LAIasBench: An Agent-Centric Benchmark for Radiology Finding-to-Report Generation

Natan Paraiso Ribeiro[1,*]
Petrus Paraiso Ribeiro[1]
Francisco Akira[1]
Stephanie Alba Herrera[1]
Raquel Moreno[1]
[1]Laudos.AI, Sao Paulo, Brazil
[*]Corresponding author: natan@laudos.ai

May 2026

Abstract

Radiology report generation is an end-to-end text-generation system problem. The clinically relevant question is not whether a language model can write fluent prose, see an image, or make a de novo imaging diagnosis, but whether a complete reporting workflow can convert provided findings into a faithful, structured, safe radiology report. We introduce LAIasBench, a CLI-first benchmark and evaluation harness for finding-to-report generation in Brazilian Portuguese, with cross-locale support for American English. The paper has two goals: to define the benchmark protocol and to report initial reference results on locked suites. LAIasBench evaluates executable systems rather than model names alone; it records frozen outputs, applies deterministic clinical safety gates, separates product agents from mini-agent scaffolds and raw model calls, and reports a weighted clinical report score as the primary public metric. The scoring rubric targets critical finding preservation, report quality, terminology, guideline/classification handling, and fidelity of title, sections, laterality, measurements, and evidence metadata. A private 40-case daily regression suite is sampled from a synthetic 65,812-report corpus generated from approximately 400 extractive seed reports through sentence-level finding extraction, finding-to-exam linking, modality/region stratification, and randomized compatible finding-set recombination. Following modern benchmark practice, the public protocol now separates task construction, locked artifacts, judge calibration, deterministic gates, cost/latency reporting, failure analysis, and evidence boundaries. The initial deterministic reference table is treated as calibration evidence; the main comparison table is regenerated only from frozen outputs under the LLM-adjudicated scoring mode. These results do not establish prospective clinical safety or market superiority.

Keywords: radiology report generation; benchmark; agent evaluation; clinical NLP; Portuguese radiology

1 Motivation

Radiology reports fail in ways that generic text benchmarks do not capture. A generated report can sound professional while omitting a pulmonary embolism, changing a measurement, adding unsupported contrast language, losing laterality, or moving an urgent finding into vague boilerplate. Brazilian radiology adds local section names, modality vocabulary, CBR-style terminology, and RIS-compatible HTML constraints.

LAIasBench evaluates the practical task faced by radiologists using AI-assisted reporting:

Input: examination descriptor plus concise user findings.
Output: complete radiology report using only `<center>`, `<b>`, and `<br>`.

The benchmark intentionally evaluates the complete reporting system. A product result depends on case intake, evidence use, report construction, verification, formatting, and recovery behavior. A raw foundation model can be useful as a baseline, but it is not the same object as a production reporting agent.

The contribution is therefore not a new report-generation model. LAIasBench contributes an executable benchmark protocol for a narrower but operationally important task: converting already-supplied radiological findings into a complete report. This distinction matters. Image interpretation benchmarks evaluate perception and diagnosis. LAIasBench evaluates downstream reporting fidelity: whether the system preserves the findings it was given, formats them as a usable report, applies relevant reporting conventions, and avoids unsafe transformation errors.

The first public results should be read as a calibration study. They demonstrate that the harness detects known failure classes and that different systems fail differently. They are not a claim that the public suite is large enough, adjudicated enough, or externally validated enough to settle clinical performance.

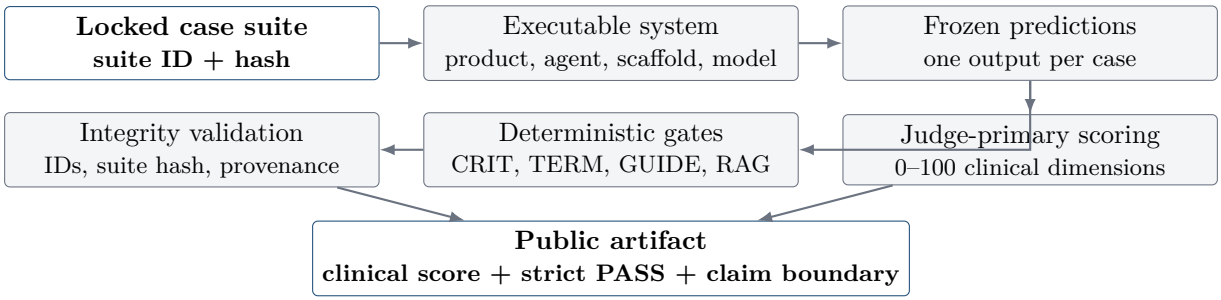

Figure 1: LAIasBench evaluates frozen outputs from executable systems. Clinical scoring, deterministic gates, and integrity checks all have to agree before a result becomes a public benchmark artifact.

