ASSESSABLE BUT UNASKED: EVIDENCE-GROUNDED CLINICAL REVIEW FOR MEDICAL AI PAPERS

Anonymous authors
Paper under double-blind review

ABSTRACT

The clinical dimensions that a medical AI (MedAI) paper makes assessable do not fully overlap with those raised as substantive concerns or requests in its peer review. A nine-dimensional diagnostic analysis shows that clinical factors, especially clinical safety, are less frequently made explicit in written review and are not fully represented by a single overall-quality axis.

We introduce Evidence-Grounded Clinical Review (EGCR), a claim-level review system that converts clinical claims into evidence questions, retrieves submission-time external evidence, assesses source applicability, and aligns external requirements with manuscript evidence to produce structured, inspectable findings.

We evaluate EGCR on 168 MedAI papers from ICLR and Nature-series journals. For 107 papers, EGCR identifies at least one clinical dimension that it marks as assessable but that is not raised as a substantive concern or request in the held-out human review. Human-review dimension coverage is 73.38%, while the union coverage of human review and EGCR is 89.80%.

A traceability audit shows 97.1% manuscript-evidence recoverability, complete preservation of external provenance chains, and 97.5% submission-time source availability. Paired case analysis against direct review further illustrates how evidence grounding can produce more specific validation questions and better align evidence requests with the paper's stated claim scope.

EGCR is not an accept/reject predictor; it provides an inspectable clinical evidence layer that makes the basis of review findings easier to trace and examine.

1 INTRODUCTION

Medical artificial intelligence (MedAI) has advanced rapidly over the past decade, and its evaluation is no longer limited to benchmark accuracy. Deep learning systems have reached or approached expert-level performance in tasks such as skin lesion classification, diabetic retinopathy grading, retinal referral, and breast cancer screening (Esteva et al., 2017; Gulshan et al., 2016; De Fauw et al., 2018; McKinney et al., 2020). More recent work has extended to medical question answering, multimodal generalist architectures, and diagnostic dialogue systems (Singhal et al., 2023; Moor et al., 2023; Tu et al., 2025).

At the same time, an increasing number of MedAI studies are appearing in general machine learning conferences. We examined accepted papers from ICLR and ICML between 2024 and 2026 and found a rapid increase in clinical MedAI papers: the combined number of papers from the two venues increased from 61 in 2024 to 119 in 2025 and further to 278 in 2026 (Figure 1). Because we used conservative screening criteria and counted only accepted papers rather than all submissions, these counts provide a conservative indicator of the scale of clinical MedAI work entering general ML review pipelines.

MedAI paper volume increased substantially over this period, while incomplete clinical coverage remained in the sampled reviews. Applying the observed coverage rates to the accepted-paper counts yields illustrative uncovered review-equivalent estimates of 21.4, 28.6, and 63.9 for 2024–2026, respectively. Figure 1 also previews where this missing coverage lies: clinical dimensions such as clinical significance and especially clinical safety are made explicit much less often than common

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

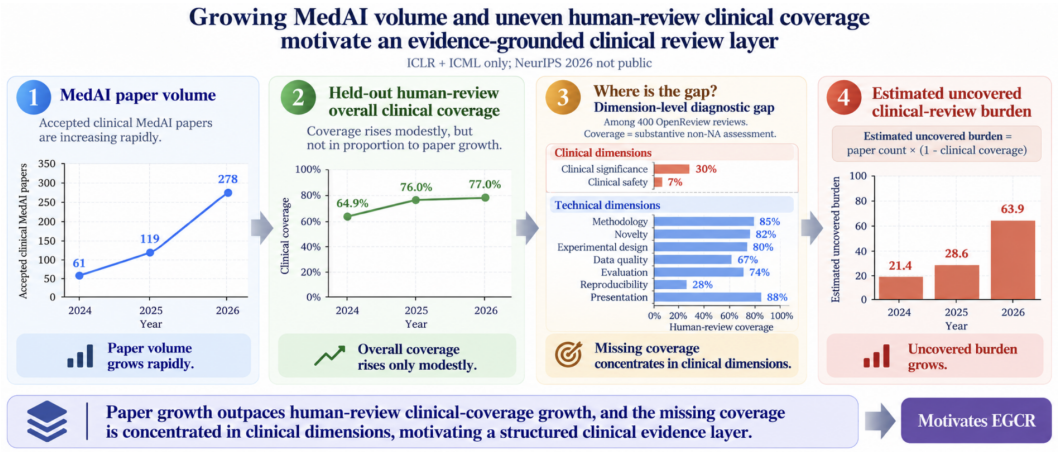


Figure 1: Growth in clinical MedAI paper volume and uneven written clinical-review coverage. The figure combines accepted-paper counts, observed held-out review coverage, and an illustrative uncovered review-equivalent estimate. Dimension-level coverage is lowest for clinical significance and clinical safety.

technical dimensions in written human reviews. As a result, an increasing number of clinical claims are being evaluated within review processes designed primarily for machine learning contributions.

This distinction matters because strong benchmark performance alone is not sufficient to establish a clinical claim. Claims about disease detection, risk stratification, patient triage, or treatment support also depend on whether the clinical problem and decision point are clearly defined; whether labels and endpoints are clinically meaningful; whether the evaluation cohort matches the target population; whether the validation design supports the strength of the claim; whether reported metrics can be interpreted at clinically meaningful operating points; and whether harms, subgroup disparities, and deployment conditions are considered (Kelly et al., 2019; Vollmer et al., 2020). These concerns are not hypothetical. Documented examples include racial bias introduced by proxy outcomes (Obermeyer et al., 2019) and performance degradation under dataset shift across institutions and over time (Subbaswamy & Saria, 2020; Finlayson et al., 2021).

Medicine has already developed systematic evidence frameworks for these issues, including governance frameworks for safety, effectiveness, equity, and lifecycle monitoring, as well as reporting and appraisal standards for different types of clinical claims (World Health Organization, 2021; U.S. Food and Drug Administration, 2021; Liu et al., 2020; Vasey et al., 2022; Collins et al., 2024; Mongan et al., 2020). In general machine learning conferences, however, this clinical evidence layer cannot be consistently guaranteed.

Our diagnostic analysis in Section 2 further shows how this issue appears in written reviews. Among 400 MedAI OpenReview reviews, clinical significance was substantively discussed in only 30%, and clinical safety in only 7%. In a separate blinded paper-level audit of 100 ICLR MedAI papers, clinical safety also received the lowest score among the nine dimensions, and the score structure suggests that the clinical dimensions retain information not fully represented by a single overall-quality axis.

Automated reviewing offers a possible path toward scalability. Existing large language model-based systems can generate fluent paper-level critiques and are commonly evaluated by similarity to human reviews or by author-rated helpfulness (Kang et al., 2018; Liang et al., 2023; Lu et al., 2024; D'Arcy et al., 2024). However, these systems are generally not designed around clinical evidence questions. They typically review the paper as a whole rather than individual claims, rely more heavily on parametric knowledge than on claim-specific reporting standards, guidelines, and reference-standard definitions, and rarely model which external evidence was available at the time of submission. As a result, "the manuscript does not support this claim" and "no applicable standard exists" may be compressed into similar critical comments. What is missing is not more review text, but a mechanism that connects each clinical claim to inspectable evidence.

We therefore introduce Evidence-Grounded Clinical Review (EGCR), which treats the clinical claim as the basic unit of review. EGCR extracts clinical claims from a paper, converts them into evidence questions based on a seven-dimensional clinical rubric, retrieves clinical standards and external evidence that were publicly available at the time of submission, assesses the applicability of each source to the claim, aligns external evidence with manuscript evidence, and produces structured findings. Throughout this process, EGCR preserves the evidence path underlying each final finding, including manuscript evidence when available, external requirements, applicability judgments, and explicit evidence-gap states.

This design allows EGCR to test a direct proposition: some clinical dimensions that the system identifies as assessable from a manuscript are not raised as substantive concerns or requests in its written peer review. If every dimension assessable by EGCR already appeared in reviewer concerns or requests, there would be less room for dimension-level complementarity. Among 168 papers, EGCR identified 107 with at least one clinical dimension that it marked assessable from manuscript content and submission-time evidence but not raised as a substantive concern or request in the held-out review. We use the term "assessable but unasked" carefully: it means an inspectable EGCR judgment did not appear in the written review record, not that a reviewer missed an error.

Our main contributions are:

1. **We identify and quantify a gap between system-identified assessable clinical dimensions and written peer review coverage.** Across 168 ICLR and Nature-series papers, EGCR finds substantial non-overlap between dimensions it marks assessable and those raised as substantive concerns or requests in written reviews.
2. **We introduce a system that makes this gap measurable.** EGCR decomposes papers into clinical claims, evidence questions, applicability decisions, and evidence-grounded findings, while separating insufficient manuscript evidence from cases in which no applicable external standard was identified.
3. **We preserve a recoverable evidence path.** Findings retain their supporting manuscript evidence when available, together with external requirements, applicability judgments, and explicit evidence-gap states.
4. **We provide a controlled paired case analysis against direct review.** Using the same model, rubric, and output budget, we examine how evidence grounding changes the clinical questions surfaced to reviewers.

