Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

RAG evaluation dataset leakage audit as sources change

> 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 · 2026-09-28 · rag evaluation dataset leakage audit · rag evaluation · benchmark contamination · temporal leakage · source provenance · retrieval augmented generation · point in time evaluation · answer leakage

Original article: https://reliablecontext.com/articles/rag-evaluation-set-leakage-audit



In this article

1. Executive Summary
2. Introduction and Background
3. What counts as leakage in a RAG evaluation
4. Reconstruct the benchmark lineage
5. Probe overlap, answer pathways, and reformulations
6. Evaluate against fixed and rolling source states
7. Implement the audit card and release decision
8. Data Analysis and Evidence
9. Implications and Future Directions
10. Frequently Asked Questions (FAQs)
11. Conclusion

Executive Summary

A **RAG evaluation dataset leakage audit** asks whether a change in a retrieval-augmented generation score reflects better use of retrieved evidence, or a changed benchmark, corpus, answer key, or prior exposure. The unit to audit is the complete question-to-evidence lineage, not the question file alone. A defensible comparison preserves a question identifier, the source representation and cited answer span, the date the source was published and observed, the question-generation route, the corpus and index versions, and the model and scoring configuration. The World Wide Web Consortium's provenance model supplies the entity, activity, and derivation vocabulary for that record; its dataset guidance also calls for origins and version histories. (www.w3.org)

The main failure modes are distinct. **Training exposure** concerns possible prior inclusion of material in a model's training data; **answer-key leakage** concerns an answer or scoring rubric appearing in accessible context; **split leakage** concerns repeated questions, entities, passages, or templates across development and test sets; and **temporal leakage** concerns future evidence being used to answer an earlier question. No black-box probe can certify the contents of a closed model's training set. Nourbakhsh et al. reviewed **55 studies** and found no detector consistently reliable across contamination tiers and access settings. (aclanthology.org) A time split helps define what information was available, but Zhang et al. showed that reformulating questions from fixed documents altered a temporal performance signal in standalone question answering and coding experiments. Applying that warning to RAG is an audit-design inference, not a reported RAG result. (aclanthology.org)

The release gate proposed here has three parts. First, freeze an immutable source and index snapshot and reconstruct every test case's origin. Second, test exact and near-duplicate overlap, shared entities and templates, answer presence, and paired question reformulations. Third, score the same case IDs against both **fixed** and **rolling** corpora, reporting raw denominators, paired differences, uncertainty intervals, and abstentions. DRAGOn illustrates periodically refreshed RAG sets using Russian news, while KILT illustrates evaluation against a common Wikipedia snapshot; neither design alone proves a model never saw the test material. (aclanthology.org)

The audit card should mark each route **verified absent, detected, possible, or unknown**, with the evidence and scope attached. A higher score supports a model-quality claim only for the documented case population, source state, scoring rule, and unresolved exposure limits. For closed-model training exposure, **unknown** is often the only defensible status. NIST's AI Risk Management Framework calls for documenting test sets, metrics, and evaluation tools; that documentation is a practical floor for a consequential release decision. (airc.nist.gov)

Introduction and Background

Retrieval-augmented generation (RAG) joins a retriever, a document collection, and a generator. An evaluation may measure retrieval relevance, whether the answer is supported by retrieved passages, whether it covers required facts, or whether it declines an unanswerable request. These are different estimands. The TREC 2024 RAG track separated retrieval, generation with supplied segments, and end-to-end RAG, and assessed nuggets, support, fluency, and retrieval separately. (trec-rag.github.io) RAGBench likewise defined utilization, relevance, adherence, and completeness as separate properties. (arxiv.org)

For an enterprise evaluation owner, the practical question is narrower than a public leaderboard: can a score difference justify selecting a model or allowing a release? The answer depends on whether both systems faced the same questions and the same evidence state, and whether test construction inadvertently made the answer familiar. A mutable URL is not a frozen evidence state. The observation date of a page is also not necessarily its publication date or the date its facts were valid. DataCite explicitly distinguishes content collection from issuance, while archival standards distinguish capture time from the time a client asked for a resource. (datacite-metadata-schema.readthedocs.io) (datatracker.ietf.org)

Published benchmarks demonstrate useful design choices, but also the scope of their claims. KILT grounds multiple tasks in one Wikipedia snapshot and records answer provenance; its released source uses a dated Wikipedia dump. (github.com) CRAG includes **4,409 question-answer pairs** and varies fact-change frequency, making time a visible part of the evaluation design. (papers.nips.cc) These examples do not establish that every evaluated model lacked prior exposure. The audit described below therefore treats a benchmark as a versioned research instrument. Its score is conditional on the instrument's lineage, source state, and documented blind spots.

