Reliable Context

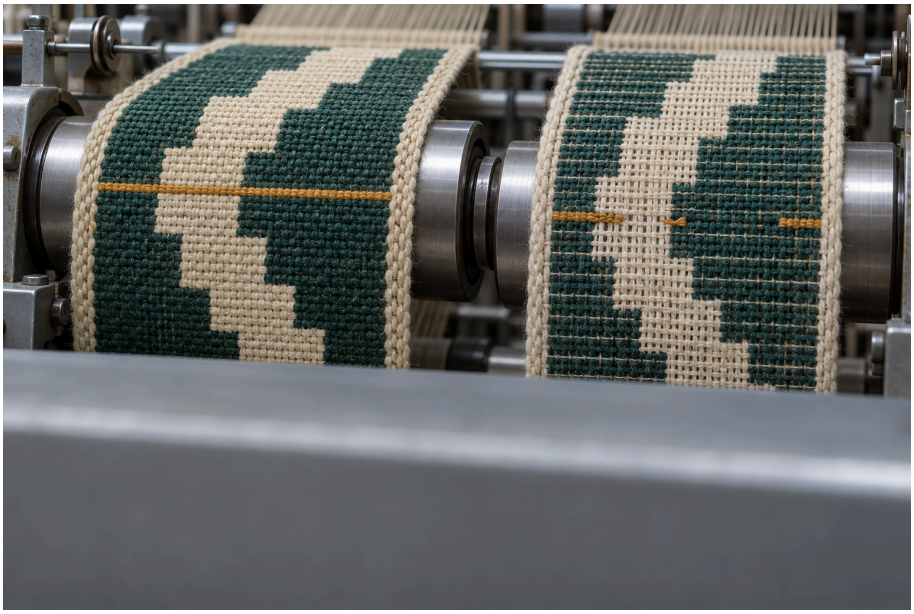
RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

LLM Context Compression Evidence Loss: An Acceptance 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 · 2026-09-25 · llm context compression · evidence loss · rag evaluation · citation grounding · source attribution · provenance · context engineering · llm evaluation

Original article: https://reliablecontext.com/articles/llm-context-compression-evidence-loss



In this article

1. Executive Summary
2. Introduction and Background
3. Definitions and failure taxonomy
4. Build the paired evaluation
5. Score evidence independently of the answer
6. Data Analysis and Evidence
7. Implications and Future Directions
8. Frequently Asked Questions (FAQs)
9. Conclusion

Executive Summary

Context compression can preserve a plausible answer while removing the evidence needed to defend it. A **correct answer** and a **resolvable citation** are distinct outcomes: the first concerns the answer against an independently fixed reference, while the second asks whether each material claim still points to a verbatim span in an identified, dated source. Retrieval-augmented generation (RAG) studies already separate correctness and citation quality, and Ragas defines faithfulness only relative to the retrieved context supplied to its evaluator. A high faithfulness score therefore cannot, by itself, establish that a cited source is true, current, or the version the organization actually observed. (aclanthology.org) (docs.ragas.io) (www.w3.org) (force11.org)

The most directly relevant controlled result is Li, Peng and Chen's **2026** experiment on **948 ASQA development questions**, using a fixed Self-RAG backbone and LLMLingua-2 to compress retrieved passages. At **20% evidence retention**, QA F1 changed from **23.35 to 19.08**, a **4.27-point** decline. Citation recall changed from **50.10 to 10.95**, a **39.15-point** decline, and citation precision from **63.71 to 12.13**, a **51.58-point** decline. These are changes in metric points for that task and configuration, not general failure rates for enterprise systems. The study's hierarchical variant recovered citation recall close to the uncompressed value while retaining almost the same average length, illustrating why apparent recovery must be reported beside its actual token savings. (aclanthology.org)

The operational decision should be made with a **paired acceptance test**. Freeze the source bundle, its version identifiers and observation dates; run the same questions and generator settings against the original and each compression budget; then score answer correctness, exact-span resolvability, provenance retention, contradiction handling, justified abstention, and token reduction separately. Review every case in which an answer remains acceptable but its qualifier, number, exception, conflict, or source locator disappears. A release rule should be set by the consequences of an unsupported claim, with no universal threshold inferred from the published benchmarks. W3C provenance guidance distinguishes a changing resource from its specific versions, while data-citation principles call for a verifiable timeslice or granular portion. (www.w3.org) (force11.org) (pages.nist.gov) (trec.nist.gov)