2 Clinical Evidence Gaps in Medical AI and Related Work

Medical AI papers are still machine learning papers, and their reviews therefore need to evaluate technical dimensions such as novelty, methodological soundness, experimental design, evaluation quality, reproducibility, and presentation. At the same time, clinical AI has developed a body of reporting and appraisal frameworks that go beyond benchmark performance. Work on safe and responsible health AI emphasizes risk, harm, and clinical evaluation (Wiens et al., 2019), and international consensus guidelines now extend these concerns to trustworthy and deployable clinical AI (Lekadir et al., 2025). Reporting guidelines for AI clinical trial protocols and reports (Cruz Rivera et al., 2020; Liu et al., 2020) and for the early-stage clinical evaluation of AI decision support systems (Vasey et al., 2022) ask authors to specify intended use, clinical pathway, user interaction, and error analysis, while reporting checklists and guidelines for clinical AI modeling and machine learning predictive models impose minimum information requirements (Norgeot et al., 2020; Luo et al., 2016). Reporting and appraisal tools for prediction models, diagnostic accuracy studies, and medical imaging studies further emphasize validation, outcomes, reference standards, data, and clinical context (Collins et al., 2024; Moons et al., 2019; Bossuyt et al., 2015; Whiting et al., 2011; Mongan et al., 2020).

Figure 1 previews both increasing MedAI paper volume and incomplete written clinical-review coverage, as well as the uneven distribution of that coverage across dimensions. To examine this gap more systematically, we design a nine-dimensional diagnostic audit that asks two questions: which clinical factors are explicitly discussed in written reviews, and whether these factors are already represented by a single overall-quality axis.

To answer these questions, we designed a nine-dimensional diagnostic audit that places clinical significance and clinical safety in the same evaluation frame as seven common technical dimensions: data quality, experimental design, evaluation quality, methodological soundness, novelty, reproducibility, and presentation. This nine-dimensional rubric is used only for diagnosis rather than as the final output space of EGCR. It tests whether clinically relevant dimensions are explicit in existing human reviews and whether clinical significance and clinical safety are already captured by general paper quality. Appendix A provides the full rubric definition and experimental setup, and Appendix Table 2 summarizes the nine diagnostic dimensions.

We first examine how written human-review coverage is distributed across the nine diagnostic dimensions. OpenReview human reviews are mapped to the nine rubric dimensions, and a dimension is counted as covered only when the review text supports a substantive high, mid, or low assessment rather than NA. Across all 400 reviews, clinical significance is explicitly covered in 30% of reviews, while clinical safety is covered in only 7%; in contrast, technical dimensions such as methodological soundness, experimental design, novelty, and presentation appear much more frequently in written reviews (Figure 1). Additional bucket-stratified results are reported in Appendix A.2. Because this analysis observes only the final written review, it does not show that reviewers never considered these issues internally. It shows that clinically relevant factors, especially clinical safety, are not consistently made explicit in the review record. Low written-review coverage alone, however, is not sufficient to establish the need for a separate clinical evidence layer. Another possibility is that clinical significance and clinical safety, even when not discussed separately, are already reflected in general paper quality.

To test this possibility, we conducted an independent nine-dimensional paper-level audit of 100 ICLR 2024–2025 MedAI papers. For each paper, the reviewer assigned a 1–5 score and a short critique to each of the nine dimensions based on the full text and main figures. During scoring, the true accept/reject decision, human ratings, and human reviews were withheld to avoid influencing the nine-dimensional assessments. Detailed score distributions are reported in Appendix A.3.2.

We then used PCA to test whether the nine dimensions could be explained mainly by a single overall-quality factor. If clinical significance and clinical safety were simply another expression of general paper quality, the nine-dimensional scores should largely collapse onto a common latent direction.

This was not the case. The first principal component explains only 46.7% of the variance, and four principal components are required to reach 80% cumulative explained variance. Further loading and correlation analyses show that clinical significance and clinical safety are not fully absorbed by the main technical-execution direction. Thus, these clinical dimensions are related to technical quality, but retain information not fully represented by a single overall-quality axis. The PCA procedure, scree plot, loading interpretation, and correlation check are reported in Appendix A.3.3.

As a result, these diagnostic analyses further explain the aggregate gap shown in Figure 1. Human-review coverage analysis shows that clinically relevant dimensions are not consistently made explicit in written review records, and that this missing coverage is highly uneven across dimensions. PCA further suggests that these clinical dimensions retain information not fully represented by a single overall-quality axis. These results motivate a complementary clinical evidence layer that directly examines the clinical claims made by a paper and the evidence supporting them.

3 Evidence-Grounded Clinical Review Architecture

The diagnostic analysis in Section 2 shows that clinically relevant factors are not consistently covered in written reviews and are not fully captured by a single overall-quality axis. We therefore formulate clinical review for MedAI papers as a claim-level evidence review problem. For each clinical claim, the review needs to determine what evidence is required, what support the manuscript provides, and whether the resulting judgment can be traced back to the manuscript and external evidence. This formulation follows the emphasis in clinical reporting and appraisal tools on intended use, population, reference standards, validation, applicability, and risk of bias (Vasey et al., 2022; Collins et al., 2024; Moons et al., 2019), and is also consistent with evidence-centered work on scientific claim verification and rationale preservation (Wadden et al., 2020; DeYoung et al., 2020).

The nine-dimensional space in Section 2 contains two clinical dimensions and seven technical dimensions. It is used mainly to diagnose whether clinical factors are already captured by conven-

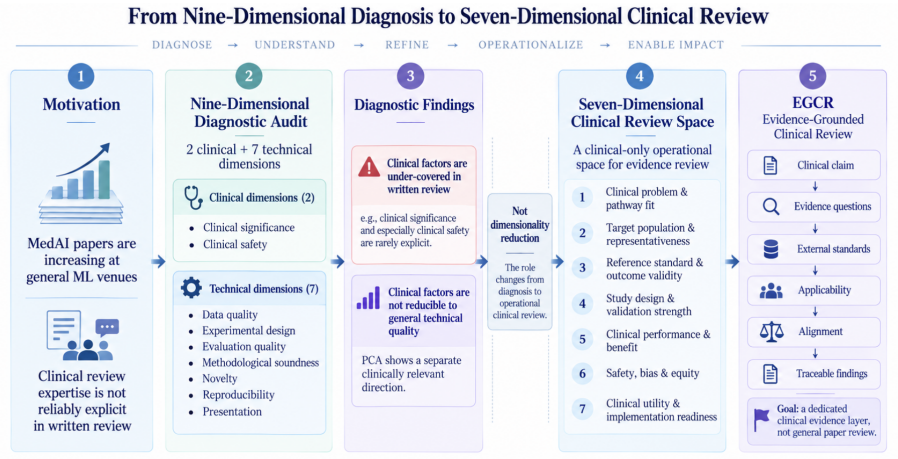


Figure 2: From the nine-dimensional diagnostic audit to the seven-dimensional clinical review space and EGCR. The nine-dimensional space is used to identify gaps in clinical evidence review, while the seven-dimensional space turns these diagnostic findings into claim-level clinical evidence review.

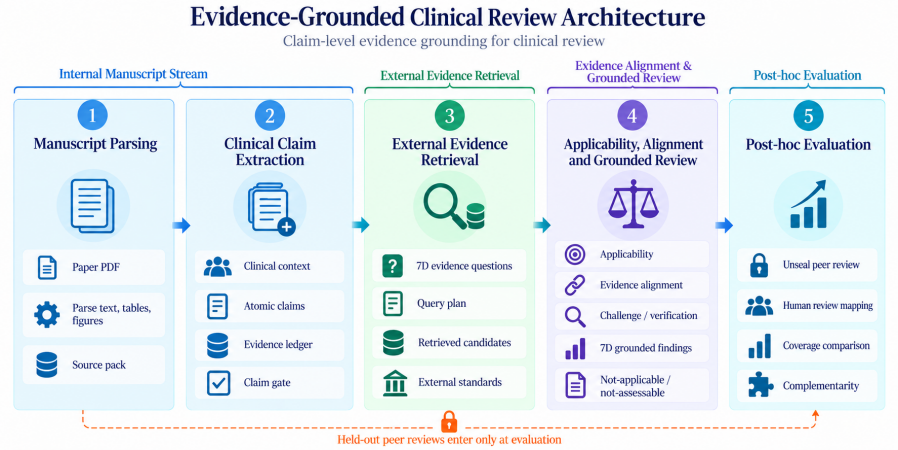


Figure 3: EGCR architecture. Clinical claims are reviewed through internal and external evidence streams, with applicability assessment and evidence alignment producing traceable findings.

tional technical review. The results show that clinical factors have distinct evidence requirements. EGCR therefore does not directly reuse this nine-dimensional space. Instead, we define a seven-dimensional operational space for clinical evidence review. This is not dimensionality reduction in the mathematical sense, but a shift from diagnosing gaps in existing review to carrying out clinical evidence review. As shown in Figure 2, the seven dimensions are clinical problem and pathway fit, target population and representativeness, reference standard and outcome validity, study design and validation strength, clinical performance and benefit, safety, bias, and equity, and clinical utility and implementation readiness.

