# Verifiable Evaluation Infrastructure and Governed Publication for AI Behaviour

**Technical Paper (v1.1)**

Status: Draft for external technical review

Audience: Technical reviewers, research partners, regulators, evaluation teams

# 1. Abstract

Current approaches to evaluating AI system behaviour often produce outputs that are difficult to verify longitudinally, reproduce precisely, or audit under changing operational conditions. In many workflows, measurement completion leads directly to publication, while governance over what enters public metrics remains limited, informal, or external to the evaluation system itself.

This paper describes the **RI Safety Layer**, an infrastructure architecture that separates behavioural measurement from publication governance.

The system combines:

- sealed, replayable evaluation evidence
- deterministic measurement outputs
- explicit, signed governance decisions
- controlled inclusion in public rollups without altering underlying results

The architecture is implemented as two distinct modules:

- **Module 1:** behavioural measurement and evidence sealing
- **Module 2 (Phase 1):** post-hoc publication governance over verified evidence bundles

The resulting pipeline provides a verifiable basis for evaluation integrity and governed disclosure. It does not introduce a new model safety mechanism. It introduces infrastructure for trustworthy evaluation operations.

# 2. Problem

## 2.1 Fragile integrity of evaluation outputs

In many evaluation environments:

- outputs are stored without strong integrity guarantees
- prompts, configurations, and scoring logic are not tightly bound to results
- historical outputs are difficult to verify against the run that produced them

As a result, evaluation history may change without clear attribution.

## 2.2 Limited reproducibility over time

Even structured pipelines often depend on implicit operational state.

Common issues include:

- incomplete provenance for model versions or scoring logic
- inconsistent judge configurations
- difficulty isolating the cause of historical changes
- weak replay capability across time

This constrains longitudinal comparison and technical confidence.

## 2.3 Automatic publication

In typical workflows:

- completed sessions flow into metrics automatically
- dashboards reflect outputs immediately
- no explicit governance step exists between measurement and exposure

This couples execution directly to publication.

## 2.4 Informal governance

Where governance exists, it is often:

- external to the evaluation system
- manually applied
- weakly attributable
- difficult to reconstruct later

This reduces trust in both the outputs and the decisions surrounding them.

# 3. Architectural Position

The RI Safety Layer addresses a common structural problem: many evaluation systems combine **measurement, judgement, publication, and governance** within a single opaque flow.

The RI approach separates these concerns into explicit layers:

- behavioural measurement
- evidence integrity
- publication governance
- public aggregation

This allows each layer to be independently inspected, tested, and improved.

# 4. System Overview

The RI Safety Layer implements a two-module architecture comprising an operational measurement layer and an implemented first governance layer.

## 4.1 Module 1 — Behavioural Measurement (Operational)

Module 1 is a deterministic behavioural measurement system that:

- executes structured probe families
- records per-turn and per-session evidence
- computes interpretable metrics
- produces a sealed evidence bundle for each session

Each bundle may contain:

- interaction trace (turns.jsonl)
- summaries and metrics
- manifests and configuration references
- judge provenance and rationale
- canonical bundle hash and signature

These bundles are designed to be:

- replayable
- attributable
- tamper-evident
- operationally durable

Module 1 establishes a stable measurement substrate.

## 4.2 Module 2 — Publication Governance (Phase 1 Implemented)

Module 2 introduces a deterministic post-hoc governance layer operating on completed and verified sessions.

It:

1. verifies the sealed evidence bundle
2. extracts declared governance signals
3. evaluates a containment profile
4. emits a signed governance decision
5. determines rollup eligibility

Module 2:

- operates only on verified evidence
- preserves underlying measurement outputs
- produces auditable governance artefacts
- controls publication without changing results

## 4.3 Separation of Responsibilities

The system enforces a strict boundary:

- **Module 1:** what occurred and how it was measured
- **Module 2:** whether the measured session is eligible for publication

Governance does not rewrite evidence. Measurement does not encode governance.

