Radiologist Adjudication and Inter-Rater Agreement for Automated Radiology Report Scoring

A Locked Study Protocol

Natan Paraiso Ribeiro, MD Petrus Paraiso Ribeiro, MD Stephanie Alba Herrera, MD
Francisco Akira, MD Raquel Moreno, MD

Laudos.AI Research • research@laudos.ai

Protocol v1.0 – 2026 (pre-registration; no results reported)

Abstract

LAIBench scores finding-text-to-report generation with deterministic clinical checks, a hard critical-finding gate, and an optional frozen judge. To date its labels are an internal data-quality process, not an independent third-party validation. This protocol pre-registers, in full, a blinded multi-reader study to estimate (i) inter-rater agreement among board-certified radiologists on the report-faithfulness constructs LAIBench measures, and (ii) the agreement between the LAIBench score and a radiologist consensus. All design parameters — sample size, stratification, reader panel, rating rubric, disagreement rule, agreement statistics, and acceptance thresholds — are fixed here, before any data is read. **No results are reported.**

1 Objectives

1. Estimate inter-rater reliability (Fleiss κ, Krippendorff α) among radiologists for five constructs: critical-finding preservation, clinically relevant omission, hallucinated/unsupported finding, factual contradiction, and impression actionability.

2. Estimate scorer–human agreement: correlation and rank concordance between the LAIBench per-case score (and its binary CRIT gate) and the radiologist consensus.

3. Identify systematic scorer failure modes (cases where the deterministic checks and the human panel disagree) to drive calibration.

2 Design

Sample size and stratification. A **locked subset of $N = 60$ report instances** is drawn from the controlled-eval suite and frozen by suite hash before reading begins. The subset is stratified by modality and over-samples the safety-critical extremes:

Modality stratum	Cases	of which critical-present
CT (chest/abdomen)	18	8
MRI (brain/spine)	12	5
Radiograph (chest)	12	4
Ultrasound (abdomen/thyroid)	10	2
Mammography	8	3
Total	**60**	**22**

In addition to the 22 critical-present cases, ≥ 12 are confirmed negative controls (no actionable finding) so that gate *specificity* is estimable. $N = 60$ is chosen for estimation, not hypothesis testing: for an observed κ in $[0.6, 0.8]$ it yields a 95% bootstrap CI half-width below ≈ 0.12, adequate to distinguish "substantial" from "poor" agreement, while remaining feasible for a 3-reader fully-crossed design ($60 \times 3 = 180$ reads).

Reader panel. Three board-certified radiologists (≥ 5 years post-residency), of whom **at least one is not affiliated with Laudos.AI**. All three read all 60 cases (fully crossed). Readers are blinded to the generating system's identity, to one another's ratings, and to the LAIBench score and gate.

Rating instrument. For each case each reader records, on a rubric agreed and frozen before reading, with a mandatory free-text justification per non-trivial judgment:

- **Critical-finding preservation** — binary: every critical finding present in the source is stated, and none is fabricated (PRESERVED / VIOLATED).
- **Clinically relevant omission** — binary (ABSENT/PRESENT).
- **Hallucinated / unsupported finding** — binary.
- **Factual contradiction** — binary.
- **Impression actionability** — 3-point ordinal (ACTIONABLE / PARTIAL / NOT ACTIONABLE).
- **Global judgment** — binary: "clinically acceptable report?" (YES/NO), the primary human reference label.

Disagreement-resolution rule. For each construct, the consensus label is the $\geq 2/3$ **reader majority**. Cases with a non-majority split, or any case flagged "VIOLATED" on critical-finding preservation by at least one reader, are escalated to an independent fourth adjudicating radiologist; the adjudicated label is then fixed by structured consensus discussion. Both the raw per-reader labels and the adjudicated consensus are retained.

3 Analysis plan

Inter-rater reliability. Report Fleiss κ and Krippendorff α per construct with 95% bootstrap confidence intervals (10,000 resamples over cases), interpreted on the Landis–Koch scale. Ordinal constructs use weighted (quadratic) κ.

Scorer–human agreement. Report Spearman ρ and Pearson r between the continuous LAIBench per-case score and the consensus (proportion of readers rating "acceptable"), plus Kendall τ rank

concordance. The binary LAIBench CRIT gate is cross-tabulated against the consensus critical-finding judgment in a 2×2 table, reporting **sensitivity, specificity, PPV, NPV** with Wilson 95% CIs.

Calibration. A reliability diagram bins the LAIBench score and plots binned score against the empirical probability of human "acceptable", summarized by Brier score and calibration slope/intercept.

Pre-specified acceptance thresholds. The scorer is declared to *track radiologists* only if **all** of the following hold on the locked subset:

1. inter-rater Fleiss $\kappa \geq 0.61$ on critical-finding preservation (otherwise the human reference is too noisy to validate against, and the study reports the human-agreement ceiling instead);
2. CRIT-gate **sensitivity** ≥ 0.95 against the consensus critical-finding label (missing a true critical is the unacceptable error);
3. scorer–consensus Spearman $\rho \geq 0.70$.

These thresholds are fixed here and will not be revised post-hoc; any deviation will be reported as a protocol amendment with its date.

4 Data management and governance

The locked subset is sealed by suite hash and **never used to tune or debug the scorer** — doing so would retire it as a validation holdout. Reader identities are pseudonymized; raw ratings, justifications, and the adjudicated consensus are stored as signed adjudication records conforming to the schema already implemented in the LAIBench repository (`src/adjudication.ts`), tied to the exact suite hash and scorer hash under test.

5 Limitations

This is a measurement study of agreement and calibration, not a clinical trial. It does not establish clinical safety, diagnostic accuracy, generalization, or regulatory fitness, and it is not image interpretation. It validates whether an automated *report-faithfulness* score agrees with radiologist judgment on a controlled set; external validity across sites and equipment is the subject of a separate hidden-holdout protocol. Until this study reports, the correct public claim for LAIBench remains *technical benchmark framework*, not *clinically validated benchmark*.