RESEARCH SERIES — EVIDENCE, AUTHORSHIP, AND INTELLECTUAL SUBSTANCE

What Evidence of AI Involvement *Can Establish*

How three papers on detection, provenance, authorship, and substance fit together, and how to apply them

DETECTION / PROVENANCE / AUTHORSHIP / SUBSTANCE

An essay in the author's research series on evidence, authorship, and intellectual substance.

Edward Meyman

FERZ, Inc. · September 2026

FERZ INC.

The question

Suppose there is no dispute that AI helped produce a document. What has actually been settled?

Less than it seems. A reviewer still needs to know who directed the work, whether the assistance was permitted or had to be disclosed, whether the claims are supported, whether the work has intellectual substance, and who answers for it. Confirmed involvement does not answer those questions by itself. And involvement is not always confirmed: a detector score is an estimate, not a record.

Three papers examine where these inferences fail: from a detector's estimate to a record of production, from that record to authorship, and from authorship to intellectual substance. This article puts their results together so they can be applied, and links each paper for its full argument and sources.

Three results

Detection is not provenance

A detector score and a provenance record are different kinds of evidence, and the difference does not depend on detectors being inaccurate. That is the argument of *Detection Is Not Provenance*. A text-based detector applies a model to finished text and returns a score or a label. The output depends on more than the document: it also depends on the detector, the model version, the training distribution, the decision threshold, and the calibration. A provenance record, in the sense used here, is an event-linked history of how a document was created and modified. A more accurate detector yields stronger evidence about production, but that evidence remains an estimate made afterward from the finished text. No gain in accuracy turns an estimate into a record.

The difference has practical consequences. In the published evaluations the paper draws on, the binary detectors evaluated performed substantially better on fully AI-generated text than on AI-edited text, which is the mixed category a policy permitting limited assistance needs to distinguish. Where a rule does draw that distinction, a detector label cannot reach it, because the permission lives in the policy while the detector was trained on text. Watermarking differs in kind, because part of its evidence is created at generation. Even so, a detected watermark does not ordinarily record the production history, show how much a model contributed, or bind the signal to any institution's rule. Records captured as the work proceeds, such as authoring-tool revision histories or Content Credentials defined by the Coalition for Content Provenance and Authenticity (C2PA), are evidence of a different kind, with published limits of their own.

Provenance does not decide authorship

Now grant a reliable record: an authentic production record, sufficiently complete for the events at issue. Depending on its schema and capture scope, it may identify the participating tool, the timing, and the recorded inputs, outputs, and transformations. As *Provenance Does Not Decide Authorship* shows, it does not, by itself, decide who directed the work intellectually, whether the assistance was permitted, whether disclosure was owed, or who answers for what the work asserts. Reading weak evidence as a record is one error. Reading a reliable record as a verdict is another. Better evidence does not repair invalid inference.

What does bear on authorship, then? The paper proposes a criterion for institutional and normative purposes. Intellectual directorship is "the exercise of judgment that controls the work's governing thesis, assertions, evidence, reasoning, revision, and final form." Under the proposed criterion, authorship rests on directorship paired with accountable adoption: a person affirmatively places the work into the record as theirs and agrees to answer for it. The criterion rejects two opposite positions: that machine participation displaces human authorship, and that accepting responsibility is enough without directing the work. Permission and disclosure are separate questions. Each is a question of rule, and a finding of noncompliance requires a rule that preceded the conduct. The criterion is not a copyrightability test.

Authorship does not decide substance

Grant both results: AI participation is reliably recorded, and a human intellectually directed the work and accountably adopted it. Neither fact establishes whether the work contains intellectual substance. A work can have a known origin and a responsible author and still be hollow. It can be produced with extensive assistance and still contain determinate, independently examinable content.

For that question, *Authorship Does Not Decide Substance* proposes an outcome-based framework of eight tests in four tiers: Object and Boundary (Tier I); Constraint and Consequence (Tier II); Operationalization, Independent examination, and Corrigibility (Tier III); and Transfer and durable use (Tier IV). The tests carry no weights, and under an explicit decision rule they yield ordered findings rather than scores. The decision rule distinguishes an unassessable product from one whose substance has not been established, specifies when hollowness is the warranted finding, and separates a substantive product from a substantive asset.

