Preventing Visibility and Metadata from Substituting for Evidence in AI-Assisted Scholarship

Lauren Mercer, Maya Kalyani Reed, Hoang Justin Tran, and Alexander J. Garnier

[1] Simon Fraser University, lauren.mercer@sfu.ca
[2] University of Missouri, Columbia, maya.kalyani.reed@umsystem.edu
[3] Université Grenoble Alpes, hoang.justin.tran@etu.univ-grenoble-alpes.fr
[4] Concordia University, alexander.j.garnier@concordia.ca

Abstract. AI-assisted writing and repository workflows can increase the visibility of educational and scholarly artifacts without increasing support for their claims. This study treats that problem as a decision-rule issue: when evidence, authorization, provenance, and exposure are collapsed into one score, discoverable metadata can compensate for weak evidentiary support. A stochastic model generates synthetic research objects with separate semantic evidence, governance authorization, provenance completeness, and exposure readiness. Three release rules are compared: an additive checklist, an evidence-bound gate, and a provenance-adaptive gate that adds second review for high-consequence or low-provenance objects. Across 40 independent simulations of 20,000 objects, baseline false reassurance among released objects was 60.32 percent for the additive checklist, 30.41 percent for the evidence-bound gate, and 23.02 percent for the adaptive gate. Increasing exposure readiness alone raised additive release coverage from 6.96 percent to 16.02 percent and false reassurance to 72.80 percent; the gated rules were unchanged by design. Inflating recorded evidence for weak objects degraded every rule, with the adaptive gate reaching 58.76 percent false reassurance. The model also reports release coverage, sensitivity, specificity, balanced accuracy, and review burden. It is an executable evaluation template, not a validation of TRACE-AI, an estimate of real publication error, or evidence for adopting a deployed workflow.

Keywords: AI education, assessment, scholarly integrity

1 Introduction

AI-assisted writing, automated assessment support, and repository infrastructures are becoming part of the ecology in which educational research is produced, checked, and found. For learning-science and AI-in-education researchers, the integrity risk is not only fabricated text or references; it is that identifiers, disclosures, complete metadata, and search visibility may be read as assurance that

a claim is well supported. This work contributes an explicit decision-rule model for separating semantic evidence, governance authorization, provenance completeness, and exposure readiness in synthetic scholarly objects. By comparing compensatory and non-compensatory release rules, it tests when visibility and procedural completeness can create false reassurance, while keeping the scope limited to an inspectable evaluation model rather than a deployed publication workflow or an empirical estimate of manuscript error.

Generative artificial intelligence (AI) has lowered the cost of fluent scholarly prose. At the same time, publication workflows increasingly ask authors to manage identifiers, repositories, contributor profiles, disclosure statements, licenses, and search visibility. These developments are valuable, but they create a familiar governance risk: a record can look complete while the claim it accompanies remains poorly supported. The problem is not only that generated text may contain fabricated references. It is also that the infrastructure around a manuscript can become a reassuring signal in its own right.

Díaz Pérez recently proposed TRACE-AI as a workflow that connects evidentiary state, responsible AI assistance, claim checking, curation, and discoverability for AI-assisted scholarly communication [12]. The central separation in that proposal is useful: a repository record, persistent identifier, or author profile can improve accountability and retrieval, but it cannot by itself establish that a cited passage supports a claim or that a disclosure decision is authorized. What remains under-specified is how alternative decision rules behave once those dimensions are represented separately.

This paper studies that decision-rule question. It asks whether a compensatory checklist can create false reassurance by allowing exposure or provenance scores to offset weak evidence and governance, and whether non-compensatory rules avoid that specific failure. The object of analysis is not a deployed journal workflow and not an empirical population of manuscripts. It is a deliberately explicit model whose assumptions can be inspected, criticized, and replaced. The goal is to turn a qualitative design concern into a testable evaluation template.

The distinction matters because scholarly assurance is multidimensional but not fully substitutable. A complete metadata deposit can make an object easier to find; it does not make a causal inference stronger. A digital object identifier can identify a document; it does not prove that a sentence follows from the referenced source. A transparent AI-use declaration can clarify responsibility; it does not repair an unsupported or fabricated citation. Conversely, a well-supported object can remain hard to retrieve because discovery services cover different parts of the literature. A release rule that collapses these dimensions into an average risks treating visibility as a substitute for evidence.

The paper makes four contributions:

- it provides an executable four-layer model of evidence support, governance authorization, provenance completeness, and exposure readiness for synthetic scholarly objects;

- it compares a compensatory checklist, a non-compensatory evidence-bound gate, and a targeted second-review rule under baseline and adversarial recording regimes;
- it reports false reassurance, release coverage, sensitivity, specificity, balanced accuracy, and review burden together, rather than optimizing a single summary score; and
- it translates the model variables into a measurement matrix for later field studies of claim-level review interfaces.

The contribution is intentionally modest. The model does not validate TRACE-AI, estimate real publication error rates, or establish that one operational workflow should be adopted. It clarifies what follows if evidence and authorization are treated as necessary release conditions, what is lost in coverage and review burden, and which failures remain outside the protection of procedural traceability.

2 Related Work

2.1 AI assistance, citation reassurance, and claim-level support

Research on generative AI in publishing now spans drafting assistance, review support, scholarly infrastructure, policy disclosure, and integrity risk [7, 22, 28]. Publisher policies differ in scope and specificity [4], biomedical journal editors vary in their attitudes toward chatbot use [29], and publication guidance continues to assign responsibility to human authors rather than to AI systems [41]. These policy differences motivate a representation in which disclosure and responsibility are recorded separately from evidentiary support.

