Witnessed Independence: Recording Whether a Verifier Graded Its Own Work

Zain Dana Harper
Seattle, WA | ORCID 0009-0001-7175-5393
https://github.com/HarperZ9

July 2026

Abstract

A verifier's verdict is only as trustworthy as its independence from what it judges: a check written by the same party that produced the artifact can pass without meaning anything. In practice this property is usually an unrecorded coding-discipline assumption—the reviewer trusts that the judge and the producer are different, but nothing in the verdict says so. We describe a small, shipped mechanism that turns this assumption into recorded, re-checkable state. A verification operation carries an optional provenance identity for the party that authored the judging criterion and for the party that produced the artifact; from the pair it resolves a three-valued independence annotation—WITNESSED-INDEPENDENT, SELF-AUTHORED, UNWITNESSED—and, at an opt-in trust boundary, downgrades a decided verdict to UNVERIFIABLE when independence is not positively witnessed. The design is deliberately conservative: absence of a witness is reported as UNWITNESSED, never assumed independent (soundness over completeness), and the default mode leaves an existing call's verdict content byte-for-byte unchanged, adding only the honest label UNWITNESSED, so the annotation is additive. We situate the mechanism inside a single verification operation (*perceive, judge against a criterion it did not author, carry a re-checkable certificate*) and give two instantiated criteria—grounded creative novelty against a witnessed corpus, and composition-sound provenance—that exercise the same operation. We are explicit about scope: this is a tested library primitive (the independence mechanism and its criteria pass a named unit suite), not a validated deployment across a tool fleet; adoption beyond the components shown is stated as future work, not claimed.

1 Introduction

The recurring failure of automated verification is not that a check is wrong but that it is not *independent*. When the party that produces an artifact also writes the check that clears it, the verdict is a tautology dressed as evidence: the model that answers grades its own answer, the pipeline that emits a result attests to its own result, the agent that acts certifies its own action. The machine-learning literature has re-encountered this directly in the LLM-as-judge setting, where a model evaluating its own or a sibling's output exhibits self-preference and can be steered into approving what it should reject [4]. The classical response—proof-carrying code [1] and, more broadly, verdicts that travel with re-checkable evidence—raises the bar from "trust the producer" to "re-derive the check." But re-derivability alone does not establish independence: a producer can ship a perfectly re-derivable check that it also authored, and a re-runner will faithfully reproduce a self-graded pass.

The gap we address is narrow and concrete. In a verification system we built, the design principle was stated as *judge an artifact against a criterion it did not author*. Yet nothing in a verdict recorded whether that principle held. A criterion carried no provenance, and the operation never

recorded who produced the artifact, so a self-graded verdict was byte-indistinguishable from an independent one. The intent was real; its satisfaction was unobservable.

This paper describes the mechanism that closes that specific gap, and only that gap. We do not propose a new notion of independence, a new cryptographic primitive, or a learned evaluator. We propose recording, in the verdict itself, a witness of whether the judging criterion and the artifact were authored by the same party, and refusing—at an explicit trust boundary—to launder a self-graded or unwitnessed decision into a positive verdict.

Contributions.

1. **Witnessed independence** (Section 3): a three-valued, default-preserving annotation that makes "a criterion it did not author" a recorded field of a verdict rather than an assumption, with two opt-in downgrade modes and a stated soundness-over-completeness bias.
2. **The host operation** (Section 2): the annotation is a property of one verification operation—perceive, judge, witness—that is fail-closed (any perception or judging failure degrades to UNVERIFIABLE) and, in default mode, verdict-faithful (the criterion's own verdict is passed through unchanged), so the annotation composes without changing existing behavior.
3. **Two exercised criteria** (Section 4): grounded creative novelty against a witnessed corpus (an anti-mode-collapse check) and composition-sound provenance (resolving mutually-contradicting authenticity signals), each a total function that reduces to the same operation.
4. **An honest boundary** (Section 5): a named unit suite substantiates the mechanism and its criteria; fleet-wide adoption is explicitly future work, and we identify what would falsify the broader claim.

2 The host operation

The mechanism lives inside a single operation we will call *reconcile*: given an artifact, *perceive* it into a witnessed form together with the bytes that form was derived from; *judge* that form with a criterion; and return one *observation* carrying the criterion's verdict, its cited evidence, a content hash of the claim, and a witness of the payload bytes. Two properties make it a safe host for an added annotation. First, it is *fail-closed*: any exception in perception, judging, or result construction (including a malformed certificate from an untrusted criterion) collapses to an UNVERIFIABLE observation rather than propagating. Second, in its default mode it is *verdict-faithful*: it never alters the criterion's verdict; the recorded verdict is the criterion's own, passed through. The opt-in guards of Section 3 are the sole, deliberate exception, and they only ever downgrade a decided verdict toward UNVERIFIABLE, never manufacture one. The operation therefore adds accountability structure around a verdict without ever inventing one.

3 Witnessed independence

The judging criterion optionally carries an `author` identity—who wrote the judge—and the operation optionally accepts a `producer` identity—who produced the artifact. From the pair, a pure resolver assigns one of three values:

- WITNESSED-INDEPENDENT: both present *and* different.
- SELF-AUTHORED: both present *and* equal.
- UNWITNESSED: either absent (the default).

Three design commitments follow.

Default-preserving. When independence is UNWITNESSED, the author and producer identities are not recorded at all. Two verdicts that differ only in an unwitnessed identity are byte-identical, exactly as before the mechanism existed. The annotation is strictly additive: an existing call site is unchanged in content and only gains the honest label UNWITNESSED.

