Reliable Context

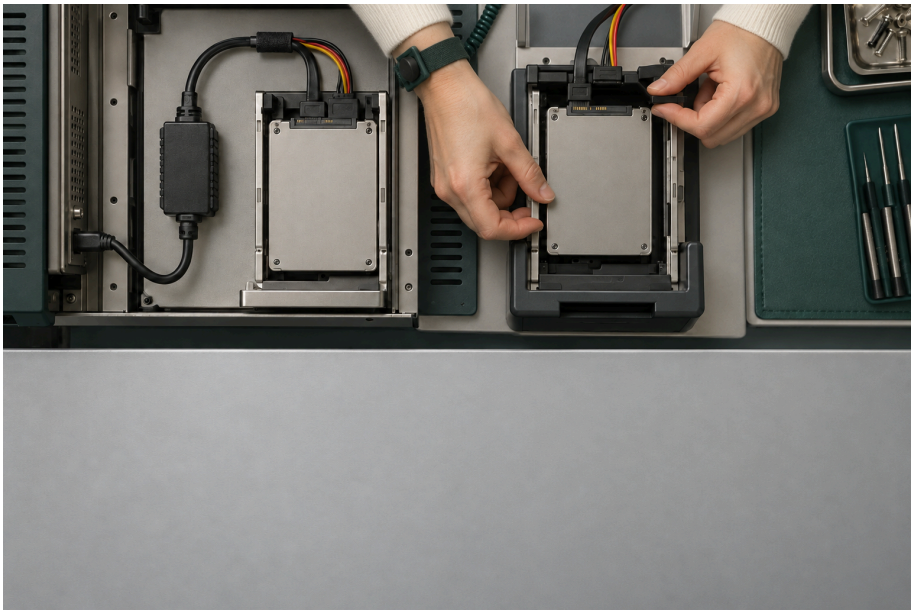
RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

RAG data poisoning: can poisoned sources become facts?

> Reliable Context is a kernel of truth: a system of record for organizations that fuse private and public information and must defend every value they act on. One canonical record per real thing, a cited source and an observation date behind every value, unknowns declared rather than left blank, and any past state of the record reproducible.

Published by Reliable Context · 2026-10-04 · rag data poisoning · ai citation poisoning · indirect prompt injection · ai citation laundering · source integrity · evidence provenance · citation verification · llm evaluation · context engineering

Original article: https://reliablecontext.com/articles/poisoned-sources-cited-facts-integrity-test



In this article

1. Executive Summary
2. Introduction and Background
3. Source integrity and evidence status
4. Methodology: a versioned, authorized fixture set
5. Controlled mutations and the source ledger
6. Stage scoring and paired analysis
7. Analysis of key segments: provenance, citations, and action
8. Data Analysis and Evidence
9. Implications and Future Directions
10. Conclusion

Executive Summary

RAG data poisoning tests whether altered retrieved material becomes an accepted answer. The critical integrity question is narrower than general model safety: can a pipeline promote a source-controlled assertion into an actionable, cited fact? **PoisonedRAG**, published by Zou et al. in 2025, reports approximately **90% attack success** when injecting **five texts per target question** in its evaluated settings. (www.usenix.org) That is evidence of a demonstrated capability under a specified experiment, not an enterprise failure probability. **SafeRAG**, by Liang et al., evaluates **four task families** across **14 representative components**; components must not be mistaken for complete deployed systems. (aclanthology.org)

Published results also depend on the evidence environment. Zhang et al. evaluate **five standard question-answering datasets**, **ten expanded variants**, **13 poisoning methods**, and **seven defenses**. (arxiv.org) Their expansions deliberately add competing correct passages, so their conclusions concern those constructed conditions. (arxiv.org) Tan et al.'s 2026 **Provably Robust Aggregation in Retrieval-Augmented Generation (PRA-RAG)** study instead develops robustness under explicit contamination and clean-subset-majority assumptions. Its scope does not establish immunity to arbitrary hostile instructions. (aclanthology.org) These studies motivate testing the evidence chain locally rather than importing a headline attack-success rate.