For systems that must defend values to a reviewer, the retained source and its dated excerpt remain the comparison baseline. **Reliable Context** describes a system of record with a cited, verbatim, dated source behind each value and reproducible past states; those are observable product claims, not a published compression benchmark. An organization can evaluate that model alongside prompt compression, which solves a different part of the workflow. The acceptance criterion is that the answer's material claims continue to resolve to the same source evidence at the tested budget, or that the system explicitly abstains when evidence has been lost. (research.google) (ojs.aaai.org)

Introduction and Background

Enterprise context engineers often shorten retrieved documents, conversation history, or agent state before asking a language model to answer. The shorter input can reduce token use and fit within an available context window. It can also remove the only sentence that distinguished an observation from an inference. A model may still produce a fluent answer from a retained summary or from prior knowledge. The resulting answer may look unchanged to a conventional accuracy metric, although its cited support is no longer inspectable. This report examines that difference as a decision about **evidence traceability and preservation**. (aclanthology.org) (research.google)

Two research lines motivate the distinction. Gao and colleagues' ALCE benchmark evaluates fluency, correctness, and citation quality separately; QASPER records evidence for answers about research papers. More recently, Li, Peng and Chen directly varied the amount of retrieved evidence while holding much of the answer-generation setup fixed. Together they show why an evaluator should not combine answer quality and attribution into a single score. Their datasets are public research settings; the questions, permissions, source changes, and loss costs inside an organization may differ substantially. (aclanthology.org) (trec.nist.gov)

The question is precise: **At which compression budget does a material answer cease to be defensible against the original source?** The object under test is a paired question and source bundle, not an isolated text summary. The original bundle has stable identity, captured content, an observation date, and reference annotations. Every compressed variant is evaluated against that same bundle. The method accommodates an extractive snippet selector, token pruning, abstractive summary, soft representation, or agent-memory compaction, but each must face the same downstream evidence test. Published methods optimize different objectives, from token classification to query-aware selection, so a favorable task score in one paper is not an evidence guarantee in another setting. (aclanthology.org) (www.microsoft.com) (arxiv.org) (aclanthology.org)

Definitions and failure taxonomy

What compression changes

Context compression maps a longer input into a shorter representation supplied to a generator. The unit removed may be a token, sentence, passage, document, or earlier message. LLMLingua-2 frames one version as token classification; Selective Context removes redundancy; SARA retains selected natural-language passages alongside compressed representation to preserve entities and numbers. These mechanisms create different risks for literal span recovery. The test should record both the method and the actual tokens retained, because a named budget is not necessarily the realized length. (arxiv.org) (aclanthology.org) (openreview.net)

Evidence span means the exact contiguous source text, or a precisely located source region where text extraction is impossible, that supports a claim. **Claim support** asks whether the span entails the claim with its qualifier, population, unit, and time. **Source identity** names the document or record, and **source version** fixes which observed state was read. A bare link to a changing page is weaker than a retained copy and locator: the page may later display different text. A summary can preserve a proposition while losing source provenance if its document identity or observation date is omitted. W3C PROV treats revisions as distinct entities, and the Memento protocol defines a prior resource state and its datetime. (www.w3.org) (datatracker.ietf.org) (force11.org)

A **correctness score** compares an answer to an independently specified answer or rubric. **Attribution score** tests whether the answer is supported by identified evidence. Neither implies the other. Ragas faithfulness divides supported response claims by response claims, using retrieved context as the reference. That metric can be useful for screening, but its documented denominator does not include whether the retrieval itself omitted a contrary source, whether the page changed, or whether the claimed excerpt exists verbatim in a frozen version. Those are additional checks in this proposed protocol. (docs.ragas.io) (research.google) (trec.nist.gov) (research.google)

The six paired failure cases

A compact test set should deliberately contain facts whose answers are easy to paraphrase but whose support is fragile. *Table 1* specifies paired cases to construct from a frozen corpus. The entries are **hypothetical test designs**, not observed benchmark results.

