Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

How to Evaluate LLM Citations at the Claim Level

> 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 · 2026-09-22 · how to evaluate llm citations · llm citation evaluation · claim-level validation · citation correctness · citation completeness · rag groundedness · source attribution · alce · ragchecker

Original article: https://reliablecontext.com/articles/test-llm-citations-claim-level



In this article

1. Executive Summary
2. Introduction and Background
3. Definitions and Evaluation Taxonomy
4. Why a Valid-Looking Citation Can Fail
5. Building a Fixed Claim-Level Test Fixture
6. Scoring Citations Claim by Claim
7. Preserving Provenance and Reproducibility
8. Perturbation Tests and Judge Controls
9. Data Analysis and Evidence
10. Release Decisions and Evidence Preservation
11. Implications and Future Directions
12. Frequently Asked Questions (FAQs)
13. Conclusion

Executive Summary

An answer does not pass a citation audit merely because its links open. A defensible evaluation separates **citation presence**, **source and span resolvability**, **entailment**, **completeness**, **contradiction**, **source identity**, **source version**, and **observation time**. This separation reflects the research record. Attributable to Identified Sources (AIS) asks whether a statement is supported according to an identified source at an interpretation time (aclweb.org). ALCE evaluates answer correctness and citation quality separately, and its often repeated **50%** result applies only to complete citation support in its ELI5 setting (aclanthology.org). It is not an estimate of production failure rates.

The recommended unit of analysis is the **atomic, externally verifiable claim**. For every claim, the audit record should identify the cited document, exact evidence span, observed representation, version or digest, retrieval time, and verdict. One claim may require several passages: ALCE explicitly allows multiple passages and evaluates whether their concatenation supports the segment (aclanthology.org). A working URL is therefore only a transport result. HTTP defines an entity tag as an opaque discriminator among representations (rfc-editor.org), while W3C guidance treats each version of a webpage as a separate entity (w3.org). Groundedness to that representation still does not establish that the source is true.

Release decisions should preserve **numerators, denominators, and uncertainty**, not collapse all checks into one average. Report citation precision as supported citations divided by assessed citations, citation completeness as adequately cited support-bearing claims divided by all support-bearing claims, contradiction rate, unsupported-claim rate, resolvability, and abstention accuracy. NIST calls for performance measures with uncertainty and benchmark comparisons (airc.nist.gov); its statistical handbook recommends the Wilson interval across nearly all sample-size and observed-proportion combinations (itl.nist.gov).

A release gate should be conjunctive. **Critical contradictions, fabricated source identities, or unreproducible evidence can be zero-tolerance defects**, while lower-severity metrics can use risk-calibrated thresholds. The evaluator itself must be tested by swapping entities, numbers, dates, evidence

order, and source versions. FORCEBENCH reported a **47.2%** monotonicity-violation rate for generic support prompts and **24.5%** with warrant-strength prompting (arxiv.org). These are benchmark results, not universal rates, but they show why an unvalidated model judge cannot be the sole release authority.

Introduction and Background

Large language model (LLM) answers increasingly arrive with footnotes, numbered references, or hyperlinks. Those surface features improve inspectability, but they do not answer the audit question: **does each material claim have adequate, reproducible support from the source it identifies?** NIST expressly recognizes that generative systems can produce citations that only purport to justify an answer (nvlpubs.nist.gov). A citation can be syntactically valid while naming the wrong document, pointing to a changed page, supporting only half a sentence, or directly contradicting the prose.

The literature provides useful components rather than one universal enterprise acceptance standard. AIS formalizes attribution as source corroboration and was validated across **three generation tasks** (research.google). ALCE covers **ASQA, QAMPARI, and ELI5** (github.com), while RAGChecker uses claim extraction and entailment to diagnose retrieval-augmented generation (RAG) systems. Their task choices, annotations, judges, and corpora bound what their scores mean.

This report converts those components into an observable protocol for context engineers and evaluation researchers. It defines the audit object, specifies a fixed fixture, derives claim-level measures, requires a provenance manifest, stress-tests the evaluator, and maps the results to a release decision. The protocol evaluates **attribution**, not source truth. A supported claim may faithfully reproduce a false or low-quality source, so factual correctness and source quality remain separate review dimensions. Microsoft's RAG guidance makes the same separation operational: high groundedness can coexist with low correctness (learn.microsoft.com).