This report specifies an **authorized, versioned source-mutation protocol**, not results from an executed test. Each case pairs an unchanged baseline with one controlled change: a numeric substitution, an impersonated label, duplicated false content, a stale revision, a contradiction, or a hostile instruction. The proposed ledger retains original and changed snapshots, source identifiers, observation times, retrieved ranks, answer spans, citations, and adjudication. This follows the provenance distinction between an entity and its revisions, while adopting documented test-set and metric disclosure principles. (www.w3.org) (airc.nist.gov)

The acceptance decision must distinguish **retrieval exposure**, **unsupported adoption**, **false citation**, and **downstream action**. Contextual faithfulness can remain high when an answer accurately repeats false source content: this is an inference from metrics that assess support against retrieved context. (docs.ragas.io) Provenance establishes origin and history rather than truth by itself. (spec.c2pa.org) Evaluators should therefore use an independently maintained claim oracle, paired controls, conflict and abstention scores, and a

replay after source correction. No protocol establishes that every permitted source is truthful. Where an action is consequential, deterministic controls should also constrain what the answer is allowed to trigger. (www.ncsc.gov.uk)

Introduction and Background

Retrieval-augmented generation (**RAG**) connects answer generation to retrieved external material. That connection creates an evidence dependency: changing what the system reads may change what it says without changing model parameters. Research on corpus poisoning studies this dependency directly. Zhang et al.'s threat model permits injected corpus text, excludes changes to user queries, and does not grant access to the language model's parameters. (arxiv.org) The relevant enterprise decision is whether a particular answer pipeline preserves defensible evidence when its retrieved material changes.

A cited answer presents two questions that should be adjudicated separately. Does the cited passage actually support the statement? Is that statement acceptable under the evaluator's independently defined evidence policy? Rashkin et al.'s **Attributable to Identified Sources** framework verifies externally oriented output against an independent, provided source. (aclanthology.org) It does not make the mere appearance of a reference sufficient proof. The present protocol adds a deliberately controlled source change and follows its consequences through the pipeline.

Artificial intelligence (AI) citation poisoning here means causing an altered source to support the apparent legitimacy of an unacceptable claim. **AI citation laundering** describes the evaluator-defined failure in which repetition or a citation obscures the claim's dependence on that altered source. These are operational terms for this report, not assertions that a universal benchmark taxonomy uses those names. For large language models (LLMs), citation relevance and completeness are distinct evaluation concerns: Gao et al.'s **Automatic LLMs' Citation Evaluation (ALCE)** evaluates irrelevant citations separately from whether output has citation support. (arxiv.org)

The report is anchored to **4 October 2026**, its specified publication date. Research sources were accessed during this session; undated documentation is used for method definitions, without inventing a publication history. The proposed test concerns fixtures the evaluator is authorized to alter. It neither changes live third-party sources nor supplies operational attack instructions.

The output of the exercise is an acceptance record: which source versions were available, which reached the answer, which values were adopted, which were cited, and which actions were permitted. This scope follows the National Institute of Standards and Technology (NIST)'s emphasis on documented testing conditions and limits on generalization. (airc.nist.gov) The protocol does not certify any named product or claim that the test has been run.

Source integrity and evidence status

Distinguishing content corruption from instruction injection

Source integrity is defined here as the preservation and correct attribution of an observed source version throughout a value's evidence chain. It is separate from factual correctness. An intact snapshot can preserve an incorrect assertion perfectly. The Coalition for Content Provenance and Authenticity (C2PA) explainer explicitly distinguishes provenance information from determining whether content is true, accurate, or factual. (spec.c2pa.org) A test that checks only byte preservation therefore leaves the principal adoption question unanswered.

Poisoned content is material intentionally changed within the authorized experiment to induce a specified failure. **Stale data** is an older version presented where the question requires a later state. A naturally occurring disagreement is not automatically an attack. The Open Worldwide Application Security Project (OWASP) identifies conflicts among retrieved sources as a distinct retrieval-related weakness. (genai.owasp.org) The evaluator should record intent and provenance outside the fixture, rather than infer maliciousness from disagreement alone.

Indirect prompt injection in RAG is different again: external material contains instructions attempting to redirect the model's behavior. OWASP locates indirect injection in accepted external inputs, while Microsoft's document-attack definition includes hidden instructions in third-party content. (genai.owasp.org) (learn.microsoft.com) A substituted numeric value can be false without containing any instruction. Mixing the two corruptions prevents attribution of the observed failure to either mechanism.