# 5. Core Concepts

## 5.1 Sealed evidence bundle

A sealed bundle is the canonical representation of a completed evaluation session.

It:

- aggregates relevant artefacts
- is hashed and signed
- provides a stable replay reference

The bundle hash acts as the primary integrity anchor.

## 5.2 Deterministic evaluation

Determinism in this system refers to evaluation behaviour, not necessarily model generation behaviour.

Given identical canonical artefacts, profile versions, and execution posture, the system produces identical semantic outputs.

This applies to:

- metric derivation
- summary generation
- governance decisions
- rollup inclusion logic

Variation is limited to non-semantic envelope metadata where appropriate (e.g. timestamps or signatures).

## 5.3 Governance decision

For each evaluated session, Module 2 may produce a signed decision containing:

- publication status (released / held)
- actions (e.g. flag, hold, escalate)
- triggers and rationale references
- integrity references to the underlying bundle

The decision is replayable, attributable, and verifiable offline.

## 5.4 Governed rollups

Public metrics are derived only from sessions with valid release state.

Examples:

- released sessions are eligible for inclusion
- held sessions are excluded
- ambiguous or invalid governance states may be excluded under strict posture

This creates explicit control over what enters aggregate views.

## 5.5 Override release path

A steward may issue a signed override that:

- references a specific held decision
- binds to the underlying evidence identity
- releases the session for inclusion

Overrides are explicit and auditable.

# 6. System Properties

## 6.1 Integrity

- evidence is sealed and attributable
- tampering invalidates verification
- governance decisions bind to evidence identity

## 6.2 Reproducibility

- sessions are replayable from canonical artefacts

- governance decisions are deterministic under fixed conditions
- rollups are rebuildable from governed evidence history

## 6.3 Auditability

Each session can carry:

- evidence bundle
- governance decision
- optional override path

This enables reconstructable decision history.

## 6.4 Separation of concerns

- governance does not alter measurement outputs
- measurement does not conceal governance state
- both layers remain independently inspectable

## 6.5 Controlled publication

Publication is no longer assumed to follow execution automatically.

Instead, inclusion depends on explicit governance state.

# 7. Non-Goals (Current Phase)

The current implementation does not attempt to:

- intervene in real-time model responses
- rewrite outputs at inference time
- fine-tune model behaviour
- act as a broad policy engine
- redefine correctness of outputs

Its role is intentionally narrower:

post-hoc governance of evaluation publication.

# 8. Implications

## 8.1 Evaluation becomes a governed process

Evaluation becomes two-stage:

1. measure behaviour
2. determine publication eligibility

This introduces operational governance between execution and exposure.

## 8.2 Trust in evaluation history

Because evidence is sealed, decisions are signed, and rollups are rebuildable, historical outputs become:

- attributable
- inspectable
- more resistant to silent drift

## 8.3 Model-agnostic deployment

The system can operate externally to model internals:

- no architecture changes required
- compatible with hosted or API models
- independent of training pipelines

## 8.4 Foundation for future control layers

Because governance is separated from measurement, later extensions can be added without redesigning the evidence substrate.

## 8.5 Institutional utility

The architecture may support:

- regulator inspection workflows
- internal evaluation controls
- controlled benchmark publication
- evidence-backed governance processes
- third-party technical review

# 9. Limitations

Current boundaries include:

- governance is post-hoc rather than real-time
- containment profiles are intentionally constrained
- steward workflows remain minimal
- the system does not claim universal correctness of outputs

These are deliberate design constraints intended to preserve determinism, auditability, and clear separation of responsibilities.

# 10. Conclusion

The RI Safety Layer separates measurement from publication governance in AI evaluation systems.

By combining:

- sealed evidence
- deterministic evaluation
- signed governance decisions
- controlled rollup inclusion

it creates an evaluation pipeline in which:

- results are verifiable
- decisions are explicit
- publication is governed
- historical outputs are more trustworthy to inspect over time

The core proposition is simple:

trust in AI evaluation should arise not from assertion, but from verifiable evidence, explicit judgement, and governed disclosure.