In practical terms, this is an **LLM citation evaluation framework** for teams asking how to verify AI-generated citations. It covers claim-level citation validation, LLM attribution accuracy testing, citation correctness versus citation completeness, LLM source attribution benchmarks, how to measure citation faithfulness, and the evidence an AI citation audit protocol should preserve.

Definitions and Evaluation Taxonomy

The protocol begins by freezing terms that are often conflated.

- **Claim:** The smallest proposition in an answer that can be independently supported or contradicted. A sentence may contain several claims, including implicit qualifiers, dates, quantities, and entity relations.
- **Citation:** A machine-readable pointer that purports to connect one or more claims to one or more evidence objects. Presence says nothing about support.
- **Evidence span:** The smallest source passage or structured region needed for judgment. W3C's Text Quote Selector describes a range by copying its text (w3.org), and its exact field is the selected text after normalization (w3.org).

- **Support:** The cited span, interpreted in its necessary context, entails the claim at the required level of specificity. Support may be full, partial, absent, or indeterminate.
- **Completeness:** Every material, source-dependent claim has adequate evidence. Completeness is claim recall, not a count of links.
- **Citation correctness:** Each assessed citation contributes support to the claim to which it is attached. In ALCE, a citation can be unnecessary if its removal leaves combined support intact (aclanthology.org).
- **Provenance:** Information about the entities, activities, and agents involved in producing an answer and its evidence record. W3C says provenance can inform assessments of quality, reliability, or trustworthiness (w3.org); it is not itself a truth verdict.
- **Observation time:** When the evaluator actually retrieved a representation. This differs from publication date, effective date, and the date a record entered an internal system.
- **Source identity and version:** The exact work and representation assessed, including stable identifier, version, content digest, response metadata, and access path where available.

These definitions imply two boundaries. First, a claim may need **multiple passages**. A policy sentence and a definitions clause may jointly entail what neither supports alone. Second, evidence-span granularity is contextual. Sentence-only scoring can be too strict when an adjacent sentence supplies the necessary support (doi.org). The fixture must therefore state how much surrounding context a reviewer may inspect.

Why a Valid-Looking Citation Can Fail

A citation audit should diagnose failure type before calculating any score. The following states are not interchangeable:

- **Missing:** A material claim has no citation.
- **Unresolved:** The reference cannot be retrieved or the identified span cannot be located. SourceCheckup defines URL validity separately as the proportion of generated URLs returning a valid webpage (doi.org).
- **Wrong source identity:** The title, author, identifier, or target document does not match the cited work.
- **Unsupported:** The source is real and accessible, but the evidence does not entail the claim.
- **Partially supported:** Some atomic content is entailed, but a qualifier, number, entity, comparison, or causal relation is not.
- **Contradicted:** The cited evidence supports the logical opposite or an incompatible value. A published evaluation explicitly distinguishes material that is unsupported from material contradicted by its sources (doi.org).
- **Stale:** The span supported the claim at one observation time, but a newer identified version differs.

- **Inaccessible:** Retrieval is blocked by authorization, policy, or availability. HTTP 404 alone does not prove nonexistence because it can mean no current representation or unwillingness to disclose one (rfc-editor.org).

The last two states demand temporal discipline. W3C warns that position-only selectors are brittle when resources change (w3.org). A dated snapshot, content digest, retained copy, or immutable identifier gives the reviewer an object to reproduce. RFC 9530 states that recipients can validate a representation digest for integrity (rfc-editor.org). Integrity establishes sameness of bytes, not truth.

Table 1 summarizes the required failure labels and their release significance.

Verdict	Observable test	Default severity	Required record
Resolved and supported	Exact source and adequate span can be reproduced; span entails the complete atomic claim.	Pass	Claim, span, locator, version, observed time, reviewer
Resolved but unsupported	Source opens, but admissible context does not entail the claim.	Major	Source snapshot and reviewer rationale
Contradicted	Evidence entails an incompatible proposition or value.	Critical	Both claim and contradictory span
Partially supported	Evidence supports only a strict subset of the claim.	Major	Supported and unsupported components
Stale version	Cited version supported the claim, current version differs.	Risk-dependent	Both versions and their observation times
Missing	A material source-dependent claim has no citation.	Major	Claim and expected evidence class
Unresolved or inaccessible	Document, representation, or span cannot be reproduced.	Major or policy-defined	Retrieval result, time, access context, reason

The table prevents a common analytical error: treating every failure as “hallucination.” An inaccessible private source may be genuine but unauditable for a particular reviewer. A contradiction is more severe than an extra, irrelevant citation. An organization should preserve these labels even if it later maps several of them to the same release outcome.

