Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

RAG citation provenance: did generated context become evidence?

> 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 · 2026-10-24 · rag citation provenance · source attribution · hyde · contextual retrieval · citation faithfulness · original evidence · context engineering · provenance evaluation

Original article: https://reliablecontext.com/articles/generated-rag-context-citation-provenance



In this article

1. Executive Summary
2. Introduction and Background
3. Methodology: define the evidence boundary
4. Observable artifacts and implementation inspection
5. Frozen fixtures and expected answers
6. Paired exposure tests with fixed originals
7. Resolve citations to retained original spans
8. Analysis of key segments and support decisions
9. Data Analysis and Evidence
10. Implications and Future Directions
11. Conclusion

Executive Summary

RAG citation provenance requires a trace from an answer's attributed claim to retained source material, including the transformations between retrieval and generation. Generated search aids may improve retrieval while remaining ineligible as quotations from the documents they help locate. **HyDE**, introduced by Gao et al. (2023), generates hypothetical relevance text and retrieves real documents through its embedding. (aclanthology.org) **Contextual retrieval** places generated explanatory text beside document chunks. (www.anthropic.com) These interventions occur on different sides of retrieval; neither description establishes what a particular application's answer generator receives.

The decision is therefore about **source origin**, rather than choosing a universally safer retrieval method. Haystack's implementation documentation describes averaging hypothetical-document embeddings and using retrieved documents downstream; it does not establish that hypothetical text must enter the answer prompt. LlamaIndex separately documents controls over metadata exposed to embeddings and to the language model. A deployment must inspect those actual views before testing whether generated material becomes evidence. (docs.haystack.deepset.ai) (developers.llamaindex.ai)

Published retrieval results establish a reason to preserve useful aids, not a citation-origin success rate. Anthropic's provider experiment reports top-20 retrieval failure falling from **5.7% to 2.9%** with contextual embeddings and contextual BM25, a **49% relative reduction**, and to **1.9%** with reranking. (www.anthropic.com) Its article explicitly advises distinguishing explanatory context from the chunk during response generation. (www.anthropic.com) Separately, the ALCE benchmark reports incomplete citation support in **50%** of its best models' ELI5 answers. (aclanthology.org) These are different endpoints and must not be combined into an estimate of generated-context attribution errors.

This report proposes a frozen fixture, an artifact ledger, and a paired exposure test; **no pipeline execution or new benchmark result is reported**. First run the intended pipeline unmodified. Then hold retrieved originals fixed while omitting, labeling, exposing, or deliberately mislabeling generated additions. Resolve each citation against the original snapshot, and assess unsupported claims even when no citation appears. World Wide Web Consortium (W3C) provenance and annotation models supply vocabulary for derivations and text

selectors; they do not make a generated paraphrase a verbatim quotation. (www.w3.org) The proposed release decision rejects falsely attributed quotations, separately evaluates warranted derivations, and requires explicit unknown outcomes when retained originals do not answer the question. Answerability must itself be evaluated, as the SQuAD no-answer task illustrates. (rajpurkar.github.io)

Introduction and Background

Retrieval-augmented generation (RAG) combines retrieval with language-model generation. (docs.haystack.deepset.ai) Its citation problem becomes more specific when the retrieved representation contains text generated by another model. An evaluator may find that the answer agrees with its supplied context yet still need to ask whether that context was authored by the cited source. **Ragas** defines faithfulness against retrieved context. The methodological implication is that the evaluator must identify which parts of that context are original evidence before interpreting a high score as source support. (docs.ragas.io)

The literature supplies complementary distinctions. Gao et al.'s ALCE work evaluates fluency, correctness, and citation quality separately. Rashkin et al.'s Attributable to Identified Sources (AIS) asks whether statements are supported by identified underlying sources. Wallat et al. distinguish citation correctness from citation faithfulness, including whether the cited reference was actually used. The test proposed here adds an origin boundary to these existing questions: does the purported supporting passage belong to retained original material or to a generated retrieval aid? (aclanthology.org) (research.google) (arxiv.org)

The intended reader is an enterprise context engineer deciding what may be indexed, passed to generation, and cited. This is a **problem and evaluation report**, not a comparison of vendors or a claim about any publisher's implementation. Its acceptance unit is the original-versus-generated text boundary. Source poisoning, truth assessment, and causal attribution remain related but separate tasks.

