External Validation of Finding-to-Report Faithfulness Scoring Across Sites and Equipment

A Locked Hidden-Holdout 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

Public synthetic cases are useful for harness inspection but are weak evidence for clinical generalization: they are contaminable and not drawn from real reporting distributions. This protocol pre-registers, in full, an external-validation study for the LAIBench scorer and for the systems it ranks, using a *hidden holdout* that is never published and never used for tuning. All parameters — holdout construction, cross-site/cross-vendor stratification, frozen-submission and contamination-audit procedures, the generalization analysis, and the claim-acceptance rule — are fixed here. **No results are reported.**

1 Motivation

A score on a contaminable public set, or on a controlled set used during development, cannot support a strong external claim. LAIBench separates four tiers: **public-smoke** (contaminable, demo only), **controlled-eval** (gated, first-party, aggregate-only), **calibration-set** (scorer tuning only), and **hidden-holdout** (never seen, never tuned). Only the hidden-holdout can sustain a strong generalization claim. This protocol governs that tier.

2 Design

Holdout construction and size. A held-out set of $N \geq 150$ **report instances** is assembled to span ≥ 3 **institutions** and ≥ 2 **equipment vendors** per modality, fully disjoint from any case used for development, demonstration, or scorer calibration. The set is sealed (suite hash recorded), access-controlled, and versioned; each version is used **at most once per submitting system**. Per-stratum minimum cell size is ≥ 20 cases; strata with fewer cases are reported but excluded from per-stratum claims.

Stratification axis	Levels	Min cases / cell
Modality	CT, MRI, radiograph, US, mammography	20
Institution	≥ 3 distinct sites	—
Equipment vendor	≥ 2 per modality where available	—
Criticality	critical-present vs negative-control	20 each

Frozen submissions. The leaderboard accepts only runs produced by the official runner, or validated frozen predictions. Each submission is a JSONL that is **content-hashed, timestamped, and pinned to a scorer version**; resubmission against the same holdout version is not permitted. The submission hash, scorer hash, and suite hash are published with the result so any number can be re-derived.

Contamination audit (per run). Every run records and publishes:

- declared training-data cutoff, where known;
- the exact public prompt / scaffold id used;
- the holdout suite hash and scorer hash;
- a canary-token scan result;
- whether the model/provider could have seen the public repository;
- whether outputs echo case identifiers or benchmark boilerplate.

These reduce to a single published flag: **contamination scan = pass / fail / unknown**. A run flagged FAIL is not eligible for a ranked claim.

3 Analysis plan

Report scorer and system performance overall and per stratum (site, vendor, modality, criticality), each with 95% bootstrap CIs. Quantify the **generalization gap** as the difference between controlled-eval and hidden-holdout clinical score, with its CI. Where the radiologist-adjudication protocol is run on a holdout subset, estimate scorer–human agreement on *never-tuned* data (the strongest available evidence). Critical-finding gate sensitivity/specificity are reported on the holdout against the adjudicated label.

Pre-specified claim-acceptance rule. A generalization claim for a system is *supported* only if, on the sealed holdout: (i) **contamination scan = pass**; (ii) every claimed stratum has ≥ 20 cases; and (iii) the holdout clinical-score 95% CI does not fall more than **10 points** below the system's controlled-eval score (a larger drop is reported as a generalization failure, not smoothed over). Absent all three, the published claim is restricted to *first-party, controlled, aggregate-only.*

4 Governance and limitations

The golden rule is enforced operationally: any suite used for debugging or scorer tuning is retired as a holdout. This study measures generalization and contamination control; it does not by itself establish clinical safety or regulatory fitness, and it is not image interpretation. Until it reports, LAIBench claims remain first-party and aggregate, with external validation explicitly *pending.*