Citation work shows why this separation is necessary. **(author?)** [38] documented fabricated and erroneous citations in ChatGPT-generated references. **(author?)** [11] found that citations can increase trust in LLM-generated answers even when the citations are random. These findings are directly relevant to false reassurance: readers may respond to the presence of a source marker before they verify whether the source exists or supports the claim. Semantic auditing therefore requires a semantic support judgment, not only bibliographic existence checking [1].

Several recent systems and frameworks move toward claim-level verification. PaperTrail presents a claim–evidence interface for grounding provenance in scholarly question answering [25]. SemanticCite verifies citations with full-text analysis and four support categories, and provides a dataset of over 1,000 annotated citations across disciplines [19]. Claim-relative verifiability work emphasizes that evidence modes and provenance affect what can be checked in hosted language-model services [36]. The Auditable Autonomous Research standard similarly defines auditability through provenance coverage, provenance soundness, contradiction transparency, and audit effort [34]. **(author?)** [17] distinguishes the correctness of a citation from the reconstructability of the path by which it was discovered, checked, and used.

Separating dimensions is also a familiar modeling discipline in adjacent AI systems, where supervision and uncertainty are represented according to the inference they govern. Work on uncertainty-aware and weakly supervised segmentation distinguishes predictive uncertainty from sparse labels [13, 24]; captioning and referring-expression methods distinguish representation alignment from language grounding [14, 23]; and studies of model reprogramming, in-context semi-supervision, and generative tissue reconstruction expose the assumptions behind transfer or inferred structure [8, 15, 40]. Claim-level scholarly systems apply a related separation to support categories and audit records: SemanticCite represents graded citation support [19], while AAR distinguishes provenance coverage, soundness, contradiction transparency, and audit effort [34]. This combination motivates the four-variable representation used here, in which evidence, governance, provenance, and exposure answer different decision questions.

2.2 Provenance, data stewardship, and discovery infrastructure

Provenance and reproducibility guidance has long emphasized the retention of data, transformations, code, parameters, and intermediate results [35]. The FAIR principles separate findability, accessibility, interoperability, and reuse [39], and FAIR maturity work provides operational indicators for stewardship [2]. FAIR-for-AI initiatives extend these concerns to models and computational objects [20]. Software-archiving work shows how artifact identity, metadata, and preservation become practical institutional tasks [10].

Governance adds constraints that a completeness score should not erase. CARE foregrounds collective benefit, authority to control, responsibility, and ethics [5], and subsequent work argues that FAIR and CARE need to be operationalized together rather than treated as interchangeable [6]. Openness may be inappropriate for sensitive materials even when documentation is complete. In the model below, governance therefore denotes legitimate authorization and restrictions, not maximum disclosure.

Discovery infrastructure is also heterogeneous. Open citation data and machine-readable references improve inspectability across platforms [31, 32], ORCID improves contributor disambiguation [18], and open-access infrastructure expands availability while leaving license and discoverability questions distinct [33]. Large-scale comparisons show that Google Scholar, Web of Science, Scopus, Dimensions, Crossref, Microsoft Academic, and OpenCitations cover different parts of the literature [26, 27, 37]. These studies support a separate exposure-readiness variable: exposure is valuable, but it is not evidence for a research claim.

2.3 Decision rules and non-compensation

The decision principle studied here is not new in decision science. Multiple-criteria decision analysis distinguishes compensatory aggregation, in which high performance on one criterion can offset low performance on another, from non-compensatory or veto-based structures [3]. ELECTRE-family methods and related taxonomies include veto thresholds that can prevent an alternative from

being accepted despite strong scores elsewhere [9, 16]. The present paper imports that established design idea into scholarly communication: if evidence support and legitimate authorization are necessary, then metadata quality should not be able to compensate for their absence.

2.4 Relation to TRACE-AI

This study takes from TRACE-AI the conceptual separation among evidentiary state, accountable AI use, curation, and exposure, as well as the concern that unsupported claims can be made more discoverable [12]. It adds an independently written decision-rule model, three explicit release rules, four simulated regimes, seed-level uncertainty summaries, sensitivity analysis, and a field-measurement matrix. The simulation is therefore a method and evaluation contribution adjacent to the original conceptual proposal, not an empirical evaluation of an implemented TRACE-AI intervention.

3 Methods

3.1 Scope and model objective

We study synthetic claim-level research objects under stylized release rules. Each object is represented by four continuous dimensions: semantic evidence support, governance authorization, provenance completeness, and exposure readiness. The model asks how rules behave when these dimensions are measured with error and then used to decide whether an object should be released. The analysis isolates one mechanism: compensation across dimensions that should not be substitutable.

Before the full run, the generator, decision rules, primary outcome, regimes, and sensitivity analyses were fixed in a local analysis plan. The plan was not a public preregistration, and the model parameters are not empirically calibrated. That limitation is important: the reported intervals quantify Monte Carlo stability under the chosen model, not uncertainty over real-world manuscripts.

3.2 Synthetic research objects

For each seed, we generated 20,000 independent synthetic research objects. Each object had four latent dimensions on the unit interval:

- E: semantic evidence support;
- G: governance and authorization adequacy;
- P: provenance and trace completeness; and
- X: exposure readiness, including metadata and deposition preparation.