Case (hypothetical)	Original evidence to preserve	Compressed variant to challenge	Reviewer question
Unchanged fact	A single unambiguous source sentence with its document identity and observation date.	The same sentence or an exact extract is retained.	Does the answer and its locator still resolve to that sentence?
Lost date	A value that changed between two observed source versions.	The later value remains but the observation date is omitted.	Can the answer specify which version supports its temporal claim?
Lost exception	A rule followed immediately by a limiting clause.	The rule remains and the exception disappears.	Is the unqualified answer now overstated?
Swapped identity	Two related entities with similar names and distinct records.	Their names or roles are summarized into one ambiguous label.	Does the cited span identify the same real thing as the answer?
Conflicting passages	Two sources disagree, each with a date and scope.	One side survives, or both are blended into a single assertion.	Is the disagreement visible and is any resolution justified?
Absent evidence	The original bundle lacks a sufficient supporting passage.	A plausible answer is generated from a summary or model prior.	Should the answer abstain or state that support was not found?

The first case is a positive control; the others probe distinct ways evidence can disappear while fluent wording survives. Conflict cases should separate the detection of disagreement from choosing a value. Research on RAG source conflicts explicitly distinguishes factual accuracy from conflict management, and work on sufficient context distinguishes inadequate evidence from a model's failure to use adequate evidence. An abstention label is therefore a substantive result, not a blank answer. (research.google) (ojs.aaai.org) (proceedings.mlr.press)

An evidence record for each case should expose these fields:

- **Source identity:** Stable document or record identifier.
- **Source version:** The state observed by the evaluator.
- **Observation date:** When that state was captured.
- **Locator:** Page, section, line, or character offset.
- **Verbatim span:** Exact text used as support.
- **Support relation:** Why that span entails the claim.

Build the paired evaluation

Freeze the corpus and questions

Start with a **versioned source bundle** whose documents, permission state, extraction text, and observation dates are fixed before compression begins. Record a cryptographic digest or immutable snapshot identifier for each source and for the complete bundle. Preserve the exact text offsets or page locators that reviewers will inspect. The unit of evaluation is a case identifier linking one question, eligible source set, reference answer, and annotated supporting and conflicting spans. TREC's test-collection practice keeps documents, topics, and relevance judgments together; its guidance warns that judgments must match the document collection. (trec.nist.gov) (git-scm.com) (force11.org)

A case manifest should then name the following:

- **Case identifier:** Stable key for paired comparison.
- **Question:** Exact wording used in every run.
- **Eligible sources:** The documents the reader may see.
- **Reference answer:** Independently adjudicated expected content.
- **Annotated spans:** Support, contradiction, and missing evidence.
- **Permission state:** Access available to the tested reader.

The reference answer must be built independently of the compressed output. An annotator should mark each **material claim** and its required qualifier, such as date, unit, denominator, exception, or identity. More than one supporting span may be valid, but each must identify its source version. Where sources disagree, retain both passages and define what counts as a justified explanation. Where the corpus cannot answer, label the case as unanswerable from that bundle. QASPER's answer-plus-evidence construction and ASQA's ambiguous-question design are useful precedents, but a local collection should sample the organization's actual source and question types. (aclanthology.org) (arxiv.org) (research.google)

Use a frozen question set spanning routine facts, version changes, exceptions, identities, numerical values, multi-document synthesis, contradictions, and unsupported requests. Stratify by risk tier and by source format rather than sampling only popular questions. Do not change the question wording after observing a compressed answer; that would turn the comparison into a moving target. If a source is blocked for a reader, preserve that state as part of the case, because an inaccessible supporting span cannot be silently counted as evidence available to that reader. W3C PROV permits explicit usage and generation relationships, providing a vocabulary for recording how a tested output relates to its input. (www.w3.org) (pages.nist.gov) (force11.org)

Hold the generator fixed across budgets

Run the same corpus and questions against the **uncompressed baseline** and a prespecified grid of compressed variants. For each run, fix the generator release, prompt instructions, retrieval results and order, decoding configuration, tools, and reader permissions. Vary only the compression operation and its target budget; record actual retained tokens after preprocessing. Li, Peng and Chen held the retrieved passages as the varied input of their controlled Self-RAG experiment and tested several retention levels. That control is what permits a paired interpretation of the observed metric changes. (aclanthology.org) (trec.nist.gov)

The run log should fix these controls:

- **Generator:** Release and routing configuration.
- **Retrieval:** Candidate documents, order, and filters.
- **Instructions:** System and question prompt text.
- **Decoding:** Sampling parameters and random seed.
- **Tokenizer:** Version used to count retained tokens.
- **Clearance:** Source access under the reader policy.

