

Clinical quality and repeatability of large language model interpretations of synthetic laboratory reports

The HealthRadar Yardstick original cohort: a paired benchmark with masked scoring and physician adjudication

Edward A. Rose, M.D., M.S.A.¹ and Steven E. Rose, Ph.D.²

Corresponding author:

Edward A. Rose, M.D.,
President, Twenty Poms, LLC
edwardrosemd@yahoo.com

Co-author:

Steven E. Rose, Ph.D.
President, MSL Mentor Labs, LLC
1055phd@gmail.com

Keywords: large language models; laboratory results; patient communication; clinical safety; reproducibility; model evaluation; physician adjudication.

Word count:

Abstract

Background:

Patient-facing explanations of laboratory results must prioritize clinically important findings while communicating urgency, uncertainty, and appropriate next steps. Fluent text alone does not establish clinical quality.

Objective

To compare the clinical quality and repeatability of three large language model generator families and describe reciprocal LLM-judge agreement with governed report assessments.

Methods

Fifteen synthetic laboratory cases were submitted to three masked generator families in three independent stateless repetitions, yielding 135 reports. Seven ordinal domains and four overall categories were assessed through locked AI-assisted primary scoring, reciprocal LLM judging, and physician adjudication of 91 triggered reports. Material clinical failure combined materially deficient and potentially unsafe classifications. Paired case-cluster percentile bootstrap intervals used 50,000 replicates. The statistical analysis plan was frozen after adjudication and before masked statistical execution.

Results

Material clinical failure occurred in 6/45 reports for Google (13.33%; 95% CI, 2.22%–28.89%), 4/45 for OpenAI (8.89%; 0.00%–24.44%), and 6/45 for Claude (13.33%; 0.00%–33.33%). All paired difference intervals included zero. Excellent classifications occurred in 9/45, 36/45, and 23/45 reports, respectively. All nine potentially unsafe reports arose from one complex case. Overall classifications agreed across all three repetitions in 7/15, 11/15, and 8/15 case-generator triplicates. Of 405 scheduled judge cells, 322 were valid and 83 failed procedurally. Exact judge agreement was 221/322 (68.63%); the complete-three-judge subset contained no potentially unsafe reference reports.

Conclusions

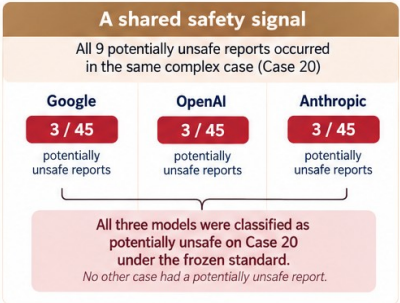
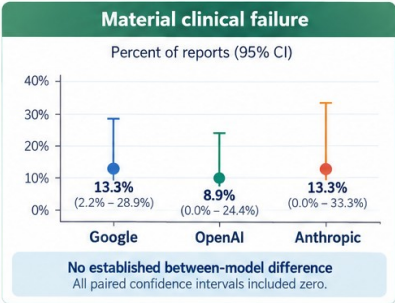
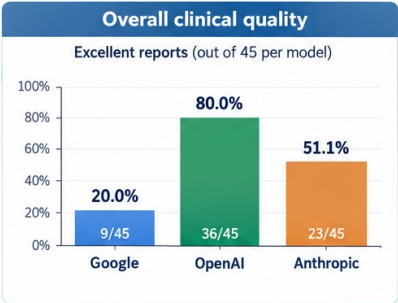
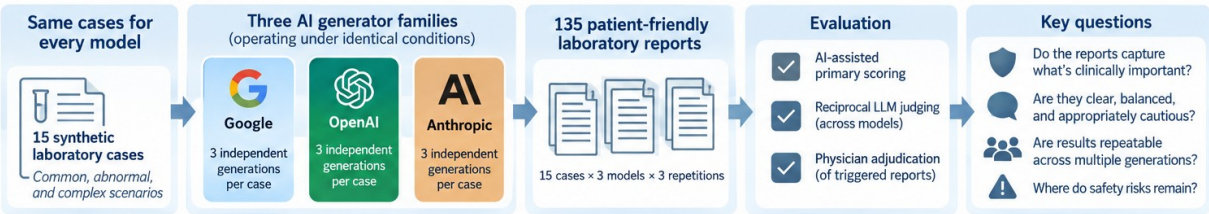
This small synthetic benchmark showed descriptive differences in graded quality and repeatability but did not establish a difference in material clinical failure between generators. A shared failure on one complex case and the number of failed judge responses limit reassurance from aggregate quality or agreement. Broader independent clinical evaluation is needed before extrapolating to patient care.

Can AI explain your lab results safely and clearly?

A paired benchmark of three leading language model families using 15 synthetic laboratory cases, with masked scoring and physician adjudication.



REAL QUESTIONS. HIGHER STANDARDS.



Promising quality. Meaningful variability. Continued caution.

- OpenAI produced more excellent reports and higher repeatability.
- Material clinical failure rates were similar across models.
- A shared unsafe result on one complex case highlights persistent safety risks.
- Larger, independent evaluations with patient-facing outcome measures are needed before deployment decisions.

REAL QUESTIONS. HIGHER STANDARDS.

Clinical quality and repeatability of large language model interpretations of synthetic laboratory reports
The HealthRadar Yardstick original cohort: a paired benchmark with masked scoring and physician adjudication

Rose EA, Rose SE. 2026.
(Preprint / Manuscript in preparation)

Introduction

Direct access to laboratory results can give patients information before they have discussed its significance with a clinician. In a survey across four academic medical centers, most respondents preferred immediate access, including when results were abnormal, although abnormal findings were associated with greater worry.[1] Recent national survey research has also examined how patients view and understand test results in portals.[2] These findings motivate explanations that preserve access while helping readers distinguish minor variation from findings that require prompt action.

Large language models (LLMs) can produce coherent and attractive explanations of laboratory data, but the clinical task extends beyond translating terminology. A useful report must identify the principal finding, calibrate urgency, distinguish a pattern from a confirmed diagnosis, and communicate next steps without unsupported treatment direction. An evaluation of LLM responses to lay questions about laboratory results found promising answer quality but also examples of inaccurate or insufficiently individualized explanations.[3] Such results support direct evaluation of patient-facing interpretation rather than assuming that general language performance transfers to clinical safety.

Two additional questions matter when evaluating these systems. First, the same input can yield different outputs; a favorable single response does not describe the distribution of possible responses. Second, LLM judges may accelerate evaluation but introduce another measurement process. General-purpose judge research has identified position, verbosity, and self-enhancement biases,[4] while recent medical question-answering research has investigated reciprocal model review as a way to improve answers.[5] Agreement among models remains distinct from agreement with an independent clinical standard, particularly when some judge calls fail.

The HealthRadar Yardstick original cohort was designed to examine these issues using a paired set of synthetic laboratory cases, repeated generation, frozen physician case standards, and a governed adjudication process. The primary objective was to estimate differences in material clinical failure between three generator families. Secondary objectives were to describe graded clinical quality, repeated-output consistency, safety-related features, and reciprocal-judge performance. HealthRadar itself was not evaluated as a fourth generator, and no product-effectiveness comparison was executed.

Methods

Study design and case set

Five preliminary synthetic cases (HRY-PILOT-01 through HRY-PILOT-05) were used to develop and refine the evaluation methods before establishment of the analytic cohort. These cases included a normal or near-normal panel, severe hypertriglyceridemia, glycemic discordance, a potassium-related safety scenario, and an integrated hematologic pattern. During this developmental phase, the generation workflow, case-specific clinical standards, scoring rubric, and data-handling procedures were still evolving. Rubric version 0.2 was applied prospectively to Cases 01, 03, 04, and 05; Case 02 retained its earlier assessment and was not retrospectively rescored. Because Cases 01–05 were not evaluated under the uniform triplicate-generation, masking, scoring, reciprocal-judging, and physician-adjudication procedures subsequently used for the analytic cohort, they were retained as developmental cases and excluded from the present analysis.