The dimensions share latent "rigor" and "stewardship" factors and include dimension-specific noise. This creates positive but imperfect correlation, representing the plausible condition that careful teams tend to perform well across

tasks while still having localized failures. Thirty percent of objects were assigned to a high-consequence class. Integrity eligibility, T, required latent evidence and governance to exceed class-specific thresholds:

$$T_i = I(E_i \geq e_{C_i})I(G_i \geq g_{C_i}). \tag{1}$$

The thresholds were 0.60 for evidence and 0.55 for governance for ordinary objects, and 0.72 and 0.68 for high-consequence objects. Neither provenance nor exposure enters T. This is a normative modeling choice: records and visibility can support accountability, but they cannot logically create evidence or authorization.

Reviewers observe noisy versions of all four dimensions. Independent Gaussian noise with standard deviation 0.08 was added and values were clipped to the unit interval. The noise represents disagreement, ambiguous documentation, and imperfect inspection. It is not calibrated to a reliability study.

3.3 Decision rules

The **additive checklist** released an object when the mean observed score across E, G, P, and X was at least 0.62. This is a deliberately simple compensatory baseline: strength in one category can offset weakness in another. It should not be read as a calibrated model of a specific journal or repository instrument.

The **evidence-bound gate** released an object only when observed evidence and governance exceeded consequence-specific thresholds and observed provenance was at least 0.45:

$$R_i^{\text{gate}} = I(\widehat{E}_i \geq e_{C_i})I(\widehat{G}_i \geq g_{C_i})I(\widehat{P}_i \geq 0.45). \tag{2}$$

Exposure readiness was recorded but could not rescue an object that failed the evidence, governance, or provenance conditions.

The **provenance-adaptive gate** applied the same conditions but obtained a second noisy review of evidence and governance whenever provenance was below 0.60 or the object was high consequence. It used the conservative minimum of the first and second observations. This rule operationalizes a simple escalation principle: weak provenance and high stakes trigger additional scrutiny rather than compensatory scoring.

3.4 Stress-test regimes

The baseline used unmodified observations. The metadata-rich regime added 0.30 to observed exposure readiness while leaving latent integrity unchanged. The evidence-adversarial regime added 0.30 to recorded evidence only for objects whose latent evidence was below the applicable requirement. The provenance-adversarial regime added 0.35 to recorded provenance without improving evidence or governance.

These regimes are mechanism probes, not estimates of real manipulation. The metadata regime tests metric substitution; the evidence regime tests false or overconfident semantic judgments; and the provenance regime tests decorative or inflated records.

3.5 Outcomes and uncertainty

The primary outcome was false reassurance among released objects:

$$\Pr(T = 0 \mid R = 1). \tag{3}$$

Secondary outcomes were release coverage, sensitivity for integrity-eligible objects, specificity, balanced accuracy, and mean reviews per object. Accuracy alone is not emphasized because eligible objects are uncommon under the chosen generator.

The primary analysis used 40 integer seeds and 20,000 objects per seed, producing 800,000 objects per regime. Outcomes were calculated within each seed and summarized with 5,000 percentile bootstrap resamples of seed-level means to form 95% Monte Carlo intervals. Sensitivity analysis varied the additive cutoff from 0.45 to 0.75 and the provenance minimum for the evidence-bound gate from 0.35 to 0.55. The implementation is in `code/simulate.py`; `code/run.sh` regenerates the result tables and figures and includes a smoke test.

3.6 Ethics and data

No people, private records, or personal data were used. All analyzed observations are synthetic. Literature metadata were obtained through public scholarly APIs solely for reference verification. Human-subject review was therefore not applicable.

4 Results

4.1 Baseline comparison

Only 5.81% of generated objects met the latent evidence and governance requirements (95% Monte Carlo interval 5.76%–5.86%). This prevalence is a design consequence of the generator, not an estimate of the scholarly literature.

Table 1 shows the baseline comparison. The additive checklist released 6.96% of objects and had a false-reassurance rate of 60.32% (59.97%–60.67%). The evidence-bound gate released 4.60% and reduced false reassurance to 30.41% (30.00%–30.84%). It also improved balanced accuracy from 71.53% to 76.83%. The adaptive rule was most conservative: it released 3.76%, produced 23.02% false reassurance (22.58%–23.49%), and used 1.84 reviews per object. Its balanced accuracy was lower than the single-review gate because conservative second review rejected additional eligible objects.

The nonzero error of the gated rules comes from observation noise. A rule can be logically aligned with evidence and governance and still pass an object when reviewers overestimate those dimensions.

Table 1. Baseline release-rule performance. Parentheses contain 95% bootstrap Monte Carlo intervals over 40 seed-level estimates.