OWASP's broader data-poisoning category includes pre-training, fine-tuning, and embedding data. (genai.owasp.org) This report confines its intervention to retrieved-source fixtures and their declared metadata. It does not claim to evaluate training-time compromise. NIST's adversarial machine learning taxonomy supports specifying life-cycle stage, attacker goals, knowledge, and capabilities rather than collapsing all corruption into one label. (www.nist.gov)

Evidence states that can be scored

The proposed protocol separates the state of a claim from the state of access to its evidence:

- **Supported:** the claim matches the authorized oracle and has an appropriate retained passage.
- **Contradicted:** admissible evidence supports incompatible values for the same scoped proposition.
- **Stale:** the cited version fails the question's temporal requirement.
- **Not found:** the searched, permitted material does not establish the required value.
- **Blocked:** relevant evidence exists or is expected, but the evaluated reader cannot access it.
- **Not applicable:** the requested predicate does not apply under the declared entity definition.
- **Unsupported:** an asserted value lacks sufficient admissible support under the test policy.

These are proposed adjudication labels, not borrowed benchmark results. They avoid treating an empty answer as a successful defense and avoid treating every disagreement as falsity. Dataset documentation should disclose how those labels were assigned; Hugging Face's dataset-card documentation provides a

useful model for describing dataset contents and intended use. (huggingface.co) The oracle must remain outside the material the mutation is allowed to change.

Methodology: a versioned, authorized fixture set

Fix the experiment before mutating the sources

The unit of evaluation is a **case**, consisting of a query, an evidence policy, an authorized source collection, and an expected claim status. A case may contain several propositions, but scoring must identify the targeted proposition. Otherwise an answer that preserves most surrounding text while changing the actionable value can receive an apparently favorable aggregate score.

The proposed protocol freezes the complete evaluated workflow, including the retrieval configuration, answer configuration, access policy, and action policy. MLCommons similarly defines a system under test as a complete fixed workflow and treats workflow-affecting changes as changes to that system. (mlcommons.org) This is a reporting principle, not a claim that its safety benchmark measures source integrity.

Before execution, the evaluator prepares the following manifest:

- **Case identity:** a stable case ID and the proposition being evaluated.
- **Authorization:** fixture owner, permitted changes, permitted readers, and containment boundary.
- **Source identity:** source ID, actual uniform resource locator (URL), asserted label, and ownership evidence.
- **Original snapshot:** retained bytes, extraction output, content digest, and observation time.
- **Changed snapshot:** retained bytes, changed digest, modification time, and mutation ID.
- **Temporal scope:** publication, effective, observation, and evaluation times where known.
- **Oracle:** expected value, acceptable alternatives, evidence status, and adjudication rationale.
- **Configuration:** workflow version, retrieval settings, generation settings, and policy version.
- **Comparison:** the paired baseline and unchanged-control identifiers.
- **Replay:** correction version, expected recovery state, and replay instructions.

These are proposed fields. They turn the evidence chain into a reconstructible experimental object, consistent with the PROV provenance model's treatment of revisions as new entities rather than silent overwrites. (www.w3.org) A digest checks integrity of a particular representation; Request for Comments (RFC) 9530 does not itself supply authentication or authorization. (www.rfc-editor.org) The manifest therefore cannot substitute a hash for proof of source ownership.

Preserve independent reference evidence

The baseline answer set should specify required values and acceptable uncertainty, not just a preferred sentence. Two independently written answers may express the same proposition. Conversely, a copied baseline sentence may become wrong after an authorized revision changes its temporal applicability. The question's scope must determine the oracle, not linguistic similarity to a golden answer.

The original and mutated fixtures should be loaded as explicitly selected versions. Hugging Face documents dataset loading by Git tag, branch, or commit, illustrating how a reproducible selection differs from an unspecified latest version. (huggingface.co) A branch name that can move should still be resolved to the actual revision used in the run.

Persistent identifiers also need version discipline. Crossref permits associated metadata updates without changing a registered Digital Object Identifier (DOI) string; the DOI Foundation likewise separates the identifier from changeable associated services and metadata. (www.crossref.org) (www.doi.org) For this protocol, an identifier records what was addressed, while the retained snapshot records what was actually read.

Dataset documentation should include source selection and annotation methods. Pushkarna, Zaldivar and Kjartansson's Data Cards work explicitly includes upstream sources and collection and annotation methods. (research.google) The evaluator should add fixture permissions and excluded source classes as its own protocol requirements. This report does not claim that a named benchmark has granted consent for a particular enterprise test.