Based on this seven-dimensional clinical review space, we build Evidence-Grounded Clinical Review (EGCR). As shown in Figure 3, EGCR treats the clinical claim as the basic unit of review and operates through an internal evidence stream and an external evidence stream. The internal stream identifies the clinical claims made by the manuscript and the support provided for them. The external stream generates evidence questions for each claim across the seven clinical dimensions and retrieves relevant clinical standards, guidelines, and other external evidence. The system then assesses source applicability and aligns external requirements with manuscript evidence to produce structured and traceable review findings. Separating retrieval from applicability is important because reporting and risk-of-bias tools distinguish the presence of a source from its applicability to the population, setting, outcome, and study design of a particular claim (Moons et al., 2019; Wolff et al., 2019; Whiting et al., 2011).

All external evidence is subject to a submission-time availability check. Sources with reliable dates are used only when they were publicly available by the paper-level cutoff; sources whose availability dates cannot be determined reliably are retained as undetermined rather than treated as verified available. Retrieved sources are not treated as valid evidence by default. Their applicability to the specific claim is assessed before they are used in the review. When a reliable judgment cannot be made, the system records the case as an evidence gap, not applicable, or not assessable rather than forcing a definite conclusion. Each final finding retains the corresponding clinical claim, supporting manuscript evidence when available, external evidence, and their alignment. Human peer reviews remain held out throughout generation and are introduced only after EGCR has completed its review, for subsequent comparison and complementarity analysis.

4 EXPERIMENTS

We evaluate EGCR on 168 MedAI papers, including 135 ICLR papers from 2024–2026 and 33 Nature-series journal papers. The ICLR set is screened from conference papers with clinical MedAI tasks, while the journal set is smaller because MedAI papers with available review material are much less common. All papers are used for structured clinical evidence analysis, evidence traceability evaluation, and complementarity analysis with held-out human reviews. Direct-review comparison is limited to paired case analysis. We further examine EGCR's results across the seven clinical dimensions, how evidence grounding changes the questions raised during review, and how human review coverage of different clinical dimensions changes across years.

4.1 STRUCTURED CLINICAL REVIEW ARTIFACTS

We first examine whether EGCR can convert a full paper into independently reviewable clinical claims while preserving the clinical context associated with those claims. The system consistently extracts a substantial number of clinical claims from both publication settings, allowing downstream review to operate at the claim level rather than as a single paper-level comment. At the same time, as shown in Figure 4a, Nature-series papers contain richer clinical context overall, while ICLR papers more often lack information about the target population, care or acquisition setting, decision or action, and intended clinical use. This pattern is consistent with the two publication settings: medical journals typically describe clinical populations, settings, and intended use more explicitly, whereas general machine learning conferences place greater emphasis on methods and technical results. Because missing context may also reflect extraction errors, we treat this comparison as descriptive rather than as a direct judgment of paper quality. The anchor categories mirror the kinds of clinical context requested by early evaluation, imaging, and prediction-model reporting guidance (Vasey et al., 2022; Mongan et al., 2020).

Figure 4b further summarizes EGCR's structured evidence-processing pipeline across the 168-paper corpus. The value at each stage is the mean number of artifacts produced per paper, rather than a fixed generation ratio between adjacent stages. Starting from accepted clinical claims, EGCR generates evidence questions and expands them into query plans and retrieved candidates. This early expansion reflects the fact that a clinical issue may require multiple searches and multiple external sources. The system then extracts relevant clinical standards from the retrieved candidates, assesses their applicability, and aligns them with manuscript evidence to produce structured review findings. The overall pattern is therefore one of expansion followed by filtering. Rather than generating a free-form review directly from the full paper, EGCR builds a sequence of stored and inspectable intermediate artifacts that connect clinical claims to external evidence and final review findings.

Across the two publication settings, Nature-series journal papers produce more clinical review artifacts per paper at nearly every stage than ICLR conference papers. This is consistent with the clinical-context analysis above: journal papers generally contain richer clinical content and more complete clinical context, and are associated with more downstream review artifacts during evidence-question generation, external retrieval, and evidence alignment.

We further audit whether EGCR preserves an inspectable evidence path. As shown in Table 1, manuscript evidence is recoverable for 97.1% of recorded spans, all final findings retain complete external provenance chains, and 97.5% of external source uses are verified as available by the submission-time cutoff; the remaining 2.5% have undetermined source dates.

Table 1: Traceability audit across the 168-paper EGCR corpus.

Audit check	Passed / Total	Rate
Manuscript evidence recoverability	3,164 / 3,260	97.1%
External evidence provenance	2,961 / 2,961	100.0%
Sources available by submission	2,887 / 2,961	97.5%
Source availability undetermined	74 / 2,961	2.5%



(a) Clinical-context anchors.

(b) Evidence-processing pipeline.

Figure 4: Panel a shows the proportion of Nature-series journal papers and ICLR conference papers in which each clinical-context anchor is identified. Panel b reports the mean number of artifacts produced per paper at each stage, showing the expansion-and-filtering process from clinical claims to structured findings.

4.2 Seven-Dimensional Clinical Evidence Scores and Assessability

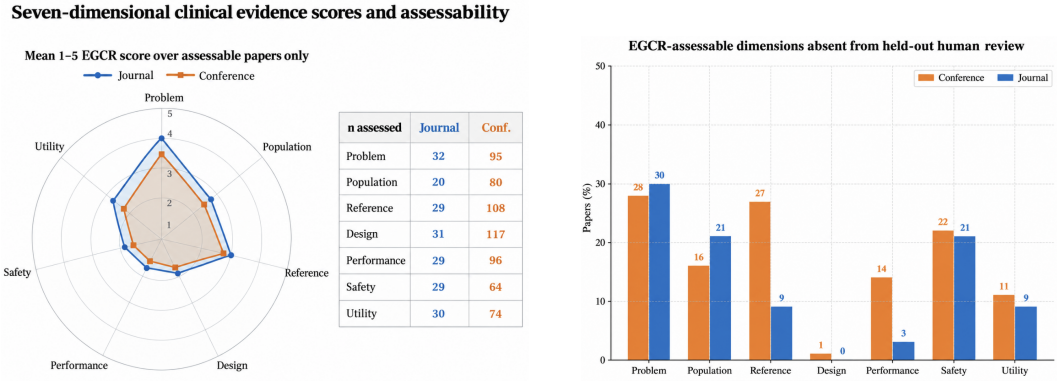
We further examine assessability and evidence strength across the seven clinical dimensions. EGCR does not assign scores to every dimension for every paper. Instead, it first determines whether a dimension contains enough information to be assessed; only assessable dimensions receive a score from 1 to 5. The score is an ordinal evidence-strength anchor: 1 indicates minimal support, 3 indicates partial but clinically incomplete support, and 5 indicates strong support for the stated claim and evidence stage rather than proof of deployment readiness. Figure 5a therefore reports both the mean score for each dimension and the number of papers that contribute to that score, allowing us to distinguish between two cases: evidence is available but weak, and information is insufficient for a reliable assessment. This pattern is consistent with the earlier clinical-context anchor analysis. Nature-series papers generally provide more complete information on target populations, clinical settings, decision processes, and intended use, so more clinical dimensions can be assessed. Among dimensions that EGCR marked as assessable, the journal subset shows higher or similar mean evidence-support scores. In contrast, ICLR papers more often lack sufficient information in some dimensions, especially safety and clinical utility and implementation.

The score profiles also show a similar pattern across the two publication settings. Clinical problem and pathway fit tends to receive stronger support, while clinical performance, safety, and clinical utility and implementation receive lower scores. This suggests that having enough information to assess a dimension does not necessarily mean that the paper provides strong clinical evidence for it, especially for dimensions closer to real-world clinical use. Figure 5a therefore separates information completeness from evidence strength and avoids treating missing information as weak evidence, a distinction also reflected in appraisal tools that separate reporting completeness, applicability, and risk of bias (Moons et al., 2019; Wolff et al., 2019; Whiting et al., 2011).

4.3 Complementarity with Held-Out Human Review

The previous section distinguishes **assessability** from **evidence strength**: insufficient information is recorded as an evidence gap or not-assessable state rather than as a low score.

Based on this setting, we further examine the complementarity between EGCR and held-out human review across the seven clinical dimensions. Human reviews remain isolated throughout EGCR generation and are mapped to the same seven-dimensional clinical space only after the system has completed its review. Figure 5b shows dimensions for which EGCR identified an inspectable evidence-



(a) Scores and assessability.

(b) Assessable-but-unasked dimensions.

Figure 5: Seven-dimensional EGCR evidence profiles and complementarity with held-out human review. Panel a reports mean 1–5 scores over assessable papers only, with counts shown alongside; non-assessable dimensions are not treated as low scores. Panel b shows the proportion of papers for which a dimension is EGCR-assessable but not raised as a substantive concern or request in held-out human review.

based judgment, while the held-out human review contains no corresponding substantive concern or request, denoted as $E \setminus H$. Overall, 107 of 168 papers (63.7%) contain at least one clinical dimension that EGCR marks as assessable but that is not raised as a substantive concern or request in the held-out human review. For conference papers, this non-overlap is concentrated mainly in clinical problem definition, reference standards, and safety. For journal papers, it appears more often in clinical problem definition, target population, safety, and clinical utility. In contrast, study design is close to zero in both settings, indicating that when this dimension is assessable by EGCR, it usually already appears as a substantive concern or request in held-out human review. These results suggest that EGCR does not simply add review content across all dimensions, but provides more complementary information in clinical dimensions that are less consistently raised as substantive concerns or requests in written human reviews.