The framework applies in the same way to human-produced, AI-assisted, collaborative, and institutional work. Under it, production method, polish, prestige, and volume count as evidence for other questions, never as verdicts on substance. Substance also stays distinct from validity, originality, and utility: a substantive product can be wrong, and a correct one can fail the tests.

A finding of hollowness is a conclusion reached by applying the tests to the work as stated, never a label with which to begin. Under the proposed rule, a finding of hollowness requires satisfaction of Object and Boundary, material failure of Constraint or Consequence, and the presence of markers of intellectual substance. Other failures may leave substance unestablished without warranting that finding.

Default marking of generated text leaves this analysis intact. It changes whether a production signal is available, not what the signal establishes, which is evidence about production and nothing further without more. What changes is the cost of the proxy error: when the signal is present and inexpensive to test, substituting it for evaluation becomes the path of least resistance.

Where to begin

Where an evaluation begins depends on what is being decided. If the question is whether a rule on AI assistance was broken, begin with the rule and the production evidence it makes material. If the question is whether a work contains intellectual substance, begin with the work.

The papers' two procedures reflect this. The second paper sets out six questions that confirmed involvement presents to a reviewer, with production first: involvement is its premise, and the list separates the one question a reliable record can answer directly from the five it cannot answer by itself. The third paper evaluates a product, and production history invites substitution for examining it. Its ten-step procedure therefore warns against beginning with "Was AI used?" and reaches production history only at its last step, for a separately identified question to which it is material. Both apply the second paper's own instruction: identify which question is actually being decided, and answer it with evidence that bears on it.

Begin with the question being decided. Evaluate substance through the product; examine production history when a separately identified question makes it material.

Keep the questions about the work separate. Reliability, as used in *Provenance Does Not Decide Authorship*, asks whether a work's claims are supported. *Authorship Does Not Decide Substance* divides the evaluation of the work further, distinguishing substance from validity, originality, and utility: whether the work satisfies the substance tests, whether its

claims are correct or adequately supported, whether it adds something to prior work, and whether applying it improves a desired outcome. Here, work-level reliability concerns whether claims are supported; instrument reliability concerns the performance of the method producing the evidence.

A worked application: default marking

Suppose a model provider marks the text its models generate by default, and a detection method is available to parties with access to it. Grant the strongest version of the case: the mark is present, detection is reliable, and there is no doubt that the model participated. This removes any objection based on detector accuracy. What does it settle?

Production: answered in part. Under the stipulated conditions, the mark establishes model participation in the text's production or processing. It does not establish how much. As the first paper observes, a watermark surviving in a document that a person wrote and a model lightly edited shows that the model touched it, not how much or whether that was allowed. The mark also has a limit of reach. Where some systems mark text and others do not, absence of a detected mark does not establish absence of AI involvement.

Compliance: depends on the rule. Where a rule prohibits assistance of any kind, a reliable mark bears directly on whether the rule was followed. The first paper notes that under such a rule the definitional gap closes; the stipulated reliability addresses much of the evidentiary gap; and a finding still requires a rule that preceded the conduct. Where a rule permits some assistance and prohibits more, the mark does not reach the distinction, for the same reason a detector label does not: the permission lives in the policy. A stronger signal does not move the line the rule draws.

Disclosure: depends on the rule. Where a rule requires disclosure of AI use, a reliable mark bears on whether that duty was triggered. A marked document submitted without a declaration establishes that no declaration was made. Absent a prior rule making the production method material, it does not establish misconduct.

Authorship: open. A mark shows tool participation, the first row of the second paper's inference table, and tool participation does not establish the absence of human authorship. Whether the named person qualifies as the author under the proposed criterion depends on evidence of intellectual directorship and accountable adoption.

Reliability and validity: open. Whether the claims are supported, or correct, is answered by the cited evidence, the reasoning, the treatment of contrary evidence, and the claims themselves. The mark alone establishes neither the correctness nor the evidentiary support of the claims.

Substance, originality, and utility: open. The mark bears on production and on nothing further without more. The substance tests apply to the work as stated, and originality and utility remain separate questions.

Responsibility: open. Who answers for the work is settled by accountable adoption and any applicable rule allocating responsibility.