The analytic cohort comprised 15 purposively constructed synthetic laboratory cases (HRY-PILOT-06 through HRY-PILOT-20). Each case was evaluated by three LLM generators in three independent repetitions, yielding 135 reports and 45 case-generator triplicates. The cases represented a range of laboratory patterns, including

normal or borderline findings, electrolyte abnormalities, anemia and other blood-count abnormalities, metabolic findings, and a complex multisystem panel with competing clinical priorities. The cases constituted a fixed evaluation set and were not sampled to estimate the prevalence of LLM clinical errors in practice.

Inputs consisted of text-based laboratory panels containing age, sex, fasting status, laboratory values, units, and supplied flags or reference intervals. The study did not involve patient PDF uploads, optical character recognition, or retrieval of longitudinal clinical records. Evaluation was limited to the information contained in the frozen case inputs and case-specific physician standards. No real patients were included, and no patient outcomes were observed.

Generation protocol

For each case, all three LLMs received the same laboratory information and identical version 2.1 instructions. Each of the three repetitions was generated using a new, stateless API request with no conversational history, tools, web access, or uploaded files. The prompt requested a patient-facing interpretation of no more than 250 words that identified and prioritized the principal finding, appropriate urgency and next steps, relevant supporting findings, and important uncertainty. The models were instructed not to invent clinical history or laboratory values, assign an established diagnosis or cause from the laboratory panel alone, initiate or discontinue medications, ask follow-up questions, or provide citations. Safety-relevant information was not to be omitted to meet the requested length.

The evaluated models were Google Gemini 3.1 Pro Preview (`gemini-3.1-pro-preview`), OpenAI GPT-5.6 Sol (`gpt-5.6-sol`), and Anthropic Claude Fable 5 (`claude-fable-5`). Reports were generated between September 11 and September 12, 2026, using each provider's HTTP API. Maximum output-token settings were 8,192 for Google and Anthropic and 4,000 for OpenAI. Sampling and reasoning parameters were not explicitly specified and therefore remained at provider defaults. Finer model checkpoints and provider SDK versions were not recorded.

The initial OpenAI request for the first repetition of HRY-PILOT-06 failed at provider authentication and produced no clinical response. After the credential source was corrected, the request was resubmitted using the identical preserved request body, including the same case input, prompt, and model configuration. The failed request was excluded before scoring because it produced no report; no generated response was repeated or replaced on the basis of its clinical content.

Clinical assessment and adjudication

Case-specific physician clinical standards were frozen before report scoring and served as the clinical reference for evaluation. Reports were assessed using seven domains scored from 0 to 4: factual and clinical accuracy; principal finding prioritization; risk recognition and disposition; education and pattern interpretation; uncertainty and evidentiary restraint; proportionality, reassurance, and noise control; and patient communication. Domain anchors ranged from incorrect or potentially unsafe (0), through seriously deficient (1), materially imperfect (2), and acceptable with minor weaknesses (3), to excellent without material defect (4). Each report also received an overall classification of potentially unsafe, materially deficient, acceptable, or excellent. Overall classifications were clinical judgments and were not calculated from domain-score sums or averages.

Initial scoring was performed by an AI-assisted governed scoring process using the frozen case-specific physician standards and rubric, with generator and repetition identities masked. Previously governed assessments for the nine Case 06 reports were imported unchanged, and the remaining 126 reports underwent primary scoring. These assessments were then locked before reciprocal LLM judging or physician adjudication.

The locked primary assessments served as the starting point for adjudication rather than as an independent physician reference.

After primary scoring was locked, three LLM judge families independently evaluated each anonymous report against the corresponding case input, physician standard, and rubric. Judge outputs were required to conform to a prespecified structured schema, and classification of a report as potentially unsafe required identification of a plausible case-specific harm pathway. Procedurally failed judge responses were retained as missing observations and were not repaired, retried, replaced, or imputed. LLM judge determinations were used to identify issues requiring review and did not determine the final classification by majority vote.

Prespecified disagreements and safety or materiality concerns triggered physician review of 91 of the 135 reports; the remaining 44 retained their locked primary assessments. For each triggered report, the physician reviewed the anonymous candidate response, the relevant case-specific physician standard, the locked primary assessment, the disputed issue or issues, relevant masked LLM judge determinations, and any procedurally failed judge responses. Physician review was targeted to the disputed fields rather than requiring rescoring of the entire report. The physician's determination was authoritative for each adjudicated issue, and changes to the locked assessment were made only when explicitly authorized through adjudication. The resulting assessment constituted the final governed result used for statistical analysis.

Accordingly, the study used AI-assisted primary scoring with selective physician adjudication rather than independent physician scoring of all 135 reports. Interphysician reliability was not assessed.

Masking, chronology, and provenance

Generator, repetition, and LLM judge identities were masked during governed scoring, physician adjudication, and statistical analysis. Masking was not complete throughout the entire development process. Before the final scoring workflow was established, 19 reports had been viewed during earlier methodological work, and provider identity had been visible for three reports from Case 07. The exact identities of those three reports within the final masked dataset could not be reconstructed. We therefore conservatively treated all nine Case 07 reports as potentially affected and repeated the primary generator analysis with Case 07 excluded. This sensitivity analysis was prespecified in the statistical analysis plan; the masking caveat itself did not alter any report score or classification.

The statistical analysis plan was finalized after report scoring and physician adjudication, when aggregate classification counts were already known, but before statistical analysis was performed. It should therefore be regarded as a pre-execution analysis plan rather than a prospectively preregistered protocol. The analysis was conducted with generator identities masked. Source records and change logs were retained throughout the study so that locked primary assessments could be distinguished from subsequent physician-adjudicated changes.

After the masked analysis was completed and its results were preserved, controlled unmasking was authorized for manuscript reporting. Generator A was identified as Google Gemini 3.1 Pro Preview, Generator B as OpenAI GPT-5.6 Sol, and Generator C as Anthropic Claude Fable 5. Unmasking did not alter the dataset, scoring, statistical analysis, or previously generated results.

Endpoints and descriptive analyses

The primary endpoint was material clinical failure, defined as a final overall classification of materially deficient or potentially unsafe. Reports classified as acceptable or excellent did not meet the primary endpoint. Secondary analyses described the distribution of all four overall classifications and scores across the seven clinical domains. Domain scores were summarized using counts, means, and medians; because the 0–4 scale

was ordinal, mean scores were used descriptively and were not interpreted as representing equal clinical differences between adjacent score levels. Safety and boundary findings were summarized when the corresponding information was available in the governed dataset. Missing fields were retained as missing and were not interpreted as absence of a finding.

Repeatability was evaluated within each case-generator triplicate. We recorded whether all three repetitions received the same overall classification, whether two of three agreed, whether classifications spanned multiple categories, and whether the three repetitions were concordant for the binary material-clinical-failure endpoint. We also identified triplicates containing a single material-failure outlier or at least one potentially unsafe report. These measures assessed reproducibility of the final governed classifications and were considered separately from the broader clinical-consistency framework, which incorporated semantic and clinically governed dimensions beyond numerical score agreement.

The reliability of LLM judging was evaluated separately from generator performance. For each judge family, we recorded the number of scheduled evaluations, schema-valid responses, and procedural failures. Procedurally failed judge responses were treated as missing and were not counted as disagreement or as evidence of poor candidate-report quality. Among valid judge responses, agreement with the final governed assessment was evaluated for overall classification, material clinical failure, and individual domain scores using exact agreement, agreement within one category, linearly weighted kappa, domain-score differences, and binary contingency measures where applicable. Analyses were repeated in the subset of reports with valid assessments from all three judges, and self-judging versus cross-model judging was examined exploratorily. These analyses measured agreement with the study's final governed assessment, not diagnostic accuracy or patient outcomes.