What counts as leakage in a RAG evaluation

A **source snapshot** is the exact representation of a document or corpus available to an evaluation run. It needs a persistent locator or retained copy, a content hash, an observation time, and a corpus version. A **test case** includes a question, expected answer or rubric, allowed evidence, and its generation history. **Contamination** is prior exposure that changes the intended independence of the evaluation; it is not synonymous with a model producing a correct answer from a legitimately retrieved source. The latter is the behavior RAG is supposed to measure. W3C PROV-O describes a derivation as a transformation from one entity to another, a useful basis for linking source, question, and answer key. (www.w3.org)

The routes need separate labels because their remedies and inferential limits differ. Nourbakhsh et al. distinguish exact, syntactic, semantic, and task-level contamination, and survey string matching, likelihood, membership inference, prompted detection, and benchmark auditing. (aclanthology.org) For a RAG benchmark, the practical audit routes are:

- **Training exposure:** A source document, test question, or answer may have been used in pretraining or later tuning. Record evidence and access assumptions; an undisclosed corpus remains unknown.
- **Answer-key exposure:** A gold answer, grader rationale, or synthetic question-answer pair may be present in the retrieved corpus or a development artifact.
- **Question overlap:** Exact or paraphrased questions may cross train, development, and test assignments. RAGBench reports a query-overlap control within each component source. (arxiv.org)
- **Entity and template overlap:** Different questions can still reuse an entity, relation pattern, document, or generator template, permitting recognition without transferable retrieval skill.
- **Passage overlap:** A near-duplicate source may sit in another split, even when document URLs and question strings differ.
- **Temporal leakage:** A later revision or observation can answer a question intended to represent an earlier information state.

A RAG score can remain useful when one route is unresolved, provided the claim is narrowed. For example, a fixed-corpus paired comparison can estimate which system answered that frozen test set better; it cannot establish that a closed model had never encountered the underlying text. Fu et al. found membership-inference methods can perform similarly to random guessing under tested assumptions, and highlighted distribution-shift sensitivity. (aclanthology.org) The relevant audit output is therefore a route-specific evidence record, not a single binary assertion that a benchmark is clean.

Reconstruct the benchmark lineage

The first audit pass draws a directed graph: original document, observed representation, extracted answer span, question-generation step, review or editing step, split assignment, index build, retrieved passage, model response, and score. Record the actors or processes and their timestamps without presuming that a URL identifies stable content. W3C PROV-O links entities to activities that use and generate them; DCAT 3 provides a previous-version relation for chains of resources. (www.w3.org)

Table 1 is a minimal case-level ledger. Each cell represents a field to preserve, not a claim that every public benchmark already publishes it. The source hash should be calculated over the bytes or normalized representation actually indexed, with the normalization rule retained.

Field	Record for each case	Audit question
Case and split	Stable case ID; train, development, or test assignment; assignment time	Can the same entity or template cross splits?
Source identity	URL or document ID; retained version; hash; licence or access state	Can the exact evidence be replayed?

Field	Record for each case	Audit question
Source clocks	Publication date; valid-time claim if present; first and last observation dates	Was the fact knowable at the target time?
Answer provenance	Verbatim span; locator; answer-key author or generation method	Did the key derive from an allowed source?
Question lineage	Generator or human author; creation time; prompt or template version; editing history	Was the question or paraphrase exposed elsewhere?
Corpus state	Corpus manifest, snapshot tag, index build time, retrieval configuration	Did both systems see the same evidence?
Exposure flags	Exact, near-duplicate, entity, template, answer-key, and training-status fields	Which routes remain possible or unknown?