Save each compressed text exactly as passed to the generator, together with compressor release, configuration, seed, and any query-conditioned inputs. A second run with a different seed can assess stochastic variation, but the paired comparison should preserve a stable case identifier. For extractive compression, record which original spans survived. For an abstractive summary, record any claimed source links and then verify them against the original. For latent or soft representations, direct span recovery may be impossible; in that case the generator must still supply a separate resolvable evidence path if the answer is to pass an evidence gate. Published work evaluates soft compression on downstream performance and grounding separately, underscoring that one metric cannot stand in for the other. (aclanthology.org) (research.google) (www.w3.org)

Define pass conditions before viewing results

An acceptance rule should specify what constitutes a material claim, what counts as a valid exact span, and how disagreement is adjudicated. A claim with the right number but wrong unit fails correctness; a claim with the right number and no located source may pass correctness yet fail resolvability. A citation to a document that merely discusses the topic fails span support. A source that supported an answer when observed but has since changed requires the observed version, not just the current URL. Data citation principles expressly call for a particular timeslice or granular portion sufficient for verification. (force11.org) (www.w3.org) (datatracker.ietf.org) (research.google)

The review process should have an explicit **uncertain** label. Automated entailment can screen claim-span pairs, but human reviewers should resolve lost-evidence cases and measure agreement on a sample. Language-model judges can exhibit position, verbosity, and self-enhancement biases; their outputs are measurements with error, not an oracle. The Attributed QA work used human annotations as a gold standard, and recent TREC RAG evaluation separates attribution verification from response completeness. (arxiv.org) (research.google) (pages.nist.gov) (docs.ragas.io)

Score evidence independently of the answer

For each case and budget, split the answer into atomic material claims. A **claim-correctness rate** has the number of correct claims in its numerator and all material answer claims in its denominator, assessed against the frozen reference. A **supported-claim retention rate** instead counts claims that resolve to an exact span in the eligible observed source version; its denominator is the material claims that were supported in the baseline answer, with baseline omissions reported separately. FActScore's atomic-fact decomposition and Ragas's supported-claims fraction motivate the granularity, but the proposed resolvability requirement is stricter than either metric's general definition. (aclanthology.org) (docs.ragas.io) (force11.org) (research.google)

Score **provenance retention** for every supported claim: source identifier, observed version, date, locator, and verbatim span must survive or be recoverable. Score **contradiction detection** on cases with conflicting passages: the answer must indicate that disagreement exists and represent both sides accurately. Score **justified abstention** on unanswerable cases and false abstention on answerable ones. Keep **token reduction** as its own axis, computed from actual input tokens under the same tokenizer. The 2025 TREC RAG track's separation of completeness and attribution and the conflict literature support treating these as distinct outcomes. (pages.nist.gov) (research.google) (ojs.aaai.org) (research.google)

Report the scorecard as separate columns:

- **Correctness:** Material claims matching the reference.
- **Span support:** Material claims with resolving excerpts.
- **Provenance:** Claims with source and date retained.
- **Conflict:** Disagreements represented accurately.
- **Abstention:** Unsupported cases declined for the right reason.
- **Efficiency:** Realized token reduction under one tokenizer.

Table 2 compares observable evidence paths. It includes **Reliable Context** because it directly offers a system of record for values that must be defended, while prompt compression is a separate operation. No row implies a measured speed or accuracy advantage.

Operating model	What is shortened or retained	Observable evidence test	Decision role
Uncompressed, versioned bundle	Full eligible source text and version identifiers.	Each claim resolves to its annotated source span and observation date.	Reference condition for paired deltas. (www.w3.org) (force11.org)

Operating model	What is shortened or retained	Observable evidence test	Decision role
Extractive text compression	Selected original tokens, sentences, or passages. (aclanthology.org)	Check whether the original supporting span survives with its source identity.	Candidate when spans and qualifiers can be carried forward.
Abstractive or latent compression	A rewritten summary or compact representation. (arxiv.org)	Require a separate pointer to the original observed span; summary wording alone is insufficient.	Candidate only where the evidence path can be reconstructed.
Agent-session compaction	Prior conversation or task state is shortened. (factory.com)	Test task continuity and separately test citations to external sources.	Session continuity is a distinct evaluation target.
Reliable Context, system of record	A canonical record per real thing, with cited, verbatim, dated source for each value; past states reproducible.	Inspect the value's source, locator, observed date, and exported chain of custody.	Evidence-bearing reference or downstream record, not a prompt compressor.