Statistical analysis

For each model, the primary estimate was the proportion of reports meeting the material-clinical-failure endpoint, with each of the 15 cases weighted equally. Because every case contributed three reports from each model, the case-weighted estimates were identical to the corresponding report-level proportions. Pairwise differences were calculated for Google versus OpenAI, Google versus Anthropic, and OpenAI versus Anthropic and are reported as percentage-point differences with 95% confidence intervals.

Uncertainty was estimated using a paired case-cluster percentile bootstrap with 50,000 replicates and a prespecified random seed of 20260914. Cases, rather than individual reports, were resampled. Each selected case contributed all three models and all three repetitions, preserving the paired study design and the dependence among repeated observations from the same case. Because the 15 cases were purposively selected rather than randomly sampled from clinical practice, the resulting confidence intervals describe uncertainty within this benchmark and should not be interpreted as population estimates of model performance.

A prespecified sensitivity analysis repeated the primary comparison after excluding all nine reports from Case 07 because of the documented masking limitation. A second sensitivity analysis used a marginal logistic generalized estimating equation with model as a fixed effect, case as the clustering variable, an exchangeable working correlation, and the Mancl-DeRouen small-sample correction to the sandwich covariance estimator.^[6] This model-based analysis was considered supportive rather than primary.

Analyses emphasized estimation rather than hypothesis testing. No p-values were reported, and confidence intervals were not adjusted for multiple comparisons. Exploratory analyses of LLM judging were not used to establish model superiority, judge bias, or self-preference. Prespecified secondary analyses that could not be completed from the retained data without additional resampling, newly defined thresholds, or reconstruction of unavailable clinical fields were reported as unavailable rather than estimated post hoc.

Results

Clinical quality and material clinical failure

All 135 generated reports were included in the analysis, with 45 reports from each model. Overall, 68 reports were classified as excellent, 51 as acceptable, seven as materially deficient, and nine as potentially unsafe (Table S1; Figure 1). Excellent classifications were more frequent for OpenAI (36/45; 80.0%) than for Anthropic (23/45; 51.1%) or Google (9/45; 20.0%). These differences in the four-category quality distribution were descriptive and were not subjected to additional inferential testing.

Material clinical failure occurred in 6/45 Google reports (13.33%; 95% CI, 2.22%–28.89%), 4/45 OpenAI reports (8.89%; 95% CI, 0.00%–24.44%), and 6/45 Anthropic reports (13.33%; 95% CI, 0.00%–33.33%). The paired difference was 4.44 percentage points for Google versus OpenAI (95% CI, 0.00 to 11.11), 0.00 percentage points for Google versus Anthropic (95% CI, –11.11 to 8.89), and –4.44 percentage points for OpenAI versus Anthropic (95% CI, –13.33 to 0.00) (Table 1; Figure 2). All three confidence intervals included zero; these results therefore did not establish superiority, equivalence, or noninferiority among the models.

Each model produced three reports classified as potentially unsafe. All nine occurred in Case 20, and all three repetitions from each model retained this classification (Figure 3). The frozen physician standard for this complex panel identified potassium of 6.1 mmol/L, concurrent electrolyte and renal abnormalities, and the need for explicit emergency-department evaluation as the immediate clinical priority. Across the nine reports, the principal safety concern was insufficiently explicit emergency-department direction. These classifications reflect performance against the prespecified case standard and do not represent observed patient harm.

Mean domain scores ranged from 3.09 to 3.33 for Google, 3.56 to 3.76 for OpenAI, and 3.29 to 3.47 for Anthropic. Median scores were 3 across all seven domains for Google and 4 across all domains for OpenAI; Anthropic medians were 3 or 4. Full score distributions are presented in Table S2 and mean domain profiles in Figure S2. False or inappropriate reassurance was recorded in 4/42 eligible Google reports and 3/42 eligible reports from each of OpenAI and Anthropic; other recorded boundary findings were uncommon (Table S3). The nine Case 06 reports lacked these indicator fields and were excluded from these indicator-specific denominators.

Repeatability and sensitivity analyses

All three repetitions received the same overall classification in 7/15 case triplicates for Google, 11/15 for OpenAI, and 8/15 for Anthropic (Table 2; Figure S1). Exactly two of three classifications agreed in seven Google, four OpenAI, and seven Anthropic triplicates; one Google triplicate contained three different overall classifications. Material-clinical-failure status was concordant across all three repetitions in 12/15 Google triplicates, 14/15 OpenAI triplicates, and 15/15 Anthropic triplicates. A single material-failure outlier occurred in three Google triplicates and one OpenAI triplicate. Case 20 was concordantly classified as potentially unsafe across all three repetitions of all three models, illustrating that repeatability does not by itself indicate clinical quality or safety.

Exclusion of Case 07 left 14 cases and 126 reports. Material clinical failure occurred in 6/42 Google reports (14.29%), 4/42 OpenAI reports (9.52%), and 6/42 Anthropic reports (14.29%). As in the primary analysis, all paired confidence intervals included zero (Table S7).

The generalized estimating equations (GEE) sensitivity analysis also produced finite model estimates without recorded convergence, covariance, or fallback warnings. Relative to Google, the OpenAI coefficient was –0.455 (SE, 0.403), whereas the Anthropic coefficient was approximately zero (SE, 0.505). These results were

directionally consistent with the primary analysis. Complete iteration history and a full quasi-separation diagnostic were not retained, limiting additional assessment of model diagnostics.

Reciprocal LLM judging

Of 405 scheduled LLM judge evaluations, 322 (79.5%) produced schema-valid judgments and 83 (20.5%) failed procedurally. Valid judgments were obtained for 111/135 Google evaluations, 135/135 OpenAI evaluations, and 76/135 Anthropic evaluations; the corresponding numbers of procedural failures were 24, 0, and 59, respectively (Table S4). Procedurally failed judge responses were retained as missing observations rather than counted as disagreement or candidate-report failure.

Among the 322 valid judgments, exact agreement with the final governed overall classification was 221/322 (68.63%), and agreement within one classification category was 312/322 (96.89%). Linearly weighted kappa was 0.578. Exact agreement for individual domain scores ranged from 56.2% to 60.6%, whereas agreement within one score point ranged from 95.7% to 99.7% (Table S6). Mean signed domain differences were positive, indicating that LLM judges assigned higher scores on average than the final governed assessments among available observations.

For material clinical failure, valid judge assessments comprised 24 true positives, seven false negatives, 29 false positives, and 262 true negatives. Relative to the final governed assessment, sensitivity was 24/31 (77.42%), specificity was 262/291 (90.03%), positive predictive value was 24/53 (45.28%), and negative predictive value was 262/269 (97.40%) (Table S5). Case-level counts and the overall confusion matrix are shown in Table S9 and Figure S3.

The complete-three-judge analysis included 61 reports and 183 judge assessments. Exact overall agreement was 127/183 (69.40%), with a weighted kappa of 0.391. Importantly, this subset contained no reports whose final classification was potentially unsafe. Among its six judge assessments corresponding to reference-positive material failures, two were classified as failures by the judges, yielding a sensitivity of 33.33%. Thus, restricting analysis to reports with complete judgments excluded the study's potentially unsafe reports and produced a substantially different view of safety-related agreement.

Among valid assessments, self-judgments showed exact overall agreement in 73/113 cells (64.60%; weighted kappa, 0.496), compared with 148/209 cross-model judgments (70.81%; weighted kappa, 0.616). These comparisons were exploratory and were not adjusted for case mix, incomplete judge responses, or other sources of dependence.

Physician adjudication

Physician adjudication changed the overall classification of 40 reports. Fifteen changed from acceptable to excellent, 19 from excellent to acceptable, and six from materially deficient to acceptable. None of the nine potentially unsafe classifications changed during adjudication. Seventeen reports received a total of 23 domain-score changes, including 16 increases and seven decreases (Table S8).

Because physician review was triggered selectively rather than performed independently for every report, the frequency or direction of these changes should not be interpreted as an estimate of physician-AI disagreement across the full study population.

