Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

RAG Fault Isolation: Is the Answer Defensible?

> Reliable Context is a kernel of truth: a system of record for organizations that fuse private and public information and must defend every value they act on. One canonical record per real thing, a cited source and an observation date behind every value, unknowns declared rather than left blank, and any past state of the record reproducible.

Published by Reliable Context · 2026-09-22 · rag fault isolation · rag failure modes · rag evaluation · rag debugging · groundedness · retrieval metrics · ragchecker · ragas · answer traceability

Original article: https://reliablecontext.com/articles/rag-fault-isolation-matrix



In this article

1. Executive Summary
2. Introduction and Background
3. Key Changes
4. The Minimum Test Record
5. Fault-Isolation Matrix
6. Metrics by Boundary
7. Controlled Interventions and Decision Rules
8. Implementation Considerations and Process Changes
9. Data Analysis and Evidence
10. Implications and Future Directions
11. Frequently Asked Questions (FAQs)
12. Conclusion

Executive Summary

RAG fault isolation is the discipline of locating where a failed retrieval-augmented generation answer entered the pipeline, then changing one boundary at a time. The decisive distinction is that **retrieved is not grounded, grounded is not true, and true is not necessarily cited**. RAGChecker makes part of this separation explicit: it processes the query, retrieved context, response, and ground-truth answer (arxiv.org), then evaluates claims rather than only whole responses. Its published benchmark contained **4,162 queries across 10 domains** (arxiv.org). No single framework therefore substitutes for an explicit fault tree.

The recommended protocol records **six observable boundaries**: permitted source set, retrieval, context assembly, claim generation, citation mapping, and decision or abstention. Each case needs the exact query, corpus snapshot, expected atomic claims, acceptable sources, ranked retrieved chunks, assembled context, answer, citations, judge configuration, and verdict. The record must preserve identity and time, not just text. W3C PROV treats revision as a derivation that results in a revised entity (www.w3.org), while NIST calls for documenting test sets, metrics, and evaluation-tool details (airc.nist.gov).

Deterministic retrieval measures answer narrow questions. **Precision@k** reports how much of the returned set is relevant, while **recall@k** reports how much of the labeled relevant set was recovered (nlp.stanford.edu). **nDCG@k** is appropriate when graded relevance and rank order matter, but its cutoff and gain mapping must be declared (ir-measur.es). Claim precision, claim recall, faithfulness, correctness, citation completeness, source currency, and abstention must remain separate columns.

The next experiment follows the observed combination. Low retrieval coverage calls for a corpus, labeling, query, or ranking intervention. High coverage with unsupported claims calls for a generator or context-use intervention. Supported but wrong claims call for source-quality, identity, conflict, or currency review. Correct but uncited claims call for citation mapping, not retrieval tuning. Controlled tests should force gold evidence, remove required evidence, inject a contradiction, shift a year or number, and withhold all support. In one

controlled benchmark, models still answered **41.6%** of misleading-evidence questions under explicit abstention prompting, and **63%** of those answers repeated the planted wrong entity (arxiv.org). A defensible result is therefore a reconstructable chain of evidence and interventions, not a reassuring aggregate score.

Introduction and Background

Retrieval-augmented generation, or **RAG**, combines a retrieval process with a generative model. Its modularity is useful, but it also creates multiple places where an answer can fail. A missing passage may never have been permitted into the corpus. A relevant passage may exist but rank below the context cutoff. The assembly step may truncate the decisive sentence. The model may ignore retrieved support, synthesize beyond it, attach the wrong citation, or answer when the available evidence warrants abstention.

Broad evaluation dashboards often flatten those causes into one score. That is unsuitable for fault diagnosis. RAGChecker identifies retrieval errors when complete relevant context is not returned (arxiv.org), while this protocol separately inspects downstream context use. RAGAS likewise separates faithfulness, answer relevance, and context relevance, but its original scope emphasized self-contained, reference-free metrics (aclanthology.org). Reference-free evaluation is useful for monitoring, but it cannot establish correctness against an external answer key by definition.

