens of “Un”, “real” and “!”. Token usage is the count of input and output tokens per query. Cost per query translates this usage into dollar terms, which allows health systems to budget for LLM-based tool deployment. Latency is the total time delay between submitting a prompt and receiving an answer. It can determine whether a system can be integrated into real-time clinical workflows such as point-of-care decision support. Broader system measures for reliability and scalability include the frequency of system outages or downtime, error and failure rates, and behavior under load.<sup>65</sup>

The operational characteristics may also be specified to a model including the maximum output tokens, which can determine the length of responses. Two additional modifiable parameters are temperature and Top-P.<sup>66</sup> Temperature adjusts the randomness in generating an output: a lower temperature provides more deterministic outputs and is usually necessary for healthcare applications, while a higher temperature helps with tasks that benefit from creativity like writing fiction. Top-P reflects the underlying LLM methodology of considering a set of options for outputting the next token when generating responses: a lower Top-P constrains these options and therefore provides more predictable outputs. Given temperature and Top-P influence outputs in similar ways, typically only one is adjusted.

## **Clinical context evaluation**

Following capability evaluation, a key question is whether LLMs perform in the intended clinical context outside the structured framework of benchmarks.

### Accuracy of free text outputs

Many LLM tasks produce free text like a patient message response or a discharge summary.

This text may be possible to reduce to categorical or numeric form such that traditional statistical methods can be used. Equally, the text may need to be considered more holistically, so exact-match scoring against a reference standard breaks down.

### *Natural Language Processing metrics*

Natural Language Processing (NLP) has developed methods for quantifying free text accuracy over many decades. Overlap metrics like BLEU,<sup>67</sup> ROUGE,<sup>68</sup> and METEOR<sup>69</sup> score the generated text by counting how many short word sequences (known as n-grams) are shared with reference text. They often require exact matches and may penalize a correct answer that is simply phrased differently. Embedding-based metrics like BERTScore<sup>70</sup> and cosine similarity<sup>71</sup> convert text into numerical vectors to capture its meaning, then compare vectors with those from reference text. They are better at crediting paraphrasing, but can remain insensitive to clinically important details like negations, numbers and named entities. Both approaches were designed for shorter, more constrained text, and are now widely regarded as inadequate for evaluating the longer, more open-ended outputs typical of LLMs.

NLP also uses readability metrics to characterize text. The Flesch-Kincaid Grade Level<sup>72</sup> and Flesch Reading Ease<sup>73</sup> estimate reading difficulty from average sentence length and average number of syllables per word. They measure form, not accuracy or meaning, and a readable

passage can still be wrong; they should be reported alongside an accuracy metric. They matter most for patient-facing text including education materials or plain-language summaries.

#### *Acceptable answers by human review*

Free-text outputs may be scored by human subject matter experts using structured scoring rubrics. This approach allows healthcare specific considerations to be explicitly evaluated such as clinical safety, data privacy and ethical alignment. Instruments such as QUEST provide literature-derived dimensions for structured human review, replacing ad hoc rubrics that otherwise differ from study to study and prevent pooling of results.<sup>74</sup> They also make rater reliability measurable: with a defined instrument, inter-rater agreement can be reported and calibration performed. In one prospective deployment, physicians reviewing AI-generated discharge summaries completed a structured feedback instrument covering type of error, harm severity and likelihood using the AHRQ Common Format Harm Scale, demonstrating how a structured instrument can operationalize human review of free-text clinical output.<sup>75</sup> Human review of model outputs within a clinical workflow is the gold standard in evaluation but is limited by scaling. Crowd-sourced evaluations have been proposed as a possible solution.<sup>76</sup>

Domain-specific frameworks recognize that evaluation criteria are not always portable across contexts. For example, principles suited to acute care transfer poorly to consumer wellness applications, which commonly involve no clinician in the loop, the user as the decision-maker, and personalization as a design goal. The SHARP (safety, helpfulness, accuracy, relevance, personalization) framework was developed specifically for consumer health and wellness LLMs.<sup>77</sup>

#### *Acceptable answers by LLM-as-a-judge review*

An LLM-as-a-judge can assess outputs against predefined evaluation criteria and scoring frameworks.<sup>78,79</sup> This approach scales more easily than human review and is particularly useful when evaluating adherence to a clinical rubric or qualitative characteristics such as relevance. The LLM should be provided with clear instructions and, where appropriate, reference standards or exemplars. Several approaches can help ensure consistency and fairness. The LLM should remain fixed across a study to avoid introducing variation. It should also ideally differ in family from the evaluated models, since models from the same family can exhibit a strong preference for their own outputs.<sup>80</sup> When using an LLM-as-a-judge, it is important to also conduct a sample of the assessments with human review.

