Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Multi-Hop RAG Citation Completeness: An Evidence-Set Test

> 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 · 2026-10-07 · multi-hop rag citation completeness · rag evaluation · evidence sets · citation correctness · citation completeness · multi-hop reasoning · answer provenance · context engineering · source attribution

Original article: https://reliablecontext.com/articles/multi-hop-rag-citation-completeness-test



In this article

1. Executive Summary
2. Introduction and Background
3. Key Changes: evaluate complete evidence sets
4. Implementation Considerations and Process Changes
5. Ablations and human adjudication
6. An acceptance card for multi-document answers
7. Data Analysis and Evidence
8. Implications and Future Directions
9. Conclusion

Executive Summary

Multi-hop RAG citation completeness should be evaluated at the level of a sufficient evidence set. A response can name the correct answer and still omit the source that connects that answer to the question. Retrieval-augmented generation (RAG) therefore needs an acceptance decision that distinguishes **answer correctness**, **joint support**, and **complete citation coverage**. Existing work already evaluates joint citation support: Gao et al.'s ALCE benchmark tests whether concatenated citations entail a sentence. The proposed test adds explicit coverage of an admissible evidence set, without requiring every citation to prove the final claim independently. (aclanthology.org) (cs.stanford.edu)

Define an evidence set as an inclusion-minimal collection of versioned source spans jointly sufficient to entail an atomic answer claim under recorded inference assumptions. Every member must be necessary within that selected set; another valid set may provide an alternative route. FEVER accepts at least one complete annotated set, while the Evidence Ranking literature explicitly accommodates multiple sufficient sets. A system should pass when its citations contain any complete admissible route, rather than reproduce a preferred reference list. (fever.ai) (aclanthology.org)

Published benchmarks supply useful starting material, but their metrics have different units. Tang and Yang's **MultiHop-RAG** repository describes **2,556 queries**, with evidence for **non-null queries** distributed across **two to four documents**, and specifies an **ODC-BY license**. (github.com) Null queries have no matching evidence in the knowledge base. (arxiv.org) Brehme, Ströhle and Breu's **CARE** evaluation supplies the full context list when judging an individual passage; its synthetic multi-hop experiments select instances with exactly two relevant contexts. Those context-relevance results are not measurements of complete cited evidence chains in production. (arxiv.org)

The proposed acceptance card requires the claim, alternative sufficient sets, exact source versions, observation dates, reader permissions, inference rule, contradiction decision, and verdict. **Complete-set recall** counts answerable cases with a fully retrieved valid set. **Citation-set completeness** counts independently supported output claims whose cited spans contain a complete valid set. These are proposed operational definitions, distinct from reference-claim recall and generated-claim faithfulness documented by Ragas. (docs.ragas.io) A hypothetical ledger below yields **50% complete-set recall**, **33.3% conditional citation-set completeness**, and **75% answer correctness**. These are worked calculations, not empirical findings. Accept

a consequential claim only after its admissible cited evidence jointly supports it and its indispensable dependencies are visible. Source truth still needs separate adjudication: attribution to a source does not establish that source's accuracy. (aclanthology.org)

Introduction and Background

The central failure is easy to overlook: an answer cites the document containing the final value but omits the document establishing why that value answers this question. A reviewer can open the citation, find the named entity, and still be unable to reproduce the join. This report calls that a **missing-hop citation failure**. It can occur even when the answer matches a reference answer exactly.

The relevant problem is evidentiary, rather than a choice of RAG framework. **HotpotQA** supplies supporting-fact supervision for multi-hop question answering, and **MultiHop-RAG** publishes queries, answers, and evidence for reasoning across documents. They establish useful evaluation resources, but a deployment-specific acceptance gate must also identify the source states and dependencies eligible for its particular decision. (hotpotqa.github.io) (github.com) The author-hosted MultiHop-RAG viewer exposes fields for queries, evidence lists, question types, and answers; these are inputs for constructing such a gate, rather than a substitute for its policy. (huggingface.co)

Citation evaluation already separates distinct questions. The Stanford verifiability study measures the fraction of verification-worthy statements supported by their citations separately from the fraction of citations supporting associated statements. (cs.stanford.edu) The multi-hop COLING study likewise separates gold citation-ID coverage, essential key-fact coverage, and sentence-level support against concatenated citations. (aclanthology.org) This report extends that separation into a reviewable acceptance protocol.