A reliable mark answers the participation question within the stipulated case, while leaving the extent and history of involvement unresolved. Together with an applicable rule, it may also bear on permission and disclosure. Authorship still requires evidence interpreted against an authorship criterion. The product still requires examination for reliability, substance, validity, originality, and utility. Responsibility remains a question of accountable adoption and any applicable rule allocating it. This is why the series was built not to depend on detector error. Had the first paper rested on the inaccuracy of detectors, a reliable mark would have overtaken it. Because it rested on the difference between kinds of evidence and the questions each can answer, a more reliable signal leaves each result standing. It makes the ordering discipline more important, because the signal is easier to reach for. In the third paper's terms, the tests do not move; the temptation to skip them does.

> **Source note (verified September 18, 2026).** The occasion for this application is a published provider announcement. Anthropic's explanation of its text watermark, dated August 14, 2026 and updated September 1, 2026, describes a watermark in text generated by its models, applied globally at launch, and states that other major providers will implement their own marking. The page describes the mark's limits in its own terms: detection yields a likelihood that the provider's model was involved at some point; it cannot distinguish text the model wrote from text the model heavily edited, and light proofreading may leave too little to register; it cannot confirm that text was human-written; and it says nothing about ownership or authorship. As of the update, the page describes a detection interface in private preview for eligible organizations. The page is unversioned and may change.

What each kind of evidence can establish

The same analysis extends to the other kinds of evidence the papers discuss. For each, the matrix below states what the evidence supports, what further criterion or evidence a finding requires, and what the evidence cannot establish alone. The middle column matters most, because evidence can contribute to a finding without deciding it. Section references use P1 for *Detection Is Not Provenance*, P2 for *Provenance Does Not Decide Authorship*, and P3 for *Authorship Does Not Decide Substance*. Where the papers do not address a combination, the matrix leaves it out rather than supply an answer the series does not give.

EVIDENCE (SOURCE)	WHAT IT SUPPORTS	WHAT A FINDING ALSO REQUIRES	WHAT IT CANNOT ESTABLISH ALONE
Detector score from finished text (P1 §§1 to 4)	An inference about production, whose strength depends on the detector, model version, training distribution, threshold, and calibration	For a finding under a rule: the rule, plus evidence of reliability, calibration, and fit to the screened population. For a record of the process: a record captured as the events occurred	How the document was produced, as a matter of record; the difference between permitted and prohibited assistance; a policy violation
Continuous estimate of AI edit magnitude from finished text (P1 §§3 to 4)	An estimate of editing extent against the evaluated proxy measures	Evidence of performance and applicability to the evaluated text; for compliance, a governing rule and evidence bearing on its distinctions	An event-linked production history; authorship; a policy violation
Detected watermark (P1 §§1, 3, 5; P3 §8)	Evidence that a compatible generation process produced or preserved the signal	For extent: evidence capable of distinguishing degrees of contribution. For history: event-linked records with sufficient capture scope. For permission: the governing rule and evidence bearing on its distinctions	The extent of contribution; a production history; authorship; compliance; intellectual substance
Provenance record, such as C2PA Content Credentials (P1 §5; P2 §2)	Participation within the record's capture scope: the tool, timing, and recorded inputs, outputs, and transformations the schema captures	Established capture scope and completeness for the events at issue; an authorship criterion; the governing rule for permission or disclosure	Authorship; permission; a disclosure duty; responsibility; steps taken outside instrumented tools; the truth of the content
Authoring-tool revision history (P1 §5)	Process evidence for the editing events the tool captured	A rule, declared in advance, for how histories will be captured, verified, and weighted	Whether an external tool was used outside the captured process; its own authenticity or completeness
No provenance record available to the reviewer (P1 §5; P2 §6)	That the reviewer lacks production-record evidence	A prior recordkeeping or production requirement, and evidence establishing whether it was breached	AI use, where capture was optional or unavailable; that no record exists elsewhere
Nondisclosure of AI use (P2 §§3, 6)	That disclosure did not occur	An applicable rule, preceding the conduct, that created a duty to disclose	Misconduct; authorship; improper production
Final approval by a named person (P2 §3)	Final approval	Evidence of intellectual directorship, and accountable adoption	Directorship or accountable adoption by itself
Documented human direction of thesis, claims, evidence, revisions, and final form (P2 §§3, 4)	Evidence of intellectual directorship	Accountable adoption	Exclusive human production of every element; complete authorship