Assessability in this analysis is determined by the EGCR pipeline rather than by an independent clinician gold standard. The non-overlap result should therefore be read as system-identified review complementarity, not as verified human-review omission.

4.4 Evidence-Grounded Review versus Direct Review

In addition to the complementarity analysis with held-out human review, we compare EGCR with direct clinical review in a controlled paired case analysis. EGCR and direct clinical review use the same model, the same seven-dimensional clinical review space, and the same output limit. The direct reviewer reads the manuscript text directly, whereas EGCR uses a structured evidence packet containing clinical claims, manuscript evidence, external standards, applicability judgments, and evidence alignments. We use this comparison as a qualitative mechanism analysis rather than an aggregate quality metric, in contrast to automated-review work that typically evaluates generated reviews at the paper level (Kang et al., 2018; Liang et al., 2023; D'Arcy et al., 2024). We selected two cases to illustrate pre-specified mechanisms: question specificity and claim-scope calibration. Appendix Table 6 provides the paired examples. The examples are not used to estimate average review quality.

These cases illustrate two ways evidence grounding can change clinical review feedback. It can surface a more specific validation question when external requirements apply to a paper's label or outcome construction, and it can separate what is required for the paper's current claim from what would be required for a stronger deployment or clinical-readiness claim.

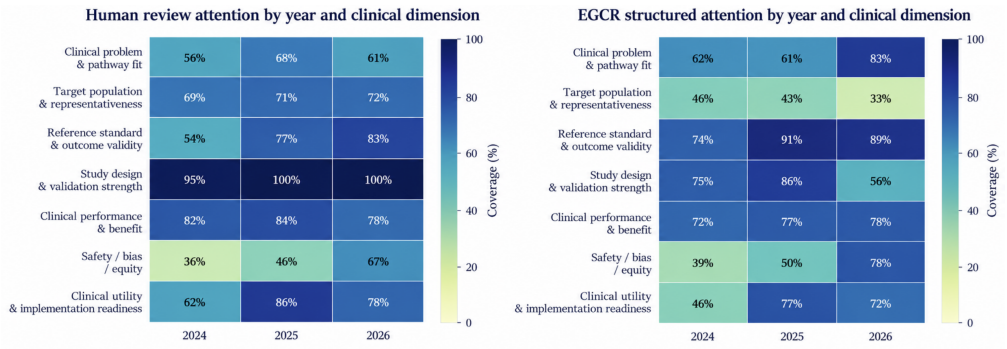


Figure 6: Dimension-level coverage of held-out human review and EGCR from 2024 to 2026. The heatmaps show how coverage across the seven clinical dimensions varies by publication year.

4.5 Clinical Review Coverage Across Years

Finally, we compare dimension-level coverage in human review and EGCR across publication years. Figure 6 shows coverage patterns of human review and EGCR from 2024 to 2026. Overall, the two approaches partially overlap, but their coverage profiles are not identical. Study design is consistently one of the most strongly covered dimensions in human review, while EGCR shows lower coverage of this dimension. In contrast, reference standards and safety-related evidence receive higher coverage in EGCR. Clinical performance remains highly covered in both approaches, suggesting that this dimension is a common focus of both human review and EGCR.

These differences remain visible across years. Human review consistently shows high coverage of study design, while EGCR shows relatively higher coverage of reference standards and safety. Descriptively, safety- and reference-standard-related coverage is higher in later-year samples. Overall, EGCR does not simply reproduce the coverage pattern of human reviewers; it forms a complementary evidence-grounded profile within the same seven-dimensional clinical space.

5 Conclusion

We introduce EGCR, an evidence-grounded clinical review system for MedAI papers that links claim-level manuscript evidence to applicable external clinical evidence through applicability assessment and evidence alignment. Experiments on ICLR and Nature-series papers show that EGCR forms inspectable judgments from available clinical evidence, identifies missing information, and surfaces complementary clinical dimensions relative to held-out human review. Paired case analysis further illustrates how evidence grounding can make review questions more specific and better aligned with the paper's claim scope. This complementarity measures EGCR-identified assessability rather than an independent clinician gold standard, so it should not be read as verified human-review omission.

These results support treating clinical review as claim-specific evidence assessment rather than a single aggregate score. Different clinical claims require different forms of evidence regarding population, reference standards, validation, safety, and implementation, and these requirements and their supporting evidence need to remain explicit and inspectable. EGCR is therefore not intended to replace human peer review or to automate accept/reject decisions. Instead, it provides an inspectable clinical evidence layer that helps reviewers see more clearly what clinical evidence a paper already provides, what evidence is still missing, and how the final review judgments are formed.

Ethics Statement

Clinical review assistance can surface safety, equity, population-mismatch, and reference-standard issues earlier in the publication process. It can also be misused if treated as an automatic decision mechanism. EGCR is therefore designed to preserve claim, source, and audit links, to mark unavailable-source and not-assessable states explicitly, and to leave final interpretation to human reviewers.

Reproducibility Statement

We provide the main system design, evaluation protocols, source-handling rules, and model and retrieval configuration in the appendix. EGCR also preserves structured intermediate artifacts linking claims, manuscript evidence, external evidence, applicability judgments, and final findings.

Exact reproduction may depend on provider-side model versions and low-level runtime configurations. The submission-time cutoff constrains retrieved external evidence but cannot control information already encoded in model pretraining.

AI Use Statement

Generative AI tools were used for language editing and proofreading during manuscript preparation, including improving grammar, clarity, and phrasing. Separately, the research system evaluated in this paper uses the language model described in Appendix C; that use is part of the reported methodology rather than writing assistance. AI-assisted writing tools were not used to determine the scientific conclusions of this work. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content of the paper.

REFERENCES

Patrick M. Bossuyt, Johannes B. Reitsma, David E. Bruns, Constantine A. Gatsonis, Paul P. Glasziou, Les Irwig, Jeroen G. Lijmer, David Moher, Drummond Rennie, Henrica C. W. de Vet, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ*, 351:h5527, 2015. doi: 10.1136/bmj.h5527.

Gary S. Collins, Johannes B. Reitsma, Douglas G. Altman, and Karel G. M. Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*, 350:g7594, 2015. doi: 10.1136/bmj.g7594.

Gary S. Collins, Karel G. M. Moons, Paula Dhiman, Richard D. Riley, Andrew L. Beam, Ben Van Calster, Marzyeh Ghassemi, Xiaoxuan Liu, Johannes B. Reitsma, and Maarten van Smeden. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385:e078378, 2024. doi: 10.1136/bmj-2023-078378.

Samantha Cruz Rivera, Xiaoxuan Liu, An-Wen Chan, Alastair K. Denniston, and Melanie J. Calvert. Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. *Nature Medicine*, 26(9):1351–1363, 2020. doi: 10.1038/s41591-020-1037-7.

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. MARG: Multi-agent review generation for scientific papers. *arXiv preprint arXiv:2401.04259*, 2024. doi: 10.48550/arXiv.2401.04259.

Jeffrey De Fauw, Joseph R. Ledsam, Bernardino Romera-Paredes, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9):1342–1350, 2018. doi: 10.1038/s41591-018-0107-6.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, 2020. doi: 10.18653/v1/2020.acl-main.408.

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017. doi: 10.1038/nature21056.

Samuel G. Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S. Kohane, and Suchi Saria. The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3):283–286, 2021. doi: 10.1056/NEJMc2104626.

Varun Gulshan, Lily Peng, Marc Coram, Martin C. Stumpe, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402–2410, 2016. doi: 10.1001/jama.2016.17216.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, 2018. URL https://arxiv.org/abs/1804.09635.

Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17:195, 2019. doi: 10.1186/s12916-019-1426-2.

Karim Lekadir, Alejandro F. Frangi, Antonio R. Porras, Ben Glocker, Celia Cintas, Curtis P. Langlotz, Eva Weicken, Folkert W. Asselbergs, Fred Prior, Gary S. Collins, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*, 388:e081554, 2025. doi: 10.1136/bmj-2024-081554.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, et al. Can large language models provide useful feedback on research papers? a large-scale empirical analysis, 2023.

Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie J. Calvert, and Alastair K. Denniston. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nature Medicine*, 26(9):1364–1374, 2020. doi: 10.1038/s41591-020-1034-x.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery, 2024.

Wei Luo, Dinh Phung, Truyen Tran, Sunil Gupta, Santu Rana, Chandan Karmakar, Alistair Shilton, John Yearwood, Nevenka Dimitrova, Tu Bao Ho, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. *Journal of Medical Internet Research*, 18(12):e323, 2016. doi: 10.2196/jmir.5870.

Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, et al. International evaluation of an artificial intelligence system for breast cancer screening. *Nature*, 577(7788):89–94, 2020. doi: 10.1038/s41586-019-1799-6.

