

PACKAGE E · FUTURE VALIDATION PLAN

R-BED Validation Plan

Studies that would test whether R-BED™ v2.1 improves evaluation and decisions

Plan version 2.1 | 29 October 2026 | **No study described here has been run. No result is reported.**

Status	Proposal. Every outcome, sample size and threshold below is to be fixed in a preregistered protocol before data are collected. Nothing in this document is evidence for R-BED.
Purpose	Convert the hypotheses H1–H7 (H7: consequence-aware monitoring allocation) into testable questions, so that the claim level of R-BED can move from “specified” toward “empirically supported” only on the strength of results.
Not in scope	Any NIST submission material. This plan is separate from the public framework and crosswalk.
Harness	A synthetic failure-injection harness in the reference implementation lets an independent researcher exercise the H7 analysis pipeline before real data exist. It is not evidence: its output depends only on invented parameters.

1 Hypotheses

ID	Hypothesis (each could be false)
H1	Trained reviewers who apply the SES rules agree on evidence status and on the high-consequence classification of the same record more than chance.
H2	Reconstructing documented incidents as SES records exposes the control, evidence or authorization gap that investigators later reported, more often than an unstructured review of the same material.
H3	Independent (I2 or I3) control challenge finds failures or control gaps that I0 self-assessment of the same system misses.
H4	Intervention-contrast estimates of control effectiveness predict containment observed later in monitored operation.
H5	Teams can complete a record at a cost they consider acceptable, and completing it changes at least one decision or safeguard they would otherwise not have changed.
H6	The authorization rules block decisions that expert reviewers independently judge to be unsupported, without blocking a large share of decisions they judge supported.
H7	(new in v2.1) Under a fixed monitoring-and-validation budget, allocating effort by R-BED’s ordinal consequence-aware tiers avoids more expected harm than allocation by observed failure frequency or by uniform allocation, in settings where severity and observed frequency diverge and where consequence is classified with adequate reliability; and R-BED does no better, or worse, in specified other settings.

2 Studies

2.1 S1 — Inter-rater reliability (H1)

Design. Reviewers independently complete the same set of SES records and assign, per material claim, an evidence status (four values), the high-consequence classification, and the independence level of each evidence item. **Materials.** Synthetic records with known planted defects (for example silent-zero residual risk, uncredited control, expired authorization), and, with permission, de-identified real records. **Analysis.** Chance-corrected agreement (Cohen's or Fleiss' kappa, or Krippendorff's alpha) with confidence intervals; agreement reported per field, because a single overall figure would hide weak fields. **Decision rule set in advance.** Fields with lower-bound agreement below a threshold fixed in the protocol are redefined before further use. **Threats.** Planted defects are easier than real ones; reviewers trained by the author may share the author's reading.

2.2 S2 — Retrospective incident reconstruction (H2)

Design. Select documented AI incidents from public databases and investigation reports by a stated sampling rule fixed before reading outcomes. Reconstructors see only the pre-incident description and build a record of the configured system, claims, authority grants, controls and evidence; a second team, blind to the investigation findings, scores whether the record flags the control or evidence gap reported afterwards. A comparison arm reviews the same pre-incident material without the framework. **Analysis.** Detection proportion per arm with intervals; paired comparison by incident. **Threats.** Hindsight bias, selection of incidents that were investigated in depth, and incidents whose cause was not a control gap. **Note on prior research.** The author's earlier incident research, if published, may be cited as motivation; its findings are not restated here and are not used as results for this plan.

2.3 S3 — Practitioner pilot (H5)

Design. Consenting teams complete a record for a real or realistic system. Measure time, omissions found by an independent reviewer, questions that the template could not answer, and whether any decision, control or test changed. **Analysis.** Descriptive, with qualitative coding of interviews; no claim of effectiveness from a pilot without a comparison. **Ethics.** Data handling and consent agreed with each organization; records protected where they describe security weaknesses.

2.4 S4 — Independence effect (H3)

Design. The same system is evaluated twice, once by the builders (I0) and once by an independent group (I2 or I3) using the same claims, then both findings are reconciled. Order and assignment are randomized across systems where possible. **Outcome.** Count and severity of failures or control gaps found by only one arm. **Analysis.** Paired difference with an interval; severity rated by a panel blind to the arm. **Threats.** Independent reviewers have different skills as well as independence; the design cannot separate the two without a third arm of builders from another team.

2.5 S5 — Prediction of control effectiveness (H4)

Design. For controls in monitored pilots, record the pre-registered intervention-contrast estimate and its conservative ratio before launch; compare with observed containment in operation. **Analysis.** Calibration of predicted against observed containment, with the number of exposures stated; rare events may not occur, in which case the study reports an upper bound only and does not claim validation. **Threats.** Operation differs from tests; adversaries adapt; controls change.