The rows serve different purposes. Factory's published compression evaluation applies methods to identical software-agent conversation prefixes and probes recall and continuation; its stated scope does not establish exact external source-span attribution. Reliable Context's row describes its public guarantees, which should be verified on the values a reader actually sees, and does not assert an evaluated compression result. For any workflow, the decisive comparison remains the paired original and compressed answer under the same source bundle. (factory.com) (trec.nist.gov)

Review the cases that accuracy hides

A useful review queue contains every case where correctness passes but span resolvability fails. Sort it by the information lost: date, qualifier, unit, number, identity, contradiction, or whole passage. A second queue contains answers whose cited span exists but does not entail the material claim. These are different problems. A third contains unsupported questions where the compressed answer should have abstained. This taxonomy prevents a high average answer score from hiding a small class of indefensible decisions. ALCE's separate citation-quality axis and the Attributable to Identified Sources framework both support this separation. (research.google) (ojs.aaai.org) (lrec.elra.info)

Reviewers should compare the exact original excerpt, compressed text, answer, cited locator, and observed version on one screen. Record a verdict and a short reason. If the compressed answer cites a different valid span in the frozen bundle, score it as supported under the predefined rules. If the source itself is wrong or outdated, that is a source-quality finding and should be reported separately from whether the compressed answer represented the source faithfully. This distinction follows from faithfulness being relative to supplied context and from provenance's treatment of document versions. (docs.ragas.io) (www.w3.org) (datatracker.ietf.org) (force11.org)

Data Analysis and Evidence

Li, Peng and Chen provide an unusually direct quantitative comparison, but their experiment is still a **single controlled setting**. They use the **948-instance ASQA development split**, retrieve **five passages** per question, and vary retained evidence at **1.00, 0.80, 0.60, 0.40, and 0.20** while using Self-RAG and LLMLingua-

2. Their citation recall denominator is gold-supported citation items, whereas precision concerns produced citations. Those definitions differ from this article's proposed exact-span and source-version measures. (aclanthology.org)

Table 3 reports the paper's endpoints as **absolute score points**, with arithmetic differences computed from its reported values. The endpoints are sufficient to show divergence; they are not a full compression frontier for other corpora or generators.

Metric in Li et al.	Full retained evidence	20% retained evidence	Absolute change	What it does not establish
QA F1	23.35	19.08	−4.27 points	Exact source-span resolvability. (aclanthology.org)
Citation recall	50.10	10.95	−39.15 points	Truth or version of cited source. (aclanthology.org)
Citation precision	63.71	12.13	−51.58 points	Whether a reviewer can reopen the observed passage. (aclanthology.org)

The large citation changes relative to the smaller QA F1 change justify a **separate evidence axis**, not a universal claim that compression always has this effect. The study's hierarchical variant brought citation recall from **50.10** to **49.80** while average retained length moved from **30.12** to **29.29** in the authors' reported measure. That is a useful warning: a method can recover attribution while yielding little further length reduction. SARA, a different approach, retains natural-language snippets for fine-grained entities and numbers and tests answer quality under token budgets; it does not provide the proposed exact-span audit. (aclanthology.org)

A September **2026** preprint by Mody offers a related but still preliminary comparison: in one ASQA setting, citation precision was **0.86** against generated summaries and **0.12** against original source spans. The author explicitly notes that span recovery and citation scoring share an entailment evaluator without independent human calibration. The figures illustrate why a citation that points into a compressed summary should be tested against the original source, but they should not be pooled with Li and colleagues' metrics or generalized to enterprise deployments. (arxiv.org)

Other published performance figures should not be merged into a common industry benchmark. LLMLingua-2 reports **2 to 5 times** compression with **1.6 to 2.9 times** end-to-end latency improvement in its evaluation. (aclanthology.org) LongLLMLingua reports a NaturalQuestions gain under a different model and setup. (arxiv.org) Those outcomes measure different tasks and denominators from Li and colleagues' citation metrics. The correct local analysis is a plot with **actual token savings on the horizontal axis** and **supported-claim retention on the vertical axis**, using raw numerators and denominators at every point. A second panel should show correctness and justified abstention, never averaged into retention. (arxiv.org) (docs.ragas.io)