The scope is **multi-document answer evaluation**, including answers that combine private and public sources. It is not an implementation tutorial or a claim that a particular system passes the proposed test. The literature and documentation were checked for publication on **October 7, 2026**. Definitions, hypothetical examples, formulas, and recommended decision rules introduced here are this report's proposals. Published benchmark observations retain their original experimental scope.

The intended decision is concrete: may a response be accepted when it mentions the right answer but leaves an indispensable evidence hop uncited or attributes that hop to the wrong source? Under this protocol, the affected claim remains incomplete until the missing dependency is supplied or an alternative sufficient set is adjudicated.

Key Changes: evaluate complete evidence sets

Change the unit from documents to atomic claims

An **atomic claim** is an assertion whose truth conditions can be assessed separately. “Entity R holds the rights, and the package supports interface Q” contains separable assertions. Splitting them prevents support for one assertion from covering the other by proximity.

An **evidence hop** is an indispensable dependency in the justification of a claim. It is not synonymous with a document, a search result, or a model reasoning sentence. A document can contain several dependencies, while repeated documents can express the same dependency. REVEAL evaluates reasoning steps for relevance, evidence attribution, and logical correctness, illustrating why logical dependence deserves its own assessment. (research.google) Its step-level formulation treats a reasoning chain as correct only when every step is correct. (yonatanbitton.github.io)

Require joint sufficiency and test minimality

Let **E** be a set of admissible source spans. E is sufficient when its contents, together with explicitly permitted inference rules, entail claim **c**. It is inclusion-minimal when removing any member makes that particular set insufficient. Inclusion-minimal does not mean the smallest possible number of spans across all alternative sets.

The phrase “jointly necessary” must therefore be interpreted within a selected sufficient set. A source can be indispensable to one route and absent from another. A provenance chain should expose the dependency without pretending that one source path is uniquely correct. W3C PROV-O represents derivation between entities and supports chains comprising both activities and entities. (www.w3.org)

Minimality is an annotation discipline, not a demand that the response cite the fewest possible documents. ALCE expressly permits supporting redundancy rather than requiring a minimal citation set. (aclanthology.org) An answer may cite additional relevant material and still pass, provided at least one complete admissible set is present.

Admit alternative valid sets

For claim c, retain a family **G(c)** of adjudicated sufficient sets. The retrieval or citation condition is satisfied if any member of that family is contained in the evaluated spans. The family is a record of known routes, not proof that no other route exists.

FEVER represents alternatives as a list of evidence sets and accepts a complete annotated set. (fever.ai) The Evidence Ranking task similarly regards a ranked prefix containing an entire gold evidence set as sufficient. (aclanthology.org) These precedents support an existential acceptance rule: require a complete route, rather than the union of every annotated route.

A newly discovered route needs review before scoring. Preserve the original annotation version, add the adjudicated alternative, and report any changed score. Otherwise benchmark incompleteness can be confused with system failure.

Separate support from truth

Joint entailment asks whether the source contents support the conclusion. **Source truth** asks whether those contents should be believed for the decision. A false source can support an equally false answer. The Attributable to Identified Sources (AIS) framework evaluates attribution to supplied corpus parts and avoids treating that evaluation as an absolute factuality judgment. (aclanthology.org)

Record source authority, applicable date, unresolved disagreement, and the basis for accepting the premise separately. NIST also states that transparency does not itself establish accuracy or other trustworthiness properties. (airc.nist.gov) A beautifully reproduced evidence chain can still depend on an outdated or disputed premise.

Keep metric names attached to their denominators

Ragas faithfulness divides supported generated-response claims by all generated-response claims; its LLM context recall instead evaluates support for reference-answer claims. (docs.ragas.io) Its ID-based recall counts matched reference context identifiers, which is a different unit again. (docs.ragas.io)

These measures can be useful alongside complete-set acceptance. They should retain their documented names and denominators. A passage-level average cannot establish that every answerable case has a complete evidence route, and a correct answer string cannot establish that the delivered citations expose that route.

