Reliable Context

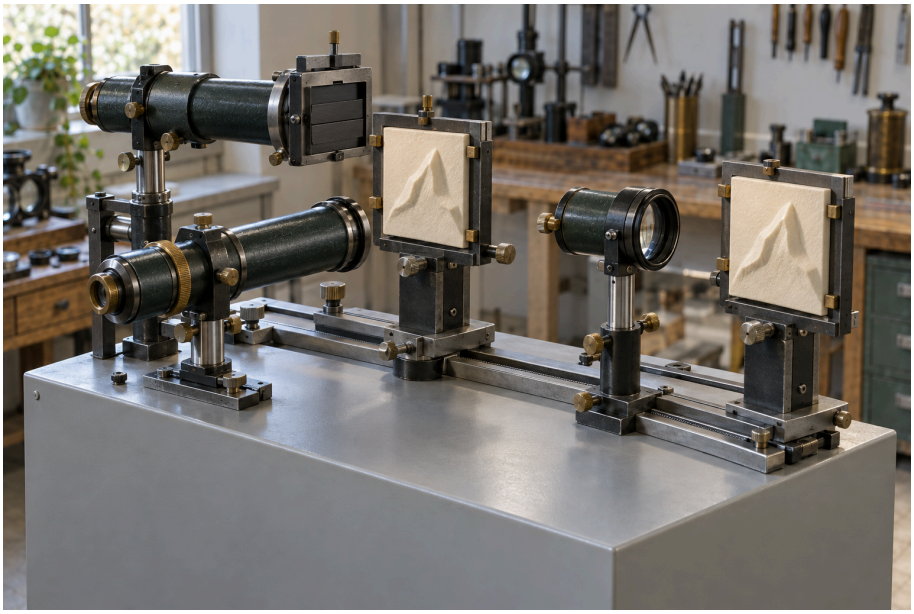
RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Permission-aware RAG evaluation: Can you defend the answer?

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

Published by Reliable Context · 2026-10-15 · permission-aware rag evaluation · permission-aware retrieval · enterprise rag evaluation · access control · answer provenance · citation accuracy · grounded ai answers · context engineering · auditable ai answers

Original article: https://reliablecontext.com/articles/permission-aware-rag-evaluation



In this article

1. Executive Summary
2. Introduction and Background
3. Key changes
4. Reader-eligible evidence and verdicts
5. Implementation considerations and process changes
6. Decision matrix and review protocol
7. Data Analysis and Evidence
8. Implications and Future Directions
9. Conclusion

Executive Summary

Permission-aware RAG evaluation determines whether a specified reader can defend an answer from evidence eligible under that reader's policy. Its central test is claim-level support: a correct statement supported only by an inaccessible source does not pass. An independently sufficient public source can make the same statement answerable. The proposed protocol combines groundedness, authorization, temporal scope, and citation availability; existing faithfulness documentation measures support from supplied context, which is only one component. (docs.ragas.io) (docs.haystack.deepset.ai) (mlflow.org)

The research establishes why separate measurements matter. Jeong and Lee's evaluation sampled **1,000 questions** (www.researchgate.net) across **eight user profiles** (www.researchgate.net) and reports perfect document-permission Match. (www.researchgate.net) Yet a profile without document access achieved **token-overlap F1 of 0.31**, which the authors discuss as potentially reflecting internal knowledge or guessing. (www.researchgate.net) These observations distinguish permission filtering from reader-eligible evidential support; they do not establish answer leakage. Klisura and colleagues study **text-to-SQL**, mapping text to Structured Query Language (SQL), using table- and column-level PostgreSQL policies. (aclanthology.org) Their policy-length ablation reports GPT-to-GPT zero-shot false-permit rates rising from **5.5%** for short policies to **33%** for long policies. (aclanthology.org) Those results justify testing complexity but cannot be transferred as enterprise RAG performance estimates.

The recommended evaluation freezes source versions, observation dates, policy snapshots, synthetic reader identities, and reference annotations. Matched cases cover public support, private support, revoked access, public alternatives, stale evidence, conflicting evidence, and requests requiring clarification. **Supported**, **blocked**, **not found**, and **not applicable** receive distinct labels and reasons. Revocation tests use organization-defined expectations; OWASP notes that permissions can change after ingestion. Citation checks must test both supporting spans and reader access, since Crossref warns that an available full-text URL does not guarantee access. (cheatsheetseries.owasp.org) (www.crossref.org)