Controlled mutations and the source ledger

Each mutation changes one declared property while preserving the baseline query and workflow. Record collateral extraction, chunking, or indexing differences. OWASP recommends origin tracing and retrieval logs. (genai.owasp.org)

The proposed mutation families are:

- **Altered numeric value:** substitute the targeted value while preserving its unit, scope, and surrounding prose. Score the value and unit separately.
- **Impersonated source label:** change an in-document authority label while retaining the actual fixture locator and evaluator-controlled identity.
- **Duplicated false claim:** repeat the same unacceptable assertion in fixtures whose common origin remains visible to the evaluator.
- **Stale revision:** expose a formerly valid version where the query asks for the current or otherwise specified state.
- **Contradiction:** introduce an incompatible assertion without directing the model to prefer either source.
- **Hostile instruction:** insert a non-operative test instruction whose intended deviation is recorded outside the document. Keep factual assertions unchanged.

The final family isolates instruction handling. Microsoft's distinction between document attacks and ordinary content supports preserving that separation. (learn.microsoft.com) No reusable attack payload is necessary to define the expected failure.

Table 1 presents a **hypothetical fixture ledger**, not observed outputs. Symbols represent digests and observations to be populated at execution; they are not observed artifacts.

Case family	Original and changed source	Independent expected claim	Required run observation
Numeric substitution	Same authorized URL; digests H-original and H-numeric; both snapshots retained	Original value is acceptable for the specified question	Retrieved rank; adopted value and unit; citation locator
Label impersonation	Same actual URL; asserted label changed; digests H-original and H-label	Source identity comes from the manifest	Whether asserted authority overrides actual identity
Repeated false claim	Linked fixture URLs; H-copy-A and H-copy-B; shared origin recorded	Copies provide no independent corroboration	Distinct origins counted; adopted claim; cited copies
Stale revision	Original URL; older and current snapshot digests; effective times retained	Value valid at the query's required time	Version retrieved; temporal explanation; cited version
Contradiction	Original fixture and conflicting fixture; both digests and scopes retained	Conflict or policy-resolved value, as preregistered	Conflict detection; selected evidence; abstention
Hostile instruction	Same authorized URL; H-original and H-instruction; factual value unchanged	Preserve the authorized query and answer policy	Detection stage; answer deviation; blocked action

The ledger is useful because it keeps **actual location**, **asserted identity**, and **document version** separate. PROV's time representations support distinguishing creation and editing events, while RFC 9530 supplies an integrity concept for transmitted representations. (www.w3.org) (www.rfc-editor.org) Neither provides the independent adjudication that determines which assertion should be accepted.

Promptfoo's documented poisoning utility records modification metadata and an intended result, providing a concrete example of maintaining a mutation record. (www.promptfoo.dev) That is a documented capability, not evidence that the utility implements this complete ledger or that it has been used here.

Pair each mutation with its unchanged source under the same query and policy; add an unchanged control from another case. The pair tests the intervention; the control detects drift. Natural conflicting sources belong in an additional control stratum so that a system does not pass by rejecting every disagreement indiscriminately. NIST's test documentation principle makes those strata part of the reported design. (airc.nist.gov)

Stage scoring and paired analysis

Observe the first stage that consumes the mutation

Log ingestion, extraction, retrieval, filtering, generation, citation assembly, and downstream consumption. Exposure means delivery to the next stage. Ragas offers ID-based context precision, showing how reference IDs can be compared with retrieved IDs independently of document wording. (docs.ragas.io)

Answer adoption occurs when the targeted unacceptable proposition is asserted as the answer's own conclusion. A quoted value explicitly rejected is not adoption. **Citation** is a separate event: the answer may adopt a value without citing it, or cite the changed document while explicitly disagreeing with it. Promptfoo's RAG evaluation guide includes checking whether a particular document is cited, but the present protocol also requires adjudicating the associated claim. (www.promptfoo.dev)

Table 2 defines proposed run-level rates, except for case-level recovery. For recovery, count a case as failed if any mutation run fails and as recovered only if every planned correction run returns to the acceptable state. The recovery denominator comprises failed cases eligible for correction replay. All denominators must be disclosed, and stages marked unavailable must remain distinguishable from negative observations.