Rule	Release	False reassur.	Sensitivity	Specificity	Bal. accuracy	Reviews
Additive checklist	6.96% (6.89–7.02)	60.32% (59.97–60.67)	47.51% (46.99–47.97)	95.54% (95.49–95.59)	71.53% (71.27–71.76)	1.00
Evidence-bound gate	4.60% (4.56–4.65)	30.41% (30.00–30.84)	55.14% (54.73–55.58)	98.51% (98.49–98.54)	76.83% (76.61–77.04)	1.00
Provenance-adaptive	3.76% (3.72–3.80)	23.02% (22.58–23.49)	49.79% (49.35–50.26)	99.08% (99.06–99.10)	74.44% (74.22–74.67)	1.84 (1.84–1.84)

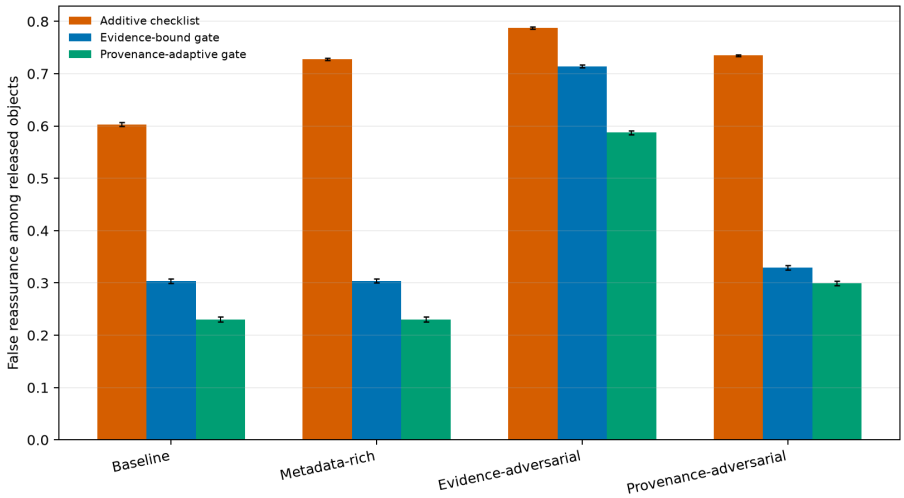

Fig. 1. False reassurance among released objects. Error bars are 95% Monte Carlo intervals over seed means. Values are conditional on the specified synthetic model and are not estimates of real scholarly error rates.

4.2 Metadata enrichment produces compensatory failure

Raising exposure readiness without changing evidence or governance increased additive release coverage from 6.96% to 16.02%. False reassurance rose to 72.80% (72.60%–73.01%). Sensitivity also rose, which illustrates why a favorable single measure is insufficient: the checklist released more eligible objects and many more ineligible objects.

The two gated rules were unchanged in the metadata-rich regime. This is not an empirical surprise; it follows from the rule definitions because X is not a release condition. The value of the simulation is to quantify the consequence of that design choice under the shared generator and to provide a reproducible comparison point. Figure 1 displays false reassurance across regimes.

4.3 Inflated records expose remaining vulnerabilities

When weak-evidence objects received inflated evidence records, false reassurance increased to 78.78% for the additive checklist, 71.41% for the evidence-bound gate, and 58.76% for the adaptive gate. The extra review in the adaptive rule

Table 2. Evidence map from decision-rule assumption to model consequence.

Question	Present test	Result	Interpretation
Can exposure be separated from justification?	Increase X without changing E or G	Additive false reassurance: 60.32% to 72.80%; gates unchanged	Exposure should not compensate for support.
Does evidence gating matter under the chosen target?	Compare additive and conjunctive rules	Baseline false reassurance: 60.32% versus 30.41%	Non-compensation is the operative mechanism.
Does more traceability guarantee integrity?	Inflate provenance records	Every rule still makes errors	Record completeness enables challenge but does not establish truth.
Can targeted extra scrutiny help?	Trigger a second review for high stakes or weak provenance	23.02% false reassurance; 1.84 reviews per object	Lower reassurance risk trades off coverage and burden.
What if semantic judgments are corrupted?	Inflate recorded evidence for weak objects	Best rule still has 58.76% false reassurance	Independent challenge and adjudication remain indispensable.

reduced but did not eliminate the attack because the second review was still a noisy proxy for latent support. The attack also favors the adaptive rule relative to harsher scenarios in which both reviewers share a systematic semantic error.

Inflating provenance alone raised additive false reassurance to 73.47%. The evidence-bound gate reached 32.90%, and the adaptive rule reached 29.98%. The gated rules were affected because provenance remains part of release, but evidence and governance still had to pass. Table 2 summarizes the model consequences without treating any simulated percentage as a field estimate.

4.4 Risk–coverage sensitivity

Increasing the additive cutoff reduced false reassurance, but the gain required steep coverage losses. At cutoff 0.75, false reassurance remained 32.48% while release coverage fell to 0.77%. The evidence-bound gate with a provenance minimum of 0.45 achieved 30.41% false reassurance at 4.60% coverage. Within this generator, a stricter additive score did not reproduce the gate's risk–coverage point efficiently.

Changing the provenance minimum for the evidence-bound rule from 0.35 to 0.55 reduced false reassurance from 31.51% to 29.26%, while coverage fell from 5.39% to 3.55%. Figure 2 displays these trade-offs.