2.6 S6 — Rule audit (H6)

Design. Expert panels, blind to the rule outcome, rate whether a set of historical or synthetic Deploy and Restrict decisions were adequately supported; compare with the pass or fail result of the eight authorization rules. **Analysis.** Sensitivity and specificity of the rules against panel judgment with intervals, and a listing of disagreements to refine or remove rules. **Threat.** Panel judgment is not ground truth.

2.7 S7 — Consequence-aware monitoring allocation (H7)

Maturity: hypothesis [H]. Nothing in this subsection has been run.

Question. For a fixed budget of monitoring and independent-validation effort across a set of pathways, does the R-BED allocation (monitoring priority tiers by rules TR1–TR6, lexicographic funding, with a random audit floor) lead to less expected harm than frequency-based or uniform allocation, and under what conditions does it not?

Comparators (allocation policies, same budget, same detection and response technology).

- BASE-01.** *Uniform:* the same effort per pathway, or random choice among pathways when the budget cannot cover all.
- BASE-02.** *Performance-only (frequency-based):* effort by descending observed failure or error rate on the pathway.
- BASE-03.** *Exposure-only:* effort by descending number of exposures.
- BASE-04.** *Practitioner baseline:* the allocation that the host organization's existing risk function would make without R-BED, elicited before the study and sealed.
- BASE-05.** *R-BED tiered:* monitoring priority tiers assigned from ordinal profiles by a person trained in the rules who has not seen outcomes; allocation by the specified order with the audit floor.
- BASE-06.** *Oracle (reference only):* allocation by true expected loss. It is unattainable in practice and is shown to bound what any allocation could achieve; it is not a comparator for any claim.

Units. A pathway is the unit of allocation (a claim or pathway on a configured system, with a consequence class). Effort is counted in a stated, comparable unit (for example reviewer-hours plus engineering-hours); policies are held to the same total. Cost per tier is fixed in advance from the organization's own figures.

Designs. (a) *Controlled failure injection* (primary): failures of known kind and timing are injected into pathways in a sandbox or shadow deployment, so that every policy faces the same injected set; the outcome is injected failures detected, interventions completed in time, and containment before the specified harm. (b) *Randomized allocation in a real portfolio:* where an organization is willing, pathways are randomly assigned to be monitored under the R-BED or the baseline allocation within strata; this is the only design that supports a causal statement about real harm. (c) *Retrospective replay* of logged incidents under each policy: association and sensitivity analysis only. (d) *Simulation* with explicit parameters (the harness): pipeline verification and sensitivity only.

Failure injection. A pre-registered library of failure types per pathway (drift, tool misuse within the envelope, out-of-envelope action, cross-account disclosure, prompt injection into a tool call, silent control failure), each with a known onset time, a harm-onset window and a true severity assigned by an independent panel before injection. Injection times are randomized within the horizon. For agentic pathways the injections are trajectory-level (a hidden out-of-envelope step in a multi-step task); for non-agentic pathways they are input or data faults.

Scenarios (all pre-registered).

Scenario	Setting	Expected direction for R-BED
Aligned	Severity and consequence class track true expected loss; classification is reliable; exposures vary moderately.	Better than BASE-01 to BASE-03.
Misclassified	The consequence class assigned by the R-BED user is wrong for a material share of pathways (for example a high-severity pathway rated low).	Expected to do worse than performance-only if the error is large enough; the study estimates how large.
Frequency-dominated	Harm is driven by high-frequency, low-severity events; severe events are rare and already well controlled.	Expected to do worse than performance-only or exposure-only.
Stale evidence	Assurance ratings are out of date and the tier rules therefore under-monitor.	Uncertain; tests the evidence-age rule.
Adaptive adversary	An attacker shifts to the pathways with the least monitoring.	Uncertain; the audit floor is the only protection.
Low budget	Budget covers fewer pathways than there are highest-tier ones.	Uncertain; tests the lexicographic order.

Outcomes. *Primary:* expected harm avoided relative to BASE-02 in the Aligned scenario, in the stated severity units, measured as injected-failure containment weighted by independently assigned severity. *Secondary:* (i) detection rate and time to detection; (ii) verified interventions (detection that led to action in time) separately from detection alone; (iii) harm reduction (containment before harm), reported separately from (i) and (ii); (iv) effort spent and unfunded pathways; (v) false-positive interventions and their cost; (vi) incidents on low-tier pathways (the cost of exploration); (vii) tier assignment agreement among raters. **Detection, verified intervention and harm reduction are not interchangeable:** a policy can increase detection while reducing nothing.

Statistics. Paired comparison of policies on the same injected set; primary estimate the mean paired difference in severity-weighted harm with a bootstrap interval clustered by pathway; secondary rates with Clopper–Pearson bounds. Sample size is set by the precision target in the protocol (the harness provides a required-pairs calculation from a pilot variance; no number is stated here). The Aligned scenario has the pre-registered primary hypothesis; the other scenarios are estimated, not tested for superiority, and the multiplicity across them is declared.