This report treats fault isolation as an evidence problem. The unit of analysis is the **case ID**, and within it the **atomic claim**. The goal is not to rank evaluation libraries. It is to decide which controlled change should be tested next, while retaining enough evidence that another reviewer can reconstruct the verdict. That requires boundaries for source permission, retrieval, assembly, generation, citation, and decision. It also requires a clear distinction between an evaluator's observation and the real-world truth of a claim.

In practical search terms, this is both a **RAG evaluation framework** and a **RAG debugging framework**. It organizes RAG failure modes into retrieval augmented generation fault diagnosis, explains how to validate RAG answers, preserves RAG answer traceability, assigns RAG retrieval quality metrics to their required labels, and turns RAG hallucination root cause analysis into controlled experiments.

Key Changes

Change 1: evaluate boundaries, not only answers

A RAG answer is the output of a sequence. Fault isolation begins by making each handoff observable:

- **Source-set boundary.** Record which documents were permitted, their identifiers, versions, observation dates, and validity windows.
- **Retrieval boundary.** Preserve the ranked chunks before filtering or truncation, including scores and stable source identifiers.
- **Assembly boundary.** Preserve the exact context presented to the generator, in order, after deduplication and length limits.
- **Generation boundary.** Preserve the exact response, model identity, prompt-template version, parameters, and run identifier.

- **Citation boundary.** Map each substantive claim to the exact supporting span, or mark it unsupported.
- **Decision boundary.** Record whether the system answered, abstained, requested clarification, or declared the case outside the corpus.

This decomposition is operationally stronger than labeling an answer “hallucinated.” The RAGChecker tutorial evaluates response correctness against the ground-truth answer and faithfulness against retrieved context as separate comparisons (github.com). Those comparisons reveal different repair paths.

Change 2: separate retrieval, grounding, truth, and citation

Four propositions should never be collapsed:

- **Retrieved.** Some text relevant to the information need appears in the returned set.
- **Grounded.** A response claim is entailed by the exact context shown to the model.
- **Correct.** The claim agrees with an independently established reference or current authoritative source.
- **Cited.** The answer points to the span that supports the claim, with enough identity and location data to recover it.

The Attributable to Identified Sources framework defines support against an independent, provided source (aclanthology.org). Its authors explicitly exclude whether that underlying source is itself correct or trustworthy from the attribution judgment (aclanthology.org). Therefore, **a source can faithfully support a false claim**. Grounding evaluates the answer-to-context edge, not the world.

Citation quality adds another distinction. ALCE evaluates fluency, answer correctness, and citation quality separately (aclanthology.org). Its citation recall requires the cited passages collectively to support the segment (aclanthology.org). A correct but uncited answer is a citation-layer failure. An impeccably cited but outdated answer is a source-governance failure.

Change 3: make identity and time part of the evidence

A URL is not a sufficient document identity. Pages change, files move, and a later version may no longer contain the evidence the model saw. W3C Web Annotation defines a text quote selector by copying the selected text (www.w3.org) and provides states for recovering an intended prior representation of a changing resource (www.w3.org). W3C PROV also distinguishes generation and invalidation times, including when an entity became available for use (www.w3.org and when it ceased to be usable (www.w3.org).

The practical record needs both **observation time** and **validity time**. Observation time states when the evaluator saw the source. Validity time states when the source's claim is intended to hold, if known. A retrieval snapshot should additionally preserve corpus version and query time. The CRAG benchmark pairs questions with retrieval results captured at the same query time (arxiv.org). Currency is therefore a testable input, not a prose caveat.

The Minimum Test Record

The minimum record should allow a reviewer to replay the evidence path without guessing what changed. *Table 1* defines the required fields and the boundary each one protects.