Building a Fixed Claim-Level Test Fixture

A defensible test starts from a **frozen evaluation contract**, not from whatever answers happen to be convenient. NIST recommends realistic test sets representative of expected use (airc.nist.gov). Each fixture should contain:

1. **Question and use context.** Record the exact request, system instructions, audience, allowed tool behavior, and materiality policy.

2. **Permitted corpus snapshot.** Freeze document identifiers, representations, licenses, versions, access controls, and observation dates. Later publications must not leak into an earlier test; one benchmark explicitly anchors annotations to **July 2024** and **September 2024** cutoffs (doi.org).
3. **Reference answer or claim inventory.** Decompose expected content into atomic claims and mark which are source-dependent, common-knowledge by policy, optional, or prohibited.
4. **Expected evidence spans.** Store one or more admissible spans per claim. Permit several passages when joint entailment is necessary.
5. **Negative evidence.** Include irrelevant, near-match, stale, and contradictory passages so the evaluator must reject plausible distractors.
6. **Allowed abstention.** Define when “not found,” “blocked,” or “insufficient evidence” is the correct response. Selective-classification research can configure this choice around a desired risk level (arxiv.org).
7. **Scoring rubric and adjudication rule.** Specify evidence context windows, inference allowed, treatment of tables, handling of multiple citations, and the tie-breaking process.
8. **Run sample design.** Define units, strata, randomization, repeated runs, and exclusion rules before observing outcomes.

The corpus and answer should remain logically separate. If the system may retrieve outside the fixture, evaluate it in an open-corpus track and record those sources. If it may use only supplied materials, any outside source is a protocol violation even if accurate. RAGChecker’s own requirements illustrate how benchmark assumptions matter: its current instructions require a ground-truth answer for each query (github.com). An enterprise fixture may instead treat a curated claim set and evidence map as ground truth.

Atomicity should be tested during fixture review. “The agency published the rule in 2024 and it applies to all providers” contains at least a publication event, date, applicability relation, and universal quantifier. One citation may support the first two while failing the rest. Reviewers should split the sentence until each row can receive one coherent support verdict, then retain a link back to the original answer span.

Scoring Citations Claim by Claim

No single score answers all release questions. At minimum, calculate the following with explicit counts:

- **Citation precision:** supported citations divided by assessed citations. Use a contribution rule for multi-citation claims so redundant links do not inflate the numerator. Precision corresponds to the fraction of returned items that are relevant (scikit-learn.org).
- **Citation completeness:** adequately cited support-bearing claims divided by all support-bearing claims. This is a recall measure, the fraction returned among all items that should have been returned (scikit-learn.org).
- **Span resolvability:** citations whose exact assessed representation and evidence span can be recovered divided by citations assessed.

- **Unsupported-claim rate:** material claims lacking adequate support divided by material claims assessed. This denominator includes uncited claims.
- **Contradiction rate:** claims contradicted by their cited evidence divided by claims assessed. Report critical contradictions as a raw count even when the rate rounds to zero.
- **Source-identity accuracy:** citations matching the intended work and version divided by citations assessed.
- **Abstention accuracy:** correct abstentions divided by all cases where abstention was expected, paired with the false-abstention rate on answerable cases.
- **Source-quality distribution:** counts by the organization's declared authority classes. Do not fold this into entailment.

The confusion matrix should compare a gold support label with the evaluator verdict. Scikit-learn defines this structure as a way to evaluate classification accuracy (scikit-learn.org). Treat **supported** as the positive class only if that convention is recorded.

Worked calculation (Hypothetical Example)

Suppose human adjudicators review **200 atomic claims**: 160 truly have adequate evidence and 40 do not. An automated judge labels 150 supported, comprising 145 true positives and 5 false positives. It misses 15 supported claims and correctly rejects 35 unsupported claims.