Pre-registered decision criteria. H7 is *supported for a scenario* only if the lower bound of the interval for the primary difference exceeds zero by a margin set in the protocol and the effect is not removed by the sensitivity analyses below. H7 is *refuted for a scenario* if the upper bound lies below the margin. Otherwise it is inconclusive. H7 is reported as supported overall only if it holds in Aligned and the report states, for each of Misclassified and Frequency-dominated, the estimated difference with its interval, including when R-BED does worse. A finding that R-BED does worse in those scenarios is the expected outcome of the design, not a failure of the study; a finding that it does not do worse would be reported as surprising and checked for an error in the scenario.

Negative controls. (1) A permuted-tier policy that assigns the R-BED tiers at random to pathways: it should show no benefit; benefit under it indicates an artifact of the harness or the severity assignment. (2) A policy given the true classes but random within-tier order. (3) The oracle, to show the ceiling.

Ablations. R-BED without the audit floor; without the evidence-age rule; with the tier rules reversed; with consequence class only (no exposure); with a numeric score in place of the ordinal rules (to test the claim that ordinality suffices); with misclassification rates of 10%, 25% and 50%.

Sensitivity. Detection, timeliness and containment probabilities varied over stated grids; harm-onset window length; severity scale (the unit-free ranking is rank-correlated with the panel's); correlated bursts of failures that exhaust response capacity; failure injection independent of tier so that exploration is not credited by design.

Independence and bias controls. The raters who assign consequence classes and tiers do not see injection assignments or outcomes. The severity panel is separate from the raters. The analysts are independent of the author. The R-BED rules and defaults are frozen before data collection; any change is recorded as a protocol deviation. Injected failures are not drawn by the same people who designed the tier rules.

Replication. A second site or an independent team runs the pre-registered protocol on a different portfolio; the claim is not upgraded on a single site. Code, synthetic record sets and parameter files are published.

Threats to validity. Injected failures may be easier or less diverse than real ones; severity assigned by a panel is a judgment; organizations that volunteer may be better governed; the horizon may be too short for rare severe events; harm is observed imperfectly and with delay (the ascertainment adjustment of the CSE Specification is a bound, not a measurement); effort units may not compare across policies; and the gaming of ordinal ratings once people know the tier rules is not observed in a short study.

Boundary. A result would speak to allocation under the studied conditions. It would not show that R-BED reduces harm in operation without design (b), and it would not show that R-BED is better than other governance practices, which requires a designed comparison with independent teams.

Protocol-verification harness. The reference implementation (command `rbed h7`) runs the pipeline on synthetic pathways for three named scenarios (aligned, misclassified, frequency-dominated), with fixed seeds and a paired difference with an interval. It exists so that the analysis code, the preregistered decision rule and the negative-control logic can be checked before data exist. Its output is determined by invented parameters; in the two negative-control scenarios it is constructed so that the R-BED-tiered policy has higher synthetic harm than the performance-only policy, and that behavior verifies the pipeline can return an unfavorable result. No harness number may be quoted as a result about R-BED.

3 Cross-cutting requirements

- **Preregistration.** Hypotheses, outcomes, sampling rules, exclusion rules and decision rules are registered before data collection.
- **Comparative evaluation.** A comparison with other resources (for example AI Verify or NIST TEVV-Athlon applied to a seeded failure by independent teams) is a separate future study; it requires teams who have not designed R-BED and must not be presented as showing superiority unless it does.
- **Reporting.** Negative and null results are reported with the same prominence. The framework's claim levels are updated only from reported results.
- **Independence of the analysts** from the author for S2, S4, S6 and S7, or a documented plan to mitigate its absence.
- **Data and code.** The synthetic record set and analysis code are published so that others can re-run them. Rights follow the separation stated in the framework (document, software, name, other materials).
- **No outcome selection.** Sample sizes follow a stated precision target for the primary interval, fixed in the protocol; no number is proposed here because none has been derived.

4 Sequence and dependencies

S1 first, because the later studies depend on the records being reliably interpretable. S2 and S6 can run in parallel using public and synthetic material. S3 follows once the template has been revised from S1. S4 and S5 need access to real systems and are the most resource-intensive. S7 can begin with failure injection in a sandbox once S1 has shown that tiers can be assigned reliably (agreement on consequence class is a precondition for H7), and its randomized-allocation design needs real portfolios. Each study ends with a short report and a decision to revise, retain or remove the part of the framework it tested.

5 What a positive result would and would not show

A result for S1 shows that records can be read consistently, not that they are correct. A result for S2 shows retrospective detection, not prevention. S3 shows feasibility. S4, S5 and S7 are the nearest to evidence of benefit, and S7 only in the scenarios in which its design can support it. None shows that R-BED is better than alternative practices; only a designed comparison could.

Contact: Vishal Kumar Srivastava · vshfrm@gmail.com