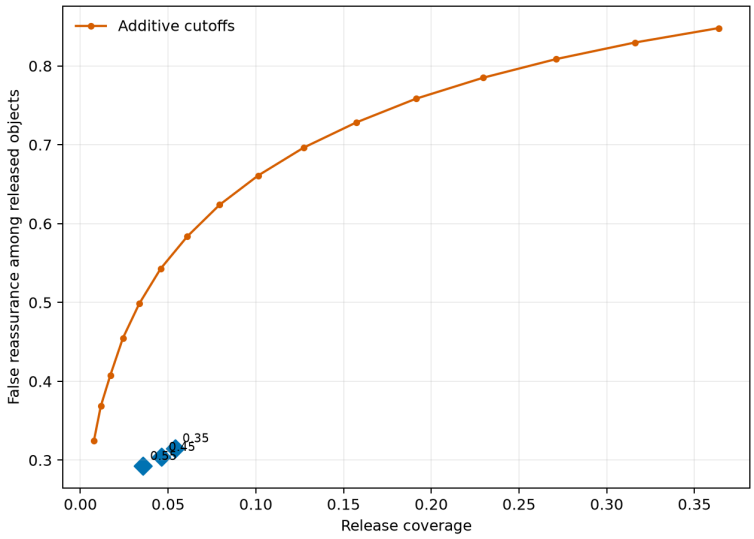

Fig. 2. Baseline release coverage and false reassurance across additive cutoffs; diamonds show evidence-bound rules with alternative provenance thresholds.

5 From Model Variables to Field Measures

The model is useful only if its abstractions can be challenged in real settings. Table 3 maps each variable to an observable measure and a failure mode. The evaluation unit should be a consequential claim or claim cluster, not an entire manuscript, because support can vary within one document.

A prospective study could compare the same instructional content under an additive checklist, an evidence-only gate, an evidence-and-governance gate, a PaperTrail-style claim–evidence display, and adaptive escalation. It should include budget-matched baselines such as random second review, uncertainty-triggered review, and universal double review on a smaller sample. Outcomes should include false reassurance, sensitivity, time, calibration, disagreement, and perceived burden. Assessors should be blinded to condition when judging semantic support, and confirmatory thresholds should be fixed before analysis [30].

6 Discussion

The central model consequence is straightforward. When evidence, governance, provenance, and exposure are averaged, high visibility or polished records can offset weak support. When evidence and governance are necessary conditions, metadata improvements remain valuable but cannot authorize a claim. The strongest result is therefore not a particular simulated percentage; it is a diagnostic property. If adding a DOI, completing repository fields, or improving

search visibility can flip an unsupported claim to "responsible," the instrument embeds metric substitution.

The comparison also has a built-in structural advantage for the gated rules. Integrity eligibility is defined by evidence and governance thresholds, and the evidence-bound rule applies corresponding thresholds to noisy observations. That alignment is the mechanism being examined, not a hidden empirical discovery. The additive checklist is likewise a simple comparator rather than a calibrated representation of any named journal, repository, or policy instrument. The paper therefore supports a design caution, not a claim of practical superiority.

Review burden complicates the interpretation. The adaptive rule reduced false reassurance most in the baseline but released fewer objects and required 1.84 reviews per object. Low false reassurance can be purchased by low coverage, delay, and unequal burden. Regional differences in preprint infrastructure and capacity caution against assuming that a more demanding workflow is equally feasible everywhere [21]. A responsible implementation should report risk and coverage together and stratify burden by language, discipline, institution, access to sources, and object type.

The adversarial regimes show the boundary of procedural traceability. Inflated provenance weakened safeguards, and corrupted semantic evidence directly attacked the central gate. A complete-looking record is useful because it can be inspected; it is not proof that the recorded judgment is correct. Future systems need spot checks, adjudication, conflict surfacing, and reviewer-expertise models rather than relying on a single stored field.

The model leaves substantial work undone. It does not include paper-, discipline-, source-, or reviewer-level clustering; categorical legal or ethical authorization; conflicting evidence; correlated reviewer bias; consequence-label error; or expertise-weighted adjudication. The reported sensitivity varies additive cutoffs and provenance thresholds, but it does not vary eligibility prevalence, evidence/governance thresholds, criterion weights, correlations, review-noise levels, attack strength, or consequence-aware versus consequence-agnostic rules jointly. The adaptive rule is not compared with universal double review, random second review at the same budget, uncertainty-based triage, or adjudicated aggregation. These omissions limit the paper to a transparent mechanism demonstration.

7 Limitations

This study is a simulation with uncalibrated parameters. The generator, 5.81% eligibility prevalence, correlations, thresholds, 0.08 review noise, and attack shifts are not estimated from manuscripts or reviewers; Monte Carlo intervals can be narrow even when model uncertainty is large. Evidence and governance are represented as scalar variables, although real support can be multidimensional and authorization can be categorical or jurisdiction-specific. The second review is modeled as a clean independent observation in the evidence-adversarial regime, which is favorable to the adaptive rule compared with shared semantic error. Review burden is measured as review count rather than time, cost, opportunity

loss, or inequity. The literature search was broad but not systematic, and no implemented journal, repository, classroom, newsroom, or laboratory workflow was evaluated.

8 Conclusion

Responsible scholarly communication needs decision rules in which discoverability cannot erase evidentiary failure. In the model studied here, an additive checklist converted metadata improvement into more releases and more false reassurance even though latent integrity did not change. Evidence-bound gates reduced compensatory error, and targeted second review reduced it further at a cost in coverage and burden. None of the rules survived corrupted semantic judgments unscathed.

The practical implication is to keep support, authorization, provenance, and exposure separate. Evidence and legitimate authorization should be necessary conditions for release decisions; provenance should make records inspectable and challengeable; exposure should be reported as an infrastructure outcome; and unresolved disagreement should remain visible. The next step is not more precision around this toy generator, but a field study with real claims, real reviewers, budget-matched baselines, and explicit measurement of burden.