The research snapshot is **October 24, 2026**. Documentation findings describe inspected pages, not installed packages or executed applications. A reproducible implementation must pin its own code, model configuration, prompt, and document versions. The original HyDE paper establishes the retrieval method, while reproducibility for an application also requires the exact implementation revision. (aclanthology.org) The report's fixtures, inserted facts, result fields, and gates below are explicitly proposed. They must not be read as observed defects in HyDE, contextual retrieval, or any named framework.

Methodology: define the evidence boundary

Source origin means the authorship and transformation status of the particular text offered as support. It is more granular than a document title or a URL. The W3C PROV Data Model describes provenance through entities, activities, and people; it distinguishes general derivation from the more specific relation of quotation. That distinction supports the proposed operational vocabulary. (www.w3.org)

- **Original evidence:** retained source bytes and an identified source representation, before generated additions.
- **Quoted excerpt:** text copied from that representation, with a locator and declared extraction policy.

- **Generated retrieval aid:** hypothetical text, questions, or explanatory context produced to help locate or rank evidence.
- **Derived claim:** a calculation, synthesis, or inference with identified inputs and an explicit warrant.
- **Source snapshot:** the particular version retained for replay, including its observation date and representation.
- **Evidence boundary:** the recorded transition between original text and generated or otherwise transformed text.

These are proposed evaluation categories. A generated prefix may accurately summarize the original document, but its exact wording remains generated. The distinction between an original and a derived artifact is consistent with DataCite's separate derivation relationship. (support.datacite.org) Conversely, a generated question can lead to a genuinely supported answer. Neither origin status nor retrieval usefulness alone decides whether the final claim is supportable. Quotation deserves the narrower copying test because PROV explicitly models it as a particular derivation. (www.w3.org)

HyDE is query-side. Gao et al. (2023) describe generating a hypothetical document, encoding it, and retrieving similar real documents. (aclanthology.org) The method is not a requirement to cite the hypothetical document. Haystack's guide describes generating **five** hypothetical documents and averaging their embeddings; the inspected page carries a **Version 3.4** documentation label. (docs.haystack.deepset.ai) Neither observation establishes an installed patch version.

Contextual retrieval is document-side. Anthropic describes generated chunk-specific context, usually **50 to 100 tokens**, prepended before indexing. (www.anthropic.com) Those statements motivate a boundary test rather than an assumption that prefixes are intrinsically unsafe.

Generated questions are another artifact class. LlamaIndex documents a QuestionsAnsweredExtractor that produces questions each node can answer. The same guide chains text splitting and metadata extraction as explicit ingestion transformations. (developers.llamaindex.ai) Such questions describe an access path to a node; their existence does not establish that any answer wording is present in the source. (developers.llamaindex.ai) The proposed policy therefore permits retrieval use while requiring a separate source-span or derivation warrant for evidence-bearing output.

Observable artifacts and implementation inspection

A method name cannot reconstruct a generation request. Haystack's inspected HyDE cookbook connects a hypothetical embedding to a retriever and shows retrieved documents without an answer-generation stage. Its embedder guidance separately distinguishes retrieval embeddings from documents subsequently provided to a language model. A test must inspect the application that connects these components. (haystack.deepset.ai) (docs.haystack.deepset.ai)

Table 1 specifies a proposed source-origin inventory. Identifiers are illustrative labels, not records from an executed pipeline.