Implementation Considerations and Process Changes

Freeze the eligible evidence universe

Construct the test before inspecting system outputs. Record the corpus boundary, observation cutoff, question population, answerability policy, and reader class. The **eligible evidence universe** contains only source versions and spans permitted for that evaluation.

W3C PROV-DM defines entities through fixed aspects and treats activities as occurring over time. (www.w3.org) For this protocol, a retained document version is an evidence entity; the fetch, annotation, and evaluation events are dated activities. A later page revision should not silently replace the version used to adjudicate the answer.

Observation time is when evidence was captured. Effective time is when its proposition applies. Publication time is what the source reports. Keep these distinct, and declare when an effective date cannot be established. A source observed today can describe an earlier state without supporting a claim about the present.

The frozen test specification should include:

- **Claim boundaries:** Store the assertion, entities, qualifiers, and applicable time.
- **Source identity:** Preserve the original address and a stable version identifier.
- **Source bytes:** Retain the permitted snapshot and its integrity digest.
- **Span locator:** Record the exact excerpt and its location within that snapshot.
- **Observation date:** Record capture time separately from source-stated dates.
- **Reader permissions:** Identify who may inspect each premise and excerpt.
- **Inference assumptions:** State joins, arithmetic, temporal rules, and identity mappings.
- **Alternative sets:** Preserve each reviewed route without collapsing it into one list.
- **Annotation status:** Record disagreements, unresolved dependencies, and adjudication history.

Web Annotation's TextQuoteSelector includes normalized selected text; its Time State can point to a persistent copy of a resource version. (www.w3.org) Those facilities help identify evidence, but this protocol additionally retains the original excerpt so normalization does not obscure the underlying bytes.

Distinguish a source record from an answer gate

Table 1 compares evidence artifacts and the additional judgment needed for an answer. Reliable Context appears as a source-record option because the article concerns defensible provenance; this is not a RAG framework comparison.

Artifact or option	Observable information	What the answer gate must add
Live URL alone	An address that a reviewer can attempt to open.	Retained version, span, date, admissibility, and dependency coverage.
W3C Web Annotation	Selected text and state information identifying a resource version. (www.w3.org)	Claim-level sufficient sets and the validity of the inference.
Amazon Bedrock citation object	Generated-text spans and metadata about cited sources. (docs.aws.amazon.com)	Whether the cited evidence contains every necessary dependency.
Reliable Context source record	The publisher states that values have cited, verbatim, dated sources and reproducible past record states.	Independent evaluation of joint support and complete evidence sets for a generated answer.
Proposed evidence-set acceptance card	Explicit claims, admissible spans, alternative routes, and verdicts.	Source-truth adjudication and domain-specific decision authority.

A citation object or reproducible record gives the evaluator inspectable material. Neither property alone demonstrates that the inference is valid. The distinction also avoids treating source custody as an answer-accuracy benchmark.

Make reader clearance part of answerability

A source that exists but cannot be shown to the intended reader is not automatically admissible for that reader's acceptance test. Google documents access-controlled search that respects an employee's existing Confluence permissions. (docs.cloud.google.com) This is a concrete example of permission-aware retrieval, not evidence that any citation chain is complete.

Maintain separate labels for **not found**, **blocked**, and **not applicable**. A blocked indispensable premise makes a case unanswerable under the reader-specific evidence policy, even if a privileged evaluator can answer it. Evaluate that state separately from retrieval failure.

For mixed private and public evidence, retain source versions inside the permitted boundary and export only what the reader may inspect. The public URL cannot stand in for the private mapping that completes the claim. Conversely, a private assertion does not make a public reference irrelevant merely because the final value is repeated internally.

Ablations and human adjudication

Run interventions that isolate the dependency

Ablation tests ask whether an answer remains supported after controlled changes to the evidence. ERASER distinguishes prediction tests that remove selected rationale tokens from tests that retain only those tokens. (www.eraserbenchmark.com) Those tests motivate useful interventions, but changes in a model's prediction are not, by themselves, a proof of factual necessity.