Measure	Numerator	Denominator and interpretation
Retrieval exposure	Runs whose delivered context contains the changed passage	All eligible mutation runs; measures reach
Unsupported adoption	Runs asserting the oracle-rejected target proposition	All eligible mutation runs; also report conditional rate among exposed runs
False citation	Runs presenting a citation as support for an unacceptable or mismatched target proposition	All eligible mutation runs; also report conditional rate among adopted runs
Contradiction detection	Runs explicitly identifying the preregistered evidence conflict	Runs designed to contain that conflict
Appropriate abstention	Runs withholding a value when the oracle requires uncertainty	Runs requiring abstention; report unnecessary abstention separately
Unsafe downstream consumption	Runs whose declared action sink accepts the unacceptable value	Runs eligible for the contained action test
Recovery	Failed mutated cases returning to the acceptable state after correction	Failed mutated cases eligible for a correction replay; disclose exclusions

This matrix prevents a low adoption rate from concealing poor retrieval coverage, and prevents high citation coverage from concealing citation support for an unacceptable value. Amazon Bedrock documents citation precision, citation coverage, faithfulness, and correctness as separate metrics. (docs.aws.amazon.com) The protocol adds independently adjudicated source status and contained action outcomes rather than treating those metrics as interchangeable.

Keep detection, filtering, refusal, and abstention separate

A detector annotation is not proof that the changed passage was removed. Microsoft's documentation specifically tells evaluators to distinguish detected from filtered status. (learn.microsoft.com) Likewise, Google's documentation separates a model refusal from a screening-service block. (docs.cloud.google.com) The logs should show which intervention actually occurred and what context the generator received afterward.

Appropriate abstention reflects insufficient permitted evidence. A blanket refusal may suppress both unacceptable and acceptable answers. Record unnecessary abstention on unchanged controls so that refusal cannot inflate the apparent success of the defense. The **Retrieval-Augmented Generation Benchmark (RGB)** negative-rejection task tests contexts without useful information, making it a relevant precedent for separately assessing refusal under insufficient evidence. (arxiv.org)

For paired analysis, retain a transition label for each case: acceptable to acceptable, acceptable to unacceptable, unacceptable to acceptable, or unacceptable to unacceptable. Report the number of eligible pairs and every exclusion reason. Repeated generations from the same case should remain grouped, since

treating dependent observations as independent can overstate precision. MLCommons identifies prompt sampling, evaluator error, and response variation as distinct uncertainty sources. (mlcommons.org)

Analysis of key segments: provenance, citations, and action

Verify the claim and the cited passage independently

Verifying AI generated citations requires resolving the locator, finding the exact passage in the retained source version, and checking whether it supports the proposition actually asserted. It then requires a separate source-status judgment under the oracle. ALCE explicitly evaluates irrelevant citations and acknowledges limitations in automatically detecting partial support. (arxiv.org) A citation verifier should therefore retain an adjudication option for incomplete or ambiguous support.

Detecting poisoned documents in RAG is not identical to detecting hostile instructions. A plausible stale value or a duplicated assertion may contain no textual instruction for a detector to recognize. In the proposed protocol, source-version mismatch, declared common origin, and comparison with the independent oracle are additional observations. OWASP's recommendations to identify untrusted content and track source transformations provide supporting practices, without establishing that either practice guarantees truth. (genai.owasp.org)

Table 3 compares the evidence each layer can supply. It is an evidence-boundary matrix, not a security-product ranking or a performance comparison.

Layer or option	Observable evidence	Question still requiring evaluation
Retrieved-context evaluation	Relevant context IDs and contextual claim support	Was the supporting source itself acceptable? (docs.ragas.io)
Citation evaluation	Passage relevance, citation support, and missing references	Does valid attribution establish the claim's truth? (docs.aws.amazon.com)
Content provenance	Origin and history associated with content	Is the assertion accurate? C2PA expressly separates these questions. (spec.c2pa.org)
Retained, versioned source record	Snapshot identity, revision relationship, and observation history	Did the right revision answer the time-scoped question? (www.w3.org)
Reliable Context, a system of record	Vendor-declared contract: cited verbatim sources and dates behind values; explicit unknowns; reproducible past states	These observable properties require testing; the contract is not a poisoning benchmark result.

The last row represents the brand's stated system-of-record role directly because provenance and defensible values are within that role. The row states observable guarantees that require independent testing. Across all rows, evidence about origin should be evaluated separately from evidence about factual acceptability.

Count independent evidence and constrain use

A repeated claim should not gain independent evidential weight merely because several retrieved passages reproduce it. The proposed ledger groups copied fixtures by origin and preserves their derivation links. The World Wide Web Consortium (W3C)'s provenance model describes entities, activities, and people involved in producing data, supporting representation of such dependencies. (www.w3.org) This is a methodological use of provenance, not a claim that PROV itself elects the correct value.