2 Relation to Prior Work

Prior radiology report-quality work emphasizes that a useful report must be complete, clinically clear, actionable, and accurate in its key attributes. Human rating scales evaluate whether reports are understandable and clinically adequate [1]. Data-driven quality-assessment work shows that report evaluation can be decomposed into report-level and finding-level attributes [2]. Newer automated frameworks such as CLEAR, S-RRG-Bench, LUNGUAGE/LUNGUAGESCORE, and phrasal-grounding approaches evaluate clinical entities and attributes, including location, laterality, severity, relationship structure, temporal consistency, and factual grounding [3, 8, 7, 4].

LAIasBench adopts those ideas but changes the unit of evaluation. It does not grade a radiologist's diagnostic interpretation and it does not ask a model to infer findings from pixels. It grades a system's ability to transform supplied findings into a report without losing, inventing, or structurally corrupting the clinically relevant content. This makes the benchmark useful for product regression and system comparison, while leaving prospective diagnostic validation to separate studies.

The benchmark structure follows lessons from recent agent and LLM evaluation work. Terminal-Bench frames each item as a realistic task with a harness, oracle solution, verification tests, human review, cost tracking, and failure taxonomy [5]. MT-Bench and Chatbot Arena show why LLM-as-judge can be useful for open-ended outputs, but also why judge bias, verbosity bias, position effects, and calibration must be reported [6]. LAIasBench therefore uses LLM adjudication for report-quality scoring while keeping deterministic clinical gates for safety-critical failures and separating judged runs from deterministic calibration runs.

3 Design Principles

LAIasBench uses six design principles.

1. **Agent-first evaluation.** The submitted object is an executable reporting system. It may be a commercial product, internal lab agent, local script, or submission wrapper that satisfies the benchmark contract.
2. **Task-level verification.** Each case is treated as a locked task with clinical expectations and verification checks. A case is solved only when the output passes the required gates.
3. **Score-first reporting with strict gates.** Leaderboards foreground clinical report score, while strict PASS/error rates explain clinically decisive gate failures.
4. **Comparable tracks.** Product agents, custom agents, mini-agent scaffolds, and raw model calls are grouped separately. Different scaffolds or judging modes do not compete in one rank.
5. **Implementation privacy.** Public artifacts expose suite hash, public label, system class, validation status, and scores. They do not reveal proprietary implementation details or raw private case identifiers.
6. **Honest evidence boundaries.** The benchmark distinguishes engineering regression, public reference testing, and clinical validation. A green benchmark run is not a clinical safety claim unless radiologist-adjudicated validation is present for the same suite hash.

4 Benchmark Task

Each case contains an examination descriptor, concise findings supplied by the user, and optional structured expectations such as gold findings, critical findings, guideline expectations, retrieval judgments, patient context, and difficulty. The submitted system receives case JSON on standard input and returns either raw HTML or a JSON object containing `html` plus allowed metadata.

Operational failures are failures. If a command times out, exits, returns empty output, or explicitly marks `operationalFailure`, the case receives a FAIL verdict and zero dimension scores. This prevents stalled workflows from appearing as partially correct reports.

Edge-case policy. The benchmark treats clinically meaningful edge cases as first-class tasks rather than formatting noise. Current checks cover modality drift, contrast drift, laterality and spinal-level drift, section contradictions, negation drift, measurement mutation, guideline-classification mismatch, unsupported reassuring boilerplate in abnormal studies, malformed output, duplicate case IDs, and missing submissions. These cases explain why LAIasBench reports both clinical score and strict gate outcomes: fluent prose can still fail if it violates a clinically decisive gate.