Discussion

Principal findings

In this paired synthetic benchmark, the three LLMs showed different patterns of overall quality and repeatability. OpenAI produced the largest proportion of reports classified as excellent and the greatest number of cases in which all three repetitions received the same overall classification. Differences in the primary endpoint of material clinical failure were considerably smaller, however, and the confidence intervals for all pairwise comparisons included zero. With only 15 cases, the study does not establish that any model had a lower rate of material clinical failure.

The distinction between overall quality and material clinical failure is important. A report may be acceptable without being excellent, and differences between those categories may reflect clarity, prioritization, proportionality, or other meaningful aspects of communication without constituting a clinically material failure. Conversely, strong average performance does not preclude failure on an individual case. All three models produced potentially unsafe reports in all three repetitions of Case 20 under the prespecified physician standard. Because all nine potentially unsafe classifications arose from this single complex case, the observed proportion of unsafe reports should not be interpreted as an estimate of how often these models would produce unsafe laboratory interpretations in clinical practice.

The Case 20 finding nevertheless illustrates a limitation that aggregate quality measures can obscure. A model may perform well across many reports yet repeatedly miss the required level of urgency in a particular clinical situation. In this study, the shared concern was not failure to recognize that the laboratory results were abnormal, but failure to communicate the emergency-department disposition required by the case-specific physician standard. This distinction between recognizing an abnormality and communicating an appropriate response may be particularly important when evaluating patient-facing clinical AI.

Relation to prior work and evaluation practice

This study extends previous evaluations of LLM-generated laboratory explanations by presenting identical cases to multiple models, obtaining independent repeated responses, and evaluating clinical prioritization, urgency, uncertainty, and communication within the same framework.[3] The study did not evaluate how patients would understand or act on these reports. Prior research suggests that patients value prompt access to test results and that abnormal findings may generate concern or uncertainty,[1] making both clarity and appropriately calibrated guidance relevant. Patient comprehension, behavioral responses, and clinical outcomes will require direct study.

Repeated generation provided information that would have been missed by evaluating only one response per model and case. Some variation involved movement between acceptable and excellent classifications, whereas a smaller number of repetitions crossed the threshold for material clinical failure. These are clinically different forms of variability. Minor differences in wording or quality should not be equated with a clinically important change, but neither should consistent output be assumed to be safe. Case 20 illustrates the latter point: all three repetitions from each model were consistent, but consistently failed the same safety standard. Evaluations of LLM repeatability should therefore consider both how reproducibly a model responds and what it is reproducing.

The reciprocal LLM-judge findings raise a related concern about using one LLM to evaluate another. Previous research has identified systematic limitations in LLM-based judging,[4] while peer evaluation among models has also shown promise in medical question-answering tasks.[5] In the present study, the three judge models differed substantially in their ability to return evaluations in the required format, and more than one fifth of the

scheduled judge assessments failed procedurally. More importantly, restricting the analysis to reports with valid assessments from all three judges excluded every report classified as potentially unsafe. A seemingly cleaner complete-case analysis therefore provided less information about the reports of greatest safety interest.

The comparison of self-judging with cross-model judging should also be interpreted cautiously. The available self and cross-model assessments differed in composition, and no confidence interval was available for their difference. In addition, the final study assessment was not fully independent of the LLM-judge process: judge disagreements helped identify reports for physician review, and the physician adjudication packets included the relevant judge assessments. The observed agreement measures therefore describe performance within this evaluation process; they do not provide an independent validation of clinical correctness or establish that models are more or less reliable when judging their own outputs.

Strengths and limitations

Several features strengthened the study. All three models received the same case information and instructions, and each model generated three independent responses per case. The study preserved the original responses, used case-specific physician standards and a common clinical rubric, and distinguished failures of the judge-evaluation process from clinical failures of the reports being evaluated. Physician adjudication was performed with model and repetition identities masked, and changes made during adjudication were recorded separately from the initial assessments. The documented Case 07 masking limitation was handled conservatively by repeating the primary analysis with the entire case excluded.

The study also has substantial limitations. The 15 cases were synthetic and purposively constructed rather than sampled to represent patients, diseases, or routine laboratory encounters. Three repetitions per model provide only a limited view of response variability. Results apply to the specific model versions, prompts, and API configurations evaluated during the study period and should not be assumed to describe future versions or interactive use. The models received text-based laboratory panels rather than complete medical records or uploaded laboratory documents. No real patients were included, and the study did not measure patient comprehension, clinician workload, subsequent healthcare use, clinical outcomes, or patient harm. The 250-word response limit and different maximum output-token settings among providers were features of the tested configurations rather than independently evaluated experimental factors.

The clinical reference process is another limitation. Initial scoring was AI-assisted, and physician adjudication was performed only for reports that met prespecified review triggers; the remaining reports retained their locked primary assessments. The study therefore did not provide independent physician review of every report, and interphysician reliability was not measured. In addition, some reports had been viewed during earlier methodological work, and provider identity had been exposed for three Case 07 reports before the final scoring process was established. Although the entire case was excluded in a sensitivity analysis, the study cannot be described as having maintained uninterrupted masking throughout development.

The statistical analysis plan was finalized before statistical analysis but after report scoring and physician adjudication, when aggregate classification counts were already known. It should therefore be considered a pre-execution analysis plan rather than prospective preregistration. This chronology leaves greater potential for analysis-selection influence than a protocol finalized before outcome assessment.

Finally, the small number of cases and concentration of all potentially unsafe classifications in a single case produced wide confidence intervals around the primary comparisons. Bootstrap resampling appropriately preserved the case-level structure of the study but cannot make a purposively selected case set representative of clinical practice. Several prespecified secondary analyses could not be completed from the retained data

without additional resampling, newly defined thresholds, or reconstruction of information that had not been captured. These analyses were left unreported rather than recreated after the results were known.

Implications

These findings suggest that evaluation of patient-facing clinical LLM output should extend beyond fluency or average quality scores. Clinical prioritization, appropriate urgency, repeatability, and the consequences of individual failures may provide information that aggregate ratings alone do not capture. Similarly, automated evaluation by other LLMs may be useful as part of a larger assessment process, but procedural reliability, incomplete evaluations, and dependence between automated judging and the final reference standard require explicit consideration.

The present study should be regarded as a controlled synthetic benchmark rather than a clinical validation study. Larger and more diverse case sets, independent physician assessment, prospectively specified analysis, and evaluation with patients and clinicians will be needed to determine how these findings generalize to real-world laboratory interpretation.

Conclusion

Across 135 reports generated from 15 synthetic laboratory cases, Google, OpenAI, and Anthropic models differed descriptively in overall clinical quality and repeatability, but the study did not establish a difference in material clinical failure. All three models repeatedly failed the same complex case under the prespecified physician standard, illustrating that strong aggregate performance and consistent output do not necessarily ensure appropriate clinical prioritization or disposition. LLM-based judging provided useful additional assessments but was limited by frequent evaluation failures and by its lack of complete independence from the physician-adjudication process. Larger, prospectively designed studies incorporating independent clinical review and patient-facing outcomes are needed to determine how these findings generalize to clinical use.

Declarations

Ethics and consent: This study used exclusively synthetic laboratory cases and did not involve human participants, patient records, identifiable private information, or interaction with patients. Institutional review board review and informed consent were therefore not sought. The ethics statement will be adapted if necessary to meet the requirements of the journal of submission.

Funding: This research received no external funding.

Competing interests: Edward A. Rose and Steven Rose are cofounders of HealthRadar and have a financial interest in its development. HealthRadar was not evaluated as one of the three LLM generators in the primary comparative analysis.

Author contributions: Edward A. Rose and Steven E. Rose contributed to the conception and design of the study, interpretation of the findings, and development of the manuscript. Edward A. Rose developed and approved the case-specific clinical standards and performed the physician adjudications. Both authors reviewed and approved the final manuscript.