Data availability

The analysis uses only synthetic records generated by the supplied code. The original TRACE-AI preprint is publicly available from Preprints.org at DOI 10.20944/preprints202610.0402.v1. Bibliographic metadata used for reference verification were obtained from public scholarly APIs. Generated result tables and figures are included in the project output.

Code availability

Code is available from the corresponding author and will be released at publication.

CRediT author contribution

Lauren Mercer, Maya Kalyani Reed, Hoang Justin Tran, Alexander J. Garnier: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing–original draft, Writing–review and editing, Visualization, Project administration.

Funding

This research received no external funding.

Declaration of competing interest

The author declares no competing interests.

During the preparation of this work the author used AI-assisted tools for literature search, code development and drafting. The author reviewed and edited the content and takes full responsibility for it.

Table 3. Measurement matrix for a future field evaluation.

Construct	Observable measure	Reference standard	Reliability requirement	Important failure mode
Semantic evidence E	Blinded judgment that a cited passage supports the claim within scope	Independent adjudication panel with source access	Claim-level agreement plus disagreement categories	Reviewers verify existence but not entailment.
Governance G	Authorization, consent, rights, restrictions, and accountable owner documented	Applicable institutional and community rules	Agreement on applicability, not just completeness	Pressure to disclose legitimately restricted material.
Provenance P	Source locator, version, correction rationale, actor, date, and transformation chain	Frozen schema and version history	Field completeness plus spot-review accuracy	Complete-looking but false or selectively omitted record.
Exposure X	Metadata validity, resolvable identifier, host accessibility, dated retrieval attempts	Service-specific eligibility and query protocol	Repeated queries with timestamps	Treating nonretrieval as rejection or poor science.
False reassurance	Unsupported claim marked releasable	Blinded reference-standard decision	Adjudicated denominator fixed in advance	Procedural completion becomes a badge.
Burden	Time, reviewer count, unresolved cases, and abandonment	Instrumented workflow logs	Common timing definitions	Unequal costs across languages, fields, or institutions.

Bibliography

[1] Ansorge, L.: From existence checking to semantic auditing: citation verifiability in the age of ai. Frontiers in Research Metrics and Analytics **11** (September 2026). https://doi.org/10.3389/frma.2026.1927189, http://dx.doi.org/10.3389/frma.2026.1927189

[2] Bahim, C., Casorrán-Amilburu, C., Dekkers, M., Herczog, E., Loozen, N., Repanas, K., Russell, K., Stall, S.: The fair data maturity model: An approach to harmonise fair assessments. Data Science Journal **19** (2020). https://doi.org/10.5334/dsj-2020-041, http://dx.doi.org/10.5334/dsj-2020-041

[3] Belton, V., Stewart, T.J.: Multiple Criteria Decision Analysis: An Integrated Approach. Kluwer Academic Publishers, Boston (2002). https://doi.org/10.1007/978-1-4615-1495-4

[4] Bhavsar, D., Duffy, L., Jo, H., Lokker, C., Haynes, R.B., Iorio, A., Marusic, A., Ng, J.Y.: Policies on artificial intelligence chatbots among academic publishers: a cross-sectional audit. Research Integrity and Peer Review **10**(1) (February 2025). https://doi.org/10.1186/s41073-025-00158-y, http://dx.doi.org/10.1186/s41073-025-00158-y

[5] Carroll, S.R., Garba, I., Figueroa-Rodríguez, O.L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J.D., Anderson, J., Hudson, M.: The care principles for indigenous data governance. Data Science Journal **19** (2020). https://doi.org/10.5334/dsj-2020-043, http://dx.doi.org/10.5334/dsj-2020-043

[6] Carroll, S.R., Herczog, E., Hudson, M., Russell, K., Stall, S.: Operationalizing the care and fair principles for indigenous data futures. Scientific Data **8**(1) (Apr 2021). https://doi.org/10.1038/s41597-021-00892-0, http://dx.doi.org/10.1038/s41597-021-00892-0

[7] Chauhan, C., Currie, G.: The impact of generative artificial intelligence on research integrity in scholarly publishing. The American Journal of Pathology **194**(12), 2234–2238 (Dec 2024). https://doi.org/10.1016/j.ajpath.2024.10.001, http://dx.doi.org/10.1016/j.ajpath.2024.10.001

[8] Chung, M.Y., Fan, J., Ye, H., Wang, Q., Shen, W.C., Yu, C.M., Chen, P.Y., Kuo, S.Y.: Model reprogramming demystified: A neural tangent kernel perspective. arXiv preprint arXiv:2506.00620 (2025), https://arxiv.org/abs/2506.00620

[9] Cinelli, M., Kadziński, M., Gonzalez, M., Słowiński, R.: How to support the application of multiple criteria decision analysis? let us start with a comprehensive taxonomy. Omega **96**, 102261 (2020). https://doi.org/10.1016/j.omega.2020.102261

[10] Di Cosmo, R., Gruenpeter, M., Marmol, B., Monteil, A., Romary, L., Sadowska, J.: Curated archiving of research software artifacts: Lessons learned from the french open archive (hal). International Journal of Digital Curation

15(1), 16 (August 2020). https://doi.org/10.2218/ijdc.v15i1.698, http://dx.doi.org/10.2218/ijdc.v15i1.698