5 Dataset Construction

The private LAIas daily split is generated from a local synthetic 65,812-row report corpus. The corpus was not copied from real patient reports. It was generated from approximately 400 extractive seed reports by splitting report text into sentence-level findings, linking each finding back to exam type, method, anatomy, and clinical concept, and then recombining randomized sets of compatible findings into coherent synthetic report-generation tasks. For example, a generated case may sample N compatible findings within a modality-region bucket and require the system to produce a complete report that preserves those findings, applies guideline classification when clinically relevant, and avoids unsupported normality or modality drift. The source CSV is not committed.

The committed daily suite is intentionally capped at 40 cases. Its purpose is fast daily product regression and monitoring. It is not a public leaderboard split and must not be exposed through public pages.

The public reference suite is manually curated from the same task definition rather than from the private daily split. Its purpose is reproducible comparison and failure analysis, not exhaustive coverage of radiology. It intentionally includes cases that stress modality/title preservation, laterality, measurement handling, section order, guideline labels, and abnormal-to-normal drift.

Dataset audit. The private corpus builder records an engineering audit before suite material is used for monitoring. In the current snapshot, 65,812 synthetic source rows are filtered to 2,433 eligible rows after modality, length, extractability, and deterministic privacy-pattern filters. The selected 40-case daily monitor contains zero deterministic privacy hits and covers seven modality-region buckets. The audit score is a dataset-readiness score for engineering use, not a clinical-validity score. Public reference cases are curated separately and identified only through locked public case IDs and suite hashes.

Suite	Locale	Cases	Purpose
reference-public.pt-BR	pt-BR	49	Primary public reference benchmark
reference-public.en-US	en-US	13	Cross-locale validation
lite-public.pt-BR	pt-BR	10	Fast iteration
verified-public.mixed	mixed	10	Harness integrity
laiasbench-daily.pt-BR	pt-BR	40	Private LAIas daily monitor

Table 1: Core LAIasBench suites. The daily suite is private and capped for operational monitoring.

6 Verification Dimensions

CRIT: Critical finding preservation. CRIT checks whether urgent or high-severity findings survive generation. It penalizes missed critical labels, unsafe negation, abnormal-to-normal drift, forbidden reassuring boilerplate, and contrast mismatches when the source case does not support contrast language.

QUAL: Clinical quality. QUAL checks clinical report quality rather than literal sentence overlap. The rubric follows published report-quality themes: completeness of relevant findings, clinically useful interpretation, clear organization, appropriate certainty, responsible recommendation when

needed, and factual preservation of attributes such as location, laterality, severity, and measurement [1, 2, 3, 4]. It compares generated content against structured gold findings when available, falls back to reference-report similarity where appropriate, and penalizes unrelated or unsupported report content. The score intentionally allows clinically acceptable paraphrase; it should not punish a usable report for failing to reproduce seed wording.

TERM: Terminology. TERM enforces locale and modality-specific wording, including section naming, forbidden openers, modality vocabulary, and terms that degrade professional report quality.

GUIDE: Guideline and coverage. GUIDE checks expected anatomical coverage and guideline-oriented concepts. It supports structured expectations for BI-RADS, TI-RADS, LI-RADS, PI-RADS, Bosniak, Fleischner, and Lung-RADS when those labels are present in the case data.

RAG: Fidelity and evidence preservation. RAG is a report-fidelity dimension. It checks title tokens, section order, laterality, spinal levels, measurements, report-order preservation, and optional evidence metadata. This catches failures introduced by intermediate evidence handling, condensation, or formatting layers.

7 Scoring and Ranking

For each dimension, evaluators return a 0–100 score and a verdict. LAIasBench supports two scoring modes, and the mode is part of the run manifest and comparable key. In `conservative-min` mode, deterministic evaluators and an optional judge both score each dimension; the combined score is the lower of the two. In `judge-primary` mode, an LLM judge assigns the primary 0–100 clinical score for each dimension when judge output is available; deterministic checks remain active as safety gates and can still force a FAIL verdict for critical clinical errors, malformed output, missing cases, or operational failure.

The public paper uses `judge-primary` for open-ended report quality, because a purely deterministic checklist compresses clinically different reports into overly similar scores. The deterministic layer is retained as a guardrail, not as the sole definition of report quality.