Artifact ID	Origin and generating activity	Snapshot, boundaries, and locator	Proposed allowed use
SRC	Original source, acquired and retained	Source version, bytes, observation time, representation	Evidence candidate, subject to support review
EXT	Extracted source text, produced by a parser	Link to SRC; extraction version; coordinate map	Quotation only under the declared extraction policy
HYP	Hypothetical query-side document, generated from a query	Separate text boundary; model and prompt revision	Retrieval feature; not a quotation from SRC
PFX	Generated chunk prefix, derived using source context	Separate prefix and original-chunk ranges	Retrieval context; generator exposure controlled
QST	Generated hypothetical questions about a node	Separate question text; originating node identifier	Retrieval feature; answers require source support
RET	Retrieved view assembled for a request	Ordered artifact IDs, scores, original ranges, added ranges	Input whose components retain their origin labels
CLM	Answer claim or derived result	Supporting originals; output range; calculation or inference warrant	Attributed claim only after resolution and support checks

The inventory separates indexed representations from cited evidence, as does Haystack's embedder guidance. (docs.haystack.deepset.ai) DataCite likewise distinguishes version relationships from derivation relationships; a new version and an artifact derived from a source express different relationships. The proposed ledger should preserve both where applicable rather than replacing either with a single mutable document ID. (support.datacite.org)

Inspect the actual generator input, not merely the fields shown in a search response. LlamaIndex states that metadata enters embeddings and language-model input by default and documents exclusions that can separate those views. Weaviate's generative-search documentation also notes that properties used in a prompt need not be among the properties returned by a query. Returned fields therefore cannot establish the complete exposure boundary. (developers.llamaindex.ai) (docs.weaviate.io)

Inspecting those views is especially important when metadata is available for embedding but excluded from the language model, a configuration LlamaIndex explicitly supports. (developers.llamaindex.ai) Collect:

- **Runtime identity:** package lock, repository revision, model identifier, and configuration snapshot.
- **Retrieval identity:** query text, generated aids, embedding inputs, retrieved original IDs, and ordering.
- **Generator view:** the fully assembled request with original and generated ranges marked.
- **Citation view:** source IDs, output offsets, evidence ranges, and the view each locator addresses.
- **Transformation record:** parser, normalization, chunking, metadata, and citation-splitting revisions.

Framework interfaces can support separation without proving it for an application. LangChain's MultiVectorRetriever documents returning parent documents from its document store after matching child representations. LlamaIndex's CitationQueryEngine constructs finer citation nodes using a configurable metadata mode. The inspected citation API defaults to MetadataMode.NONE, but the configured mode and node content still require inspection. (developers.llamaindex.ai) Verify these behaviors and node text against the pinned implementation. (reference.langchain.com) (developers.llamaindex.ai)

Frozen fixtures and expected answers

The proposed fixture freezes a query, source bytes, extraction output, exact evidence spans, generated additions, permitted derivations, and an expected answer. It also records which question is unanswerable from the retained originals. Microsoft recommends testing positive and negative search examples, including queries that indexed data cannot answer; this provides a documented rationale for the no-answer stratum. (learn.microsoft.com)

Microsoft also recommends returning source metadata and raw uncleaned content fields, useful inputs for retaining the fixture's original view. (learn.microsoft.com)

Hypothetical Example: a fictional specification says, “The inspection interval is not specified.” A generated prefix is deliberately altered to say, “The inspection interval is 73 days, effective 2031-04-17.” The number and date are **invented stress-test inputs**. They are not facts about a real document, an observed model response, or a finding about a named tool. The expected response is that the interval and effective date are unknown from the retained specification.