Record component	Required fields	Fault isolated	Retention rule
Case identity	Stable case ID, exact query, task type, expected behavior, evaluation-set version	Test-set drift or mismatched before/after cases	Never reuse an ID for a changed case
Permitted source set	Corpus ID, snapshot/version, document IDs, access policy, observation date, validity window, checksum where available	Evidence absent because it was unavailable, impermissible, stale, or misidentified	Retain the manifest and the referenced source state
Expected evidence	Atomic expected claims, acceptable sources, relevance grades, contradiction labels, answerability label	Ambiguous gold answer or incomplete relevance judgments	Version labels separately from retrieved output
Retrieval output	Rank, chunk ID, parent document ID, text, score, retriever configuration, cutoff	Candidate generation and ranking faults	Preserve the complete top-k list, not only displayed context
Assembled context	Exact ordered text sent to the model, truncation and deduplication outcomes	Context packing, ordering, or omission faults	Store the rendered context as a run artifact
Generation output	Exact answer, actual model, prompt-template version, decoding parameters, run ID	Model-use and synthesis faults	Keep raw output before post-processing
Claim and citation map	Extracted claim IDs, supporting or contradicting spans, citation targets, entailment verdicts	Unsupported synthesis, wrong citation, or uncited correct claim	Preserve extractor and judge outputs, not only their scores
Decision record	Answer, abstain, clarify, blocked, not found, or not applicable, with reason	Policy or abstention faults	Treat unknown states as explicit values

Implementation records should capture more than a seed. MLflow Tracking is designed to log parameters, code versions, metrics, and output files (mlflow.org), while its version tracking captures Git commit, branch, and repository URL (mlflow.org). OpenTelemetry's generative-AI conventions include the prompt-template version (github.com) and distinguish the model that actually generated the response (github.com). TensorFlow warns that a seeded random sequence can change across software versions (www.tensorflow.org). PyTorch likewise states that deterministic algorithms alone do not guarantee reproducibility (docs.pytorch.org). A reproducible case therefore records software and model versions alongside randomness controls.

Fault-Isolation Matrix

Diagnosis should move from observation to the next controlled experiment. *Table 2* is a compact decision matrix for the most common RAG failure modes.

Observed symptom	Evidence check	Likely boundary	Next controlled experiment
Missing evidence	Required claim is absent from every retrieved chunk	Source set, query, candidate generation, or ranking	Inject the gold passage. If the answer recovers, test corpus membership, query rewriting, and rank cutoff separately
Noisy evidence	Required claim is present, but irrelevant chunks dominate the context budget	Ranking or assembly	Hold gold evidence fixed, remove distractors, then add them back at controlled ratios
Stale evidence	Claim is supported by the retrieved snapshot but conflicts with a newer valid source	Source governance or snapshot selection	Replace only the dated source version and rerun the same case ID
Unsupported synthesis	Response claim is absent from and not entailed by assembled context	Generation	Force gold context with temperature and prompt fixed; test whether the unsupported claim persists
Contradicted claim	A retrieved span entails the negation or an incompatible value	Conflict handling, source selection, or generation	Insert one labeled contradiction while holding all other chunks and their order constant
Correct but uncited	Claim matches the reference, but no citation resolves to its supporting span	Citation mapping or rendering	Hold answer text fixed and evaluate only claim-to-span assignment
Cited but wrong	Citation entails the answer, but the source is false, stale, or about another entity	Source identity, quality, or currency	Swap to an authoritative current source without changing the question
Unjustified answer	No acceptable support exists, but the system answers	Decision or abstention	Withhold all gold evidence and score answer rate plus planted-error adoption
Over-abstention	Acceptable support is present, but the system abstains	Decision threshold or context use	Supply only concise gold evidence, then reintroduce ambiguity one factor at a time
Unstable verdict	Repeated runs produce materially different judge labels	Claim extractor, judge, prompt, or model version	Freeze extracted claims, repeat judges, counterbalance order, and compare disagreement

The matrix deliberately gives **one intervention per suspected edge**. A corpus change plus a prompt change cannot identify which change caused recovery. Likewise, a higher overall answer score after increasing top-k does not show whether improved retrieval coverage or increased context noise drove the result.

The noisy-evidence row has direct experimental precedent. RGB uses an all-negative condition to test rejection (github.com). GRAB-RAG goes further by holding questions constant across supportive, degraded, missing, and misleading contexts (arxiv.org). These are causal probes of system behavior, not merely another leaderboard score.