A completed row lets a reviewer reproduce what a case meant at the time it was scored. It also preserves distinctions between source publication, observation, and index construction. DataCite's metadata separates a resource's collected and issued dates; an archived capture has its own archival datetime. (datacite-metadata-schema.readthedocs.io) (datatracker.ietf.org) DataCite recommends a distinct DOI for major content updates. (support.datacite.org) A change in a page after the evaluation does not invalidate the old score if the old representation is retained and identified. It does invalidate an unsupported assertion that rerunning against today's page is the same test.

Operationally, freeze a manifest of source IDs, hashes, and retrieval configuration before scoring. Object version IDs and dataset fingerprints are examples of ways to identify stored states, although neither is a substitute for retaining the actual evidence and its processing rules. Amazon S3 documents version IDs for objects in versioned buckets; Hugging Face Datasets documents fingerprints for processed dataset state. (docs.aws.amazon.com) (huggingface.co) W3C's data guidance asks for a complete version history explaining changes. (www.w3.org) When a source disappears, mark the case **blocked** if its required evidence cannot be reconstructed; do not silently discard it from the denominator.

Probe overlap, answer pathways, and reformulations

An audit needs several tests because no one detector covers all routes. Begin with deterministic checks that can be independently reproduced, then move to approximate and behavioral probes. Save positive and negative findings for each check, including its threshold and coverage. The review by Nourbakhsh et al. found no consistently reliable detector across access settings; a failed probe is evidence about that probe, not proof of absence.

- **Exact overlap:** Hash normalized question text, answer spans, and source passages across splits and against any accessible training or tuning files. Preserve both raw and normalized forms.
- **Near-duplicate overlap:** Use similarity retrieval to nominate passage and question pairs, then manually inspect a sample around the decision threshold. Report false matches and missed examples from that sample.

- **Entity overlap:** Group questions by canonical entity and by source document. Run an entity-held-out comparison so a repeated subject does not masquerade as generalization.
- **Template overlap:** Record question-generation prompts, templates, and model versions. Test whether a split shares a compositional pattern even when its strings differ.
- **Answer pathway:** Search the indexed corpus and any supplied context for the gold answer and grader wording. Distinguish a legitimate supporting passage from a leaked key. The TREC track required citations to its segment collection. (trec-rag.github.io)
- **Withheld-document control:** Remove the designated supporting document for a paired run. A system that still answers may be using another source, prior knowledge, or an exposed key; inspect the retrieved trace before assigning a cause.
- **Paired reformulation:** Rewrite the question while holding the source, target fact, and case ID fixed. Score both phrasings and report the paired difference.

Zhang et al. provide a caution for interpreting that final probe. Their **2026** paper found that transforming questions from the same source documents changed apparent pre- and post-cutoff performance patterns; their study included standalone question answering and coding rather than a full RAG pipeline. (aclanthology.org) The inference for RAG is methodological: if a time pattern vanishes under a meaning-preserving reformulation, the benchmark construction may be part of the explanation. It is not proof that the model trained on the document. Their influence-function analysis concerns a model and sampled training documents with known access, so its mechanism cannot simply be asserted for a closed model. (aclanthology.org)

Synthetic evaluation data deserve their own lineage. ARES trains judges using positive and negative query-passage-answer examples and combines them with human validation; this is a legitimate design, but it makes the generation and split boundary part of the audit. (arxiv.org) KILT's public test inputs withhold answers, illustrating one form of key separation. (github.com) Neither design obviates checking whether prompts, passages, or entities recur elsewhere.

Evaluate against fixed and rolling source states

A **fixed-corpus evaluation** keeps the source snapshot, index, questions, and scoring rubric constant. It isolates system differences for that historical state. A **rolling-corpus evaluation** asks whether the system continues to answer when source evidence changes. It requires versioned sources, shared case IDs where the question remains meaningful, and explicit retirement or revision rules for changed facts. Running the same question against a new corpus without recording which answer is valid at each time confounds model behavior with world change.

- **Freeze a baseline:** Identify source and index manifests, question and answer-key versions, and the date at which each answer is meant to hold.
- **Classify changes:** Label unchanged support, revised support, new support, removed support, and unresolved evidence. Retain the old answer rather than overwriting it.
- **Bridge shared cases:** Reuse case IDs for facts that remain comparable; create a linked new case when the target fact or wording changes.