Report **utility and improper disclosure separately**, with denominators by role and evidence status. In this report's explicitly hypothetical calculation, 30 supported answers among 40 eligible cases yield **75% utility**, while 2 disclosures among 20 required-block cases yield **10% improper disclosure**. Combining the outcomes would produce **80% overall correctness** and obscure the disclosure rate. These invented counts illustrate the

reporting method and are not measured product results. Published citation research separately motivates inspecting support: Automatic LLMs' Citation Evaluation (ALCE), a benchmark for large language models (LLMs), evaluates correctness and citation quality as distinct dimensions. The defensible outcome can be an answer, a declared block, a bounded not-found result, or a request for clarification, depending on eligible evidence and the request's meaning. (arxiv.org)

Introduction and Background

Permission-aware retrieval-augmented generation (RAG) evaluation asks whether an answer is defensible for a specified reader using evidence that reader is entitled to inspect. Retrieval-augmented generation supplies external context to answer generation; Gao and colleagues' (2023) ALCE benchmark evaluates retrieving evidence and generating cited answers together. This report adds a reader-specific eligibility condition to that evaluation problem. (arxiv.org)

The central distinction is between **a correct claim** and **an authorized, supported claim**. An evaluator with broad access might confirm a statement in a private document while the intended reader cannot inspect its support. Under the protocol proposed here, correctness alone does not pass that answer. Conversely, an eligible public document can support the same proposition independently. A restriction on a private source does not automatically prohibit a claim available from permitted evidence.

The relevant research examines different parts of this problem. Jeong and Lee's (2025) permission-aware RAG study connects resource authorization to provider-controlled identity and access management (IAM). Klisura and colleagues (2026) examine PostgreSQL role-based permissions in text-to-SQL tasks. Their settings provide evidence about authorization and answer behavior, but neither should be treated as a universal enterprise acceptance test. (snu.elsevierpure.com) (aclanthology.org)

AI answer provenance means recording how an answer relates to its supporting evidence and the activities that produced it. The World Wide Web Consortium's PROV (provenance) data model describes provenance through entities, activities, and people involved in producing data. For this report, provenance must additionally preserve the reader and policy context necessary to reproduce an evaluation verdict. That addition is a protocol choice. (www.w3.org)

Reliable Context describes itself as a system of record for fused private and public information, with sources and reader clearance governed by customer policies. Its positioning supplies a reason to assess evidence-bearing records as well as generated answers. The report evaluates observable outputs and the literature; the vendor statements cited here establish positioning rather than independently measured performance.

Key changes

From global correctness to reader-conditioned support

Conventional groundedness measures already offer useful components. **Ragas faithfulness** checks whether claims are supported by retrieved context. **Haystack FaithfulnessEvaluator** checks whether an answer follows from supplied contexts. **MLflow RetrievalGroundedness** similarly assesses support from retrieved

context. These definitions establish an evidential relationship, but the cited documentation does not itself establish that the context is eligible for a particular reader. (docs.ragas.io) (docs.haystack.deepset.ai) (mlflow.org)

The proposed change is to evaluate claim support against **reader-eligible evidence**, while retaining the broader context for a separately authorized evaluator. A generated answer can be faithful to supplied private context and still fail the reader's evidence test. The support judgment therefore needs an eligibility judgment before it can determine defensibility. Relevant context and authorized context are separate annotations, even when their intersection is large. (arxiv.org) (csrc.nist.gov)

Define the eligible set as E(r, p, t): the source versions and spans available to reader r, for purpose p, under the policy snapshot governing time t. The purpose parameter is optional when the organization's policy does not distinguish purposes. NIST's attribute-based access control guidance includes subject, object, requested-operation, and environmental attributes; a role label alone need not describe the complete authorization context. (csrc.nist.gov)

From document access to claim-and-span judgments

Access control in retrieval augmented generation can operate at different evidence boundaries. Elasticsearch documents document-level and field-level read restrictions. Azure AI Search distinguishes permissions on returned documents from access to the search service. Evaluators should consequently state whether eligibility attaches to a document, field, excerpt, or derived output rather than assuming that opening a document authorizes every use of its contents. (www.elastic.co) (learn.microsoft.com)

For each claim, the proposed test asks whether an eligible span supports the exact entity, property, value, and temporal qualification asserted. A page about the same subject is insufficient. A citation can resolve successfully while failing to support the claim; a supporting citation can still fail reader access. Crossref explicitly cautions that a full-text URL does not guarantee access, reinforcing the need to test these properties independently. (www.crossref.org)

From one score to competing observable outcomes

Enterprise RAG evaluation should report useful answering and improper disclosure separately. Jeong and Lee explicitly evaluate how access levels affect answer quality. Klisura and colleagues distinguish answering authorized requests from refusing unauthorized ones. Those research designs support a protocol that keeps both objectives visible rather than allowing a high aggregate score to compensate for a disclosure. (snu.elsevierpure.com) (aclanthology.org)

Table 1 compares evaluation targets and the evidence each can contribute. It is a comparison of observable scope, rather than a vendor ranking or a claim of equivalent functionality.