Metrics by Boundary

Retrieval measures require explicit labels

The retrieval boundary can be measured deterministically only when the denominator is defined. A classical test collection contains a document collection, queries representing information needs, and query-document relevance judgments (nlp.stanford.edu). Relevance should mean that a document addresses the information need, not merely that it shares query terms (nlp.stanford.edu).

Use the measures according to the available labels:

- **Precision@k = relevant retrieved in top k / k.** Use it when every returned item can be judged and context capacity is scarce. If three of five chunks are relevant, precision@5 is **3/5 = 0.60** (docs.aws.amazon.com).
- **Recall@k = relevant retrieved in top k / all labeled relevant.** Use it only when the relevant-set denominator is sufficiently complete. If three of six labeled chunks appear in the first five ranks, recall@5 is **3/6 = 0.50** (ir-measur.es).
- **Mean reciprocal rank, or MRR.** Use it when the first relevant result is the operational objective. It gives no credit for additional correct results (trec.nist.gov), so it is insufficient for multi-claim questions.
- **Normalized discounted cumulative gain at k, or nDCG@k.** Use it for graded relevance and ranked utility. It compares the observed ranking with the best possible ranking (docs.opensearch.org) and places greater value on useful material near the top. The cutoff must be reported because an implementation can consider only the highest k scores (scikit-learn.org).
- **Judgment coverage@k.** Report the fraction of returned items that have relevance scores (docs.opensearch.org). This stops an apparent precision gain from hiding an expanding set of unjudged chunks.

Incomplete judgments matter. Traditional pooling combines top results from several runs (trec.nist.gov), and commonly treats unjudged documents as nonrelevant (trec.nist.gov). That assumption must be declared. Otherwise, recall@k and nDCG can appear precise while measuring an incomplete label set.

Claim-level generator measures

After retrieval, score claims rather than whole responses:

- **Claim precision = supported correct response claims / all response claims.** RAGChecker defines response precision as the proportion of correct claims among response claims (arxiv.org).
- **Claim recall = expected claims recovered / all expected claims.** RAGChecker defines response recall relative to ground-truth answer claims (arxiv.org).
- **Faithfulness = response claims entailed by provided context / response claims.** Current RAGAS documentation uses a prompt that breaks a response into atomic statements (docs.ragas.io).

- **Citation completeness = claims requiring support that have adequate cited support / all claims requiring support.** Keep this separate from citation correctness, which asks whether each cited span actually entails its claim.
- **Correctness = response claims agreeing with a reference or authoritative current answer / response claims.** Do not infer this from faithfulness.

Suppose a response contains **10 atomic claims**, eight are supported by context, seven are correct, and six correspond to six of nine expected claims. Faithfulness is **8/10 = 0.80**, claim precision is **7/10 = 0.70**, and claim recall is **6/9 = 0.67**. Those raw counts are more diagnostic than their harmonic mean: two claims need grounding review, three expected claims are missing, and at least one context-supported claim is still wrong. The calculation is a reporting example, not an external benchmark.

Controlled Interventions and Decision Rules

The strongest diagnosis comes from counterfactual tests that change one evidence condition while keeping the case ID fixed.

- **Force gold evidence.** Replace retrieval output with the minimal authoritative passages. Recovery implicates source-set or retrieval boundaries. Continued failure implicates assembly, generation, or the expected-claim specification.
- **Remove one required passage.** This tests whether the answer is evidence-sensitive. A model that reproduces the missing claim may be relying on parametric memory or unsupported synthesis.
- **Add distractors.** Increase irrelevant chunks in known steps while keeping gold passages fixed. Measure faithfulness, correctness, and lost-in-context behavior separately.
- **Insert a contradiction.** Add one clearly labeled incompatible passage. Evaluate whether the answer acknowledges conflict, selects by a declared rule, or silently adopts one source.
- **Shift a date or number.** Change only a year, numeric value, or validity boundary while holding the surrounding passage constant.
- **Withhold all support.** Supply only hard negatives and measure unjustified answer rate. RGB's all-noise tests reported best rejection rates of only **45% in English and 43.33% in Chinese** among the evaluated models (arxiv.org).
- **Hold answer text fixed.** Re-evaluate only citations to isolate citation mapping from generation.
- **Hold extracted claims fixed.** Repeat judges on the same claim set to isolate judge variance from claim-extraction variance.