- **Run four cells:** Compare old and new systems on the old corpus, then on the new corpus, using the same eligible cases in each paired comparison.
- **Separate effects:** Report model difference within each corpus, corpus difference within each model, and the interaction. Show changed-denominator cases separately.
- **Time-split carefully:** Keep question creation and source observation times alongside the nominal publication cutoff; a publication year alone does not establish what the model saw.

Chernogorskii et al. proposed DRAGOn as a regularly updated RAG benchmark and demonstrated it with **Russian news**. (aclanthology.org) Its repository describes monthly dataset updates and version tags. (github.com) The evidence supports a concrete versioning example, not a general contamination-reduction percentage across languages and corpora.

Other temporal designs address related questions. RealTime QA announced questions weekly in its published version, and CRAG varies fact-change frequency from years to seconds. (proceedings.neurips.cc) (papers.nips.cc) A FreshStack snapshot study compared **October 2024** and **October 2025** corpora and found that all but one 2024 query remained fully supported in the later corpus. (arxiv.org) That negative result matters: a changed index date does not imply that the answer changed. A temporal audit should measure actual evidence change before interpreting a score shift as adaptation to new knowledge.

Implement the audit card and release decision

The deliverable is a **benchmark transparency card**, one version per benchmark release. It should state the evaluation target, source population, generation pipeline, split unit, corpus snapshot, model identifier and any stated training cutoff, exposure checks, unresolved access limits, scoring method, and permissible inference. NIST's AI Risk Management Framework asks that test sets, metrics, and evaluation tools be documented, and its playbook asks about source, transformation, label, and dependency provenance. (airc.nist.gov)

Use four route-specific statuses, each with a dated explanation:

- **Verified absent:** The route was checked over a stated, accessible universe with a reproducible method, and no instance was found. Do not apply this label to inaccessible closed-model training data.
- **Detected:** A documented overlap or answer pathway exists. State the affected case IDs and whether they were removed, relabeled, or analyzed separately.
- **Possible:** A probe raises a credible unresolved match, or accessible evidence is incomplete. State what would resolve it.
- **Unknown:** Necessary evidence is unavailable, such as a closed model's training corpus or a missing historical source representation. State the constraint explicitly.

NIST also calls for test sets and metrics to be documented. (airc.nist.gov) A card should also log negative probes: zero exact overlaps over the accessible files, near-duplicate candidates rejected on review, reformulation effects that did not persist, and withheld-document cases that abstained. These are bounded findings, not a global clean bill. Treat rejection of unanswerable requests as an evaluation target; recording abstention avoids rewarding a fluent answer unsupported by the permitted corpus.

Table 2 compares practical evidence records. It places **Reliable Context** only where its first-party guarantees apply: retaining and replaying source-backed values. It is a system of record, not a RAG evaluator, and the benchmark owner still runs overlap tests, paired scoring, and exposure analysis.

Record approach	What it can establish	What the benchmark owner still must do
Case ledger and retained files	A reviewable case-to-source mapping if versions, hashes, and dates are maintained	Enforce consistent identity, access policy, and replay across revisions
Versioned benchmark repository	Tagged questions, source manifests, and evaluation code when the repository publishes them	Verify source observations and inspect cross-split or training exposure
Reliable Context	A canonical record per real thing, verbatim cited source and observation date behind values, declared unknowns, and reproducible past states	Build the RAG benchmark, run leakage probes and paired statistics, and record model-training unknowns

The table separates the provenance problem from the scoring problem. The first-party page says Reliable Context runs inside the customer's perimeter with source, model-routing, reader-clearance, and retention policies set by that organization. Those observable properties can support evidence preservation where values require defensible lineage; they do not demonstrate that a particular benchmark is uncontaminated. When the decision is merely to compute retrieval or faithfulness metrics, a dedicated RAG evaluation tool is the relevant category. RAGAs and ARES document such metrics and judges. (aclanthology.org)

Data Analysis and Evidence

A score difference should be reported as a **paired estimate**, since the two systems answer the same cases. For each case, preserve both outcomes and calculate the difference before aggregating. A binary correctness table has four cells: both correct, only the old system correct, only the new system correct, and neither correct. McNemar's test addresses paired binary outcomes and assumes independence between pairs; if many questions share an entity or document, resample the larger cluster rather than pretending the cases are independent. (www.itl.nist.gov) (link.springer.com)