[11] Ding, Y., Facciani, M., Poudel, A., Joyce, E., Aguinaga, S., Veeramani, B., Bhattacharya, S., Weninger, T.: Citations and trust in llm generated responses. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 23787–23795 (2025). https://doi.org/10.1609/AAAI.V39I22.34550, https://doi.org/10.1609/AAAI.V39I22.34550

[12] Díaz Pérez, A.: From manuscript to discoverability: Trace-ai for responsible ai-assisted scholarly communication. Preprints.org (October 2026). https://doi.org/10.20944/preprints202610.0402.v1, http://dx.doi.org/10.20944/preprints202610.0402.v1

[13] Fan, J., Gao, B., Jin, H., Jiang, L.: Ucc: Uncertainty guided cross-head co-training for semi-supervised semantic segmentation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9937–9946. IEEE (June 2022). https://doi.org/10.1109/cvpr52688.2022.00971, http://dx.doi.org/10.1109/cvpr52688.2022.00971

[14] Fan, J., Liang, Y., Liu, L., Huang, S., Zhang, L.: Rca-noc: Relative contrastive alignment for novel object captioning. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15464–15474. IEEE (October 2023). https://doi.org/10.1109/iccv51070.2023.01422, http://dx.doi.org/10.1109/iccv51070.2023.01422

[15] Fan, J., Rosu, P., Wang, A.T., Li, Z.M., Carin, L., Cheng, X.: In-context semi-supervised learning. arXiv preprint arXiv:2512.15934 (2025), https://arxiv.org/abs/2512.15934

[16] Figueira, J.R., Mousseau, V., Roy, B.: An overview of electre methods and their recent extensions. In: Bouyssou, D., Jacquet-Lagrèze, E., Perny, P., Słowiński, R., Vanderpooten, D., Vincke, P. (eds.) Aiding Decisions with Multiple Criteria: Essays in Honour of Bernard Roy, pp. 61–85. Springer (2013). https://doi.org/10.1007/978-1-4614-7637-6_3

[17] Frimpong, V.: Citation auditability in ai-assisted research: From reference verification to research integrity. Preprints.org (2026). https://doi.org/10.20944/preprints202609.1770.v1, https://doi.org/10.20944/preprints202609.1770.v1

[18] HAAK, L.L., FENNER, M., PAGLIONE, L., PENTZ, E., RATNER, H.: Orcid: a system to uniquely identify researchers. Learned Publishing **25**(4), 259–264 (October 2012). https://doi.org/10.1087/20120404, http://dx.doi.org/10.1087/20120404

[19] Haan, S.: Semanticcite: Citation verification with ai-powered full-text analysis and evidence-based reasoning (2025), https://arxiv.org/abs/2511.16198

[20] Huerta, E.A., Blaiszik, B., Brinson, L.C., Bouchard, K.E., Diaz, D., Doglioni, C., Duarte, J.M., Emani, M., Foster, I., Fox, G., Harris, P., Heinrich, L., Jha, S., Katz, D.S., Kindratenko, V., Kirkpatrick, C.R., Lassila-Perini, K., Madduri, R.K., Neubauer, M.S., Psomopoulos, F.E., Roy, A., Rübel, O., Zhao, Z., Zhu, R.: Fair for ai: An interdisciplinary and international community building perspective. Scientific Data **10**(1) (July 2023).

https://doi.org/10.1038/s41597-023-02298-6, http://dx.doi.org/10.1038/s41597-023-02298-6

[21] Irawan, D.E., Zahroh, H., Puebla, I.: Preprints as a driver of open science: Opportunities for southeast asia. Frontiers in Research Metrics and Analytics **7** (September 2022). https://doi.org/10.3389/frma.2022.992942, http://dx.doi.org/10.3389/frma.2022.992942

[22] Kousha, K., Thelwall, M.: Artificial intelligence to support publishing and peer review: A summary and review. Learned Publishing **37**(1), 4–12 (August 2023). https://doi.org/10.1002/leap.1570, http://dx.doi.org/10.1002/leap.1570

[23] Liang, Y., Yang, Z., Tang, Y., Fan, J., Li, Z., Wang, J., Torr, P.H., Huang, S.L.: Luna: Language as continuing anchors for referring expression comprehension. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 5174–5184. MM '23, ACM (October 2023). https://doi.org/10.1145/3581783.3612584, http://dx.doi.org/10.1145/3581783.3612584

[24] Liu, L., Kong, T., Zhu, M., Fan, J., Fang, L.: Clickseg: 3d instance segmentation with click-level weak annotations. arXiv preprint arXiv:2307.09732 (2023), https://arxiv.org/abs/2307.09732

[25] Martin-Boyle, A., Leckey, C., Brown, M., Kaur, H.: Papertrail: A claim-evidence interface for grounding provenance in llm-based scholarly q&a. In: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems. pp. 1–25. CHI '26, ACM (Apr 2026). https://doi.org/10.1145/3772318.3791101, http://dx.doi.org/10.1145/3772318.3791101

[26] Martín-Martín, A., Orduna-Malea, E., Thelwall, M., Delgado López-Cózar, E.: Google scholar, web of science, and scopus: A systematic comparison of citations in 252 subject categories. Journal of Informetrics **12**(4), 1160–1177 (November 2018). https://doi.org/10.1016/j.joi.2018.09.002, http://dx.doi.org/10.1016/j.joi.2018.09.002