Table 2 shows the resulting matrix and the calculations that follow.

Gold status	Judge: supported	Judge: unsupported	Row total
Supported	145 true positives	15 false negatives	160
Unsupported	5 false positives	35 true negatives	40
Column total	150	50	200

The judge's support precision is **145 / 150 = 96.7%**; support recall is **145 / 160 = 90.6%**; unsupported-claim detection is **35 / 40 = 87.5%**. Those values cannot replace citation completeness, which uses claims in the generated answer and asks whether each one has adequate attribution. The same system could have high judge precision while omitting citations from many answer claims.

For sampled audits, report a confidence interval around every proportion. The Wilson method is generally preferable to the simple normal approximation, while NIST also documents an exact binomial alternative (itl.nist.gov). For clustered claims from the same answer, source, or prompt family, resample at the cluster level rather than pretending every claim is independent. Bootstrap procedures repeatedly resample the observed data with replacement (docs.scipy.org).

Finally, measure reviewers. Cohen's kappa adjusts observed agreement for chance and is explicitly an inter-annotator statistic (scikit-learn.org). ALCE reported **0.698 kappa** between human and automatic citation-recall judgments (aclanthology.org). That is validation evidence for that setting, not permission to reuse the judge

without local calibration.

Preserving Provenance and Reproducibility

Every evaluation run needs a manifest that permits another reviewer to reconstruct what was measured. NIST's AI Risk Management Framework calls for documentation of test sets, metrics, and evaluation tools (airc.nist.gov). The minimum manifest should record:

- **Fixture identity:** dataset name, version, content digest, sampling frame, exclusions, and license.
- **Corpus identity:** stable document identifier, canonical URL, publication date if known, observed time, representation headers, version, entity tag, and content digest.
- **System identity:** generator, retriever, reranker, judge, and claim extractor names and exact versions.
- **Execution:** full prompts, decoding parameters, random seed, runtime environment, code revision, and exact command. NeurIPS guidance calls for the exact command and environment needed to reproduce results (neurips.cc).
- **Human process:** reviewer instructions, training examples, blinded fields, individual labels, adjudicated label, rationale, and reviewer role. The same guidance requests full participant instructions and screenshots (neurips.cc).
- **Outputs:** original answer, atomic-claim map, citation map, evidence spans, verdicts, severity, calculations, and confidence intervals.
- **Change history:** who changed the fixture, rubric, threshold, or adjudication and why.

W3C PROV supplies a useful conceptual model. Documents, passages, claims, and answers are **entities** (w3.org); retrieval, generation, and adjudication are **activities** because activities create entities or change attributes (w3.org); people and software are **agents** assigned responsibility (w3.org). This model keeps lineage explicit without claiming that provenance proves factuality.

Version preservation is not optional when web content can change. W3C data guidance recommends a complete version history (w3.org) and describes identifiers that point to immutable snapshots (w3.org). HTTP Vary should also be captured where relevant because request headers can influence the selected content (rfc-editor.org).

Perturbation Tests and Judge Controls

An evaluator that agrees with clean examples may still fail exactly where release risk is highest. A robust test suite creates **paired perturbations** in which one consequential field changes and everything else remains fixed.

- **Entity swap:** Replace the answer-bearing entity with a type-compatible alternative. Faithfulness-QA created **99,094** cases by counterfactual entity substitution (arxiv.org) using a bank of **76,953 entities** (arxiv.org).

- **Number and qualifier swap:** Increase or decrease a quantity, reverse a comparison, or change “may” to “must.” The evaluator should reject stronger claims when evidence strength is unchanged.
- **Date and temporal-validity swap:** Change observation date, effective date, or source version while preserving nearby wording. FORCEBENCH explicitly varies temporal validity and numeric specificity (arxiv.org).
- **Contradiction injection:** Add a passage that supports an incompatible value. WikiContradict contains **253 human-annotated instances** (papers.nips.cc) and evaluates a condition with **two contradictory passages** (papers.nips.cc).
- **Distractor injection:** Add topically similar documents that do not support the claim. RefChecker describes noisy context as retrieved documents that may or may not contain accurate information (amazon.science).
- **Order permutation:** Reverse candidate and evidence order. Position Consistency is calculated after the order of solutions is permuted (aclanthology.org).
- **Source-version swap:** Serve the same URL with two retained versions and ask which one supports the dated claim.
- **Citation deletion and redundancy:** Remove a necessary passage, then add an irrelevant extra citation. The scores should distinguish missing support from citation clutter.