The no-answer contrast follows the paragraph-support boundary used in SQuAD's abstention task. (rajpurkar.github.io) **Hypothetical Example:** use a second fixture whose original does contain the supported answer, and another permitting an explicit calculation. For example, an answer may legitimately compute a difference from retained operands even when the resulting sentence never appears verbatim. These fixtures keep an exact-quote test from becoming an inappropriate demand that every valid derivation already exist as a source sentence. AIS supplies the broader support question; the proposed resolver adds the copying and origin checks for attributed excerpts. (research.google)

Hypothetical Example, executable fixture and resolver: the following check freezes the earlier fictional source as UTF-8 bytes, an explicit version, a digest, and an end-exclusive byte range. No derivations are permitted in this no-answer fixture. The aid is controlled test text. The query asks for the interval and effective date; both expected values are unknown. This is a minimal exact-quotation check, not a RAG benchmark or a semantic-support judge.

```
from hashlib import sha256

source = b"The inspection interval is not specified."
aid = b"The inspection interval is 73 days, effective 2031-04-17."
snapshot = "dd4de6ad830cf900b5229227b50f1582c79f4144b343df67ca70275085ede5eb"
fixture = {
    "version": "fixture-v1",
    "query": "What are the interval and effective date?",
    "expected": {"interval": "unknown", "effective_date": "unknown"},
    "permitted_derivations": [],
}
ledger = {
    "src-v1": {"kind": "original", "bytes": source},
    "aid-v1": {"kind": "generated", "bytes": aid, "parent": "src-v1"},
}
ref = {"id": "src-v1", "digest": snapshot, "range": [0, 41]}
```

```
def resolve(citation, quoted_text):
    item = ledger.get(citation["id"])
    if item is None:
        return "unresolved_id"
    if item["kind"] != "original":
        return "generated_target"
    if sha256(item["bytes"]).hexdigest() != citation["digest"]:
        return "snapshot_mismatch"
    start, end = citation["range"]
    if not 0 <= start < end <= len(item["bytes"]):
        return "unresolved_range"
    exact = item["bytes"][start:end] == quoted_text.encode("utf-8")
    return "quote_match" if exact else "quote_mismatch"

assert resolve(ref, source.decode()) == "quote_match"
assert resolve(ref, aid.decode()) == "quote_mismatch"
assert resolve({"id": "aid-v1"}, aid.decode()) == "generated_target"
assert resolve({**ref, "digest": "wrong"}, source.decode()) == "snapshot_mismatch"
assert resolve({**ref, "range": [0, 99]}, source.decode()) == "unresolved_range"
assert resolve({"id": "missing"}, source.decode()) == "unresolved_id"
```

A production adapter must validate malformed citation objects and implement its own source-native selection and normalization policy. A match establishes only that the quoted bytes occur at the cited original range. Support, context, and derived-claim warrants still require the separate adjudication described above.

Paired exposure tests with fixed originals

Run the **intended pipeline unmodified first** and retain its complete trace. Haystack's cookbook illustrates why the inspected artifact may be only a retrieval pipeline, rather than a complete answer generator. (haystack.deepset.ai) Only then introduce a generator-side ablation that holds retrieved originals, order, original evidence budget, and generation settings fixed. Reserve input space so added aids cannot truncate the frozen originals, and record their order and positions. An aid absent from the baseline generator input must be recorded as absent; exposing it later is a stress intervention, not a description of normal HyDE behavior. The Haystack workflow makes that caution material because its documented retrieval stage can feed actual retrieved documents downstream. (docs.haystack.deepset.ai)

Table 2 is the proposed paired result matrix. All outcomes remain unmeasured; the rows specify treatments and recording requirements.

Exposure condition, same originals	Unsupported inserted fact	Answer incorporation	Cited span and resolution	Human support verdict
Intended pipeline, unmodified	Only if naturally present in its input; otherwise absent	Record complete answer	Retain emitted locator and resolved original	Judge observed baseline only
Generated aid omitted	Absent	Record change from baseline	Check every emitted citation	Originals sufficient or insufficient

Exposure condition, same originals	Unsupported inserted fact	Answer incorporation	Cited span and resolution	Human support verdict
Aid visible and clearly labeled	Present as labeled test intervention	Record adoption, rejection, or unknown	Identify original versus generated target	Judge against originals
Aid visible without origin labels	Same intervention text	Record adoption and wording	Preserve exact purported support	Judge against originals
Aid deliberately mislabeled as source	Same intervention text	Record attributed adoption	Test claimed quotation and source snapshot	Experimental stress condition only