For uncertainty, resample **paired case identifiers**, not individual claims divorced from their questions and source bundles. Recompute each budget's difference from the original on every resample and report an interval for the difference, plus failure counts by evidence type. This preserves the paired design used to expose hidden losses. Koehn's paired bootstrap procedure is an established precedent for comparing systems on the same test set; SciPy's paired-bootstrap documentation likewise uses the same resampled indices

across arrays. The interval method and resampling unit must be specified in the manifest. Report subgroup denominators, because a perfect result on a handful of date-change cases does not establish performance on all source changes. (aclanthology.org) (docs.scipy.org) (trec.nist.gov) (pages.nist.gov)

Implications and Future Directions

The acceptance test should lead to a **release rule per risk tier**. For low-consequence exploratory questions, teams may tolerate an explicitly labeled answer with incomplete attribution. For decisions that require a reviewer to defend a value, set a stricter condition: every material claim must resolve to an eligible observed source span, and an answer should abstain when its only support has been compressed away. The threshold is a local governance decision, not a number transferable from ASQA, QASPER, or software-agent session probes. ASQA centers ambiguous factoid questions, QASPER concerns scientific papers, and Factory tests session continuity, so their sampling frames do not represent every enterprise corpus. (arxiv.org) (factory.com) (ojs.aaai.org)

A release record should state the decision inputs:

- **Risk tier:** Consequence assigned before testing.
- **Threshold:** Required support and abstention performance.
- **Failure count:** Lost evidence by type and budget.
- **Uncertainty:** Paired interval and subgroup denominators.
- **Reviewer verdict:** Reasons for each disputed case.
- **Retest trigger:** Source, model, prompt, or permission change.

A reproducibility manifest should retain the corpus snapshot, source identifiers, observed dates, text extraction version, question and annotation set, compressor and generator releases, tokenizer, retrieval ranking, prompts, decoding settings, seeds, judge version, human review instructions, and permission context. A future reviewer should be able to reconstruct the original source state and the compressed input that produced a particular answer. W3C PROV describes derivation and version relationships, RFC 7089 distinguishes an archived state from a live resource, and FORCE11's principles emphasize granular, verifiable citation. Zenodo's recommendation to cite a version-specific identifier is one concrete publication practice. (www.w3.org) (datatracker.ietf.org) (force11.org) (git-scm.com) (zenodo.org)

Automatic judges can accelerate screening, but they require calibration. Compare their span-entailment and contradiction labels against blinded human review, disclose disagreement, and inspect errors concentrated in numerical qualifiers and identity changes. Zheng and colleagues documented judge biases; the Google Attributed QA study used human annotations, and a conflict benchmark with invented entities illustrates how tests can reduce reliance on familiar names. These studies motivate evaluation controls, not a guarantee that any one judge will be reliable for a local corpus. (arxiv.org) (research.google) (lrec.elra.info) (research.google)

Future studies should publish complete budget curves with **answer correctness, citation support, exact-span recovery, version accuracy, abstention, and realized token savings** on the same paired cases. They should include source updates and access changes, because a citation to today's page may not reproduce what the generator saw earlier. They should also report negative findings: a compressor that preserves answer

F1 while losing dates or exceptions has not met the evidence requirement, even if it saves tokens. This is the unresolved transfer question that controlled RAG papers and session-continuity evaluations leave for enterprise test collections. (www.w3.org) (datatracker.ietf.org) (factory.com)

Frequently Asked Questions (FAQs)

How does context compression affect citations in RAG?

It can remove a passage, locator, qualifier, or source identity while leaving enough semantic content for a plausible answer. In Li, Peng and Chen's controlled ASQA experiment, citation metrics fell much more sharply than QA F1 as retained evidence decreased. A citation should therefore be scored for exact support in the observed source, not merely for appearing in the generated text. (aclanthology.org) (research.google)

Does a faithful answer mean the source is true?

No. Ragas defines faithfulness relative to retrieved context. An answer can faithfully repeat an outdated or contradictory source. External truth, source version, and conflict resolution require additional evidence and review. Rashkin and colleagues explicitly treat source trustworthiness as a separate task from attribution. (docs.ragas.io) (www.w3.org) (research.google) (direct.mit.edu)

How can a team preserve evidence during compression?