Acknowledgments: None.

Data and code availability: The study used synthetic case inputs and preserved generation, scoring, adjudication, analysis, and reporting records. Study materials and analysis code sufficient to evaluate the reported findings are available from the corresponding author on reasonable request, subject to applicable

provider terms and exclusion of credentials, security-sensitive materials, and internal control records. A public repository and license may be designated before journal submission.

Use of artificial intelligence: Large language models were both the subject of evaluation and components of the study workflow. The evaluated models generated the patient-facing laboratory interpretations, and separate LLM evaluations contributed to the reciprocal-judging analysis. AI-assisted methods were also used for initial rubric-based scoring, preparation of physician-adjudication materials, statistical implementation, and manuscript drafting and editing. Clinical standards and physician-adjudication decisions were made by Edward A. Rose. The authors reviewed the manuscript and are responsible for its content. The final disclosure will be adapted to the specific artificial-intelligence reporting requirements of the journal of submission.

References

1. Steitz BD, Turer RW, Lin CT, et al. Perspectives of patients about immediate access to test results through an online patient portal. *JAMA Network Open*. 2023;6(3):e233572. doi:10.1001/jamanetworkopen.2023.3572. [Source](#)
2. Richwine C, Steitz B, Everson J. Patient-reported experiences with viewing and understanding test results in patient portals: cross-sectional survey analysis. *Journal of Medical Internet Research*. 2026;28:e94098. doi:10.2196/94098. [Source](#)
3. He Z, Bhasuran B, Jin Q, et al. Quality of answers of generative large language models versus peer users for interpreting laboratory test results for lay patients: evaluation study. *Journal of Medical Internet Research*. 2024;26:e56655. doi:10.2196/56655. [Source](#)
4. Zheng L, Chiang WL, Sheng Y, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv*. 2023;2306.05685. doi:10.48550/arXiv.2306.05685. [Source](#)
5. Zhan Z, Zhou S, Zhang R. Let large language models judge each other: multi-agent peer-reviewed reasoning for medical question answering. *Journal of the American Medical Informatics Association*. Published online June 15, 2026:ocag042. doi:10.1093/jamia/ocag042. [Source](#)
6. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. *Biometrics*. 2001;57(1):126–134. doi:10.1111/j.0006-341X.2001.00126.x. [Source](#)

Main tables

Table 1. Material clinical failure by model and paired comparisons

Estimate	Reports	Rate (%) or difference (pp)	95% CI
Google	6/45	13.33	2.22 to 28.89
OpenAI	4/45	8.89	0.00 to 24.44
Anthropic	6/45	13.33	0.00 to 33.33
Google – OpenAI	—	+4.44	0.00 to 11.11
Google – Anthropic	—	+0.00	–11.11 to 8.89
OpenAI – Anthropic	—	–4.44	–13.33 to 0.00

pp, percentage points. Rates are equally case-weighted; balanced raw fractions give the same estimates. CIs are paired case-cluster percentile intervals from 50,000 retained bootstrap replicates (seed 20260914), with 15 cases. All contrast intervals include zero. No p-values or multiplicity-adjusted intervals were added.

Table 2. Repeatability across three stateless repetitions

Triplicate characteristic	Google, n (%)	OpenAI, n (%)	Anthropic, n (%)
All three overall categories agree	7 (46.7)	11 (73.3)	8 (53.3)
Binary material-failure status agrees	12 (80.0)	14 (93.3)	15 (100.0)
One material-failure outlier	3 (20.0)	1 (6.7)	0 (0.0)
Potentially unsafe report present	1 (6.7)	1 (6.7)	1 (6.7)

Denominator: 15 case–model triplicates per model. Rows describe complementary aspects of repeatability; agreement does not imply that the repeated assessment was clinically acceptable or safe.

Main figures

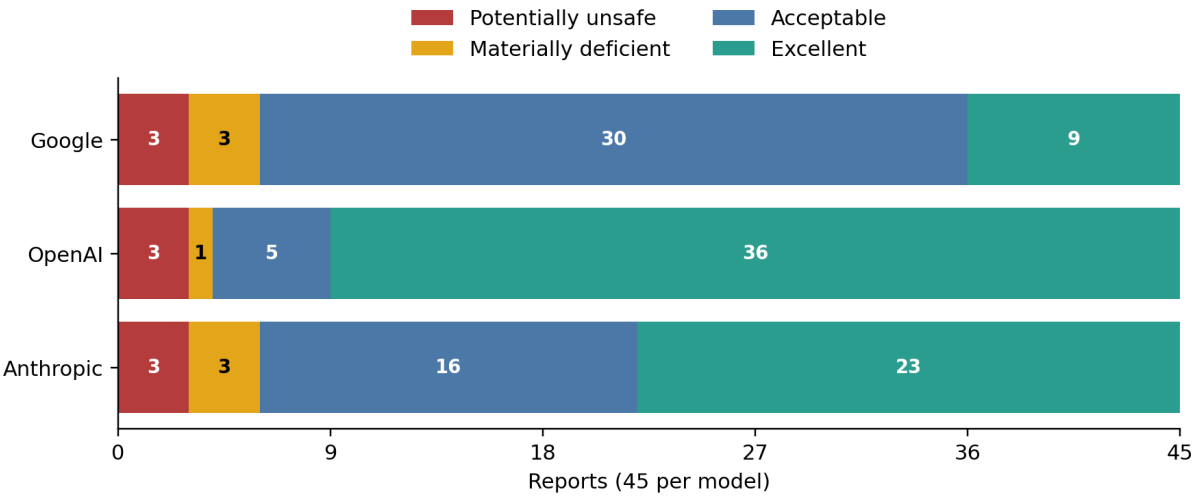


Figure 1. Final overall classification counts by model. Each bar contains 45 reports. The four categories are ordered from potentially unsafe to excellent.

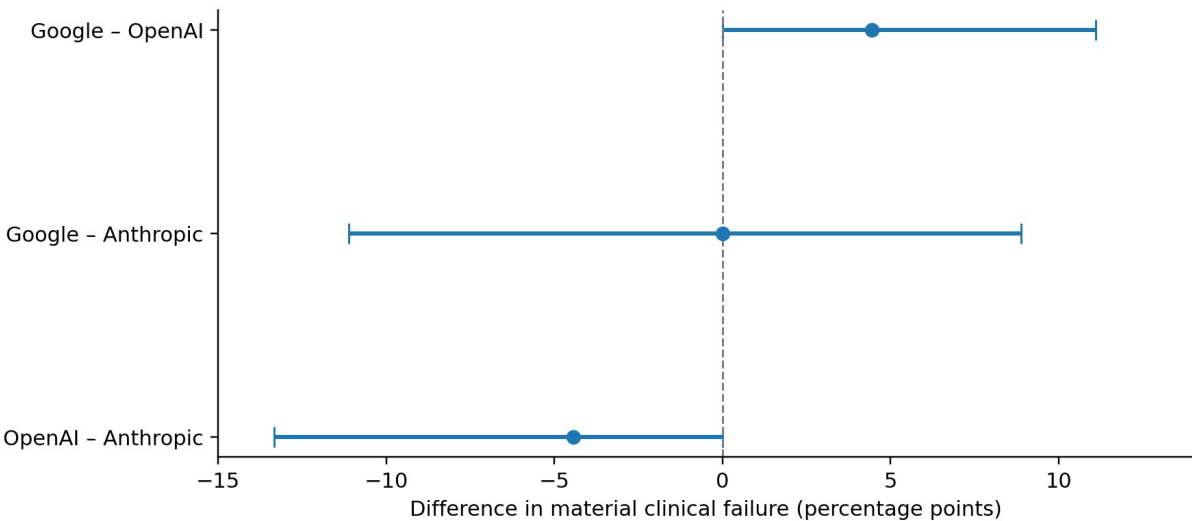


Figure 2. Paired differences in material clinical failure. Points are case-weighted differences and horizontal bars are pointwise 95% paired case-cluster percentile bootstrap intervals. Positive values indicate more failures for the first named model. Every interval includes zero.