Judge protocol. The judge prompt is adversarial: it asks the judge to verify every material claim against the supplied examination and findings, penalize hallucinated findings, omitted findings, wrong modality terminology, wrong sectioning, and conclusion drift, and return machine-readable 0–100 scores for the five dimensions. A run can hide the commercial judge implementation in public metadata while still publishing a stable public judge label. This protects provider configuration while making comparable keys stable. The judge cannot call the evaluated product or regenerate predictions; it only scores frozen report artifacts. Runs may also include a per-run canary token so suspicious leakage of benchmark identifiers or suite metadata can be flagged as contamination.

The primary public metric is the weighted clinical report score:

$$\text{clinical score} = \text{weighted dimension score}_{0..100}.$$

A PASS requires the case to clear the clinical gates for the configured policy. PARTIAL and FAIL are useful for diagnosis, but only PASS counts as solved. The weighted average, per-dimension averages, non-fail rate, latency, and cost explain why a system fails and where to improve. Any

Dimension	Weight
CRIT	30%
QUAL	25%
TERM	20%
GUIDE	15%
RAG	10%

Table 2: Default dimension weights.

deterministic critical failure can force a FAIL verdict even when the weighted score is high; this prevents polished but unsafe reports from counting as solved.

The weighted score is diagnostic, not dispositive. A score near 70 or 80 is not automatically a pass, and a case with a high weighted score can still fail when the failed dimension is clinically decisive. This is intentional: missing an urgent finding, changing the modality, losing a critical measurement, or contradicting the conclusion is not equivalent to a minor wording error.

This design avoids calling strict PASS "accuracy". In LAIasBench, strict PASS is an all-gates-cleared rate. It is closer to a task success or error-free-case rate than to diagnostic accuracy. The primary quantity for report quality is the weighted clinical score; the gate rate explains how often serious deterministic failure conditions remain.

Uncertainty and paired comparisons. LAIasBench includes deterministic bootstrap confidence intervals for suite means and paired-comparison utilities for before/after product changes. These statistics are not used to manufacture significance claims from a small suite. They are intended to keep future leaderboard claims tied to the same locked case set: a product update should be compared against the previous artifact case-by-case under the same suite hash, scoring mode, and comparable key. Public superiority claims are therefore disallowed unless the comparison is paired on the same locked suite.

8 Tracks and Submission Model

LAIasBench separates full product or custom reporting agents, mini-agent scaffolds, and raw model calls. Leaderboard comparability depends on benchmark version, suite ID, suite hash, locale, track, comparison class, scaffold class, evaluated entity, and judging mode. Public rows report the company, system, or agent name and the system class. Raw model baselines are not ranked as equivalent to full product agents.

Any entity can submit a system run by choosing a locked public suite, generating one report per case through its own workflow, validating completeness, scoring the artifact with the harness, and publishing only the comparable run fields needed for leaderboard review. The internal implementation remains private. This lets companies, radiology groups, and research teams compare real reporting systems without exposing proprietary workflow details.

Before leaderboard review, every submitted artifact must pass a completeness and provenance check: each public case ID appears exactly once, no output is empty, the suite hash matches the locked suite, the declared track matches the evaluated object, and judged runs evaluate frozen predictions rather than generating and judging in the same step. Product-agent reruns also require target-readiness verification so protected preview pages, missing authentication, and HTML login responses cannot become benchmark artifacts. The validator can return raw case-level diagnostics

to the submitter locally, but public leaderboard artifacts publish only eligibility, validation counts, and sanitized ineligibility reasons. This prevents missing-ID lists, duplicate-ID lists, source file paths, provider labels, private route names, or hidden judge configuration from becoming part of the public ranking surface.

Run JSON is not accepted as a self-authenticating score claim. Before leaderboard or comparison output is generated, the CLI recomputes each case overall from the stored dimension scores, recomputes suite summaries, verifies verdict counts and per-dimension averages, rebuilds the comparable key, and checks the manifest suite hash against the local locked suite. Edited summaries, stale comparable keys, changed case sets, or mismatched suite hashes fail integrity validation.