The matrix prevents a retrieval improvement from being confused with a generation-origin effect. It also preserves a useful null result: an unchanged answer under fixed originals says nothing conclusive about whether an aid would retrieve better originals end to end. RAGChecker's separate retrieval and generation evaluation motivates keeping component diagnostics distinct, although it does not measure the proposed fixture's outcomes. (arxiv.org)

Before accepting a generated-only label, search the retained complete original for the inserted assertion. Relevant support may lie outside the selected chunk. Record that as original-supported material with a newly identified locator, and report the retrieval omission separately. Do not use unretained pages or outside model knowledge to repair an evaluation whose defined evidence set is the frozen snapshot. Source metadata and raw content, as recommended by Microsoft, should remain available during this review. (learn.microsoft.com)

Resolve citations to retained original spans

The resolver accepts a claim, its citation object, and the frozen artifact ledger. It first identifies the cited snapshot and origin class, then obtains the claimed supporting text through a source-native locator. A page URL alone is insufficient for the proposed quotation gate. FORCE11's data citation principles require enough fixity and provenance information to verify the specific version, timeslice, or granular portion originally cited. (force11.org)

W3C's Web Annotation model offers an exact-text selector with optional surrounding prefix and suffix, as well as position selectors. These are useful implementation choices, provided their coordinate system and representation are recorded. The same specification warns that position selection is brittle when resources change and supplies state information for identifying the intended representation. (www.w3.org)

The proposed resolver procedure is:

- **Identify:** map the citation's public or local ID to the retained source version, not merely a current URL.
- **Locate:** resolve a page and text range, markup selector, table coordinate, or other source-native target.
- **Reconstruct:** replay extraction and selection against the declared representation and coordinate system.
- **Compare:** test the claimed quotation against original text under a versioned normalization policy.
- **Inspect context:** retain headings, qualifications, units, and surrounding passages needed to judge meaning.

- **Classify support:** distinguish direct support, warranted derivation, contradiction, insufficiency, and unresolved origin.
- **Record:** retain the verdict, evidence span, transformation chain, and reviewer disagreement.

Source and output coordinates must not be conflated. A TextQuoteSelector addresses quoted source text, whereas a citation API may also identify a span in the generated answer. (www.w3.org) Cohere's documentation describes output text indices and document IDs in citation objects. Amazon Bedrock separately documents a citation's generated-response segment and retrieved references containing cited text and source-location information. Cohere also allows supplied document IDs to become citation references, providing an application-controlled identifier to resolve. (docs.cohere.com) These interfaces provide useful objects to inspect; their existence does not prove the referenced text was retained unchanged from an original source. (docs.cohere.com) (docs.aws.amazon.com)

Chunk splitting needs special attention. The inspected LlamaIndex citation implementation explicitly clears inherited character offsets when producing finer citation nodes because the offsets no longer describe the new text. Supermemory's engineering guide similarly recommends preserving original character spans or source-native locators for production citations. The proposed test verifies the resulting map rather than presuming that any intermediate node offset is an original offset. (developers.llamaindex.ai) (supermemory.ai)

Normalization is a declared transformation. Preserve the raw extracted quotation and its normalized comparison view. Unicode distinguishes canonical from compatibility equivalence and warns against blindly applying compatibility normalization to arbitrary text. The distinction matters because compatibility-equivalent forms can have different appearances or behavior. (www.unicode.org) The proposed policy must therefore name permitted transformations and reject semantic rewriting masquerading as quote matching. Record whether offsets count bytes, code points, or code units, and retain a mapping when extraction changes the coordinate space. (www.unicode.org)

Finally, preserve content identity without confusing it with truth. RFC 9530 distinguishes digests of message content from digests of a representation; its mechanism is not itself authentication. (www.rfc-editor.org) Hash the specific retained artifact and state what was hashed. A matching digest supports the identity check proposed here, not the factual correctness of the source.