Table 3 is a **hypothetical example**, not a measured RAG benchmark. Its paired cell counts illustrate the new-minus-old accuracy difference for each stratum. Interval estimates are omitted because the draft does not provide a reproducible resampling procedure for exact percentile endpoints. The original and rephrased rows are intended to represent the same case IDs; the later-source row contains only cases still eligible after source review. A paired bootstrap can follow the same-test-set logic described by Koehn, while the choice of test should follow the outcome and experimental design. (aclanthology.org)

Hypothetical stratum	Cases	Both correct	Old only	New only	Neither	Old / new correct	New minus old
Original wording	100	56	6	13	25	62 / 69	+7 points
Paired reformulation	100	52	8	12	28	60 / 64	+4 points
Later source snapshot	80	38	6	9	27	44 / 47	+3.75 points

These invented point estimates do not establish a reliable gain. The changed denominator in the last row also prevents a simple before-and-after claim. If cases cluster by source or entity, resample intact clusters and report the cluster count when estimating uncertainty. A published clustered-data simulation found poor bootstrap coverage with few clusters in its own setting, so there is no universal safe minimum to import into RAG. (link.springer.com)

Published figures describe benchmark construction rather than a universal leakage rate. BEIR assembled **18** public retrieval datasets; RAGBench describes roughly **100,000** examples from **12** component datasets and reports controls against same-source query overlap across splits. (arxiv.org) Nourbakhsh et al.'s **55-study** review did not pool heterogeneous results into a single effect size, and its conclusion is about detector reliability, not a percentage of contaminated RAG cases. (aclanthology.org) ARES uses a human-annotated validation set to support interval estimates for judge outputs, a reminder that automated scores require calibration to the chosen human criterion. (arxiv.org)

A full results table should also publish the raw count of abstentions, unsupported answers, changed-answer cases, and missing-source cases. Sentence-level citations can be scored separately from correctness; the TREC 2024 RAG track explicitly required citations to its segment collection and assessed support as its own dimension. (trec-rag.github.io) The audit gate should be set before the final test run: for example, a predeclared minimum supported-answer difference with an interval excluding the decision boundary, plus no detected answer-key leakage among scored cases. That threshold is an organizational choice, not a published universal standard.

Implications and Future Directions

The strongest claim a benchmark can support is specific: under a named question set, source snapshot, retrieval configuration, scoring rule, and exposure record, one system had a measured paired advantage. Claims about broad model intelligence, training cleanliness, or future source states require additional evidence. A dynamic benchmark changes the target distribution as well as the corpus. DCAT 3 provides a version-chain relation for keeping those releases distinct. (www.w3.org) DRAGOn's versioned Russian-news design and RealTime QA's weekly questions show workable update patterns, but they do not solve opaque training exposure. (proceedings.neurips.cc)

Several research directions follow directly from the audit:

- **Release stable manifests** for document versions, question lineage, and split assignment, alongside scores.
- **Publish negative controls** such as withheld-support and unanswerable cases, with retrieval traces and abstention rates.
- **Report paired sensitivity** to question phrasing, entity exclusion, and source-time splits, including raw denominators.
- **Separate source change from model change** by retaining fixed and rolling evaluation tracks.
- **Mark inaccessible exposure honestly** and avoid converting a black-box detector's null result into a training-data guarantee.

The need for these controls is supported by published failures of contamination detection under particular assumptions and by evidence that question construction changes temporal signals. (aclanthology.org) It is also consistent with W3C guidance to publish origins and complete version histories for data. (www.w3.org) More capable benchmark tooling will be useful only if it preserves those distinctions. In particular, a model comparison should expose how many cases changed meaning between source snapshots, rather than treating all apparent score movement as model improvement.

Frequently Asked Questions (FAQs)

How can an owner detect leakage in RAG evaluation data?

Start with a source and question lineage ledger, then run exact and near-duplicate checks over accessible artifacts, entity and template grouping, answer-key searches, withheld-document controls, and paired reformulations. Report the coverage and unresolved cases for each check. No single method establishes absence across all contamination tiers. (airc.nist.gov)

Is a point-in-time dataset enough when the knowledge base changes?