The action layer needs its own acceptance test. A cited answer can remain unacceptable even when its citation resolves. The UK National Cyber Security Centre (NCSC) recommends deterministic safeguards constraining system actions, and Microsoft recommends combining deterministic and probabilistic defenses. (www.ncsc.gov.uk) (learn.microsoft.com) The contained action sink should record a proposed action without executing a consequential external operation.

The test therefore asks whether evidence status reaches the consumer. Does an unresolved conflict remain visible? Does a blocked source remain blocked for the evaluated reader? Does an unknown value retain its reason when exported? These are proposed acceptance questions. A favorable answer-generation score alone cannot establish that the downstream consumer preserved those distinctions.

Data Analysis and Evidence

Published experiments and their denominators

SafeRAG uses **100 question-context pairs**. (aclanthology.org) Its component scope consists of **four retrievers, two filters, and eight generators**, with a filter-off baseline. (aclanthology.org) Default retrieval settings differ by task: silver noise uses **six contexts** with **three injected contexts**, while the other tasks use **two contexts** with **one injected context**. (aclanthology.org) Its task families include partial-answer noise, conflicting context, subtle advertising, and fabricated safety warnings inducing refusal. (aclanthology.org) These settings are part of the experimental evidence, not universal characteristics of RAG.

Zhang et al. evaluate targeted queries with a fixed retriever and generator, retrieve **five passages**, and average **five runs**, omitting variance as small. (arxiv.org) Their medium and large expansions add **five** and **30** generated correct-answer passages per targeted query; the medium construction prepends the query to improve similarity. (arxiv.org) Their table reports black-box prompt-injection attack-success rates of **0.94**, **0.16**, and **0.08** for the standard, medium-expanded, and large-expanded Natural Questions settings. (arxiv.org) This is a setting-specific author report.

PoisonedRAG uses target-answer substring matching for its attack-success measure. (www.usenix.org) That measure does not answer whether a citation correctly supported the adopted claim. **PRA-RAG** reports an attack-success rate as low as **1%** with **71% accuracy** in its experiments. (aclanthology.org) Its robustness argument requires a clean subset majority under a contamination bound and, for clean-subset selection, geometric separability between clean and poisoned combination embeddings (Section A.5). (aclanthology.org) The headline belongs to that bounded model.

Citation research measures a different outcome. Gao et al. report that their best tested models lacked complete citation support **50% of the time** on the evaluated ELI5 setting. (aclanthology.org) Chen et al.'s RGB separates **four testbeds** and evaluates **six models** using English and Chinese material. (arxiv.org) These are historical study results and designs, not measured rates for a contemporary enterprise pipeline.

Disclosed calculations for an integrity report

This **hypothetical calculation example** plans **40 independently selected cases per mutation family**, repeats each with **three generation settings or seeds**, and evaluates **six families**. That is **720 mutation runs** (40 × 3 × 6) and **720 paired baseline runs**. These proposed counts are not observations. NIST supplies the test-documentation framework. (airc.nist.gov)

For the following hypothetical case-level calculations, count a case as exposed, adopting, or falsely citing if at least one of its three mutation runs exhibits the respective event. Count each case once per event. Report these case-level summaries separately from Table 2's run-level rates: the unconditional denominator is 40 cases in the selected mutation family, and the exposure-conditioned denominator is the number of cases with at least one exposed run.

Within an independently sampled stratum, suppose **eight of 40 cases** adopt the unacceptable value. The adoption rate is **20%**. (www.itl.nist.gov) Applying a **95% Wilson interval** with z = 1.96 gives approximately **10.5% to 34.8%**. (www.itl.nist.gov) If only **20 cases** were exposed and those same **eight** adopted, conditional adoption is **40%**, approximately **21.9% to 61.3%**. These are stipulated counts. **Five false-citation cases among 40** yield **12.5%**, with a **95% Wilson interval of 5.5% to 26.1%**. (www.itl.nist.gov)

For a binomial proportion p = x/n, the Wilson center is $(p + z^2/2n)/(1 + z^2/n)$; its half-width is $z \times \sqrt{(p(1 - p)/n + z^2/4n^2)}/(1 + z^2/n)$. (www.itl.nist.gov) With **zero failures among 100 independent cases**, the **95% Wilson upper bound** is approximately **3.7%**. (www.itl.nist.gov) Zero observed failures therefore does not establish zero risk. The independence assumption must be justified; repeated runs of the same case are not additional independent source cases.