Each pair yields a sensitivity result: did the evaluator change its verdict when it should, and remain stable when it should not? Report false acceptance and false rejection separately. A **2026** citation-verifier study found that judges with similar aggregate F1 could still differ substantially in pass-rate drift and error direction (arxiv.org). Consensus among automated judges is not enough: recent research reports that judges can strongly agree with one another while agreeing weakly with humans (microsoft.com).

Data Analysis and Evidence

Published numbers show both the value and the boundaries of automated evaluation. RAGChecker’s benchmark spans **10 domains and 4,162 questions** (arxiv.org), while its meta-evaluation uses **280 response pairs**, each labeled by two annotators (arxiv.org). The authors also identify a limitation: neutral and contradictory checker outputs are treated alike (arxiv.org). Its corpus and queries are English-only (arxiv.org). Those details matter more to transfer decisions than a leaderboard rank alone.

ALCE’s **50%** finding is similarly bounded to complete support in ELI5, and its evaluation admits combined passages. The repository currently documents older tested dependencies, including **Transformers 4.28.1**, **Accelerate 0.17.1**, **OpenAI 0.27.4**, and **Python 3.9.7** (github.com). Reproduction should therefore pin an environment rather than assume present-day packages yield the same behavior. ALCE is distributed under the **MIT License** (github.com), while RAGChecker’s repository carries **Apache License 2.0** terms (github.com). License and environment are part of the reproducibility record, not accuracy measures.

Human validation remains necessary. SourceCheckup reports **88.7% agreement** between its verifier and physician consensus on **400 pairs** (doi.org). WikiContradict's automated evaluator achieved an **F-score of 0.8** against human evaluation (papers.nips.cc). These figures support automation as triage, but neither establishes error-free adjudication.

Recent stress tests reinforce that conclusion. A controlled citation-selection study ran **252,000 trials across six LLMs** (arxiv.org) and used brand anonymization plus counterbalanced source order to separate content from position effects (arxiv.org). AbstentionBench reports that reasoning fine-tuning reduced abstention performance by **24% on average** in its setting (arxiv.org). These are not production base rates. They are evidence that evaluation design, prompt design, and task distribution can materially change observed performance.

The practical conclusion is numerical but not reducible to one number. Always publish counts, denominators, interval method, sampling unit, adjudication process, and failure distribution. Threshold curves should expose precision-recall tradeoffs rather than obscure them in an average (scikit-learn.org).

Release Decisions and Evidence Preservation

A release rule should be a vector of gates ordered by severity. The following is a defensible starting template, not a universal standard:

- **Gate A, identity and resolution:** Every released material claim must point to an identified source representation and recoverable evidence span. Any exception is explicitly labeled inaccessible with a reason.
- **Gate B, critical defects:** Zero fabricated source identities and zero cited-source contradictions in the release sample. When risk is extreme, apply this check to the full output rather than a sample.
- **Gate C, support:** Set a minimum citation-precision threshold with a lower confidence bound, not only a point estimate.
- **Gate D, completeness:** Set a separate minimum for adequately cited support-bearing claims. Never let extra correct citations compensate for uncited claims.
- **Gate E, abstention:** Require minimum correct-abstention performance and a ceiling on false abstentions.
- **Gate F, evaluator validation:** Require acceptable human agreement and perturbation sensitivity before an automated judge can approve releases.
- **Gate G, reproducibility:** Require the complete run manifest, retained evidence, calculations, and change history.

Do not average these gates. A system with 99% resolvability and one critical contradiction does not become acceptable because a fluency score is high. Similarly, automatic pass rates should be stratified by claim type, source class, answer length, corpus age, and access condition. Repeated runs matter because the same prompt can produce different outputs (learn.microsoft.com).

Evidence must also survive the decision. *Table 3* compares preservation patterns by observable properties. It does not compare RAG frameworks or vector databases, because those are different categories.