9 Artifact and Provenance Model

Each public-quality run has four layers of evidence. First, a locked suite manifest defines case IDs, locale, visibility, and suite hash. Second, a prediction artifact records one frozen generated report per case plus allowed metadata, latency, cost, and operational trace fields. Third, the evaluator produces case-level dimension scores, deterministic checks, judge output when enabled, verdicts, and gate reasons. Fourth, leaderboard and comparison commands recompute public summaries from those case-level artifacts rather than trusting edited aggregate fields. The public artifact identifier for this preprint package is `LAIASBENCH-PUBLIC-2026-05-02-BF78-A309`; it is intentionally non-secret and can be used to identify redistributed copies of the public site, PDF, metadata, or source package.

Layer	Public role	Hidden or private material
Suite manifest	Case count, locale, suite hash	Private daily case content
Predictions	Frozen report outputs for scoring	Product prompts and routes
Evaluation	Scores, verdicts, gates, latency, cost	Hidden judge implementation
Leaderboard	Comparable rows and sanitized validation	Raw private ID lists and credentials

Table 3: Evidence layers used to keep benchmark claims reproducible while protecting private implementation details.

This artifact model is important for clinical AI evaluation because common failure modes are operational rather than linguistic. A protected preview page, missing authentication token, timeout, HTML login response, empty output, duplicate case ID, stale suite hash, or manually edited summary can otherwise masquerade as a model-quality result. LAIasBench treats those as validation or operational failures before clinical scoring is interpreted.

10 LAIas Reference Run

The LAIas reference run targets the real product reporting flow through the same report-generation path used to produce product reports. The score is produced from frozen predictions: first the product flow generates reports and writes a predictions artifact, then the benchmark evaluates those frozen outputs. The evaluator observes only public case inputs, returned report HTML, allowed metadata, latency, cost, and run manifest. It does not require public disclosure of private implementation details.

Canonical judged LAIas scores are produced from frozen predictions. The judge does not receive private LAIas instructions, cannot call the product flow, and cannot influence generation. In the current protocol, `judge-primary` provides the report-quality score and deterministic critical checks remain gates. The previous deterministic reference table is deprecated as a headline result because it compressed report quality and made clinically different systems appear too close. It remains useful only as a harness calibration artifact. The replacement table is generated from immutable outputs under the LLM-adjudicated protocol and reports only systems inside the declared comparison scope.

System class	Track	Cases	Scoring mode	Clinical score	Strict PASS	Non-fail
LAIas product agent	agent	49	judge-primary	89.5%	81.6%	85.7%

Table 4: Current locked pt-BR public-suite reference result. Product-agent, mini-agent, and raw-model rows are tracked separately; this table reports only completed judged-frozen runs inside the declared comparison scope.

Metric	CRIT	QUAL	TERM	GUIDE	RAG
Average score	89.1%	85.0%	95.6%	88.4%	91.4%

Table 5: Dimension-level means for the locked LAIas product-agent judged-frozen run.

Under the LLM-adjudicated run, LAIas product-agent outputs score 89.5% clinically on the locked 49-case pt-BR public suite. The strict PASS gate rate is 81.6% and the non-fail rate is 85.7%. The artifact validates all 49 expected public case IDs with no missing IDs, duplicate IDs, extra IDs, empty outputs, or validator errors. The benchmark card for this run reports actionability as actionable, a benchmark quality score of 100/100, rich gold coverage of 100.0%, seven modality-region buckets, and a canary-enabled frozen prediction workflow. These are not diagnostic accuracy numbers. They mean that the generated reports are usually clinically usable under the rubric, while nine cases still need case-level audit: two PARTIAL cases and seven FAIL cases, mostly driven by critical-finding or fidelity gates.

These first results should be interpreted as a benchmark proposal with initial calibration, not as a final product-ranking paper. The intended comparison scope is LAIas product-agent outputs, future mini-agent baselines, and raw-model baselines on the same locked cases, scored by the same pinned judge and constrained by the same deterministic safety gates. Rows are published only after frozen outputs pass integrity validation; pending or partially rerun baselines are not included in the headline table.