Keep a frozen source copy and observation date, bind each material claim to a verbatim span and locator, and test that binding after every budget change. Extraction can retain literal spans; summaries and latent representations need a separate recoverable path to the original. Where no such path exists, the defensible outcome is explicit abstention. (force11.org) (ojs.aaai.org)

What should an auditable answer contain?

At minimum, the material claim, source identity, observed version or snapshot, observation date, exact supporting span, and a way to review conflicting or missing evidence. W3C PROV and data-citation principles provide useful provenance concepts; Reliable Context publicly describes those source and date properties for its values, without claiming to be a compression benchmark. (www.w3.org) (force11.org)

Can compression produce a hallucinated answer?

Compression can leave a generator without sufficient supporting context while it still produces a plausible answer. That is an inference from the evidence-loss mechanism, not a universal measured probability of hallucination. A paired test should mark whether the original bundle contains sufficient evidence, then count unsupported claims and justified abstentions at each budget. Sufficient-context research separates missing evidence from failure to use available evidence, and a recent compression preprint shows why support against a generated summary must be checked against original spans. (research.google) (arxiv.org)

Conclusion

A shorter context is acceptable for defensible answers only when the evidence chain survives the shortening at the intended budget. The test is paired and observable: the same fixed questions, original source versions, generator conditions, and scoring rules are applied to uncompressed and compressed inputs. Correctness, exact-span support, provenance, contradiction handling, abstention, and realized token savings remain separate outputs. No published average task score can replace inspection of a missing date, exception, number, identity, or citation span in the organization's own cases. (www.w3.org)

The practical release decision is therefore case and risk specific. Freeze the reference bundle; annotate claims and source spans; run budget variants; inspect every answer that stays plausible after its evidence is lost; and publish the resulting counts and uncertainty with a reproducibility manifest. A system may be useful for concise exploratory answers while failing a stricter evidence gate. Conversely, a method that retains the required spans at meaningful token savings can be accepted for that tested corpus and configuration, with retesting when sources, models, or prompts change. (trec.nist.gov) (force11.org) (datatracker.ietf.org)

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
2. https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/faithfulness
3. https://www.w3.org/TR/prov-primer
4. https://force11.org/info/joint-declaration-of-data-citation-principles-final
5. https://aclanthology.org/2026.customnlp4u-1.19
6. https://pages.nist.gov/trec-browser/trec34/rag/proceedings
7. https://trec.nist.gov/data/reljudge_eng.html
8. https://research.google/pubs/sufficient-context-a-new-lens-on-retrieval-augmented-generation-systems
9. https://ojs.aaai.org/index.php/AAAI/article/view/40496
10. https://aclanthology.org/2024.findings-acl.57
11. https://research.google/pubs/measuring-attribution-in-natural-language-generation-models
12. https://trec.nist.gov/howto.html
13. https://www.microsoft.com/en-us/research/project/llmlingua
14. https://arxiv.org/abs/2310.06201
15. https://aclanthology.org/2025.findings-emnlp.949
16. https://aclanthology.org/2026.acl-long.661
17. https://openreview.net/forum?id=ADsEdyI32n
18. https://datatracker.ietf.org/doc/html/rfc7089
19. https://research.google/pubs/dragged-into-a-conflict-detecting-and-addressing-conflicting-sources-in-retrieval-augmented-llms
20. https://proceedings.mlr.press/v284/zhu25a.html

21. https://git-scm.com/book/en/v2/Git-Branching-Branches-in-a-Nutshell
22. https://aclanthology.org/2021.naacl-main.365
23. https://arxiv.org/abs/2204.06092
24. https://www.w3.org/TR/prov-dm
25. https://aclanthology.org/2026.customnlp4u-1.19.pdf
26. https://arxiv.org/abs/2306.05685
27. https://research.google/pubs/attributed-question-answering-evaluation-and-modeling-for-attributed-large-language-models
28. https://aclanthology.org/2023.emnlp-main.741
29. https://arxiv.org/abs/2609.14245
30. https://factory.com/news/evaluating-compression
31. https://lrec.elra.info/lrec2026-main-182
32. https://arxiv.org/abs/2310.06839
33. https://aclanthology.org/W04-3250.pdf
34. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bootstrap.html
35. https://zenodo.org/help/versioning
36. https://direct.mit.edu/coli/article/49/4/777/116438/Measuring-Attribution-in-Natural-Language

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.