John Mongan, Linda Moy, and Charles E. Kahn. Checklist for artificial intelligence in medical imaging (CLAIM): A guide for authors and reviewers. *Radiology: Artificial Intelligence*, 2(2): e200029, 2020. doi: 10.1148/ryai.2020200029.

Karel G. M. Moons, Robert F. Wolff, Richard D. Riley, Penny F. Whiting, Marie Westwood, Gary S. Collins, Johannes B. Reitsma, Jos Kleijnen, and Sue Mallett. PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. *Annals of Internal Medicine*, 170(1):W1–W33, 2019. doi: 10.7326/M18-1377.

Karel G. M. Moons, Johanna A. A. Damen, Tabea Kaul, Lotty Hooft, Constanza Andaur Navarro, Paula Dhiman, Andrew L. Beam, Ben Van Calster, Leo Anthony Celi, Spiros Denaxas, et al. PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*, 388:e082505, 2025. doi: 10.1136/bmj-2024-082505.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, et al. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023. doi: 10.1038/s41586-023-05881-4.

Beau Norgeot, Giorgio Quer, Brett K. Beaulieu-Jones, Ali Torkamani, Raquel Dias, Milena Gianfrancesco, Rima Arnaout, Isaac S. Kohane, Suchi Saria, Eric Topol, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nature Medicine*, 26: 1320–1324, 2020. doi: 10.1038/s41591-020-1041-y.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. doi: 10.1126/science.aax2342.

Mark Sendak, Joshua D'Arcy, Sehj Kashyap, and Michael Gao. A path for translation of machine learning products into healthcare delivery. *EMJ Innovations*, 2020. doi: 10.33590/emjinnov/19-00172.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023. doi: 10.1038/s41586-023-06291-2.

Viknesh Sounderajah, Ahmad Guni, Xiaoxuan Liu, Gary S. Collins, Alan Karthikesalingam, Sheraz R. Markar, Robert M. Golub, Alastair K. Denniston, Shravya Shetty, David Moher, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nature Medicine*, 31:3283–3289, 2025. doi: 10.1038/s41591-025-03953-8.

Adarsh Subbaswamy and Suchi Saria. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics*, 21(2):345–352, 2020. doi: 10.1093/biostatistics/kxz041.

Tao Tu, Mike Schaekermann, Anil Palepu, et al. Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067):442–450, 2025. doi: 10.1038/s41586-025-08866-7.

U.S. Food and Drug Administration. Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. Technical report, U.S. Food and Drug Administration, 2021. URL https://www.fda.gov/media/145022/download.

U.S. Food and Drug Administration, Health Canada, and UK Medicines and Healthcare products Regulatory Agency. Good machine learning practice for medical device development: Guiding principles. Technical report, U.S. Food and Drug Administration, 2021. URL https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles.

Baptiste Vasey, Myura Nagendran, Bruce Campbell, David A. Clifton, Gary S. Collins, Spiros Denaxas, Alastair K. Denniston, Livia Faes, Xiaoxuan Liu, Bilal A. Mateen, Purvi Mathur, and Peter McCulloch. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature Medicine*, 28:924–933, 2022. doi: 10.1038/s41591-022-01772-9.

Sebastian Vollmer, Bilal A. Mateen, Gergo Bohner, Franz J. Kiraly, Rayid Ghani, Pall Jonsson, Sarah Cumbers, Adrian Jonas, K. S. L. McAllister, Puja Myles, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ*, 368:l6927, 2020. doi: 10.1136/bmj.l6927.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 7534–7550, 2020. doi: 10.18653/v1/2020.emnlp-main.609.

Penny F. Whiting, Anne W. S. Rutjes, Marie E. Westwood, Susan Mallett, Jonathan J. Deeks, Johannes B. Reitsma, Mariska M. G. Leeflang, Jonathan A. C. Sterne, and Patrick M. M. Bossuyt. QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine*, 155(8):529–536, 2011. doi: 10.7326/0003-4819-155-8-201110180-00009.

Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X. Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, et al. Do no harm: A roadmap for responsible machine learning for health care. *Nature Medicine*, 25:1337–1340, 2019. doi: 10.1038/s41591-019-0548-6.

Robert F. Wolff, Karel G. M. Moons, Richard D. Riley, Penny F. Whiting, Marie Westwood, Gary S. Collins, Johannes B. Reitsma, Jos Kleijnen, and Sue Mallett. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine*, 170(1):51–58, 2019. doi: 10.7326/M18-1376.

World Health Organization. Ethics and governance of artificial intelligence for health. Technical report, World Health Organization, 2021. URL https://www.who.int/publications/i/item/9789240029200.

Table 2: The nine-dimensional rubric used only for diagnostic motivation. Clinical significance and clinical safety are compared against general technical review dimensions in the same scoring frame.

Dimension	Type	Diagnostic role
Clinical significance	Clinical	Whether the paper's problem, task, or application has a clear clinical rationale.
Clinical safety	Clinical	Whether the paper discusses failure modes, risk mitigation, subgroup harms, or deployment risks.
Data quality	Technical/general	Whether the data source, quality, representativeness, and bias profile support the conclusions.
Experimental design	Technical/general	Whether comparisons, validation settings, and ablations support the empirical claims.
Evaluation quality	Technical/general	Whether metrics, statistical analysis, and error analysis support the conclusions.
Methodological soundness	Technical/general	Whether assumptions, algorithmic design, and technical arguments are sound.
Novelty	Technical/general	Whether the method or problem formulation contributes beyond prior work.
Reproducibility	Technical/general	Whether data, code, hyperparameters, and experimental details are sufficient for reproduction.
Presentation	Technical/general	Whether the writing, structure, figures, and claims are clear.

A Nine-Dimensional Diagnostic Audit

This appendix gives details for the nine-dimensional diagnostic audit summarized in Section 2. The audit answers two questions: whether clinical dimensions are explicit in written human reviews, and whether clinical significance and clinical safety are fully represented by a single overall-quality axis. It is used only for diagnosis. EGCR uses the seven-dimensional clinical review space defined in Appendix B.1.

A.1 Nine-Dimensional Rubric Definition

The rubric places two clinical dimensions next to seven common technical review dimensions. This lets us compare how often each dimension appears in review text and how strongly the dimensions overlap in a paper-level audit.

The nine-dimensional rubric is not the output space of EGCR. It is a diagnostic tool for studying gaps in existing review.

A.2 Are Clinical Dimensions Explicit in Human Review?

A.2.1 Human Review Mapping Protocol

To analyze whether clinical dimensions are already covered in ordinary review, we mapped OpenReview human review text to the nine rubric dimensions. For each review, a dimension was marked non-NA when it was substantively discussed in the text, and NA otherwise. For a dimension d, coverage is the fraction of reviews in which that dimension is non-NA. This metric measures what reviewers explicitly wrote in the final review text. It does not measure every factor a reviewer may have considered internally.

For dimensions that were discussed, we also coded reviewer stance as high, mid, or low, corresponding to positive, mixed or neutral, and negative discussion. For summary analysis, high, mid, and low were mapped to 3, 2, and 1. Figure 7 shows main-text coverage, and Appendix Figure 8 shows stance across the nine rubric dimensions.

A.2.2 Human Review Coverage Results

Across 400 human reviews, coverage varied strongly by dimension. Traditional technical dimensions such as methodological soundness, experimental design, novelty, and presentation appeared

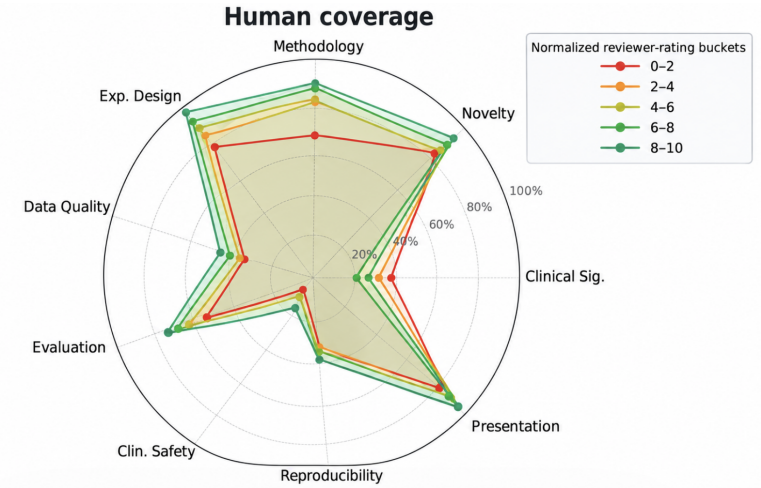


Figure 7: Human-review coverage of the nine diagnostic dimensions across normalized reviewer-rating buckets. Coverage is the fraction of reviews in which a dimension receives a substantive non-NA assessment.



Figure 8: Human-review stance over the nine rubric dimensions, computed among reviews in which the dimension was substantively discussed.

frequently in review text, while explicit coverage of the clinical dimensions was much lower. Clinical significance had 30% coverage, and clinical safety had 7% coverage. Clinical safety was the least-covered of the nine dimensions.