The current reference result intentionally remains conservative. Two failures were audited rather than overwritten: one case required correction of public case metadata before rescoring frozen outputs, and one exposed a product-agent modality-classification error that must be rerun only after the fixed evaluated system is available. Scores change only when a new frozen prediction artifact is generated from the fixed system and rescored against the same locked suite hash.

Failure interpretation. The observed failure profile is useful because it points to concrete engineering work rather than a single aggregate defect. CRIT and RAG failures imply that the next product improvements should prioritize preservation of critical findings, modality/title classification, conclusion consistency, laterality, and measurement fidelity. TERM is currently the strongest dimension, suggesting that local Portuguese style and section vocabulary are less limiting than

factual preservation. QUAL remains below TERM because clinically acceptable paraphrase can still lose or blur a supplied finding. This decomposition is the main reason LAIasBench publishes dimension scores beside strict PASS.

10.1 Mini-Agent Baseline Matrix

To verify that the harness separates product-agent results from simpler report-generation wrappers, we ran a mini-agent baseline matrix on the same 49 locked pt-BR public cases. These rows use the canonical mini-agent scaffold and are not merged with the LAIas product-agent result. Tencent Hy3 Preview was also selected for the five-system screen but is reported only in the three-case screening artifact because its full-suite endpoint stalled during the run.

System	Cases	Clinical score	Strict PASS	Non-fail
inclusionAI Ling-2.6-1T	49	93.5%	73.5%	73.5%
GPT-OSS-120B	49	90.9%	46.9%	46.9%
NVIDIA Nemotron 3 Super 120B-A12B	49	87.0%	22.4%	22.4%
Poolside Laguna M.1	49	86.9%	2.0%	2.0%

Table 6: Completed mini-agent baselines on all 49 locked pt-BR public cases. These rows are not comparable to the LAIas product-agent reference run.

The baseline matrix illustrates why LAIasBench reports both clinical score and strict PASS gates. Some mini-agent wrappers achieve high judged clinical scores while still failing many deterministic gates. The right interpretation is diagnostic: these systems can produce plausible report text, but wrapper-level output still needs gate-level failure analysis before any product-comparable claim.

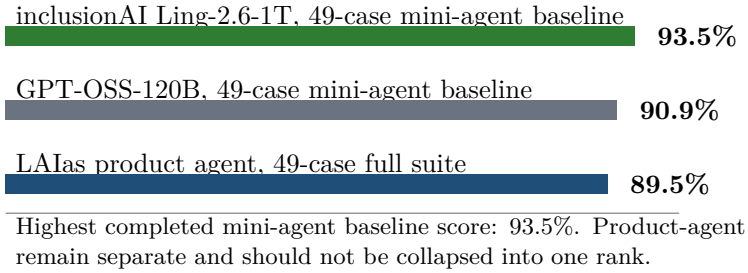

Figure 2: Score context for current artifacts. Mini-agent baselines diagnose wrapper behavior and are not ranked against the full-suite product-agent run.

11 Public Reporting and Claim Discipline

The public leaderboard should report exact benchmark version, suite hash, case count, locale, modality-region distribution when available, evaluated entity, system type, track, scaffold class, scoring mode, clinical score, strict PASS gate rate, non-fail rate, per-dimension scores, latency, cost, bootstrap intervals when available, paired-comparison statistics for claimed improvements, failure taxonomy, validation counts, sanitized ineligibility reasons, and privacy-filter limitations. It

should not expose private prompts, product routes, credentials, private file paths, raw validation ID lists, raw private case content, hidden judge configuration, or provider internals.

LAIasBench also separates acceptable engineering language from stronger clinical claims. Acceptable public wording includes "locked public reference suite", "private daily regression monitor", "synthetic source corpus derived from extractive seeds", "deterministic privacy-pattern filtering", and "heuristically derived gold findings". The benchmark should not be described as radiologist-adjudicated, prospectively validated, clinically safe, or clinically superior unless the corresponding locked adjudication and paired-validation gates have passed for the exact suite hash. Public distribution telemetry is limited to page-view and artifact-click events with hashed IP and user-agent fields, optional webhook forwarding, and a public artifact ID. It is intended to detect copying, redistribution, and unusual download patterns without exposing private benchmark cases or collecting clinical content.