Faithfulness metrics ask if generated text is supported by the source. They typically include two key components: groundedness checks that statements trace back to supporting passages,<sup>81</sup> while hallucination rate reports the frequency of model statement without such support.<sup>82</sup> ACUEval is a tool that assesses the faithfulness of summarizations with an LLM-as-a-judge approach.<sup>83</sup> It breaks down summarizations into individual facts (referred to as atomic content units) and then compares them with the source text. Similarly, VeriFact is a tool that verifies individual statements in LLM-generated documents like the hospital course in discharge summaries,<sup>84</sup>

### Clinical trial methodology

Beyond evaluating LLM outputs, traditional clinical trial methodology can assess the impact of LLMs on clinical outcomes. This methodology includes randomized clinical trials, where randomization occurs at a provider or institution level and compares those using versus not using an LLM. A challenge with any trial is balancing endpoint metrics that have clearer attribution to LLM use with metrics that have more substantial clinical meaning like mortality rate; the latter may have many confounding factors unless they are appropriately controlled.

## **Conclusion**

LLMs have introduced new capabilities to medicine and provide promise for many more to follow. They will require rigorous evaluation, which will draw on existing study methodologies alongside newer methodologies to account for considerations like agentic systems and multi-turn conversations. This primer described four complementary areas: principles of study design, statistical methods, capability evaluation and clinical context evaluation. As LLM evaluations continue to evolve, there will likely be an increasing focus on real-world performance and clinical outcomes. Ultimately, rigorous healthcare LLM evaluation should not seek a single universal metric, but a transparent and reproducible framework for selecting complementary assessments that align with the research question being asked and the clinical decision that the model intends to support.

## References

1. Teo, Z. L. *et al.* Generative artificial intelligence in medicine. *Nat Med* **31**, 3270–3282 (2025).
2. Brodeur, P. G. *et al.* Performance of a large language model on the reasoning tasks of a physician. *Science* **392**, 524–527 (2026).
3. Ferber, D. *et al.* Towards autonomous medical artificial intelligence agents. *Nature* **655**, 1282–1291 (2026).
4. Liévin, V. *et al.* Towards conversational artificial intelligence for disease management. *Nature* **655**, 1292–1299 (2026).
5. Ayers, J. W. *et al.* Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med* **183**, 589–596 (2023).
6. Vishwanath, K. *et al.* General-purpose large language models outperform specialized clinical AI tools on medical benchmarks. *Nat Med* **32**, 2405–2409 (2026).
7. Mercaldo, S., Hillis, J. & Blume, J. Evaluating the Performance and Clinical Utility of AI-driven Diagnostic Tools in Radiology. *Radiology* **317**, (2025).
8. Yan, Z. *et al.* LiveMedBench: A Contamination-Limited Medical Benchmark for LLMs with Automated Rubric Evaluation. in *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2* 10162–10173 (Association for Computing Machinery, New York, NY, USA, 2026). doi:10.1145/3770855.3817579.
9. Gallifant, J. *et al.* The TRIPOD-LLM reporting guideline for studies using large language models. *Nat. Med.* **31**, 60–69 (2025).
10. Park, S. H. *et al.* Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Healthcare (MI-CLEAR-LLM): 2025 Updates. *Korean J Radiol* **26**, 1123–1132 (2025).