The stance analysis adds another distinction. When clinical significance was mentioned, its average stance was relatively positive, with mean stance 2.39. Clinical safety had mean stance 1.31, meaning that it was not only rarely discussed, but also tended to appear when reviewers raised a weakness or risk. This suggests that the clinical motivation of medical work is often easier for reviewers to recognize, while clinical risks and safety issues are less often made explicit in a systematic way.

We do not interpret absence from review text as proof that reviewers did not consider the issue. The more cautious conclusion supported by this analysis is that clinical significance, and especially clinical safety, do not consistently appear explicitly in standard review text. This is the conclusion used in the main-text human-review analysis.

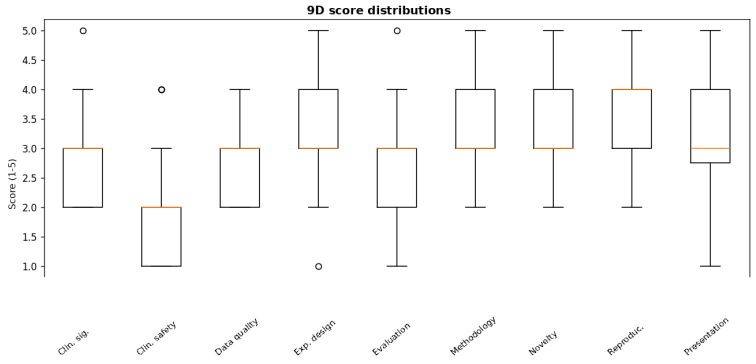


Figure 9: Paper-level score distributions across the nine diagnostic dimensions.

A.3 Are Clinical Dimensions Redundant with General Paper Quality?

Differences in human-review coverage are not enough to show that a separate clinical evidence layer is needed. One possible explanation is that reviewers do not separately discuss clinical significance or clinical safety because these factors are already captured by methodological soundness, experimental design, or overall quality. We therefore conducted a paper-level nine-dimensional audit to examine whether the clinical dimensions are redundant with general paper quality.

A.3.1 Paper-Level Audit Protocol

The paper-level audit used 100 ICLR medical AI papers: 50 from ICLR 2024 and 50 from ICLR 2025. Each year was balanced by decision, with 25 accepted and 25 rejected papers.

For each paper, the reviewer input consisted of parsed full text and main-paper figures. The reviewer assigned a 1–5 score and a short critique for each of the nine dimensions. The reviewer also produced an overall rating, confidence, and recommendation.

To avoid leakage from venue outcomes, the gold accept/reject decision, human rating, and human review text were all withheld during scoring. These fields were joined only after scoring for later analysis. The scoring run was isolated: it used no web access and no external tools. The novelty dimension was interpreted with respect to information available at submission time, so that later work did not affect the evaluation of the submitted paper.

A.3.2 Nine-Dimensional Score Distributions

Figure 9 shows the paper-level score distribution for the nine dimensions. Clinical safety had the lowest mean score, 1.83/5, and 32 of 100 papers received a clinical-safety score of 1/5. Clinical significance had mean score 2.77/5. Technical dimensions such as methodological soundness, experimental design, and reproducibility had higher score distributions.

These scores should be read as manuscript-level evidence judgments. They measure what is explicitly presented in the paper. They do not show that the underlying systems are unsafe in the real world, and they do not evaluate author intent or venue quality. A low clinical-safety score indicates weak explicit reporting of failure modes, risk mitigation, subgroup harms, clinical consequences of errors, or deployment risks.

A.3.3 PCA and Correlation Structure

We next asked whether the nine-dimensional scores are mainly explained by one overall-quality factor. If clinical significance and clinical safety were only noisy proxies for overall paper quality, then the dimensions should largely collapse onto a single latent direction.

The PCA input was a 100×9 matrix, with one row per audited ICLR 2024–2025 paper and one column per diagnostic dimension. Each entry was the 1–5 paper-level score assigned in the blinded audit. All nine dimensions were scored for each paper in this audit, so no score imputation was

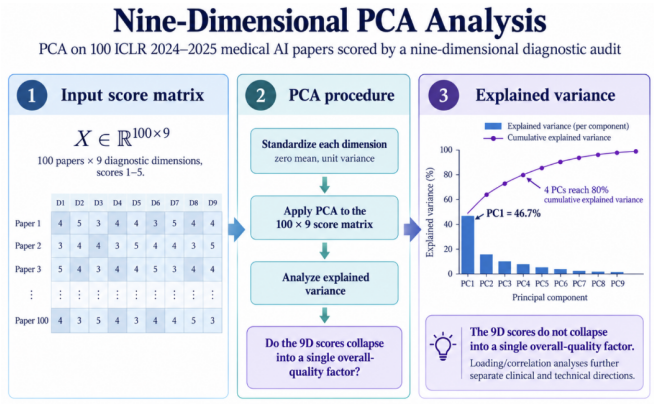


Figure 10: PCA explained variance for the paper-level nine-dimensional diagnostic scores. Scores were standardized across dimensions before PCA; PC1 explains 46.7% of variance, and four components are required to reach 80% cumulative explained variance.

used. Before PCA, each dimension was centered and scaled to unit variance, so the analysis is equivalent to PCA on the dimension correlation matrix rather than on raw-score covariances. We then computed explained variance for the resulting principal components and inspected the loading directions together with the pairwise correlation matrix.

Figure 10 reports PCA explained variance. The first principal component explained 46.7% of the variance, and four principal components were needed to reach 80% cumulative explained variance. Thus, the nine-dimensional scores did not mainly collapse to a single overall-quality axis.

The PCA loadings separated two interpretable directions. One direction was driven mainly by presentation, experimental design, methodological soundness, and evaluation quality, corresponding to technical execution and paper clarity. The other direction was driven mainly by clinical significance and clinical safety, corresponding to a clinically relevant direction. The correlation analysis led to the same interpretation: dimensions were correlated, but not redundant. In particular, the clinical dimensions were not isolated from the technical dimensions, but they also did not behave as duplicate encodings of presentation, experimental design, or methodological soundness.

This analysis does not validate the absolute accuracy of every individual score. It supports a narrower structural conclusion: clinical dimensions are related to the main technical-execution direction, but they are not fully absorbed by it. Clinical significance and clinical safety are not simply repeated encodings of the same good-paper versus bad-paper signal.

B EGCR System Details

The evidence-grounded review system is implemented as a staged pipeline with explicit intermediate artifacts. Each stage writes structured YAML or JSONL outputs, records manifests, and applies schema checks before the next stage consumes the result. The design goal is not to make every LLM call infallible. It is to ensure that errors are localized, inspectable, and auditable.

B.1 Seven-Dimensional Clinical Review Space

EGCR organizes claim-specific evidence review around seven clinical dimensions. The dimensions are a shared review space, not a fixed checklist. The concrete questions depend on the paper's claim, task, population, setting, and intended use. Table 3 defines the seven dimensions and the typical evidence used for each one.

B.2 Paper Intake and Source Pack

The input to a run is a paper PDF, metadata, and a submission or publication date used for cutoff handling. The parser extracts body text, tables, captions, and page-level anchors into a source pack. Blocks receive stable identifiers so that later claims, ledger entries, and review findings can point

Table 3: Seven clinical dimensions used by EGCR.

Dimension	Review question	Typical evidence
Clinical problem and pathway fit	Is the clinical problem and decision point clear?	Clinical task, intended user, workflow role, decision context.
Target population and representativeness	Does the data support the stated target population?	Cohort definition, inclusion criteria, data source, external or temporal validation.
Reference standard and outcome validity	Are labels, endpoints, and outcomes clinically valid?	Diagnostic criteria, pathology, adjudication, follow-up, label protocol.
Study design and validation strength	Does the design support the strength of the claim?	Retrospective or prospective design, comparators, calibration, validation setting.
Clinical performance and benefit	Can performance be interpreted clinically?	Sensitivity, specificity, predictive values, calibration, decision or benefit analysis.
Safety, bias, and equity	Are risks and subgroup harms examined?	Failure modes, subgroup performance, dataset shift, bias, risk mitigation.
Clinical utility and implementation readiness	Are utility or deployment claims supported?	Workflow integration, human–AI interaction, monitoring, prospective use.

back to verbatim manuscript evidence. When parsing omits important material, later stages may mark a claim or dimension as not assessable rather than forcing a judgment from incomplete text.

B.3 Clinical Context and Claim Extraction

The context stage extracts the clinical elements needed to interpret the paper: condition or pathway, task, population, care or acquisition setting, model input and output, intended use, comparator, outcome, reference standard, evidence stage, and submission date. The claim extractor then proposes candidate clinical claims. A claim gate accepts only claims that are clinical or biomedical, locatable in the paper, and sufficiently atomic for evidence review. Claims are repaired or split when a sentence bundles several populations, outcomes, tasks, or intended uses that require different evidence.

Accepted claims are stored with an internal evidence ledger. The ledger records supporting quotes, reported results, limitations, and any manuscript-internal uncertainty. This ledger is the internal evidence stream used by later alignment; EGCR does not rely on a final free-form summary of the paper.