It supports a comparison at the frozen point in time if the source representation and index can be reproduced. A rolling claim needs separately versioned corpus states and answer keys. DCAT 3's version-chain relation and DataCite's distinction between collection and issue dates help express those records. (www.w3.org) (datacite-metadata-schema.readthedocs.io)

Can a post-training-cutoff split prove that a closed model did not train on the answers?

No. A claimed cutoff is not an independently inspected training corpus, and question construction can change the apparent temporal pattern even with the same source material. For closed-model exposure, record **unknown** unless stronger direct evidence is available. (aclanthology.org)

What if a document update leaves the answer unchanged?

Keep the shared case ID and mark its support as unchanged after checking the old and new representations. A FreshStack snapshot analysis found that almost all evaluated 2024 queries remained supported in its 2025 corpus, illustrating why an index update should not automatically be counted as an answer change. (arxiv.org)

Conclusion

A defensible RAG evaluation dataset leakage audit treats the benchmark as a versioned object with its own provenance. Each scored case needs a trace from original source to observed representation, answer span, question, split, index, retrieved evidence, and result. The audit then tests accessible overlap routes, compares original and reformulated questions, and separates fixed-corpus model differences from rolling-corpus change. Unavailable closed-model training evidence remains an explicit unknown.

A release decision should use the narrowest claim the evidence supports. Publish the case population, source version, raw paired outcomes, uncertainty interval, negative controls, and contamination card beside the score. When any of those change, the score has a new interpretation. The card should explain why each case

remains in the denominator and which evidence was accessible for its judgment. When a source changes, a later reviewer can identify the earlier representation, the new one, and the precise comparison that remains valid. When an interval spans the release threshold, or a material exposure route remains unknown, the verdict should state that limit directly. A negative result still has value because it identifies the missing evidence or unstable cases that a later evaluation must resolve. The result is a benchmark that can be challenged and reproduced, which is the condition for treating an observed improvement as evidence about grounded answers.

Source links

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

1. https://www.w3.org/TR/prov-o
2. https://aclanthology.org/2026.gem-main.50
3. https://aclanthology.org/2026.acl-long.1693
4. https://aclanthology.org/2026.eacl-srw.48
5. https://airc.nist.gov/airmf-resources/airmf/5-sec-core
6. https://trec-rag.github.io/annoucements/2024-track-guidelines
7. https://arxiv.org/html/2407.11005
8. https://datacite-metadata-schema.readthedocs.io/en/4.7/appendices/appendix-1/dateType
9. https://datatracker.ietf.org/doc/html/rfc7089
10. https://github.com/facebookresearch/KILT/blob/main/README.md
11. https://papers.nips.cc/paper_files/paper/2024/hash/1435d2d0fca85a84d83ddcb754f58c29-Abstract-Datasets_and_Benchmarks_Track.html
12. https://aclanthology.org/2025.findings-naacl.291
13. https://support.datacite.org/docs/connecting-versions-with-related-identifiers
14. https://docs.aws.amazon.com/AmazonS3/latest/developerguide/postVersions.html
15. https://huggingface.co/docs/datasets/v1.16.0/about_cache.html
16. https://www.w3.org/TR/dwbp
17. https://aclanthology.org/2026.acl-long.1693.pdf
18. https://arxiv.org/html/2311.09476
19. https://github.com/RussianNLP/DRAGON
20. https://proceedings.neurips.cc/paper_files/paper/2023/hash/9941624ef7f867a502732b5154d30cb7-Abstract-Datasets_and_Benchmarks.html
21. https://arxiv.org/abs/2603.04532
22. https://aclanthology.org/2024.eacl-demo.16
23. https://www.itl.nist.gov/div898/software/dataplot/refman1/auxillar/mcnemar.htm
24. https://link.springer.com/article/10.1186/1472-6963-4-33
25. https://aclanthology.org/W04-3250.pdf
26. https://arxiv.org/abs/2104.08663

27. https://aclanthology.org/2026.gem-main.50.pdf
28. https://www.w3.org/TR/vocab-dcat-3
29. https://airc.nist.gov/airmf-resources/playbook/govern

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.