11. Atıl, B. *et al.* Non-Determinism of “Deterministic” LLM System Settings in Hosted Environments. in *Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems* (eds Akter, M. *et al.*) 135–148 (Association for Computational Linguistics, Mumbai, India, 2025). doi:10.18653/v1/2025.eval4nlp-1.12.
12. Wornow, M. *et al.* Zero-shot clinical trial patient matching with LLMs. *NEJM AI* **2**, Alcs2400360 (2025).
13. Unlu, O. *et al.* Manual vs AI-Assisted Prescreening for Trial Eligibility Using Large Language Models-A Randomized Clinical Trial. *Jama* **333**, 1084–1087 (2025).
14. Verga, P. *et al.* Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *ArXiv Prepr. ArXiv240418796* (2024).
15. Kottlors, J. *et al.* Guidelines for Reporting Studies on Large Language Models in Radiology: An International Delphi Expert Survey. *Radiology* **318**, e250913 (2026).
16. Qiu, L. *et al.* Bayesian teaching enables probabilistic reasoning in large language models. *Nat. Commun.* **17**, 1238 (2026).
17. Chen, S. F. *et al.* LLM-assisted systematic review of large language models in clinical medicine. *Nat. Med.* **32**, 1152–1159 (2026).
18. Park, Y. *et al.* Evaluating artificial intelligence in medicine: phases of clinical research. *JAMIA Open* **3**, 326–331 (2020).
19. Altman, D. G. & Bland, J. M. Diagnostic tests 2: Predictive values. *Bmj* **309**, 102 (1994).
20. Altman, D. G. & Bland, J. M. Diagnostic tests. 1: Sensitivity and specificity. *Bmj* **308**, 1552 (1994).
21. H. He & E. A. Garcia. Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.* **21**, 1263–1284 (2009).
22. Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan & Joachim M. Buhmann. The Balanced Accuracy and Its Posterior Distribution. (2010) doi:10.1109/icpr.2010.764.
23. Nancy Chinchor. MUC-4 evaluation metrics. (1992) doi:10.3115/1072064.1072067.

24. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **20**, 37–46 (1960).
25. Cohen, J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* **70**, 213–20 (1968).
26. Dice, L. R. Measures of the Amount of Ecologic Association Between Species. *Ecology* **26**, 297–302 (1945).
27. Jaccard, P. THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE. *New Phytol.* **11**, 37–50 (1912).
28. Tversky, A. Features of similarity. *Psychol. Rev.* **84**, 327–352 (1977).
29. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).
30. Saito, T. & Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE* **10**, e0118432 (2015).
31. Brier, G. W. Verification of Forecasts Expressed in Terms of Probability. *Mon. Weather Rev.* **78**, 1 (1950).
32. Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Mak.* **26**, 565–74 (2006).
33. Spearman, C. The proof and measurement of association between two things. *Am. J. Psychol.* **15**, 72–101 (1904).
34. Pearson, K. VII. Note on regression and inheritance in the case of two parents. *Proc. R. Soc. Lond.* **58**, 240–242 (1895).
35. Koo, T. K. & Li, M. Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J Chiropr Med* **15**, 155–63 (2016).
36. Lin, L. I. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45**, 255–68 (1989).

37. Jin, D. *et al.* What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Appl. Sci.* **11**, 6421 (2021).
38. Raji, I., Daneshjou, R. & Alsentzer, E. It's Time to Bench the Medical Exam Benchmark. *NEJM AI* **2**, (2025).
39. Goh, E. *et al.* Toward a test of medical AI superintelligence. *Nat. Med.*  
<https://doi.org/10.1038/s41591-026-04539-8> (2026) doi:10.1038/s41591-026-04539-8.
40. Bedi, S. *et al.* Holistic evaluation of large language models for medical tasks with MedHELM. *Nat Med* **32**, 943–951 (2026).
41. Liu, R. *et al.* PhysicianBench: Evaluating LLM agents in real-world EHR environments. *ArXiv Prepr. ArXiv260502240* (2026).
42. Liu, Q. *et al.* HealthAgentBench: A Unified Benchmark Suite of Realistic Agentic Healthcare Environments for Challenging Frontier AI Agents. *ArXiv Prepr. ArXiv260631179* (2026).
43. Jiang Yixing *et al.* MedAgentBench: A Virtual EHR Environment to Benchmark Medical LLM Agents. *NEJM AI* **2**, Aldbp2500144 (2025).
44. Arora, R. K. *et al.* Healthbench: Evaluating large language models towards improved human health. *ArXiv Prepr. ArXiv250508775* (2025).
45. Bai, G. *et al.* Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 7421–7454 (2024). doi:10.18653/v1/2024.acl-long.401.
46. Kwan, W.-C. *et al.* Mt-eval: A multi-turn capabilities evaluation benchmark for large language models. in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* 20153–20177 (Association for Computational Linguistics, Miami, Florida, USA, 2024). doi:10.18653/v1/2024.emnlp-main.1124.