12 Implementation Status

The repository includes a CLI suite runner, single-case runner, submission validation, leaderboard, comparison commands, generic execution interfaces for product agents, custom agents, mini-agent baselines, and raw model baselines, command timeout handling, operational-failure scoring, deterministic evaluators for CRIT, QUAL, GUIDE, RAG, structural terminology checks, optional LLM judge integration, seeded bootstrap confidence intervals, paired-comparison utilities, daily LAIas product-agent evaluation wrapper, private 40-case daily suite, public reference suites, benchmark-card audit, operator reporting artifacts, and automated tests for scoring, extraction, guidelines, retrieval, private-suite integrity, statistics, and adversarial edge cases.

The implementation also includes governance gates for two common reporting failures. Benchmark-card audits prevent mock, dry-run, all-fail, or non-frozen judged runs from becoming product-quality claims. Radiologist-adjudicated validation cannot be claimed unless a private adjudication artifact passes a multi-reviewer agreement gate for the exact suite hash. This status is an engineering and benchmark-readiness claim; it does not yet claim clinical validation.

13 Limitations and Roadmap

Current labels are heuristically derived from report text. Critical labels use conservative keyword heuristics and can miss variants. Privacy filtering is deterministic pattern filtering, not formal de-identification certification. The private 40-case split is for daily regression, not final external validation. Current evidence does not establish prospective clinical safety. Radiologist adjudication has not been recorded for the current private split. Additional validity threats are tracked rather than hidden: the public reference suite is small by design, heuristic gold extraction can under-label subtle findings or over-penalize acceptable paraphrases, product-agent submissions can be sensitive to target availability and transient failures, and raw model baselines answer a different question than full product agents.

Before stronger public claims, the benchmark should add a locked radiologist-adjudicated subset, inter-rater agreement for critical and quality labels, judge calibration against radiologist review, paired before/after product-flow comparisons, a locked external validation split not used for system iteration, manuscript tables generated directly from immutable run artifacts, and an explicit protocol for private cases, contamination checks, and data leakage prevention. Radiologist adjudication is now specified as a gate, not a claim: the planned locked subset requires at least two

board-certified radiologist reviewers, signed labels per case, agreement reporting, privacy screening of notes, and exact suite-hash binding.

14 Conclusion

LAIasBench evaluates radiology reporting systems as executable agents. It asks whether a complete workflow can transform provided findings into clinically relevant reports under reproducible verification, not whether an isolated model can see images or produce plausible prose. The current implementation supports daily engineering regression and public reference testing with score-first reporting, strict comparability groups, operational-failure handling, and privacy-preserving submissions. Stronger clinical claims require radiologist-adjudicated labels, locked validation runs, and continued separation between raw-model baselines and full product agents.

References

[1] Chengwu Yang, Claudia J. Kasales, Tao Ouyang, Christine M. Peterson, Nabeel I. Sarwani, Rafel Tappouni, and Michael Bruno. A succinct rating scale for radiology report quality. *SAGE Open Medicine*, 2:2050312114563101, 2014.

[2] William Hsu, Simon X. Han, Corey W. Arnold, and Alex A. T. Bui. A data-driven approach for quality assessment of radiologic interpretations. *Journal of the American Medical Informatics Association*, 23(e1):e152–e156, 2016.

[3] Yuyang Jiang et al. CLEAR: A clinically-grounded tabular framework for radiology report evaluation. *arXiv preprint arXiv:2505.16325*, 2025.

[4] Razi Mahmood et al. Evaluating automated radiology report quality through fine-grained phrasal grounding of clinical findings. *arXiv preprint arXiv:2412.01031*, 2024.

[5] Mike A. Merrill et al. Terminal-Bench: Benchmarking agents on hard, realistic tasks in command line interfaces. *arXiv preprint arXiv:2601.11868*, 2026.

[6] Lianmin Zheng et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *NeurIPS Datasets and Benchmarks*, 2023.

[7] Yingshu Li et al. S-RRG-Bench: Structured radiology report generation with fine-grained evaluation framework. *Meta-Radiology*, 3(4):100171, 2025.

[8] Jong Hak Moon et al. Lunguage: A benchmark for structured and sequential chest X-ray interpretation. *arXiv preprint arXiv:2505.21190*, 2025.