For each selected sufficient set, run the following variants:

- **Gold-only condition:** Supply the full selected set without irrelevant distractors.
- **Single-hop removal:** Remove one indispensable member while retaining the others.
- **Alternative-route check:** Determine whether another complete valid set survives removal.
- **Contradiction replacement:** Replace one premise with an explicitly labeled incompatible premise.
- **Citation restriction:** Keep the full context but permit citations to only a proper subset.
- **Identity perturbation:** Replace an entity mapping with a similar but different identifier.
- **Temporal perturbation:** Substitute a version outside the claim's applicable interval.
- **Permission perturbation:** Make one indispensable span unavailable to the target reader.

Classify the output as **supported answer**, **appropriate abstention**, or **unsupported synthesis**. Also retain answer correctness as a separate judgment. A model that remembers the correct name after a hop is removed can be correct and unsupported under the intervention.

Removal establishes necessity only relative to the remaining admissible evidence. If another complete set survives, a supported answer is expected. Even when the response remains unchanged, inspect its revised citations before concluding that the system used the removed premise.

Separate behavioral dependence from evidentiary sufficiency

A model might ignore a supplied source yet generate an answer supported by it. Conversely, its answer might change when irrelevant text is removed. Neither behavioral observation alone decides whether the delivered evidence entails the claim.

CARE judges individual contexts with access to the broader list, acknowledging dependencies that single-context judgments can miss. (arxiv.org) LlamaIndex's FaithfulnessEvaluator judges whether context supports a response, which is an evidentiary assessment rather than a causal intervention on the model. (docs.llamaindex.ai) The proposed protocol records both kinds of result and avoids substituting one for the other.

The citation-restriction condition is especially diagnostic for completeness. If all gold spans are available to generation but the response cites only the last hop, retrieval can pass while citation completeness fails. That is the central acceptance problem, rather than a reason to rebuild the retriever.

Use human adjudication where the inference is disputed

Attributed QA uses human annotations as its evaluation gold standard and proposes a correlated automatic metric for development. (research.google) That arrangement supports using automated judges as triage while preserving reviewed reference judgments.

Reviewers should separately decide:

- **Entailment:** Does the cited set justify the precise claim and qualifiers?
- **Minimality:** Does each selected member contribute an indispensable premise?
- **Admissibility:** Are version, time, permissions, and source policy satisfied?
- **Alternatives:** Is a proposed new route sufficient without the preferred reference set?
- **Contradiction:** Is disagreement material, and has its resolution been recorded?
- **Boundary:** Does the answer introduce an additional unreviewed claim?

AVeriTeC's quality-control annotators evaluate evidence without the original fact-checking article or claim label, and disagreements can trigger another annotation cycle. (proceedings.neurips.cc) The proposed adaptation is to retain independent initial judgments, disagreement reasons, and the final decision. Report the number of cases awaiting adjudication; do not silently discard them from difficult slices.

An acceptance card for multi-document answers

Make the final-answer-only failure reviewable

Hypothetical Example: Project Cedar is mapped to record R17 in span a1. A separate rights schedule, span a2, states that Entity R holds the rights associated with R17. The response says “Entity R holds Cedar's rights” and cites a2 alone. The answer is correct against the frozen reference, but a2 does not identify Cedar. The missing a1 citation prevents reproduction of the indispensable mapping.

This example also illustrates misattribution. Attaching a citation to a correct answer does not repair a dependency if the cited source supports a different record, time, or relationship. Amazon's citation metadata can identify retrieved references and generated-text spans; the acceptance test still has to inspect the relationship between those spans and the actual claim. (docs.aws.amazon.com)

Table 2 extends the hypothetical example into a frozen ledger. Every source label, version, date, entity, and output in this table is invented for the calculation. “Supported” means that the output claim is independently supported by eligible reference evidence, not necessarily by the delivered citations.