B.4 Evidence Questions and Query Planning

For each accepted claim, the system asks which clinical evidence questions are needed under the seven dimensions. These questions are claim-specific. A reference-standard question for a sepsis prediction claim, for example, is not the same as a reference-standard question for pathology segmentation or an omics biomarker claim. Questions are normalized into query plans that include source role, search target, date constraints, and expected evidence type.

The question stage is also where clinical-anchor insufficiency becomes visible. If the manuscript does not specify a condition, target population, care setting, endpoint, operating point, or intended use, the system may generate fewer answerable questions. This behavior is not treated as silent failure; it is part of the audit trail.

B.5 Source Acquisition, Applicability, and Alignment

Source acquisition retrieves candidate external standards and contextual sources according to the query plans. Retrieved sources are not automatically accepted. The applicability stage determines whether a source is directly applicable to the claim, contextually applicable, indirectly relevant, or not applicable. Only then does the alignment stage compare manuscript evidence with the source.

An alignment records the claim, paper evidence, external source excerpt or standard, applicability decision, and evidence state. Verification and challenge passes check whether the alignment is

supported by the cited material, whether the source role has been misused, and whether the paper evidence actually addresses the standard. Failed or incomplete alignments are repaired where possible; otherwise they remain explicit gaps.

B.6 Seven-Dimensional Review Generation

The final review stage aggregates verified alignments into one panel for each of the seven clinical dimensions. A panel records whether the dimension is assessable, the evidence state, a score when appropriate, the main rationale, and concrete revision requests. Structured findings include severity, affected claim, source links, and recommended revision. The output is therefore a review artifact with claim-source provenance, not a single holistic review essay.

B.7 Source Roles and Cutoff Handling

Source roles prevent the system from treating every retrieved document as if it answered the same kind of question. The main source families are:

- **Reporting standards**: CONSORT-AI, SPIRIT-AI, TRIPOD-AI and TRIPOD+AI, CLAIM, STARD, QUADAS, PROBAST-AI, DECIDE-AI, and related reporting or evaluation checklists (Liu et al., 2020; Cruz Rivera et al., 2020; Collins et al., 2015; 2024; Mongan et al., 2020; Bossuyt et al., 2015; Sounderajah et al., 2025; Whiting et al., 2011; Moons et al., 2019; Wolff et al., 2019; Moons et al., 2025; Vasey et al., 2022).
- **Clinical practice guidelines**: WHO, NICE, KDIGO, ADA, USPSTF, and disease- or specialty-specific clinical practice guidelines where available.
- **Reference-standard and endpoint sources**: diagnostic criteria, pathology or adjudication protocols, outcome definitions, follow-up definitions, and label-construction references.
- **Risk-of-bias and applicability tools**: sources used to evaluate study design, prediction-model bias, population applicability, and validation strength (Moons et al., 2019; Wolff et al., 2019; Moons et al., 2025; Whiting et al., 2011).
- **Regulatory and evaluation frameworks**: guidance relevant to medical device, diagnostic, or model-evaluation claims when applicable (World Health Organization, 2021; U.S. Food and Drug Administration, 2021; U.S. Food and Drug Administration et al., 2021; Wiens et al., 2019; Sendak et al., 2020; Lekadir et al., 2025).
- **Contextual biomedical evidence**: background literature used to contextualize a claim when no directly normative standard exists. Contextual evidence is not allowed to masquerade as a clinical guideline.

All source use is subject to the submission-time availability check. A source published after submission may be useful historically, but it is not used to judge whether the submitted manuscript supported its claim. When dates are unavailable or ambiguous, the source is flagged as undetermined rather than silently treated as verified available.

Applicability is recorded separately from retrieval. A source may be retrieved because it mentions a disease or task yet still be non-applicable because it addresses a different population, care setting, intervention, endpoint, or study type. The four applicability categories used in aggregation are direct, contextual, indirect, and not applicable. This distinction is central to the paper's interpretation: low coverage can reflect a genuine lack of clinical anchors or standards rather than a retrieval miss.

B.8 Prompts, Schemas, and Artifacts

Each stage is prompted to produce a structured artifact rather than an unconstrained critique. The prompts specify the stage contract, allowed fields, evidence requirements, and failure states. Outputs are validated against local schemas before being consumed downstream. Schema repair is allowed only when the repair can preserve the intended evidence and field semantics; otherwise the stage remains failed or not assessable.

The main artifacts are:

- **source pack**: parsed manuscript blocks with stable anchors;
- **clinical context**: condition, population, setting, intended use, model input/output, comparator, label, outcome, and evidence stage;
- **clinical claims**: accepted and rejected candidate claims with gate reasons;
- **claim evidence ledger**: internal paper evidence linked to each accepted claim;
- **evidence questions**: claim-specific questions over the seven-dimensional rubric;
- **query plans**: source-role-aware retrieval plans with date constraints;
- **external standards**: verified candidate standards or contextual sources;
- **applicability records**: direct, contextual, indirect, or not-applicable judgments;
- **alignment bundle**: paper evidence aligned against applicable sources;
- **verified review**: seven dimension panels and structured findings;
- **human-review dimension mapping**: held-out reviewer concerns and requests mapped to the same dimensions for complementarity analysis.

C Data and Experimental Setup

C.1 Venue Survey and Clinician-Author Coding

The venue survey in Figure 1 is used only to motivate the setting: clinical MedAI work has become common enough in general ML venues that clinical evidence review is no longer a niche concern. We screened accepted programs from ICLR and ICML for papers whose primary task, application, dataset, or claimed use case involves health, medicine, biomedicine, clinical care, diagnosis, prognosis, treatment, screening, monitoring, or biomedical measurement. The screen is intentionally conservative: ambiguous papers were excluded unless the title, abstract, or available metadata indicated a medical or biomedical target rather than a generic method with a passing example.

Clinician-author presence was additionally coded from public affiliation and profile evidence (not reported in the main-text figure). A clinical MedAI paper is counted as having clinician participation only when at least one author has clear public evidence of clinical affiliation or practice, such as a hospital, clinical department, medical school clinical appointment, or professional clinical role. Authors whose public affiliation evidence is ambiguous or purely technical are not counted as identifiable clinicians. This should not be interpreted as a credential audit; it is a descriptive signal about visible clinical participation in accepted MedAI papers. Affiliation evidence is not equally available in every venue-year: where camera-ready proceedings were published, we read affiliations from the paper's first page, and where they were not, we fell back to preprint or indexed metadata. The resulting ratio should therefore be read as a lower bound on publicly identifiable clinician participation, not as a complete census of all clinically trained authors.

The survey has three limitations. First, it uses accepted papers rather than submissions, so it does not estimate the full reviewing load. Second, it uses public metadata and therefore under-counts authors whose clinical background is not visible. Third, it does not judge paper quality or clinical adequacy. Those questions are handled by the diagnostic audit and the evidence-grounded review experiments.

C.2 EGCR Evaluation Corpus

The main EGCR experiments use 168 MedAI papers: 135 ICLR papers from 2024–2026 and 33 Nature-series journal papers with available review material. The ICLR subset is drawn from accepted conference papers with clinical MedAI tasks. The journal subset is smaller because papers with both MedAI content and usable peer-review material are less common. We include papers when the paper develops, evaluates, or makes a clinical or biomedical claim about an AI/ML system and when the manuscript and review material needed for the held-out comparison are available.

The corpus is used for structured clinical evidence analysis, traceability checks, and comparison with held-out human review. Direct-review comparison is used for the paired case analysis. Human review text is held out during EGCR generation and is used only after the system has produced its review artifacts.

Table 4: Configuration information for the reported EGCR experiments and direct-review control.

Item	Setting reported in this study
Main model family	Claude Haiku 4.5, invoked through the non-interactive Claude CLI as `claude-haiku-4-5`.
Resolved model version	Not fixed in the exported paper artifacts; CLI aliases may resolve to provider-side dated versions.
Decoding parameters	Temperature and sampling parameters were not explicitly exported in the saved artifacts; provider/CLI defaults were used.
Model-call isolation	One-turn calls with no tools, no session persistence, strict MCP configuration, and safe mode.
Runs per paper	One EGCR pipeline run per paper was used for the reported corpus-level analyses. Failed or incomplete runs were not silently imputed.
Direct-review control	Same model family, same seven-dimensional prompt space, and same output limit as EGCR; the input differs between paper text and EGCR evidence packet.
Retrieval driver	Claim-driven query plans routed to source-specific candidate retrieval and source-acquisition stages.
Retrieval limits	Candidate retrieval and source-acquisition stages used source-specific limits; automatic source acquisition used a limit of 20 candidates per item in the batch runner.
De-duplication	Candidates were de-duplicated by plan, source, canonical URL, passage identity, and related source identifiers before applicability filtering.
Truncation	Long manuscripts and retrieved material were passed through structured source packs and stage artifacts rather than a single full-context review prompt. Exact truncation boundaries were not exported as a single table.
Batch concurrency	Batch runs used sharded execution with per-paper claim workers and a global retrieval-slot limit; these affect throughput rather than the review rubric.

C.3 Direct-Review Control Setting

The direct-review comparison uses the same model, the same seven-dimensional clinical review space, and the same output limit as EGCR. The direct reviewer reads the manuscript text directly. EGCR reads the evidence packet produced by the pipeline, including clinical claims, manuscript evidence, external evidence, applicability decisions, and alignments. This comparison holds the model, review space, and output limit constant while comparing the two workflows.