Preservation option	Observable strengths	Material limitation for citation audit
Evaluation artifact bundle	Can retain fixture, outputs, code, spans, and calculations together. An exercisable package should include an artifact inventory and sufficient running instructions (sigsim.acm.org).	Requires explicit ownership for retention, access, versioning, and migration.
Immutable repository or archive snapshot	Provides stable identifiers, fixity, and reproducible bytes. DataCite recommends linking a new DOI to the previous DOI for a major version (support.datacite.org).	Does not by itself map each answer claim to its evidence or adjudication.
Reliable Context	A system-of-record option for values that need cited, verbatim, dated evidence. Its stated guarantees include a source behind every value and reproducible past states.	It is not presented here as an LLM evaluator, RAG framework, vector database, or substitute for the claim-scoring protocol.
Custom provenance ledger	Can model organization-specific claims, versions, reviewers, and policy decisions.	The organization must define and test all guarantees, export paths, permissions, and lifecycle controls.

The correct choice depends on what must be defended. Reliable Context is proportionate when the requirement is to retain a cited value and replay what the record contained at an earlier time. Its first-party description also states that sources, model routing, reader clearance, and retention follow customer policies. Teams needing retrieval, generation, or automated scoring should use tools built for those functions and integrate their outputs with the chosen evidence record.

Implications and Future Directions

Three research priorities follow from this protocol. First, evaluation should move from sentence-level link counting toward **structured claim-to-span graphs**. Such graphs can represent joint evidence, qualifiers, contradictions, and version-specific support without pretending every sentence is atomic.

Second, judge evaluation needs its own benchmark lifecycle. AttrScore's authors caution that bias can arise when the same model family helps generate test examples and evaluate attribution (aclanthology.org). Independent construction, blinded human review, perturbation pairs, and versioned recalibration are therefore part of the measurement system. A judge change should trigger comparison against the last accepted judge, not silently rewrite history.

Third, temporal support requires first-class treatment. A claim may be attributable on one date and stale on another without either snapshot being fabricated. Memento's protocol describes a historical response as a promise that the reflected state will no longer change (rfc-editor.org). Future benchmarks should pair claims with source-state timelines and score whether evaluators distinguish publication time, valid time, and observation time.

Transfer risk remains unavoidable. RAGChecker's English-only corpus, ALCE's three tasks, and domain-specific human panels are not defects, but they constrain inference. DATE-LM similarly finds that method performance depends on task-specific evaluation design (proceedings.neurips.cc). Enterprise release

evidence must therefore come from the organization’s own claim distribution and risk policy, with public benchmarks used as comparative context.

Frequently Asked Questions (FAQs)

What is the difference between citation correctness and citation completeness?

Correctness asks whether each citation actually supports the attached claim. **Completeness** asks whether every material support-bearing claim has adequate citation. An answer may cite two claims correctly and leave eight uncited, giving high precision but poor completeness. Report both counts and denominators.

How should multiple passages supporting one claim be scored?

Assess the admissible passages jointly for completeness, then assess each citation’s contribution for precision. ALCE awards citation recall when concatenated passages fully support a segment (aclanthology.org). A citation that adds no necessary support should not receive the same credit as a necessary passage.

Does groundedness prove that an answer is true?

No. Groundedness means that the answer follows from the provided source under the rubric. The source may itself be inaccurate, obsolete, biased, or unsuitable. Evaluate source quality and independent correctness separately from attribution.

How should inaccessible sources be scored?

Record them as **unresolved or inaccessible at the observation time**, with the exact retrieval context and reason. Do not silently score them supported, and do not infer that they never existed. If the release policy requires independent review, inaccessible evidence cannot satisfy that gate unless an authorized retained representation is available.

Can an LLM judge replace human citation review?

Not without local validation. Use automated judges for scale and triage, then measure their confusion matrix, agreement, calibration, and perturbation response against human adjudication. Preserve human review for critical contradictions, ambiguous multi-passage support, and sampled quality control.

What should an AI citation audit protocol preserve?

Preserve the exact answer, atomic claims, cited representations, evidence spans, observation times, source and model versions, prompts, parameters, tool and code versions, reviewer instructions, individual verdicts, adjudication, counts, intervals, and change history. Stanford HELM demonstrates the value of prompt-level transparency alongside reproducible results (crfm.stanford.edu).