47. Tu, T. *et al.* Towards conversational diagnostic artificial intelligence. *Nature* **642**, 442–450 (2025).
48. Johri, S. *et al.* An evaluation framework for clinical use of large language models in patient interaction tasks. *Nat. Med.* **31**, 77–86 (2025).
49. Wei, J. *et al.* Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **35**, 24824–24837 (2022).
50. Liévin, V., Hother, C. E., Motzfeldt, A. G. & Winther, O. Can large language models reason about medical questions? *Patterns N* **5**, 100943 (2024).
51. Zhou, S. *et al.* Automating expert-level medical reasoning evaluation of large language models. *Npj Digit. Med.* **9**, 34 (2025).
52. Qiu, P. *et al.* Quantifying the reasoning abilities of LLMs on clinical cases. *Nat. Commun.* **16**, 9799 (2025).
53. Zhou, S. *et al.* Explainable differential diagnosis with dual-inference large language models. *Npj Health Syst.* **2**, 12 (2025).
54. Mehandru, N. *et al.* Er-reason: A benchmark dataset for llm-based clinical reasoning in the emergency room. *ArXiv Prepr. ArXiv250522919* (2025).
55. McCoy, L. G. *et al.* Assessment of large language models in clinical reasoning: a novel benchmarking study. *NEJM AI* **2**, Aldbp2500120 (2025).
56. Lanham, T. *et al.* Measuring faithfulness in chain-of-thought reasoning. *ArXiv Prepr. ArXiv230713702* (2023).
57. Turpin, M., Michael, J., Perez, E. & Bowman, S. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Adv. Neural Inf. Process. Syst.* **36**, 74952–74965 (2023).
58. Tutek, M., Hashemi Chaleshtori, F., Marasovic, A. & Belinkov, Y. Measuring Chain of Thought Faithfulness by Unlearning Reasoning Steps. in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (eds