C.4 Model and Retrieval Configuration

Table 4 summarizes the configuration information used in the reported experiments. Some low-level runtime settings were not exported in a single frozen configuration file; we mark those fields explicitly rather than inferring values after the fact.

D Evaluation Protocols

This appendix gives the operational definitions used in the experiments in Section 4. It covers the clinical review space, clinical anchors, traceability checks, and held-out human-review mapping.

D.1 Evidence States and Score Anchors

EGCR assigns a 1–5 score only when a dimension is assessable from the paper and available evidence. Other states are recorded without a score: not assessable when the claim lacks enough information for a judgment, not applicable when the dimension does not apply to the claim, and evidence gap when the needed evidence is identifiable but missing.

Table 5: Ordinal score anchors used only for dimensions that are assessable. A high score indicates stronger support for the paper's stated claim at its evidence stage; it does not by itself imply clinical deployment readiness.

Score	Anchor
1	Minimal, contradictory, or largely unsupported evidence for the stated clinical claim.
2	Weak support with major missing clinical context, validation, or risk information.
3	Partial support: the dimension can be assessed, but important clinical evidence remains incomplete.
4	Mostly adequate support for the stated claim, with remaining limitations or uncertainty.
5	Strong support for the stated claim and evidence stage, without implying that the system is ready for clinical deployment unless that stronger claim is itself made and supported.

Table 6: Paired qualitative comparison between direct review and EGCR under the same model, rubric, and output budget.

Case	Direct review	EGCR	Mechanism
Cardiac pause detection	Focused mainly on agreement between adjudicators.	Asked whether adjudicators were blinded to model predictions during label reassessment, because reassessment was triggered by disagreement between model predictions and human labels.	More specific validation question with an explicit evidence path.
MRI reconstruction	Extended method-level results toward requirements for clinical readiness.	Kept the current judgment within the paper's stated method-level scope and framed stronger clinical validation as evidence needed for future clinical use.	Better alignment between evidence requests and claim scope.

D.2 Direct-Review Case Analysis

The paired direct-review comparison is used only as a qualitative mechanism analysis. We selected the two cases in Table 6 to illustrate question specificity and claim-scope calibration under the same model, rubric, and output budget.

D.3 Clinical-Anchor Definition

To describe how much context a paper provides for clinical evidence review, we define six clinical anchors. These anchors indicate whether a claim has enough clinical context to support concrete evidence questions.

Condition or pathway. The paper states the disease, clinical state, or clinical pathway under study. A dataset name or broad biomedical domain alone is not counted as a clear condition or pathway anchor.

Target population. The paper states the patient, population, or sample group targeted by the model, including major inclusion or population characteristics that affect clinical interpretation.

Care or acquisition setting. The paper states the clinical setting related to data acquisition or potential use, such as inpatient care, outpatient care, screening, emergency care, specialty clinic, or imaging workflow.

Decision or action. The paper states what clinical decision, classification consequence, intervention decision, or downstream action corresponds to the model output.

Intended clinical use. The paper states the intended clinical role of the model. Research utility, benchmark evaluation, software function, or dataset use alone is not counted as intended clinical use.

Clinical benefit. The paper states the clinically meaningful outcome or benefit the model is expected to improve. An improvement in AUROC, accuracy, or another technical metric alone is not counted as a clinical benefit anchor.

For each anchor, we mark it as present when the manuscript contains an explicit statement or when specific, non-empty clinical information can be reliably identified from the relevant text. For a paper P, the aggregate anchor score is

$$A(P) = \sum_{i=1}^{6} \mathbf{1}(a_i \text{ is present}),$$

with possible values from 0 to 6.

The anchor score describes the amount of manuscript context available for downstream clinical evidence processing. It is not a paper-quality score. A paper with more anchors is not necessarily a better paper; it simply gives the system more explicit clinical information from which to form specific evidence questions.

Association with downstream processing. The aggregate anchor score is associated with the volume of downstream evidence artifacts. In conference papers, higher scores were associated with more evidence questions ($\rho = 0.44$, $p < 0.001$), more query plans ($\rho = 0.51$, $p < 0.001$), more verified external standards ($\rho = 0.49$, $p < 0.001$), and higher EGCR dimension coverage ($\rho = 0.43$, $p < 0.001$); in journal papers the same direction holds for verified standards ($\rho = 0.68$, $p < 0.001$) and dimension coverage ($\rho = 0.35$, $p = 0.05$), with the smaller cohort not reaching significance for questions or query plans.

Because these anchors are extracted by the EGCR pipeline from the manuscript, this analysis cannot fully separate two possibilities: the manuscript may lack clinical detail, or the system may fail to extract information that is present. The associations reported above should therefore be read as internal consistency of the processing chain — the anchors and the downstream artifacts are produced by the same pipeline — rather than as an independent gold-standard measure of manuscript completeness.

D.4 Evidence Traceability Audit

A core design goal of EGCR is to preserve an inspectable evidence path from a clinical claim to a final review finding. We audit this path from three angles: manuscript evidence, external evidence, and submission-time availability.

Manuscript evidence. For each manuscript quotation or evidence span associated with a structured finding, we check whether it can be located again in the corresponding parsed manuscript. To reduce false mismatches from PDF extraction, Unicode encoding, and formatting differences, matching normalizes Unicode representation, whitespace, and non-semantic punctuation differences.

This check asks whether a reviewer can return to the manuscript evidence that the system recorded when forming a finding. It is a recoverability test, not an entailment test. A quotation can be recovered correctly without being sufficient on its own to prove the final clinical judgment.

External evidence. For findings that use an external standard, guideline, or biomedical source, EGCR stores the source identity, relevant requirement or passage, applicability decision, and alignment. The traceability audit checks whether the final finding can be traced through saved intermediate artifacts back to the external requirement.

For sources from a structured reporting or appraisal corpus, we can further locate the recorded requirement in the corresponding source text. For broader clinical guidelines, regulatory documents, or contextual biomedical evidence, the system preserves the source path through metadata

and provenance information. This check measures source-provenance preservation. It does not independently verify that the external requirement was interpreted correctly, and it is not a clinical correctness assessment.

Submission-time availability. External evidence should be used to evaluate a submitted manuscript only if it was available at the time of submission. Each external source is therefore compared with the manuscript's submission-time cutoff. When an availability date can be determined reliably, the system records whether the source was available on or before the cutoff. When the source date cannot be determined reliably, the state is kept as undetermined rather than assuming that the source was available.

This prevents post-submission knowledge from being used to judge what the submitted manuscript had already supported. It also lets the final review artifact distinguish submission-time evidence from later evidence.

Overall, the traceability audit does not ask whether the system is always clinically correct. It asks whether a final review finding preserves enough information for a reviewer to inspect the manuscript passage, external requirement, applicability decision, and time condition that led to the finding.

D.5 Held-Out Human Review Mapping

Held-out human reviews are used only after EGCR has completed all generation stages. The system cannot access these review texts while generating clinical claims, retrieving external evidence, judging applicability, forming alignments, or producing final review findings.

To compare human review with EGCR in a common space, we first split human-review text into review units. Each unit is classified by function as a substantive concern, request, praise, resolution, editorial or decision-related content, or other text. The main clinical coverage analysis uses only substantive concerns and requests. This makes the metric reflect issues that reviewers asked authors to explain, revise, or support further, rather than treating praise, editorial text, or later resolution as new review concerns.

For each paper P, let $H(P)$ be the set of clinical dimensions with at least one mapped substantive human concern or request. Let $E(P)$ be the set of clinical dimensions for which EGCR can form an assessment. Human-review dimension coverage is

$$C_H(P) = \frac{|H(P)|}{7},$$

and the union coverage of human review and EGCR is

$$C_U(P) = \frac{|H(P) \cup E(P)|}{7}.$$

The additional coverage is

$$\Delta C(P) = C_U(P) - C_H(P).$$

Human-review mapping and EGCR use the same seven-dimensional clinical space. Union coverage therefore asks whether EGCR brings additional assessable clinical dimensions into review under that shared space. A dimension in $E(P) \setminus H(P)$ does not mean that a human reviewer missed an error. It means that two conditions hold: EGCR can form an assessment for that dimension from the manuscript and available evidence, and the held-out human review contains no mapped substantive concern or request for that dimension.

An additional dimension may correspond to an evidence gap, a supportive check, a scope clarification, or other clinical information that a reviewer should inspect. The human-review comparison therefore evaluates complementarity of clinical coverage. It does not treat human reviews as a complete gold standard, and it does not define non-overlap as a true positive or false positive.

Nature-series peer-review material and OpenReview text differ in review process. For example, Nature review files may include author responses and later-round resolution, while OpenReview does not always provide comparable resolution information. To keep the publication settings comparable, we count whether a substantive concern or request was raised for a dimension, and we do not change its coverage status based on whether it was later resolved.

This design makes the main coverage comparison answer a limited question: beyond existing human-review concerns, can EGCR make other assessable clinical dimensions explicit for reviewer inspection? It is not used to judge whether EGCR is better than human reviewers overall.