Abstention needs two error rates. **Hallucination when the system should abstain** measures answers on unsupported cases, while **unnecessary abstention** measures abstentions on answerable cases. GRAB-RAG directly defines the first direction (arxiv.org). A clarification request may be correct for an underspecified query even though it is not a direct answer.

Implementation Considerations and Process Changes

An evaluation owner should turn the protocol into a controlled change process:

1. **Freeze the case set.** Assign stable IDs and preserve the exact before-state inputs.
2. **State the hypothesis.** Name one suspected boundary and the observable change expected if it is causal.
3. **Choose the intervention.** Change one corpus, retrieval, assembly, generation, citation, or decision variable.
4. **Run paired cases.** Score old and new configurations on identical case IDs. Paired comparison assumes both systems process the same test set (aclanthology.org).
5. **Retain raw counts.** Store per-case retrieved evidence, claim labels, citation mappings, and decisions, not just averages.
6. **Quantify uncertainty.** Bootstrap aligned before/after pairs, using the same resampled indices for both arrays (docs.scipy.org). Preserve the random generator for repeatability (docs.scipy.org).
7. **Review regressions by boundary.** A retrieval gain can coincide with a faithfulness loss if added context is noisy. Treat that as a tradeoff, not a net win hidden inside one composite.
8. **Promote only reconstructable results.** Retain the source snapshot, configuration, claim set, judge output, and paired-case report.

NIST describes an AI system inventory as an organized database of artifacts (airc.nist.gov) and recommends reviewing third-party release schedules and change-management plans (airc.nist.gov). Both practices belong in the run record because dependency changes can alter results without a deliberate RAG configuration change. DVC describes an experiment-versioning approach that keeps metrics, hyperparameters, and artifact information in versionable text files (dvc.org).

Automatic judging is itself a boundary. LLM judges can exhibit position, verbosity, and self-enhancement biases (arxiv.org). Repeated judgments can also have low intra-rater reliability (aclanthology.org). Decomposed binary checklist questions improved average cross-model agreement by **0.45** in one study of **12 evaluator models** (aclanthology.org). This supports narrow rubrics and replicated labels, but it does not make a judge deterministic.

Claim extraction also changes scores. Factual-support measurement has been shown to be sensitive to the decomposition method (aclanthology.org). Preserve the extracted claims, their boundaries, and extractor version. If a sentence contains two entities and three predicates, different atomic decompositions can change both numerator and denominator without any change in the answer.

Data Analysis and Evidence

The quantitative core should report raw case transitions, not only a macro score. *Table 3* maps retrieval coverage and answer support into the first diagnostic decision.

Retrieved evidence	Answer supported by shown context	Interpretation	Next measurement
Yes	**Yes**	Retrieval and grounding succeeded; correctness remains unproven	Check authoritative reference, entity identity, source currency, and citation completeness
Yes	**No**	Evidence reached the generator, but one or more claims exceeded it	Claim-level faithfulness, context order, distractor sensitivity, and generator prompt
No	**Yes**	Usually a measurement mismatch, hidden evidence path, or support from text not recorded in the retrieval snapshot	Audit context assembly and instrumentation before tuning retrieval
No	**No**	Earliest visible fault is corpus or retrieval, though downstream behavior may also fail	Force gold evidence, then repeat the claim and abstention checks

The upper-left cell is not a declaration of truth. Source verification is necessary but insufficient for broader quality. A good citation metric can therefore coexist with an inadequate answer metric.

Published datasets quantify why boundary tests matter:

- **RAGAS WikiEval.** It used **50 Wikipedia pages** about events since the start of 2022 (arxiv.org). Two annotators agreed in about **95%** of faithfulness and context-relevance cases (arxiv.org). These results describe a small synthetic construction, not universal judge validity.
- **RAGTruth.** The official repository reports **17,790 responses**, including **7,664** with annotated hallucination spans (github.com). Its research release contains annotations at both case and word levels (aclanthology.org), which supports analyses that whole-answer labels obscure.
- **ConfRAG.** Its **1,814 questions** average **9.58 retrieved paragraphs**, and **57.2%** contain explicit contradictions (aclanthology.org). Conflict handling is therefore empirically separable from basic relevance.
- **FreshQA.** Its evaluation included more than **50,000 human judgments** across questions that include fast-changing knowledge and false premises (arxiv.org).