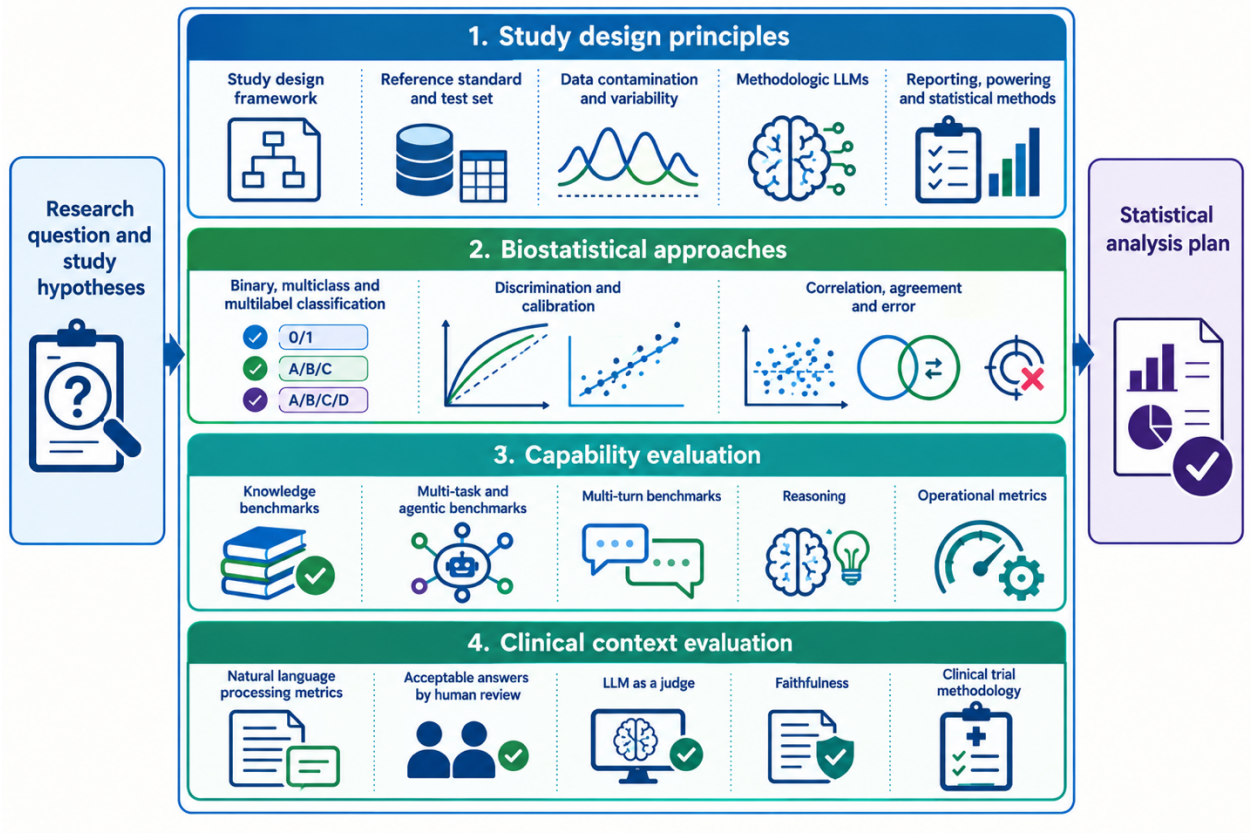
- Christodoulopoulos, C., Chakraborty, T., Rose, C. & Peng, V.) 9935–9960 (Association for Computational Linguistics, Suzhou, China, 2025). doi:10.18653/v1/2025.emnlp-main.504.
59. Gurnee, W. *et al.* Verbalizable Representations Form a Global Workspace in Language Models. *Transformer Circuits Thread* (2026).
60. Belinkov, Y. Probing Classifiers: Promises, Shortcomings, and Advances. *Comput. Linguist.* **48**, 207–219 (2022).
61. Zhang, F. & Nanda, N. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. in *International Conference on Learning Representations* (eds Kim, B. *et al.*) vol. 2024 1651–1678 (2024).
62. Oami, T., Okada, Y. & Nakada, T. A. Optimal large language models to screen citations for systematic reviews. *Res Synth Methods* **16**, 859–875 (2025).
63. Ding, Z. *et al.* Tracking the carbon footprint of global generative artificial intelligence. *Innov. Camb* **6**, 100866 (2025).
64. Nebius team. What is a token in AI? Understanding how AI processes language with tokenization. <https://nebius.com/blog/posts/what-is-token-in-ai> (2025).
65. Ranganathan, B., Zhang, M. & Wu, K. Enhancing reliability in AI inference services: An empirical study on real production incidents. *ArXiv Prepr. ArXiv251107424* (2025).
66. Raiden. Understanding ChatGPT's Parameters: Model Foundations and API Controls. <https://cyberraiden.wordpress.com/2025/08/06/understanding-chatgpts-parameters-model-foundations-and-api-controls/> (2025).
67. Kishore Papineni, Salim Roukos, Todd Ward & Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. (2002) doi:10.3115/1073083.1073135.
68. Lin, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. (2004).
69. Alon Lavie & Abhaya Agarwal. Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments. (2007).

70. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. BERTScore: Evaluating Text Generation with BERT. in (Addis Ababa, Ethiopia, 2020).
71. Reimers, N. & Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. (2019) doi:10.18653/v1/D19-1410.
72. Kincaid, P., Fishburne, R. P., Rogers, R. L. & Chissom, B. S. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. (1975).
73. Flesch, R. A new readability yardstick. *J. Appl. Psychol.* **32**, 221–233 (1948).
74. Tam, T. Y. C. *et al.* A framework for human evaluation of large language models in healthcare derived from literature review. *Npj Digit. Med.* **7**, 258 (2024).
75. Grolleau, F. *et al.* Physician-Reported Safety Outcomes of AI-Generated Hospital Course Summaries. *JAMA Netw. Open* **9**, e2616556–e2616556 (2026).
76. Hood, M. *et al.* The Healthcare AI Challenge: Grading Foundation Model Chest Radiograph Drafts Against the Radiologist Reference Standard. In submission.
77. Winslow, B. *et al.* A Principle-based Framework for the Development and Evaluation of Large Language Models for Health and Wellness. *ArXiv Prepr. ArXiv251208936* (2025).
78. Zheng, L. *et al.* Judging Llm-as-a-judge with mt-bench and chatbot arena. *Adv. Neural Inf. Process. Syst.* **36**, 46595–46623 (2023).
79. Genovese, A. *et al.* Artificial Authority: The Promise and Perils of LLM Judges in Healthcare. *Bioeng. Basel* **13**, (2026).
80. Panickssery, A., Bowman, S. R. & Feng, S. Llm evaluators recognize and favor their own generations. *Adv. Neural Inf. Process. Syst.* **37**, 68772–68802 (2024).
81. Rashkin, H. *et al.* Measuring Attribution in Natural Language Generation Models. *Comput. Linguist.* **49**, 777–840 (2023).
82. Ziwei Ji *et al.* Survey of Hallucination in Natural Language Generation. *ACM Comput Surv* **55**, Article 248 (2023).