Analysis of key segments and support decisions

Different evidence classes require different acceptance paths. A single aggregate citation score would obscure the difference between an unavailable locator, an altered quotation, and a valid calculation. Wallat et al.'s distinction between correctness and faithfulness also cautions against treating compatible supporting text as proof of actual causal reliance. The proposed exposure experiment tests a controlled input change; it does not inspect every internal reason for the model's answer. (arxiv.org)

For **direct quotations**, require resolution to retained original text and agreement under the declared quotation policy. A correct paraphrase must be labeled as a paraphrase, not formatted as words authored by the source. For **derived claims**, retain the operands or premises, units, formula or inference rule, and the connection from those inputs to the result. PROV's broader derivation concept accommodates transformations that are not quotations. (www.w3.org)

For **generated explanatory context**, assess two questions separately: is the explanation supported somewhere in the original, and did the output attribute the generated wording to the source? Supported content can still be falsely quoted. Conversely, a generated question may contain no answer claim at all. LlamaIndex's metadata extraction documentation establishes generated questions as an explicit transformation, making origin tagging useful even when the added artifact is interrogative. (developers.llamaindex.ai)

For **insufficient originals**, require an explicit unknown or abstention outcome. The SQuAD no-answer task requires recognizing when the paragraph does not support an answer and abstaining. Here the proposed assessment additionally checks whether an insufficient-evidence response identifies the missing basis, rather than silently dropping a citation while retaining an unsupported assertion. (rajpurkar.github.io)

The adjudication record should preserve:

- **Answerability:** supported, insufficient, ambiguous, or outside the frozen corpus.
- **Origin verdict:** original, generated, mixed, or unresolved at the claimed supporting boundary.
- **Quotation verdict:** exact, accepted normalization, mismatch, or not a quotation.
- **Derivation verdict:** warranted, incomplete warrant, invalid transformation, or not applicable.
- **Coverage verdict:** supported or unsupported for every assessed claim, including uncited statements.
- **Review state:** initial judgments, disagreements, adjudication, and the policy revision used.

Neighbor expansion is a separate intervention. Pinecone documents retrieving neighboring chunks after initial retrieval, potentially broadening available context. Additional retained original text can supply legitimate support; it should not be classified as generated scaffolding merely because the initial matched chunk was smaller. (www.pinecone.io) A useful report therefore stratifies results by origin transformation and answerability, rather than treating all expanded context as one category.

Data Analysis and Evidence

Published measurements provide context, not results for the source-origin test. In Gao et al.'s HyDE paper, Table 1 reports **DL19 nDCG@10 of 61.3** for HyDE versus **44.5** for Contriever, and **Recall@1k of 88.0 versus 74.6**, on the table's displayed scale. Its DL20 values are **57.9 versus 42.1** and **84.4 versus 75.4**, respectively. (aclanthology.org) These are retrieval endpoints; they do not measure quotation resolution or generated-only claims.

Anthropic's provider experiment defines retrieval failure as one minus recall@20 and reports averages across knowledge domains with its selected embedding configuration. Contextual embeddings plus contextual BM25 change failure from **5.7% to 2.9%**; adding reranking changes it to **1.9%**. The first change is **2.8 percentage points**, calculated by subtraction, or approximately **49% relative**, using the baseline as denominator. (www.anthropic.com) Retrieval failure and false source attribution have different denominators, so these results supply no provenance-error estimate.

Evaluation work provides other scopes. ALCE's **50%** incomplete-support result concerns ELI5 citation support. (aclanthology.org) RAGChecker evaluates **eight systems**, formed from **two retrievers and four generators**, on **4,162 queries across ten domains**. (arxiv.org) These studies motivate component and support evaluation

without validating this report's generated-aid acceptance gates.

Table 3 defines the proposed metrics. No result values are supplied because the frozen fixture has not been run.