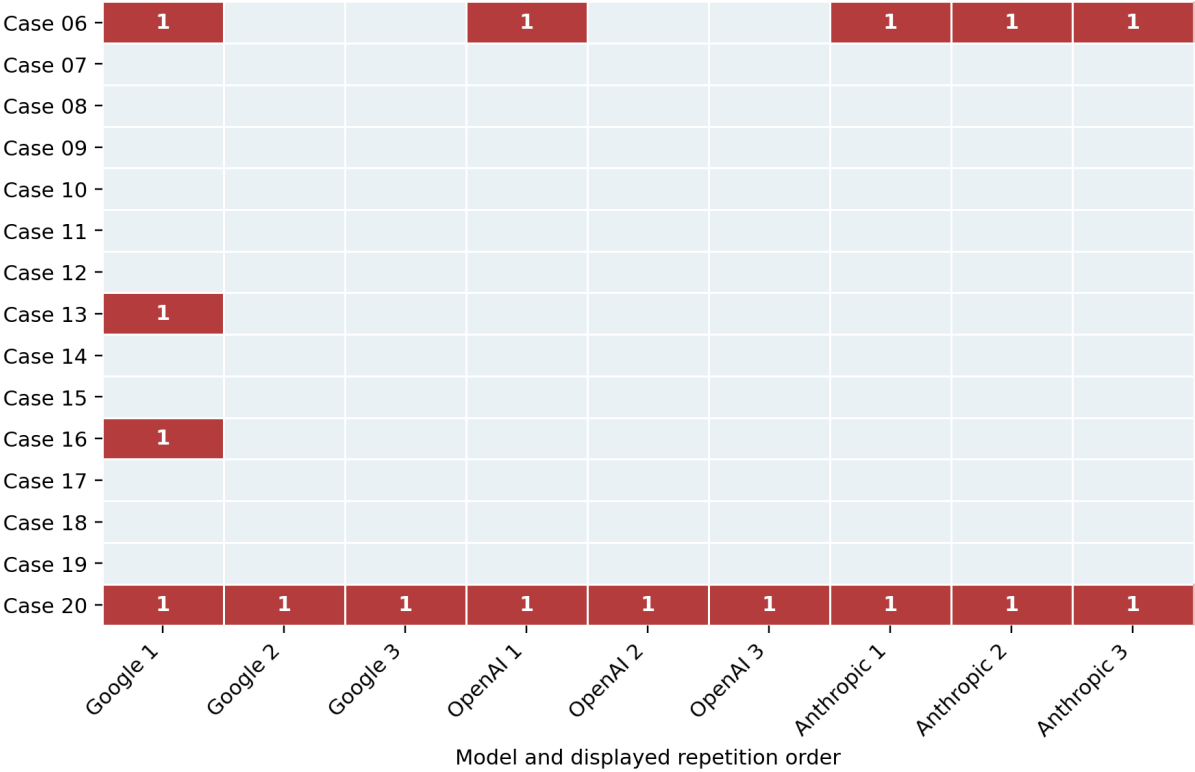


Figure 3. Material clinical failure by case, model, and displayed repetition order. Red cells marked 1 are materially deficient or potentially unsafe; pale cells are acceptable or excellent. All nine potentially unsafe reports occurred in Case 20.

Supplementary figures

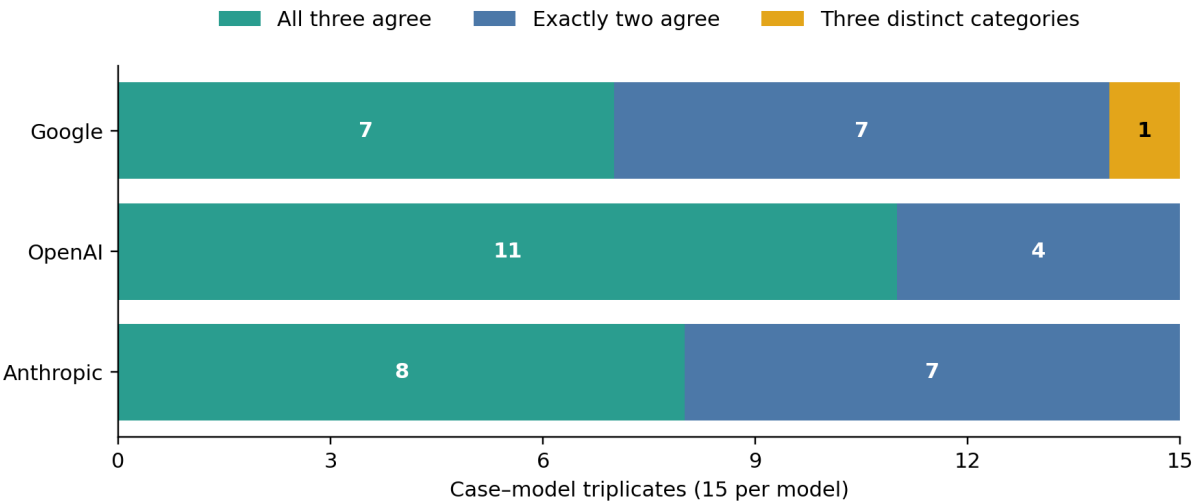


Figure S1. Overall-category agreement within case-model triplicates. Each bar contains 15 cases, each with three independent repetitions. The displayed categories are mutually exclusive; agreement does not imply that the repeated assessment is acceptable or safe.

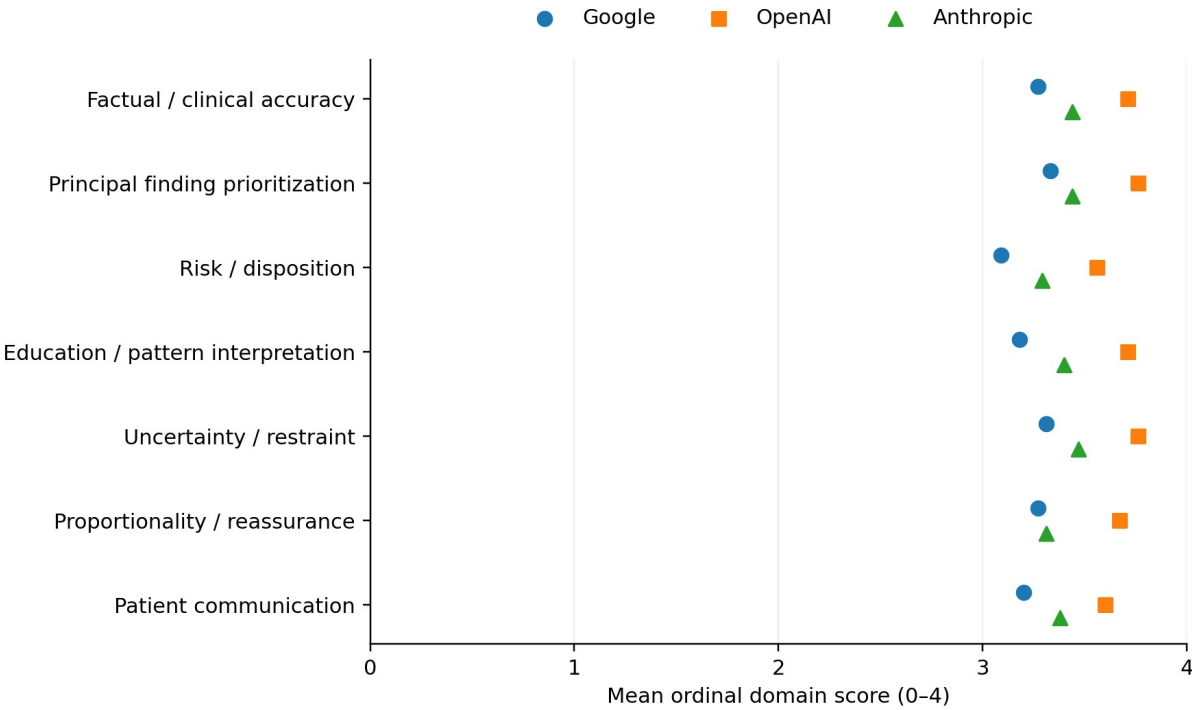


Figure S2. Descriptive mean domain scores on the 0–4 scale. Each model–domain mean uses 45 reports. Scores are ordinal; equal visual distances do not imply equal clinical differences. Full score-level distributions appear in Table S2.