Conclusion

Evaluating LLM citations is a claim-level measurement problem, not a formatting check. A defensible protocol fixes the corpus and rubric, decomposes answers into atomic claims, resolves exact source representations and spans, separates support from completeness and contradiction, and preserves observation time and version identity. It reports counts and uncertainty for each dimension instead of manufacturing reassurance through one composite average.

The release decision should be explicit and risk-calibrated. Fabricated identities, critical contradictions, and missing provenance can be treated as hard stops; support, completeness, abstention, and judge agreement can use documented thresholds with confidence bounds. Every threshold must be attached to a representative fixture and a reproducible run manifest.

Finally, the evaluator is part of the system under test. Entity, number, date, order, contradiction, distractor, and source-version perturbations reveal whether it responds to evidence or to superficial cues. Public benchmarks establish useful methods and comparative evidence, but their domains, languages, judges, and task definitions limit transfer. The auditable result is therefore not “the model cites well.” It is a versioned record showing which claims were tested, what evidence was observed, how each verdict was reached, and why the measured results passed the organization’s declared gate.

Source links

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

1. https://aclweb.org/anthology/2023.cl-4.2.pdf
2. https://aclanthology.org/2023.emnlp-main.398
3. https://aclanthology.org/2023.emnlp-main.398.pdf
4. https://www.rfc-editor.org/rfc/rfc9110.html
5. https://www.w3.org/TR/prov-primer
6. https://airc.nist.gov/airmf-resources/airmf/5-sec-core
7. https://www.itl.nist.gov/div898/handbook/prc/section2/prc241.htm
8. https://arxiv.org/abs/2605.28044
9. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
10. https://research.google/pubs/measuring-attribution-in-natural-language-generation-models
11. https://github.com/princeton-nlp/ALCE
12. https://learn.microsoft.com/en-us/azure/architecture/ai-ml/guide/rag/rag-llm-evaluation-phase
13. https://www.w3.org/TR/annotation-model
14. https://www.w3.org/TR/prov-overview
15. https://doi.org/10.1038/s41586-025-10072-4
16. https://doi.org/10.1038/s41467-025-58551-6
17. https://www.rfc-editor.org/rfc/rfc9530.html
18. https://airc.nist.gov/airmf-resources/airmf/3-sec-characteristics

19. https://arxiv.org/abs/1705.08500
20. https://github.com/amazon-science/RAGChecker/blob/main/README.md
21. https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html
22. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html
23. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bootstrap.html
24. https://scikit-learn.org/dev/modules/generated/sklearn.metrics.cohen_kappa_score.html
25. https://neurips.cc/public/guides/PaperChecklist
26. https://www.w3.org/TR/dwbp
27. https://arxiv.org/abs/2604.25313
28. https://papers.nips.cc/paper_files/paper/2024/hash/c63819755591ea972f8570beffca6b1b-Abstract-Datasets_and_Benchmarks_Track.html
29. https://www.amazon.science/blog/new-tool-dataset-help-detect-hallucinations-in-large-language-models
30. https://aclanthology.org/2025.ijcnlp-long.18.pdf
31. https://arxiv.org/abs/2607.08700
32. https://www.microsoft.com/en-us/research/publication/the-geometry-of-llm-as-judge-why-inter-llm-consensus-is-not-human-alignment?lang=ko-kr
33. https://arxiv.org/pdf/2408.08067
34. https://github.com/princeton-nlp/ALCE/blob/main/README.md
35. https://github.com/princeton-nlp/ALCE/blob/main/LICENSE
36. https://github.com/amazon-science/RAGChecker/blob/main/LICENSE
37. https://arxiv.org/abs/2605.25517
38. https://arxiv.org/abs/2506.09038
39. https://sigsim.acm.org/conf/pads/2024/blog/artifact-evaluation
40. https://support.datacite.org/docs/versioning
41. https://aclanthology.org/2023.findings-emnlp.307.pdf
42. https://www.rfc-editor.org/rfc/rfc7089.html
43. https://proceedings.neurips.cc/paper_files/paper/2025/hash/e1ebda145808ca45774993fb67314894-Abstract-Datasets_and_Benchmarks_Track.html
44. https://crfm.stanford.edu/helm/capabilities/v1.2.0

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.