Output claim and required hops	Retrieved hops	Cited spans	Exact source versions and observation date	Joint entailment of output	Contradiction	Verdict
A: Entity R holds Cedar's rights. Required set {a1: Cedar to R17; a2: R17 to Entity R}.	a1, a2	a2	mapping-v1/a1; rights-v3/a2. Both observed 2026-10-01.	Reference: yes. Citations: no.	None.	Correct and supported claim; incomplete citations.
B: Package P exposes interface Q. Required set {b1: P includes J; b2: J resolves to version V; b3: V exposes Q}.	b1, b3	b1, b3	manifest-v2/b1; lock-v4/b2; interface-v5/b3. All observed 2026-10-01.	Reference: yes. Citations: no.	None.	Correct output; missing retrieved and cited hop b2.
C: Team L maintains Corpus N. Alternatives {c1, c2} or {c3, c4}, each combining a corpus mapping and maintenance assignment.	c3, c4	c3, c4	registry-v1/c1; assignment-v2/c2; alternate-registry-v3/c3; alternate-assignment-v4/c4. All observed 2026-10-01.	Reference: yes. Citations: yes, through the alternative set.	None.	Accept, subject to source-truth policy.
D: Record U is inactive at the frozen snapshot. Expected answer is active, supported by {d1: U to registration; d2: registration active}.	d1, injected d2x	d1, d2x	identity-v1/d1; status-v2/d2. Observed 2026-10-01. Synthetic incompatible replacement status-v2x/d2x is ineligible.	Reference: no. Delivered text appears supportive but fails admissibility.	Injected d2x conflicts with eligible d2.	Reject unsupported, incorrect synthesis.

The decisive contrast is between A and C. A uses the expected final-value source but omits its mapping. C uses an alternative complete route and passes. D demonstrates why syntactic source coverage is insufficient: the incompatible replacement is not part of the eligible frozen evidence.

Specify the card before setting a release threshold

The **acceptance card** is this report's proposed per-claim record:

- **Claim:** Exact answer assertion, scope, entities, and temporal qualifiers.
- **Answerability:** Whether the reader has an eligible sufficient set available.
- **Reference routes:** Every adjudicated alternative set and its annotation version.
- **Retrieved route:** At least one complete admissible set, or the missing dependency.
- **Delivered citations:** Exact cited spans and the claim segment they support.
- **Joint support:** Entailment result with the permitted inference stated explicitly.
- **Source truth:** Authority and unresolved factual disagreement recorded separately.
- **Contradiction:** Relevant conflicting evidence and its reviewed treatment.
- **Interventions:** Gold-only, missing-hop, and contradictory-premise outcomes.
- **Verdict:** Accept, abstain, revise citations, or escalate for adjudication.

NIST calls for documenting test sets, metrics, and evaluation tools, which supports keeping this card alongside the aggregate result. (airc.nist.gov) LangSmith's official documentation also describes rubric-based human review of grouped runs and corrected reference examples. (docs.langchain.com)

For claims designated consequential by the organization's decision policy, the proposed gate requires full evidence-set coverage; a high average score does not waive an individual missing dependency. Other workloads may use different release policies, but they should disclose them before evaluation. There is no universal threshold established here.

Data Analysis and Evidence

Published quantities describe different evaluation objects

Tang and Yang's benchmark knowledge base contains **609 news articles**; its query distribution includes **301 null queries**, alongside inference, comparison, and temporal questions. (arxiv.org) The repository's licensing statement is **ODC-BY**, while HotpotQA states a different license, **CC BY-SA 4.0**. Dataset reuse should follow each artifact's own terms rather than treating benchmark availability as a uniform license. (github.com) (hotpotqa.github.io)

In CARE's synthetic setup, each multi-hop instance has exactly two relevant contexts, and the authors sample **600 relevant and 600 non-relevant contexts per dataset**. (arxiv.org) For **GPT-4.1 on HotPotQA**, the reported context-relevance F1 is **0.814 for CARE**, **0.658 for the direct approach**, and **0.474 for the indirect approach**. These are relevance-labeling results, not complete-set retrieval or citation acceptance rates; the authors identify synthetic, text-only data as a limitation. (arxiv.org)

Alternative routes are quantitatively material in the Evidence Ranking study's curated samples: the average annotated set count is **2.1 for FEVER**, **1.0 for HoVer**, and **4.9 for WICE**, on samples of **400**, **400**, and **200 instances**, respectively. Those sample means are not estimates for arbitrary enterprise corpora. (aclanthology.org) REVEAL, by comparison, reports **817 unique questions**, **1,226 reasoning chains**, and **4,207 reasoning steps**; its annotation unit is a reasoning step. (yonatanbitton.github.io)

These quantities should not be pooled into a single quality score. They describe queries, contexts, alternative sets, and reasoning steps under different sampling designs.