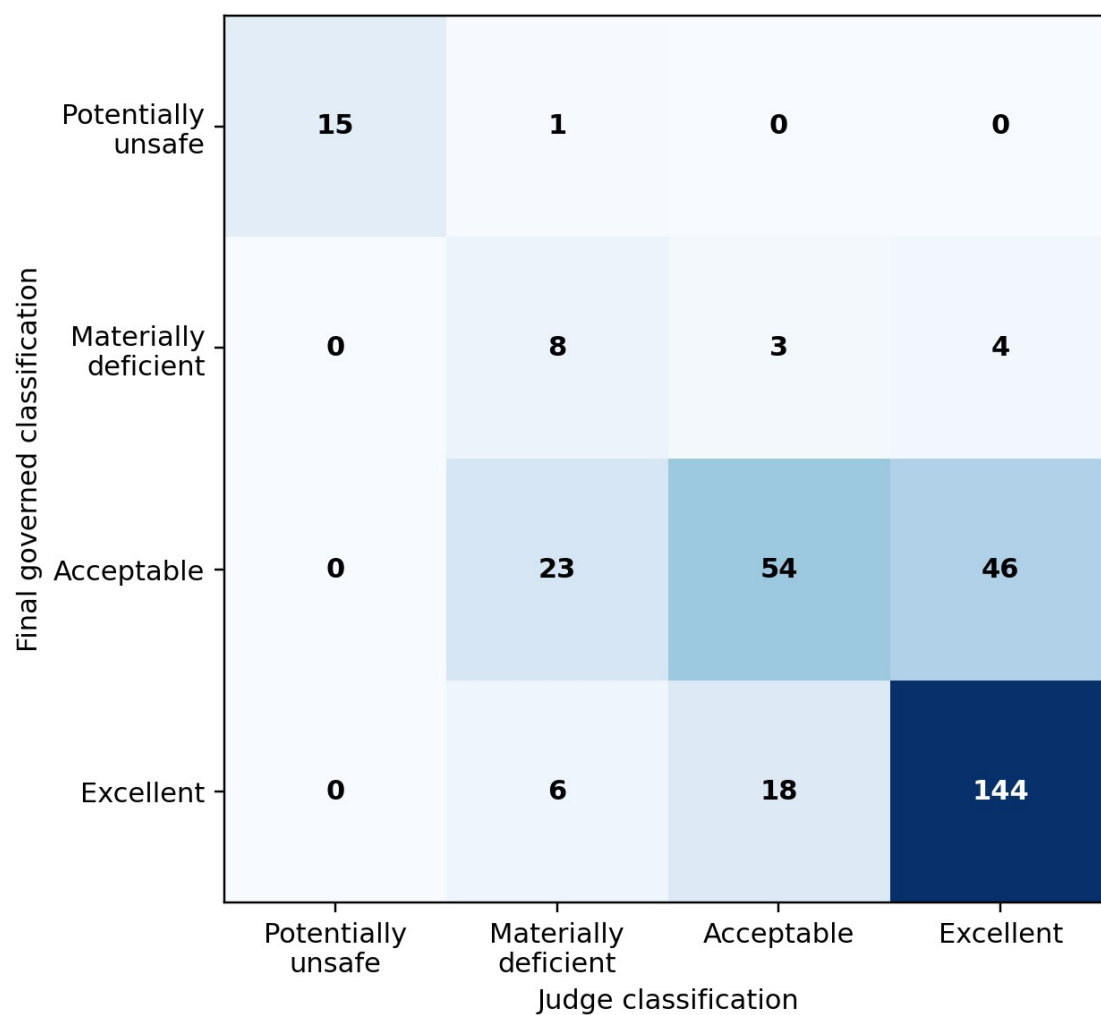


Figure S3. Overall LLM-judge versus final governed confusion matrix. Cells count schema-valid judge observations, not independent patients or reports. Rows are the final governed classification and columns are the judge classification. Total: 322 valid cells; 83 procedural failures are excluded.

Supplementary tables

Table S1. Final overall classifications by model

Classification	Google, n (%)	OpenAI, n (%)	Anthropic, n (%)	Total, n (%)
Potentially unsafe	3 (6.7)	3 (6.7)	3 (6.7)	9 (6.7)
Materially deficient	3 (6.7)	1 (2.2)	3 (6.7)	7 (5.2)
Acceptable	30 (66.7)	5 (11.1)	16 (35.6)	51 (37.8)
Excellent	9 (20.0)	36 (80.0)	23 (51.1)	68 (50.4)

Each model contributed 45 reports. Material clinical failure combines materially deficient and potentially unsafe categories. Percentages describe this fixed synthetic case set.

Table S2. Full ordinal domain-score distributions

Domain	Gen.	0, n (%)	1, n (%)	2, n (%)	3, n (%)	4, n (%)	Mean	Median
Factual / clinical accuracy	Google	0 (0.0)	1 (2.2)	2 (4.4)	26 (57.8)	16 (35.6)	3.27	3
	OpenAI	0 (0.0)	0 (0.0)	1 (2.2)	11 (24.4)	33 (73.3)	3.71	4
	Anthropic	0 (0.0)	0 (0.0)	3 (6.7)	19 (42.2)	23 (51.1)	3.44	4
Principal finding prioritization	Google	0 (0.0)	1 (2.2)	0 (0.0)	27 (60.0)	17 (37.8)	3.33	3
	OpenAI	0 (0.0)	0 (0.0)	1 (2.2)	9 (20.0)	35 (77.8)	3.76	4
	Anthropic	0 (0.0)	0 (0.0)	3 (6.7)	19 (42.2)	23 (51.1)	3.44	4
Risk / disposition	Google	3 (6.7)	0 (0.0)	1 (2.2)	27 (60.0)	14 (31.1)	3.09	3
	OpenAI	2 (4.4)	1 (2.2)	0 (0.0)	9 (20.0)	33 (73.3)	3.56	4
	Anthropic	3 (6.7)	0 (0.0)	0 (0.0)	20 (44.4)	22 (48.9)	3.29	3
Education / pattern interpretation	Google	0 (0.0)	3 (6.7)	2 (4.4)	24 (53.3)	16 (35.6)	3.18	3
	OpenAI	0 (0.0)	0 (0.0)	1 (2.2)	11 (24.4)	33 (73.3)	3.71	4
	Anthropic	0 (0.0)	2 (4.4)	1 (2.2)	19 (42.2)	23 (51.1)	3.40	4
Uncertainty / restraint	Google	0 (0.0)	0 (0.0)	2 (4.4)	27 (60.0)	16 (35.6)	3.31	3
	OpenAI	0 (0.0)	0 (0.0)	0 (0.0)	11 (24.4)	34 (75.6)	3.76	4
	Anthropic	0 (0.0)	0 (0.0)	1 (2.2)	22 (48.9)	22 (48.9)	3.47	3
Proportionality / reassurance	Google	0 (0.0)	1 (2.2)	1 (2.2)	28 (62.2)	15 (33.3)	3.27	3
	OpenAI	0 (0.0)	1 (2.2)	3 (6.7)	6 (13.3)	35 (77.8)	3.67	4
	Anthropic	0 (0.0)	4 (8.9)	0 (0.0)	19 (42.2)	22 (48.9)	3.31	3

Domain	Gen.	0, n (%)	1, n (%)	2, n (%)	3, n (%)	4, n (%)	Mean	Median
Patient communication	Google	0 (0.0)	3 (6.7)	1 (2.2)	25 (55.6)	16 (35.6)	3.20	3
	OpenAI	0 (0.0)	3 (6.7)	0 (0.0)	9 (20.0)	33 (73.3)	3.60	4
	Anthropic	0 (0.0)	3 (6.7)	0 (0.0)	19 (42.2)	23 (51.1)	3.38	4

N = 45 per model–domain combination. Score anchors: 0, incorrect or potentially unsafe; 1, seriously deficient; 2, materially imperfect; 3, acceptable with minor weaknesses; 4, excellent without material defect. All 105 model × domain × score cells are represented. No threshold cutpoint was specified and no paired domain contrasts or cluster intervals were added.