Soundness over completeness. The absence of a witness is UNWITNESSED, never WITNESSED-INDEPENDENT. An unsupplied witness is honestly unknown; it is never assumed to establish independence. The mechanism thus never over-reports independence, at the cost of under-reporting it when identities simply were not provided.

Opt-in refusal at a trust boundary. Two modes, both off by default, convert the annotation into an enforced guard on a *decided* verdict (a positive or negative decision, as opposed to UNVERIFIABLE):

- *strict*: a decided verdict from a SELF-AUTHORED criterion is downgraded to UNVERIFIABLE with a stable reason—the system refuses to launder a self-graded decision as a witnessed one.
- *require-independent*: the stronger guard—a decided verdict is downgraded unless independence is WITNESSED-INDEPENDENT, so *both* self-authored and unwitnessed decisions are refused. This encodes the principle that the mere absence of an external check must not be read as truth.

Because both modes only ever downgrade toward UNVERIFIABLE, they cannot manufacture a positive verdict; the worst they do is refuse to decide. The downgrade records its reason, so a reader can see precisely why a decision was withheld.

4 Two exercised criteria

The operation is generic over the criterion, so independent domains reduce to it by supplying a judge. We give two that were built as motivating instances.

Grounded creative novelty. A reported failure mode of naive generate–critique–regenerate loops is collapse to a handful of generic motifs when the critic is an ungrounded caption [5]. A grounded criterion judges a work's perceived signature against a *witnessed corpus* of prior work: VERIFIED-novel only if the minimum distance to every corpus signature meets a threshold, REFUTED-derivative if it collapses onto something seen, UNVERIFIABLE on an empty corpus. The criterion is generic over the signature type and distance metric; a perceptual hash with Hamming distance is the shipped instantiation, and we are explicit that this establishes non-derivativeness *in the perceived signature*, not deep semantic novelty.

Composition-sound provenance. Uncoordinated authenticity layers can each pass while jointly contradicting—a file can carry a valid "human-authored" manifest and an "AI-generated" watermark at once, because the layers never consult one another [6]. The fix is composition: map each provenance signal onto a verdict lattice and take the proven meet, so any denying or contradicting signal refutes the origin, all-affirming signals verify it, and absence attenuates to UNVERIFIABLE. The certificate records each signal's raw value and its mapped verdict, so the dominating signal—"which one refuted origin?"—is recoverable from the proof.

5 Evaluation and honest scope

What is substantiated. The mechanism and its criteria are shipped, tested code. The reconcile operation, the independence resolver, and the two downgrade modes are covered by a

named unit suite; among its cases are a demonstration that a pre-existing shipped verifier organ yields a byte-identical verdict and oracle when re-expressed as this operation under the same criterion, and two cases that assert *require-independent* downgrades both an UNWITNESSED and a SELF-AUTHORED decided verdict to UNVERIFIABLE while recording the correct independence value. The two criteria pass their own cases, including the anti-collapse guard (a degenerate zero threshold is rejected at construction) and the contradiction case (a denying signal refutes a composed origin).

What is not. This is a library primitive, not a validated deployment. The operation is imported so far by the verifier's own organs, its accountable-memory re-verification path, one external provenance-composition tool, and internal specification-verification scripts; it is *not* imported by the named flagship fleet the surrounding project describes, and we make no claim that "every verification surface is this operation." That broader claim is future work, and it is falsifiable: it fails the moment a verification surface in the fleet cannot be expressed as this operation without changing its verdicts. We also note the mechanism's limits: it witnesses *authorship* independence via declared identities, and is therefore only as meaningful as those identities are honest; it does not establish semantic correctness, and colluding-but-distinct identities would read as WITNESSED-INDEPENDENT.

6 Related work

Proof-carrying code [1] and proof-carrying results generally move trust from the producer to a re-checkable artifact; independence-witnessing is complementary—it records *who* authored the check that is being re-derived. Refinement and dependent types [2] enforce properties at the type level but do not annotate the provenance of the property's author. Content-provenance standards such as C2PA [3] carry authorship claims but, as the authenticated-contradictions result shows [6], do not by themselves compose across layers; our origin criterion is one composition rule over such signals. The LLM-as-judge literature [4] documents the self-grading pathology empirically; our contribution is not a better judge but a recorded, enforceable statement of whether a judge was independent at all.

7 Conclusion

Independence is the property that makes a verdict mean something, and it is usually the property no verdict records. We described a small mechanism that records it: a three-valued, default-preserving witness of whether the judging criterion and the artifact share an author, and an opt-in refusal to decide when independence is not positively witnessed. It is deliberately conservative and deliberately additive, and it is honest about its own reach—a tested primitive whose fleet-wide generalization is stated as a falsifiable open question rather than a result.

References

[1] Necula, G. C. *Proof-carrying code.* POPL, 1997.

[2] Freeman, T. and Pfenning, F. *Refinement types for ML.* PLDI, 1991.

[3] Coalition for Content Provenance and Authenticity. *C2PA Technical Specification.* https://c2pa.org, 2024.

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

[5] Hintze, A., Proschinger Astrom, F., and Schossau, J. *Autonomous language-image generation loops converge to generic visual motifs.* Patterns (Cell Press), 2025. DOI: 10.1016/j.patter.2025.101451.

[6] Nemecek, A., He, H., Cheng, G., and Ayday, E. *Authenticated Contradictions from Desynchronized Provenance and Watermarking.* arXiv:2603.02378, 2026.