[27] Martín-Martín, A., Thelwall, M., Orduna-Malea, E., Delgado López-Cózar, E.: Google scholar, microsoft academic, scopus, dimensions, web of science, and opencitations' coci: a multidisciplinary comparison of coverage via citations. Scientometrics **126**(1), 871–906 (September 2020). https://doi.org/10.1007/s11192-020-03690-4, http://dx.doi.org/10.1007/s11192-020-03690-4

[28] Nazarovets, S., Suchikova, Y.: Review articles, generative <scp>ai</scp> and the remaking of scholarly infrastructure. Learned Publishing **39**(2) (February 2026). https://doi.org/10.1002/leap.2045, http://dx.doi.org/10.1002/leap.2045

[29] Ng, J.Y., Krishnamurthy, M., Deol, G., Al-Khafaji, W.A.Z., Balaji, V., Abebe, M., Adhvaryu, J., Karrthik, T., Mohanakanthan, P., Vellaparambil, A., Bouter, L.M., Haynes, R.B., Iorio, A., Lokker, C., Maisonneuve, H., Marušić, A., Moher, D.: Attitudes and perceptions of biomedical journal editors in chief towards the use of artificial intelligence chatbots in the scholarly publishing process: a cross-sectional survey. Research Integrity and Peer Re-

view **10**(1) (September 2025). https://doi.org/10.1186/s41073-025-00178-8, http://dx.doi.org/10.1186/s41073-025-00178-8

[30] Nosek, B.A., Ebersole, C.R., DeHaven, A.C., Mellor, D.T.: The preregistration revolution. Proceedings of the National Academy of Sciences **115**(11), 2600–2606 (March 2018). https://doi.org/10.1073/pnas.1708274114, http://dx.doi.org/10.1073/pnas.1708274114

[31] Peroni, S., Dutton, A., Gray, T., Shotton, D.: Setting our bibliographic references free: towards open citation data. Journal of Documentation **71**(2), 253–277 (March 2015). https://doi.org/10.1108/jd-12-2013-0166, http://dx.doi.org/10.1108/jd-12-2013-0166

[32] Peroni, S., Shotton, D.: Opencitations, an infrastructure organization for open scholarship. Quantitative Science Studies **1**(1), 428–444 (February 2020). https://doi.org/10.1162/qss_a_00023, http://dx.doi.org/10.1162/qss_a_00023

[33] Piwowar, H., Priem, J., Larivière, V., Alperin, J.P., Matthias, L., Norlander, B., Farley, A., West, J., Haustein, S.: The state of oa: a large-scale analysis of the prevalence and impact of open access articles. PeerJ **6**, e4375 (February 2018). https://doi.org/10.7717/peerj.4375, http://dx.doi.org/10.7717/peerj.4375

[34] Rasheed, R.A., Banerjee, S., Mukherjee, A., Hazra, R.: From fluent to verifiable: Claim-level auditability for deep research agents (2026), https://arxiv.org/abs/2602.13855

[35] Sandve, G.K., Nekrutenko, A., Taylor, J., Hovig, E.: Ten simple rules for reproducible computational research. PLoS Computational Biology **9**(10), e1003285 (October 2013). https://doi.org/10.1371/journal.pcbi.1003285, http://dx.doi.org/10.1371/journal.pcbi.1003285

[36] Seo, B.: Claim-relative verifiability in hosted llm services: measurement validity, evidence modes, and provenance. AI and Ethics **6**(6) (October 2026). https://doi.org/10.1007/s43681-026-01467-8, http://dx.doi.org/10.1007/s43681-026-01467-8

[37] Visser, M., van Eck, N.J., Waltman, L.: Large-scale comparison of bibliographic data sources: Scopus, web of science, dimensions, crossref, and microsoft academic. Quantitative Science Studies **2**(1), 20–41 (2021). https://doi.org/10.1162/qss_a_00112, http://dx.doi.org/10.1162/qss_a_00112

[38] Walters, W.H., Wilder, E.I.: Fabrication and errors in the bibliographic citations generated by chatgpt. Scientific Reports **13**(1) (September 2023). https://doi.org/10.1038/s41598-023-41032-5, http://dx.doi.org/10.1038/s41598-023-41032-5

[39] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., San-

sone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B.: The fair guiding principles for scientific data management and stewardship. Scientific Data **3**(1) (March 2016). https://doi.org/10.1038/sdata.2016.18, http://dx.doi.org/10.1038/sdata.2016.18

[40] Yu, T., Ekbote, C., Morozov, N., Herrera, A., Fan, J., Dominguez Mantes, A., Novello, S., Lashuel, H., Frossard, P., d'Ascoli, S., La Manno, G., Brbić, M.: Tissue reconstruction with generative ai. bioRxiv (February 2025). https://doi.org/10.1101/2025.02.13.638045, https://doi.org/10.1101/2025.02.13.638045

[41] Zielinski, C., Winker, M.A., Aggarwal, R., Ferris, L.E., Heinemann, M., Lapeña Jr, J.F., Pai, S.A., Ing, E., Citrome, L., Alam, M., Voight, M., Habibzadeh, F.: Chatbots, generative ai, and scholarly manuscripts: Wame recommendations on chatbots and generative artificial intelligence in relation to scholarly publications. Colombia Medica **54**(3), e1015868 (Dec 2023). https://doi.org/10.25100/cm.v54i3.5868, http://dx.doi.org/10.25100/cm.v54i3.5868