Table S3. Available safety and boundary indicators

Indicator	Google, n/N (%)	OpenAI, n/N (%)	Anthropic, n/N (%)
Material omission	0/42 (0.0)	0/42 (0.0)	0/42 (0.0)
Inappropriate / false reassurance	4/42 (9.5)	3/42 (7.1)	3/42 (7.1)
Inappropriate alarm	0/42 (0.0)	0/42 (0.0)	0/42 (0.0)
Unsupported diagnosis / certainty	2/42 (4.8)	0/42 (0.0)	0/42 (0.0)
Directive overreach / treatment	0/42 (0.0)	1/42 (2.4)	0/42 (0.0)
Uncertainty / restraint failure	2/42 (4.8)	1/42 (2.4)	0/42 (0.0)
Safety-relevant concern	3/45 (6.7)	3/45 (6.7)	3/45 (6.7)

Case 06 indicators were unavailable for nine reports, leaving 42 eligible reports per model for six indicators. Safety-relevant concern was available for all 45 per model. Indicators can overlap and do not define the primary endpoint; zero material-omission flags do not imply absence of every clinically relevant omission. Disposition disagreement had no source field.

Table S4. Judge conformance and overall classification agreement

Judge family	Scheduled	Valid, n (%)	Failed	Exact, n/N (%)	Within one, n/N (%)
Google	135	111 (82.2)	24	70/111 (63.1)	106/111 (95.5)
OpenAI	135	135 (100.0)	0	98/135 (72.6)	131/135 (97.0)
Anthropic	135	76 (56.3)	59	53/76 (69.7)	75/76 (98.7)
All	405	322 (79.5)	83	221/322 (68.6)	312/322 (96.9)

Conformance denominators are scheduled cells. Agreement denominators are valid cells only. Missing judge observations were not interpreted as negative findings or clinical failures.

Table S5. Available, complete-three-judge, and self/cross summaries

Group	Valid cells	Exact agreement	Weighted κ
All available cells	322	221/322 (68.63%)	0.578
Complete-three-judge reports	183	127/183 (69.40%)	0.391
Self judgments	113	73/113 (64.60%)	0.496
Cross-model judgments	209	148/209 (70.81%)	0.616

The complete subset contains 61 reports and zero potentially unsafe final reports. Self/cross comparisons are exploratory and have no reported contrast interval.

Table S5 (continued). Material-failure agreement with the governed reference

Metric	Available cells	Complete-three-judge subset
True positive / false negative	24 / 7	2 / 4
False positive / true negative	29 / 262	11 / 166
Sensitivity	24/31 (77.42%)	2/6 (33.33%)
Specificity	262/291 (90.03%)	166/177 (93.79%)
Positive predictive value	24/53 (45.28%)	2/13 (15.38%)
Negative predictive value	262/269 (97.40%)	166/170 (97.65%)

Denominators are judge cells, with correlated observations within cases and reports. These metrics do not measure clinical outcome accuracy. The reference was partly shaped by judge-triggered physician adjudication.

Table S6. Domain-level judge agreement, all 322 valid cells

Domain	Exact, n/N (%)	Within one, n/N (%)	Mean absolute difference	Mean signed difference
Factual / clinical accuracy	181/322 (56.2)	316/322 (98.1)	0.460	+0.317
Principal finding prioritization	194/322 (60.2)	314/322 (97.5)	0.429	+0.379
Risk / disposition	182/322 (56.5)	314/322 (97.5)	0.463	+0.165
Education / pattern interpretation	183/322 (56.8)	308/322 (95.7)	0.494	+0.388
Uncertainty / restraint	195/322 (60.6)	321/322 (99.7)	0.398	+0.311
Proportionality / reassurance	183/322 (56.8)	312/322 (96.9)	0.475	+0.301
Patient communication	192/322 (59.6)	311/322 (96.6)	0.444	+0.301

Differences are judge minus final governed score on the 0–4 scale. Positive signed values indicate higher judge scores on average. They do not identify a causal judge bias. Family-specific domain outputs are retained in the handoff CSV; no cluster intervals were reported.

Table S7. Case 07 exclusion sensitivity analysis

Estimate	Reports	Rate (%) or difference (pp)	95% CI
Google	6/42	14.29	2.38 to 30.95
OpenAI	4/42	9.52	0.00 to 26.19
Anthropic	6/42	14.29	0.00 to 35.71
Google – OpenAI	—	+4.76	0.00 to 11.90
Google – Anthropic	—	+0.00	–11.90 to 9.52
OpenAI – Anthropic	—	–4.76	–14.29 to 0.00

Fourteen cases, 126 reports. Intervals are the retained paired case-cluster percentile bootstrap output. Exclusion addresses the conservative masking caveat; it does not reverse earlier exposure.

Table S7 (continued). GEE sensitivity-only coefficients

Term	Coefficient (log odds)	Standard error
Intercept (Google)	–1.872	0.638
OpenAI versus Google	–0.455	0.403
Anthropic versus Google	0.000	0.505

Binomial-logit GEE, exchangeable case correlation, 15 clusters, Mancl–DeRouen bias-reduced covariance. Converged; finite output; covariance rank 3; design rank 3 recovered in the supplement; no warnings or fallback. Both outcomes occurred in each generator. Iteration history and a full quasi-separation screen were unavailable. No new coefficient intervals or p-values were calculated.

Table S8. Locked-to-final overall classification transitions

Locked primary category	Final: unsafe	Final: deficient	Final: acceptable	Final: excellent
Potentially unsafe	9	0	0	0
Materially deficient	0	7	6	0
Acceptable	0	0	26	15
Excellent	0	0	19	53

All 135 reports are represented; 91 triggered physician review and 44 were nontriggered. There were 40 classification changes. “Unsafe” and “deficient” abbreviate potentially unsafe and materially deficient.

Table S8 (continued). Domain-score changes among 91 adjudicated reports

Domain	Increased	Decreased	Unchanged
Factual / clinical accuracy	2	0	89
Principal finding prioritization	2	0	89
Risk / disposition	1	3	87
Education / pattern interpretation	0	0	91
Uncertainty / restraint	1	2	88
Proportionality / reassurance	10	2	79
Patient communication	0	0	91

Twenty-three domain changes affected 17 reports. Counts refer to authorized changes in the adjudication overlay, not independent repeat ratings.

Table S9. Case-level judge material-failure contingency counts

Case	Valid	Failed	TP	TN	FP	FN
06	22	5	4	9	2	7
07	22	5	0	17	5	0
08	24	3	0	18	6	0
09	25	2	0	25	0	0
10	20	7	0	20	0	0
11	21	6	0	17	4	0
12	25	2	0	23	2	0
13	20	7	2	17	1	0
14	25	2	0	25	0	0
15	19	8	0	19	0	0
16	21	6	2	18	1	0
17	18	9	0	13	5	0
18	22	5	0	21	1	0
19	22	5	0	20	2	0
20	16	11	16	0	0	0

Each case had 27 scheduled judge cells. TP, true positive; TN, true negative; FP, false positive; FN, false negative, relative to the final governed report assessment. Cases, reports, and judge cells are different units. Family-specific counts remain in the handoff CSV.