EVIDENCE (SOURCE)	WHAT IT SUPPORTS	WHAT A FINDING ALSO REQUIRES	WHAT IT CANNOT ESTABLISH ALONE
The author's ability to explain and defend the work (P2 §§4, 7)	Evidence of where judgment resided, as one indicator of directorship	Consideration with other indicators, and accountable adoption	Authorship as a threshold result: indicators locate judgment and carry no passing score
Linguistic polish (P2 §7; P3 §2)	Bears on revision	For production or authorship: evidence of production or of directorship. For substance: the substance tests	Diminished human contribution; AI production; substance
Exclusively human production (P3 §§2, 3)	Evidence relevant to compliance with a requirement of exclusively human production	The rule	Substance
Peer review (P3 §2)	That the work passed a particular screening process under a venue's standards; may bear on validity, methodological adequacy, and editorial selection	For substance: examination of the product itself	Substance
Institutional affiliation (P3 §2)	May bear on access to resources, selection processes, and established expertise	For substance: application of the substance tests	Substance
Volume of output (P3 §§2, 5)	Bears on productivity	For cumulative development: later work that formalizes, tests, limits, applies, corrects, or extends earlier work	Substance; cumulative development
Circulation or institutional adoption (P3 §7)	That the product has moved beyond its creator	Constraint-bearing content that remains identifiable and usable when others apply or develop it	Substance; asset status

Production evidence establishes what its capture scope and reliability support. It may also bear on authorship, compliance, or disclosure, but those findings require the relevant criterion or governing rule. Evidence about screening, standing, or reach likewise cannot substitute for examining the product's substance.

The ability to explain and defend the work is an indicator of intellectual directorship, not a test of unaided recall or speed of refamiliarization. Present difficulty explaining an older work does not by itself establish that the author lacked directorship when developing it. Conversely, a successful explanation does not by itself establish that the person exercised the original governing judgment. Accountable adoption remains a separate requirement of the proposed criterion and cannot substitute for directorship.

What the series does not claim

The series does not argue against detection or provenance infrastructure. Its argument concerns what may validly be inferred from them. It does not hold that production history never matters. Production method is material wherever a separately defined rule or relationship makes it so, including attribution and rights, academic or professional assessment conditions, research-method reporting, confidentiality and protected-data handling, contractual representations about how a deliverable is produced, and regulated production or validation requirements. It offers no copyright test; legal authorship remains a separate, jurisdiction-specific question. The substance framework is a proposal open to examination, and it confers no license to call a work hollow without applying the tests to it. None of the three results depends on detectors being inaccurate.

Closing

Detection cannot establish provenance. Provenance cannot decide authorship. Authorship and production history cannot decide intellectual substance. Each kind of evidence can matter to a separately defined question, and none relieves anyone of answering the question actually being decided. Begin with that question.

About the author

Edward Meyman is founder and CEO of FERZ, Inc., which develops deterministic governance infrastructure for AI systems. This article and the three papers it discusses belong to his research series on evidence, authorship, and intellectual substance. They propose general evaluative frameworks and do not describe FERZ's products or product direction.

The papers

1. Meyman, E. *Detection Is Not Provenance: Why an AI detector score is evidence about production, not a record of the production process.* Version consulted: 1.1. FERZ, Inc., 2026. Zenodo. https://doi.org/10.5281/zenodo.21881429
2. Meyman, E. *Provenance Does Not Decide Authorship: What recorded AI involvement does and does not establish.* Version consulted: 1.0. FERZ, Inc., 2026. Zenodo. https://doi.org/10.5281/zenodo.21890923
3. Meyman, E. *Authorship Does Not Decide Substance: An outcome-based framework for intellectual products.* Version consulted: 1.1. FERZ, Inc., 2026. Zenodo. https://doi.org/10.5281/zenodo.21892669

Each DOI above resolves to the latest version of its paper. Each paper is deposited with a companion document, Sources and Verification Method, recording its sources and the verification performed on them.

Source note: Anthropic, "How Claude's text watermark works," August 14, 2026, updated September 1, 2026. https://www.anthropic.com/news/claude-text-watermark