83. Wan, D. *et al.* ACUEval: Fine-grained Hallucination Evaluation and Correction for Abstractive Summarization. (2024) doi:10.18653/v1/2024.findings-acl.597.
84. Chung Philip *et al.* Verifying Facts in Patient Care Documents Generated by Large Language Models Using Electronic Health Records. *NEJM AI* **3**, Aldbp2500418 (2026).

Figures

Figure 1: From research question to statistical analysis plan: a framework for evaluating healthcare LLMs. Image generated using ChatGPT.



## Tables

Table 1: Overall summary of evaluation techniques

| Evaluation category                | Question it answers                                                                                                 | When it applies                                                                                                          | Representative approaches                                                                                                              | Principal limitation                                                                                                            |
|------------------------------------|---------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------|
| <b>Statistical methods</b>         | Does the output match a correct answer, and how reliably?                                                           | Output is discrete or numeric and a reference standard exists (or free text has been reduced to such a form)             | Sensitivity, specificity, AUC, calibration, kappa, ICC, RMSE                                                                           | Requires a reference standard; scores the endpoint only                                                                         |
| <b>Capability evaluation</b>       | Does the model have the broad clinical knowledge, reasoning, and task-completion ability needed for healthcare use? | Assessing overall model readiness via standardized, fixed-protocol tasks, before or independent of a specific deployment | Multiple-choice/knowledge benchmarks, multi-task/agentive benchmarks, multi-turn benchmarks, reasoning benchmarks, operational metrics | Benchmark performance may not reflect real clinical workflow; aggregate scores obscure task-level variation; contamination risk |
| <b>Clinical context evaluation</b> | Does the model perform well in its intended real-world context, beyond structured benchmarks?                       | Evaluating free-text or in-context outputs for a specific clinical use case                                              | NLP metrics, human review, domain-specific frameworks, LLM-as-a-judge, faithfulness metrics                                            | Human review doesn't scale; LLM judges can inherit bias; criteria don't always apply across contexts                            |

Table 2: Statistical techniques

| <b>Evaluation Category</b>                          | <b>Metric</b>                                                                                                                                                                       | <b>Purpose</b>                                                                                                                      | <b>Example Application (Healthcare LLM)</b>                                                                         |
|-----------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------|
| <b>Binary classification</b>                        | Sensitivity (recall, true-positive rate), specificity (true-negative rate), positive predictive value (precision), negative predictive value, accuracy, balanced accuracy, F1 score | Measure classification performance given predicted labels                                                                           | Presence/absence detection (e.g., condition documentation, abnormal-finding flagging)                               |
| <b>Multiclass / ordinal agreement</b>               | Cohen $\kappa$ , weighted $\kappa$ (linear or quadratic), macro- and micro-averaged precision / recall / F1                                                                         | Quantify chance-corrected agreement and per-class performance                                                                       | Severity grading (e.g., staging, triage category)                                                                   |
| <b>Multilabel classification</b>                    | Dice coefficient, Jaccard index (intersection over union), Tversky index                                                                                                            | Measure the similarity between a predicted set of items and a reference set                                                         | Set extraction and coding (e.g., ICD-10 code assignment, problem-list extraction)                                   |
| <b>Discrimination and ranking</b>                   | Receiver operating characteristic (ROC) curve, area under the ROC curve (AUC), area under the precision–recall curve (AUPRC)                                                        | Evaluate overall ability to separate the classes, aggregated over a range of thresholds                                             | Risk stratification and triage (e.g., ranking cases by predicted probability)                                       |
| <b>Calibration (if model outputs probabilities)</b> | Brier score, calibration curve                                                                                                                                                      | Assess how closely predicted probabilities match observed event rates                                                               | Probability-based risk estimation (e.g., calibrated diagnosis or event probabilities)                               |
| <b>Clinical utility</b>                             | Decision curve analysis (DCA)                                                                                                                                                       | Plots net benefit of a decision against threshold, a minimum probability of a disease where a decision-maker chooses to take action | Decision-support tools (e.g., biopsy)                                                                               |
| <b>Correlation</b>                                  | Pearson correlation, Spearman correlation                                                                                                                                           | Measure linear or monotonic association between predicted and reference values                                                      | Score ranking and trend agreement (e.g., risk-score ordering vs. reference)                                         |
| <b>Agreement</b>                                    | Intraclass correlation coefficient (ICC), concordance correlation coefficient (CCC)                                                                                                 | Assess whether predicted and reference values coincide on the same scale                                                            | When predicted numeric values should agree closely with a reference value (e.g., ejection fraction, tumor diameter) |