For a before/after report, publish the number of cases improved, unchanged, and regressed for every boundary measure. Report mean paired change and a bootstrap confidence interval. A **95% confidence interval** is a property of repeated use of the procedure, under which about 95% of resulting intervals would cover the population parameter (www.itl.nist.gov). It is not a 95% probability that this realized interval contains the truth.

Implications and Future Directions

RAG evaluation is moving from answer scoring toward **evidence accounting**. Three implications follow.

First, teams need **test-record governance**. Google Research's Data Cards template asks maintainers to state dataset versioning criteria (sites.research.google). Hugging Face model-card metadata can link the source from which evaluation results were computed (huggingface.co). These practices should extend to corpus snapshots, retrieved chunks, and citation spans.

MLflow evaluation records expose identifiers, timestamps, actors, and source metadata to track lineage and version history (mlflow.org). OpenTelemetry evaluation events use a response identifier to connect a verdict to the specific completion evaluated (github.com). These identifiers should be joined to the stable case ID, not substituted for it.

Second, **automatic evaluation needs its own validation set**. Human review should sample disagreements, high-impact cases, and boundary conditions. One abstention study found **92.3%** average pairwise agreement among three LLM judges, but only **84%** agreement between the default judge and a human on 100 sampled unanswerable responses (aclanthology.org). Agreement among automated judges is not the same as alignment with the intended rubric.

Third, future benchmarks should connect **source validity** to **claim validity**. W3C PROV recommends tracing secondary to primary sources so reliability can be investigated (www.w3.org). Current metrics are stronger at checking text-to-text entailment than deciding whether a source is current, authoritative, about the correct entity, or valid for the requested time. That gap is precisely where grounded-but-wrong answers survive.

Frequently Asked Questions (FAQs)

What is the fastest way to diagnose a failed RAG answer?

Start with three checks: was the required evidence in the permitted corpus, did it appear in the retrieved top-k and assembled context, and does each answer claim have an entailing span? Then verify source identity, currency, and correctness. The earliest failed boundary determines the next controlled experiment.

Which metrics validate RAG answers?

No single metric validates an answer. Use precision@k, recall@k, or nDCG@k for labeled retrieval; claim precision and recall for response content; faithfulness for support by shown context; correctness against an external reference; citation completeness and correctness for traceability; and separate error rates for unjustified answers and unnecessary abstentions. Microsoft likewise distinguishes labeled retrieval evaluation requiring qrels from LLM-judged retrieval evaluation when ground truth is unavailable (learn.microsoft.com).

Is a faithful or grounded answer necessarily true?

No. Faithfulness asks whether context entails the answer. It does not establish that the context is true, current, authoritative, or about the correct entity. Source-quality and correctness checks are separate.

How should RAG evaluation owners compare a proposed fix?

Run the old and new configurations on identical case IDs, retain per-case outputs, and report paired changes with a confidence interval. Do not compare averages from different test sets. AWS documentation similarly cautions against comparing solution versions trained with different data (docs.aws.amazon.com).

What should be retained for RAG answer traceability?

Retain the query, case ID, corpus snapshot, document and chunk identifiers, observation and validity dates, relevance labels, full ranked retrieval, assembled context, response, claim decomposition, citation spans, source-quality review, judge configuration, model and prompt versions, decision, and before/after change record. Empty evidence should be represented as a reasoned unknown, not silently omitted.

Conclusion

RAG fault isolation is successful when a failed answer can be assigned to an observable boundary and followed by a controlled next experiment. The essential sequence is source set, retrieval, context assembly, generation, citation, and decision. Each transition needs a retained artifact, and each metric must be interpreted only against the labels it actually has.