Calculate set metrics with visible denominators

Use an **answerable case** consisting of a question and one required atomic answer claim. For multi-claim questions, report both claim-case and whole-answer results.

Complete-set recall = answerable cases with at least one fully retrieved valid evidence set / all answerable cases.

Citation-set completeness = independently supported output claims whose citations contain at least one complete valid evidence set / all independently supported output claims.

The second denominator is conditional. Report unsupported outputs and omitted required claims separately so that a system cannot improve its displayed score merely by producing fewer assessable assertions. An empty denominator is undefined, not perfect performance.

Table 3 calculates these metrics from the hypothetical ledger, using the full eligible reference universe to decide whether an output is independently supported.

Metric	Numerator	Denominator	Worked result
Complete-set recall	A and C have complete retrieved valid sets: 2.	Answerable cases A, B, C, D: 4.	**2/4 = 50%**
Answer correctness	Outputs A, B, C match their reference claims: 3.	Evaluated cases: 4.	**3/4 = 75%**
Citation-set completeness	C has a complete cited set: 1.	Independently supported outputs A, B, C: 3.	**1/3 = 33.3%**, rounded to one decimal place.
Unconditional complete cited support	C passes: 1.	All emitted output claims: 4.	**1/4 = 25%**
Unsupported output rate	D lacks eligible reference support: 1.	All emitted output claims: 4.	**1/4 = 25%**
Required-claim omission rate	No case omits its required answer assertion: 0.	Required answer claims: 4.	**0/4 = 0%**

Correctness exceeds complete retrieval and citation coverage in this example because A and B name the correct answer without delivering complete citations. D stays outside the conditional supported-claim denominator but remains visible in the unconditional measures. The table is illustrative arithmetic, not a benchmark measurement.

Keep aggregate comparisons reproducible

Retain raw numerator counts, denominator counts, missing annotations, and unresolved disagreements. Stratify results by hop depth, alternative-set availability, contradictory evidence, permissions, and source age. Do not combine blocked-reader cases with answerable retrieval cases.

The paper-defined MultiHop-RAG Hit@K is a fraction of evidence appearing in the retrieved set, which does not establish an all-or-nothing complete-set rate. (arxiv.org) Ragas's non-LLM context recall similarly averages context coverage against reference contexts. (docs.ragas.io) A high partial-coverage score remains compatible with every case missing an indispensable dependency.

Report sampling uncertainty for empirical estimates, preserving the question as a clustering unit when it yields several claims. This hypothetical ledger is too small and deliberately constructed to support a population inference. Its role is to make the acceptance logic reproducible.

Implications and Future Directions

The proposed gate changes the evaluation record more than the generation architecture. An organization can use a vector database or a RAG framework for retrieval and orchestration, then apply evidence-set adjudication to the delivered answers. LangSmith's tutorial separates correctness, groundedness, answer relevance, and retrieval relevance, showing that these boundaries can already be evaluated independently. (docs.langchain.com) The remaining proposal is to make dependency coverage an explicit acceptance object.

For source custody, Reliable Context's first-party description states that the document is retained as read, with its fetch time, and that sources, reader clearance, and retention follow customer policies. These are observable record guarantees, not evidence that a generated multi-hop answer passes this report's test. The comparison therefore concerns what an evaluator can inspect, without making a performance claim about the publisher.

Evidence-set annotations are necessarily revisable. An annotator may miss an alternative source, an entity mapping may remain ambiguous, or a source may later cease to support a time-qualified proposition. PROV-DM includes invalidation, under which an entity is no longer available for subsequent use. (www.w3.org) Evaluation records should preserve the earlier verdict and its applicable evidence state instead of overwriting history.

Future studies should compare alternative-aware annotation with single-route annotation, test reader-specific answerability, and measure disagreement about joint entailment. They should also determine how performance changes across private and public evidence boundaries. CARE's text-only synthetic scope does not resolve those transfer questions. (arxiv.org) A study needs to publish the source eligibility policy and annotation changes before presenting a claimed improvement.