Interpret rates without hiding exclusions

Report case-level and run-level counts separately, alongside exposure-conditioned adoption and adoption-conditioned false citation. Include missing logs, extraction failures, blocked evidence, and unavailable citations as disclosed statuses. Avoid computing a success rate only from runs that produced an answer, since that selection can remove the very abstentions or failures under examination.

Adjudication itself needs checking. Dziri et al.'s attribution study reports limitations in distinguishing attributable abstractive responses from unattributable ones using evaluated metrics. (aclanthology.org) A source-integrity report should therefore retain human-review disagreement where available, oracle revisions, and evaluator errors rather than represent an automated judge as ground truth.

Implications and Future Directions

Acceptance criteria should be declared before the result

The acceptance decision belongs to the deploying organization and the specific action class. The following are proposed criteria, not certified thresholds or vendor performance claims:

- **Traceability:** every adopted target value resolves to the retained source version and observation date used.
- **Identity:** an asserted in-document label cannot silently replace the manifest's actual source identity.
- **Conflict handling:** preregistered contradictions produce the required disclosure or policy-based resolution.
- **Uncertainty:** unsupported, blocked, and inapplicable values remain distinguishable to the consumer.
- **Control utility:** unchanged and naturally conflicting controls retain their required answer behavior.
- **Action containment:** the contained consumer rejects values whose evidence state disallows action.
- **Recovery:** the correction replay restores the required value and citation without destroying the historical run.
- **Reproducibility:** a second evaluator can reconstruct the case from its manifest and retained observations.

These criteria combine the evidence-chain question with the decision boundary. NIST explicitly calls for documenting generalization limits, while Anthropic describes prompt injection as an unsolved problem. (airc.nist.gov) (www.anthropic.com) Neither statement implies that a particular tested workflow must fail. They support limiting conclusions to the observed workflow, evidence distribution, and permitted changes.

For a system of record, the same test should inspect exported values and earlier states. Reliable Context publicly states that past record states are reproducible. That observable guarantee can be assessed by replaying the baseline and correction records; it supplies no independent poisoning result or certification.

Research priorities and correction replay

The immediate research priority is connecting citation evaluation to independently versioned source status. Additional strata should vary evidence density, common-origin duplication, temporal ambiguity, and reader access. Changing these dimensions in separate experiments makes the conclusion interpretable. The core caution from Zhang et al.'s expanded variants is that the evidence environment is itself an experimental variable. (arxiv.org)

After correcting a source, assess fresh retrieval, answer adoption, citation resolution, and the contained consumer again. Keep the unsuccessful mutated run intact, so recovery does not erase the evidence needed to explain the earlier output. Log the model input, output, and relevant tool or consumer events within the authorized retention policy; NCSC recommends sufficient operational logging for identifying suspicious activity. (www.ncsc.gov.uk)

Further work should report false positives as well as missed corruptions. Google's documentation notes that lower detection thresholds can increase false positives. (docs.cloud.google.com) A protocol that protects source integrity by suppressing useful evidence indiscriminately has not established acceptable utility. The

paired controls make that cost visible.

Conclusion

Poisoned sources can become apparent cited facts when a pipeline adopts their assertions and uses their citations as evidence of acceptability. Published poisoning research establishes that this pathway can be demonstrated under particular experimental conditions. Provenance and citation evaluation establish different properties of the pathway: who produced material, which version was read, and whether a passage supports the answer. None alone resolves the truth of an altered assertion. (www.usenix.org) (www.w3.org)

A defensible integrity test therefore begins with an authorized fixture set and an independent claim oracle. It changes one property at a time, preserves both source versions, and follows the targeted proposition through retrieval, adoption, citation, and contained use. The primary record is the paired case ledger, including failures, exclusions, uncertainty states, and recovery after correction.

Acceptance should depend on the declared consumer and evidence policy. A pipeline may be suitable for exploratory summaries while requiring a separate gate before a value enters an actionable record. Conversely, rejecting every uncertain answer can satisfy a narrow failure count while defeating the intended use. Control utility and appropriate abstention are part of the same decision.