The central rule is simple: **retrieved does not mean grounded, grounded does not mean true, and true does not mean traceable**. Retrieval metrics diagnose ranked evidence only. Claim-level faithfulness diagnoses support by the shown context. Correctness requires an independent reference or current authoritative evidence. Citation metrics test whether readers can follow the claim back to that evidence. Source identity and time determine whether the evidence is the right evidence at all.

The practical workflow is equally direct. Freeze the case IDs, preserve the full test record, state one boundary hypothesis, change one variable, and compare paired results with raw counts and uncertainty. Force gold evidence, remove it, add noise, plant a contradiction, shift a date, and test abstention without support. Preserve both the successful and failed cases. An aggregate score can summarize a run, but only a reconstructable evidence record can defend why the next change was chosen.

Source links

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

1. https://arxiv.org/html/2408.08067
2. https://www.w3.org/TR/prov-o
3. https://airc.nist.gov/airmf-resources/playbook/measure
4. https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-unranked-retrieval-sets-1.html
5. https://ir-measur.es/en/latest/measures.html
6. https://arxiv.org/html/2608.22228
7. https://aclanthology.org/2024.eacl-demo.16
8. https://github.com/amazon-science/RAGChecker/blob/main/tutorial/ragchecker_tutorial_en.md
9. https://aclanthology.org/2023.cl-4.2
10. https://aclanthology.org/2023.cl-4.2.pdf
11. https://aclanthology.org/2023.emnlp-main.398
12. https://aclanthology.org/2023.emnlp-main.398.pdf
13. https://www.w3.org/TR/annotation-model

14. https://arxiv.org/html/2406.04744
15. https://mlflow.org/docs/latest/ml/tracking
16. https://mlflow.org/docs/latest/genai/version-tracking/track-application-versions-with-mlflow
17. https://github.com/open-telemetry/semantic-conventions-genai/blob/main/docs/gen-ai/gen-ai-events.md
18. https://www.tensorflow.org/api_docs/python/tf/random/set_seed
19. https://docs.pytorch.org/docs/main/generated/torch.use_deterministic_algorithms.html
20. https://github.com/chen700564/RGB/blob/master/readme.md
21. https://nlp.stanford.edu/IR-book/html/htmledition/information-retrieval-system-evaluation-1.html
22. https://docs.aws.amazon.com/personalize/latest/dg/working-with-training-metrics.html
23. https://trec.nist.gov/presentations/TREC8/removed/tsld051.htm
24. https://docs.opensearch.org/latest/search-plugins/search-relevance/evaluate-search-quality
25. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.ndcg_score.html
26. https://trec.nist.gov/pubs/trec33/papers/overview_33.pdf
27. https://docs.ragas.io/en/stable/howtos/customizations/metrics/modifying-prompts-metrics
28. https://arxiv.org/html/2309.01431
29. https://aclanthology.org/www.mt-archive.info/EMNLP-2004-Koehn.pdf
30. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bootstrap.html
31. https://airc.nist.gov/airmf-resources/playbook/govern
32. https://airc.nist.gov/airmf-resources/playbook/map
33. https://dvc.org/blog/ml-experiment-versioning
34. https://arxiv.org/abs/2306.05685
35. https://aclanthology.org/2025.findings-emnlp.1361
36. https://aclanthology.org/2025.emnlp-main.796
37. https://aclanthology.org/2024.starsem-1.13
38. https://arxiv.org/html/2309.15217
39. https://github.com/ParticleMedia/RAGTruth
40. https://aclanthology.org/2024.acl-long.585
41. https://aclanthology.org/2026.acl-long.11
42. https://arxiv.org/abs/2310.03214
43. https://www.itl.nist.gov/div898/handbook/prc/section1/prc14.htm
44. https://sites.research.google/datacardsplaybook
45. https://huggingface.co/docs/hub/model-cards
46. https://mlflow.org/docs/latest/genai/datasets/sdk-guide
47. https://aclanthology.org/2026.eacl-long.361.pdf
48. https://learn.microsoft.com/en-us/azure/foundry/concepts/evaluation-evaluators/rag-evaluators

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.