Metric	Numerator and denominator, or measurement	Required interpretation
Generated-only citation rate	Assessed citations whose claimed supporting text exists only in generated aids / assessed citations	Origin error among citations; adjudicate support elsewhere in originals first
Generated-only unsupported claim rate	Assessed output claims matching generated aids and unsupported by retained originals / all assessed output claims	Includes uncited claims; omitting citations cannot improve this denominator
Exact-quote mismatch rate	Assessed quotations that disagree with resolved original text / assessed quotations	Report unresolved quotations separately as release failures
Original-span resolution rate	Assessed citations resolving to retained original spans / assessed citations	Resolution alone does not establish semantic support
Source-supported correctness	Claims judged correct and supported by originals / assessed output claims	Publish the correctness reference and permitted derivations
Abstention correctness	Correct answer-or-abstain decisions / assessed queries	Also separate answerable and no-answer strata
Retrieval recall at budget k	Relevant original evidence retrieved / reference relevant original evidence	Keep corpus, evidence unit, and budget fixed
Latency and actual cost	End-to-end elapsed time and incurred usage cost per query	Report distribution, retries, caching, and accounting boundaries

The citation and claim denominators must both be published, with raw counts and exclusions. A citation-free answer yields no assessed citations, so its citation rate is **undefined**, not zero; its assessed claims still enter claim-level support review. Ragas documents that context recall requires a reference, underscoring the need to define reference evidence before interpreting a retrieval score. (docs.ragas.io)

Estimate paired differences using the **query as the proposed resampling unit**, preserving each query's treatment outputs together. State the interval procedure, repetitions, missing-data policy, and any repeated model draws. This design avoids treating correlated claims as independent queries. Dror et al. explain that task and evaluation setup affect significance-test choice; Soboroff examines confidence intervals for retrieval means. Neither source establishes interval coverage for this particular proposed paired provenance metric. (aclanthology.org) (www.nist.gov)

Run a separate end-to-end retrieval comparison with and without the aid, fixing corpus, retrieval budgets, generator budget, and accounting rules. Report relevance gains, origin failures, latency, and actual cost independently. This preserves the possible retrieval benefit that the fixed-original experiment intentionally holds constant.

Implications and Future Directions

The proposed release policy keeps useful search aids available while imposing an evidence boundary on attributed output. It does not require that every generated prefix be discarded, nor does it designate HyDE as an unsafe method. Gao et al.'s distinction between hypothetical relevance text and real retrieved documents remains the appropriate starting point. (aclanthology.org)

Proposed release gates should require:

- **Resolvable originals:** every evidence-bearing citation identifies a retained original representation and span; unresolved support blocks evidence-bearing release.
- **Quotation fidelity:** no released quotation attributes generated wording to an original source.
- **Warranted derivation:** calculations and inferences identify their original inputs and the transformation that supports the result.
- **Complete claim accounting:** unsupported uncited statements remain in the evaluation denominator.
- **Correct unknown outcomes:** insufficient originals produce an explicit statement of missing evidence or abstention.
- **Separate retrieval reporting:** relevance, cost, and latency gains never substitute for origin checks.
- **Replayable decisions:** reviewers can reconstruct the source state, exposure condition, locator, and adjudication policy.

These are proposed acceptance requirements, not universal performance thresholds established by the cited papers. A team should specify its release threshold and uncertainty policy before running its test. Strict rejection of falsely attributed quotations can coexist with a measured, separately reported retrieval improvement.

Replay requires both a stable identity and a retained representation. W3C annotation state can link to a persistent copy of the resource version; FORCE11 recommends persistence of identifiers and descriptive metadata. An application should therefore distinguish deleted evidence, blocked access, and unsupported content rather than collapsing all three into an empty field. (www.w3.org) (force11.org)

Future comparisons should publish the artifact-boundary policy, query strata, generated-addition exposure, source snapshots where permitted, and adjudication rules. Portability should be tested across extraction and citation transformations, not inferred from a framework name. Citation-origin checks remain limited: they do not establish the truth of an original, exclude poisoned original content, or prove complete causal dependence on a cited passage. A recorded derivation describes how an entity was produced, using PROV's transformation vocabulary; this report does not infer truth from that record. (www.w3.org) The narrower deliverable is a reproducible answer to whether claimed support resolves to the original evidence that the application retained.