|              |                                                          |                                            |                                                                                            |
|--------------|----------------------------------------------------------|--------------------------------------------|--------------------------------------------------------------------------------------------|
| <b>Error</b> | Root mean square error (RMSE), mean absolute error (MAE) | RMSE/MAE give error in the outcome's units | Numeric extraction and measurement (e.g., lab values, dosages, size/volume quantification) |
|--------------|----------------------------------------------------------|--------------------------------------------|--------------------------------------------------------------------------------------------|

---

Table 3: Capability evaluation methods

| <b>Evaluation Category</b>               | <b>Representative approaches</b>                                                                      | <b>Purpose</b>                                                                                                                       | <b>Example Application (Healthcare LLM)</b>                                                           |
|------------------------------------------|-------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------|
| Multiple-choice and knowledge benchmarks | MedQA and similar exam-style question sets                                                            | Fixed, predetermined-protocol scoring of factual/clinical knowledge                                                                  | Baseline screening of candidate models                                                                |
| Multi-task and agentic benchmarks        | MedHELM, MIRA, HealthAgentBench, Physician Bench                                                      | Standardized, reproducible assessment across many clinical tasks/domains; agentic benchmarks add evaluation of planning and tool use | Benchmarking overall suitability before deployment; assessing an agent navigating a clinical workflow |
| Multi-turn benchmarks                    | LLM-as-a-judge scoring of fixed conversation sets; patient-actor or LLM-agent simulated dialogues     | Assess performance across a conversational sequence rather than a single exchange                                                    | Chatbot-style patient communication tools                                                             |
| Reasoning benchmarks                     | MedThink-Bench, MedR-Bench, Open-XDDx, ER-Reason; script concordance testing; interpretability probes | Evaluate the process behind an answer, test whether reasoning appropriately updates with new information                             | Distinguishing sound clinical reasoning from keyword-anchored guessing                                |
| Operational metrics                      | Token usage, cost per query, latency, uptime/error rate; temperature/Top-P as tunable parameters      | Assess efficiency, cost, and feasibility of deployment at scale                                                                      | Budgeting and integrating a tool into real-time clinical workflow                                     |

Table 4: Clinical context evaluation methods

| <b>Evaluation Category</b> | <b>Representative approaches</b>                                                                                                        | <b>Purpose</b>                                                                                                                        | <b>Example Application (Healthcare LLM)</b>                             |
|----------------------------|-----------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------|
| NLP metrics                | N-gram overlap (BLEU, ROUGE, METEOR); embedding-based (BERTScore, cosine similarity); readability (Flesch-Kincaid, Flesch Reading Ease) | Quantify surface overlap or semantic similarity against a reference                                                                   | Scoring patient education materials                                     |
| Human review               | Structured rubrics (e.g., QUEST)                                                                                                        | Structured, reproducible expert assessment of clinical safety, privacy, ethical alignment; enables measurable inter-rater reliability | Clinician panel review of LLM-generated summaries                       |
| Domain-specific frameworks | SHARP (safety, helpfulness, accuracy, relevance, personalization)                                                                       | Principle-based criteria tailored to a specific context (e.g., consumer health/wellness)                                              | Evaluating a wearable's AI-generated health recommendations             |
| LLM-as-a-judge             | Fixed evaluator model against defined criteria                                                                                          | Scale rubric-based/qualitative assessment beyond manual review capacity                                                               | Assessing adherence to a communication rubric at scale                  |
| Faithfulness metrics       | Groundedness checks, hallucination rate, atomic content unit comparison                                                                 | Assess whether generated text is supported by its source                                                                              | Evaluating if clinical note summarization is true to the clinical chart |