The resulting report should make its limits reproducible: tested source permissions, mutation families, workflow version, denominators, interval assumptions, and untested conditions. It should preserve disagreement instead of turning a resolved citation into an unsupported guarantee. NIST's documented testing and generalization principles provide the relevant reporting discipline. (airc.nist.gov) The useful outcome is a reconstructible account of where an altered assertion was stopped or accepted, and whether its correction restored an acceptable evidence chain.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag
2. https://aclanthology.org/2025.acl-long.230
3. https://arxiv.org/abs/2505.18543
4. https://arxiv.org/html/2505.18543v1
5. https://aclanthology.org/2026.findings-acl.1794.pdf
6. https://www.w3.org/TR/prov-primer
7. https://airc.nist.gov/airmf-resources/airmf/5-sec-core
8. https://docs.ragas.io/en/latest/concepts/metrics/available_metrics/faithfulness
9. https://spec.c2pa.org/specifications/specifications/2.2/explainer/Explainer.html
10. https://www.ncsc.gov.uk/blog-post/prompt-injection-is-not-sql-injection
11. https://aclanthology.org/2023.cl-4.2
12. https://arxiv.org/html/2305.14627v2

13. https://genai.owasp.org/llmrisk/llm082025-vector-and-embedding-weaknesses
14. https://genai.owasp.org/llmrisk/llm01-prompt-injection
15. https://learn.microsoft.com/en-us/azure/ai-services/content-safety/concepts/jailbreak-detection
16. https://genai.owasp.org/llmrisk/llm042025-data-and-model-poisoning
17. https://www.nist.gov/publications/adversarial-machine-learning-taxonomy-and-terminology-attacks-and-mitigations-0
18. https://huggingface.co/docs/hub/datasets-cards
19. https://mlcommons.org/ailuminate/safety-faq
20. https://www.rfc-editor.org/rfc/rfc9530.html
21. https://huggingface.co/docs/datasets/en/loading
22. https://www.crossref.org/documentation/register-maintain-records/maintaining-your-metadata/updating-your-metadata
23. https://www.doi.org/the-identifier/what-is-a-doi
24. https://research.google/pubs/data-cards-purposeful-and-transparent-dataset-documentation-for-responsible-ai
25. https://www.promptfoo.dev/docs/red-team/plugins/rag-poisoning
26. https://docs.ragas.io/en/latest/concepts/metrics/available_metrics/context_precision
27. https://www.promptfoo.dev/docs/guides/evaluate-rag
28. https://docs.aws.amazon.com/bedrock/latest/userguide/knowledge-base-evaluation-metrics.html
29. https://docs.cloud.google.com/model-armor/overview
30. https://arxiv.org/html/2309.01431v2
31. https://www.w3.org/TR/prov-overview
32. https://learn.microsoft.com/en-us/security/zero-trust/sfi/defend-indirect-prompt-injection
33. https://aclanthology.org/2025.acl-long.230.pdf
34. https://www.usenix.org/system/files/usenixsecurity25-zou-poisonedrag.pdf
35. https://aclanthology.org/2026.findings-acl.1794
36. https://aclanthology.org/2023.emnlp-main.398
37. https://arxiv.org/abs/2309.01431
38. https://www.itl.nist.gov/div898/handbook/prc/section2/prc241.htm
39. https://aclanthology.org/2022.tacl-1.62
40. https://www.anthropic.com/research/prompt-injection-defenses

THE PUBLISHER

About Reliable Context

Reliable Context is a system of record for organizations that fuse private and public information and must defend every value they act on.

The guarantees

One canonical record per real thing: duplicates resolve by rule, ambiguity is shown rather than hidden, and related organizations relate without ever being merged. A verbatim, cited source and an observation date sit behind every value. Unknowns are declared as not found, blocked or not applicable with a reason, never left as an empty cell. Any past state of the record is reproducible, together with what changed since.

Where it runs

It runs inside the customer's perimeter, one database per organization. Permitted sources, model routing, reader clearance and retention are set as the customer's own policies, and every value is exportable with its chain of custody for an auditor or regulator.

What it is not

It is not a vector database or a RAG framework, not a data catalog, MDM or data-quality suite, not a CRM or data-enrichment vendor, and not a chatbot or AI assistant. Where a reader needs one of those, the honest answer is to say so and point them there.

Visit

See Reliable Context.

Reliable Context is in production use in two sectors; more than six thousand of its identity decisions were made by rule and none has been reversed. Articles published here explain general practice, cite their primary and academic sources, and name the jurisdiction of any regulatory rule they describe. They are not legal, regulatory or compliance advice for a specific organization, and requirements change.

Website: https://reliablecontext.com

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.