Broader answer-quality evaluations remain complementary. Microsoft's GraphRAG research post reports evaluation by an LLM judge on qualities including comprehensiveness, diversity, and empowerment. (www.microsoft.com) Such judgments can assess whether an answer is useful, while a set-level gate asks whether its indispensable evidence is inspectable and sufficient. Neither result should be relabeled as the other.

A defensible research contribution would release permitted claim annotations, version identifiers, alternative-set decisions, and raw metric counts. Private source text need not be published to make the protocol and its limits explicit.

Conclusion

A correct final answer is insufficient for multi-hop citation acceptance when an indispensable source dependency is missing or misattributed. The evaluator should identify atomic claims, preserve eligible source versions, and determine whether the delivered citations contain at least one complete sufficient evidence set.

The **evidence-set test** proposed here distinguishes answer correctness, complete retrieval, joint entailment, citation coverage, and source truth. It accepts alternative valid routes and permits additional relevant citations. Existing attribution work already allows joint support, while FEVER provides an explicit complete-set

acceptance precedent. (aclanthology.org) (fever.ai) The contribution is a frozen, reviewable decision protocol that combines those ideas with version and permission checks.

The hypothetical ledger demonstrates why the distinctions matter: a correct answer can coexist with an incomplete citation route, and an apparent supporting route can fail because its premise is ineligible. Its percentages are worked examples, not claims about a deployed system or the general performance of RAG.

Context engineers should make the per-claim acceptance card the primary review object and use aggregate metrics to describe the evaluated population. Keep denominators, annotation disagreements, unsupported outputs, and missing required claims visible. When a hop is unavailable, abstain or qualify the affected assertion under the stated policy; when a new valid route is found, adjudicate it and preserve the changed annotation.

The practical decision is therefore direct: accept the claim only when its admissible cited spans jointly justify it, its indispensable dependencies are exposed, and any material source conflict has been addressed. A plausible name and an openable citation are starting evidence, not a completed justification.

Source links

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

1. https://aclanthology.org/2023.emnlp-main.398.pdf
2. https://cs.stanford.edu/%7Enfliu/papers/liu%2Bzhang%2Bliang.emnlpfindings2023.pdf
3. https://fever.ai/2018/task.html
4. https://aclanthology.org/2026.eacl-long.340.pdf
5. https://github.com/yixuantt/MultiHop-RAG
6. https://arxiv.org/html/2401.15391v1
7. https://arxiv.org/html/2604.18234v1
8. https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/context_recall
9. https://aclanthology.org/2023.cl-4.2.pdf
10. https://hotpotqa.github.io
11. https://huggingface.co/datasets/yixuantt/MultiHopRAG
12. https://aclanthology.org/2025.coling-main.285.pdf
13. https://research.google/pubs/a-chain-of-thought-is-as-strong-as-its-weakest-link-a-benchmark-for-verifiers-of-reasoning-chains
14. https://yonatanbitton.github.io/publication/chain_of_thought/2402.00559v3.pdf
15. https://www.w3.org/TR/prov-o
16. https://airc.nist.gov/airmf-resources/airmf/3-sec-characteristics
17. https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/faithfulness
18. https://www.w3.org/TR/prov-dm
19. https://www.w3.org/TR/annotation-model
20. https://docs.aws.amazon.com/bedrock/latest/APIReference/API_agent-runtime_TextResponsePart.html

21. https://docs.cloud.google.com/generative-ai-app-builder/docs/data-source-access-control
22. https://www.eraserbenchmark.com
23. https://docs.llamaindex.ai/en/stable/api_reference/evaluation/faithfullness
24. https://research.google/pubs/attributed-question-answering-evaluation-and-modeling-for-attributed-large-language-models
25. https://proceedings.neurips.cc/paper_files/paper/2023/file/cd86a30526cd1aff61d6f89f107634e4-Paper-Datasets_and_Benchmarks.pdf
26. https://docs.aws.amazon.com/bedrock/latest/APIReference/API_agent-runtime_Citation.html
27. https://airc.nist.gov/airmf-resources/airmf/5-sec-core
28. https://docs.langchain.com/langsmith/annotation-queues
29. https://docs.langchain.com/langsmith/evaluate-rag-tutorial
30. https://www.microsoft.com/en-us/research/blog/graphrag-new-tool-for-complex-data-discovery-now-on-github

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.