Conclusion

Generated RAG context may help locate evidence without becoming evidence itself. The relevant acceptance question is whether the final claim, quotation, or explicit derivation resolves to retained original material under a declared source-origin policy. A useful retrieval feature and a defensible attribution are separate properties, consistent with HyDE's distinction between hypothetical text and retrieved real documents. (aclanthology.org)

The proposed evaluation starts with the intended pipeline and inspects its real generator input. It then freezes retrieved originals for paired exposure tests, including clearly labeled experimental corruption and no-answer fixtures. A separate end-to-end comparison measures retrieval value. Keeping those experiments distinct prevents a change in retrieved evidence from being mistaken for an effect of generated text on attribution.

The resolver must identify a source version, reconstruct its text representation, validate quoted words, and retain the inputs and warrant for derived claims. Output citation markers, populated metadata, and compatible context do not replace those operations. An evaluator that checks support against retrieved context must explicitly define that context's origin boundary. (docs.ragas.io) Claim-level assessment must also include unsupported statements with no citation, so suppressing citation markers cannot create an apparent improvement.

No provenance-error rate or framework ranking follows from this report because no pipeline execution was performed. The reviewable contribution is the proposed fixture, inventory, exposure matrix, metric definitions, and release policy. They preserve the possibility of retrieval gains while requiring a concrete, replayable boundary between generated search scaffolding and the original evidence an answer asks its reader to trust.

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.acl-long.99
2. https://www.anthropic.com/research/contextual-retrieval
3. https://docs.haystack.deepset.ai/docs/hypothetical-document-embeddings-hyde
4. https://developers.llamaindex.ai/python/framework/module_guides/loading/documents_and_nodes/usage_documents
5. https://aclanthology.org/2023.emnlp-main.398
6. https://www.w3.org/TR/prov-dm
7. https://rajpurkar.github.io/SQuAD-explorer
8. https://docs.haystack.deepset.ai/docs/choosing-the-right-embedder
9. https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/faithfulness
10. https://research.google/pubs/measuring-attribution-in-natural-language-generation-models
11. https://arxiv.org/abs/2412.18004
12. https://support.datacite.org/docs/connecting-to-works
13. https://developers.llamaindex.ai/python/framework/module_guides/loading/documents_and_nodes/usage_metadata_extractor
14. https://haystack.deepset.ai/cookbook/using_hyde_for_improved_retrieval
15. https://docs.weaviate.io/weaviate/search/generative
16. https://developers.llamaindex.ai/python/framework-api-reference/query_engine/citation
17. https://reference.langchain.com/python/langchain-classic/retrievers/multi_vector/MultiVectorRetriever
18. https://learn.microsoft.com/en-us/azure/architecture/ai-ml/guide/rag/rag-information-retrieval
19. https://arxiv.org/html/2408.08067v2

20. https://force11.org/group/joint-declaration-of-data-citation-principles-final
21. https://www.w3.org/TR/annotation-model
22. https://docs.cohere.com/docs/rag-citations
23. https://docs.aws.amazon.com/bedrock/latest/APIReference/API_agent-runtime_Citation.html
24. https://supermemory.ai/blog/text-chunking-strategies-for-rag
25. https://www.unicode.org/reports/tr15
26. https://www.rfc-editor.org/rfc/rfc9530.html
27. https://www.pinecone.io/learn/chunking-strategies
28. https://aclanthology.org/2023.acl-long.99.pdf
29. https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/context_recall
30. https://aclanthology.org/P18-1128
31. https://www.nist.gov/publications/computing-confidence-intervals-common-ir-measures

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.