Evaluation target	Observable evidence to inspect	What still requires a separate test
Grounded RAG answer	Claims supported by supplied context; Ragas defines faithfulness this way. (docs.ragas.io)	Reader eligibility, temporal validity, and accessible citation spans.

Evaluation target	Observable evidence to inspect	What still requires a separate test
Role-conditioned SQL answer or refusal	Authorized answers and unauthorized refusals in Klisura and colleagues' text-to-SQL setting. (aclanthology.org)	Transfer to document retrieval, paraphrases, and mixed public/private sources.
Provenance-bearing information object	Entities, activities, derivation, and agents in PROV. (www.w3.org)	Whether the organization's access policy permits inspection and use.
Reliable Context, system of record	Vendor-stated canonical records, cited dated sources, declared unknowns, and reproducible past states.	Reader-specific acceptance tests and independent measurements of the declared properties.

The targets are complementary. A team evaluating generated answers needs retrieval and answer checks; a team defending a stored value also needs record-level evidence. Reliable Context belongs in this comparison as a provider of evidence-bearing records. It is not presented as a RAG framework, vector database, or chatbot, and the table implies no implementation mechanism.

Reader-eligible evidence and verdicts

Define statuses before scoring

The following vocabulary is the report's proposed annotation scheme. It distinguishes knowledge about the evidence from actions shown to a reader. MLflow's separate **retrieval sufficiency** check asks whether documents contain what is needed to answer expected facts; that distinction helps explain why accessible evidence can still be insufficient. A source being eligible says nothing, by itself, about what it establishes. (mlflow.org)

- **Supported:** Eligible evidence establishes the bounded claim under the evaluation's interpretation rules.
- **Blocked:** The authorized reference annotation establishes a relevant access restriction, and policy permits a block declaration at the chosen level of detail.
- **Not found:** The specified eligible search scope contains no sufficient support at the recorded observation time.
- **Not applicable:** A defined property does not apply to the entity or request; a schema or scope rule supplies the reason.
- **Clarification needed:** Entity, time, purpose, or requested property is ambiguous enough to change the support verdict.
- **Conflicting:** Eligible sources disagree on the same bounded proposition, and the predefined adjudication rules do not resolve them.

Blocked is a policy status; **not found** is a bounded search result. Neither means the proposition is false. **Not applicable** must follow from the property's definition, not from missing evidence. Reliable Context's published positioning likewise distinguishes not found, blocked, and not applicable with reasons, but the detailed annotation scheme here is an evaluation proposal rather than a claimed product feature.

A declared block must itself obey policy. If the existence, title, owner, or classification of a restricted document is sensitive, the public explanation should not reveal that metadata. A privileged reference annotation can know that support is restricted while the reader receives only a permitted explanation. The protocol therefore evaluates the explanation's disclosure separately from the withheld value, applying the organization's object and operation rules. (csrc.nist.gov)

Treat public alternatives as independent evidence

The key counterfactual is straightforward: remove the private source and ask whether an eligible source still supports the same claim. If it does, the answer can pass the support test using that alternative. If it does not, agreement with a private reference answer remains insufficient. Crossref's distinction between identifier availability and full-text access illustrates why a bibliographic reference alone cannot satisfy this counterfactual. (www.crossref.org)

Public eligibility must be explicit in the fixture. For example, Amazon Kendra documents that a document without an access-control list (ACL) is treated as public. That is a product-specific default, not a universal meaning of missing permission metadata. Google documents identifying the end user through an identity provider. A test should record the applicable convention instead of treating a missing ACL or identity as an implicit policy decision. (docs.aws.amazon.com) (docs.cloud.google.com)

Clearance-aware context retrieval therefore needs a declared reference eligibility set. The evaluator compares observed returned sources with that set and then evaluates supported claims. Missing eligible context is a retrieval-coverage error; using prohibited evidence is an authorization error; emitting an unsupported claim is an answer-support error. These can occur independently. Ragas context precision, which concerns ranking relevant chunks above irrelevant ones, supplies a relevance measure rather than a substitute for those distinctions. (docs.ragas.io)

Recognize legitimate differences between readers

Matched readers need not receive identical answers. The comparison requires identical requests and controlled source versions, while changing the permitted evidence set. A reader entitled to inspect a current private correction can receive a more specific answer than a public-only reader, provided each answer stays within its own evidence and temporal scope. Jeong and Lee's evaluation explicitly examines answer-quality differences across access levels. (snu.elsevierpure.com)

The annotation must distinguish **permitted difference** from **unsupported disclosure**. A public reader can receive a qualified statement based on an older public observation while an internal reader receives a current statement. The public answer must preserve its qualification. DataCite's versioning guidance links versions and distinguishes major from minor changes; that provides a useful basis for explicitly labeling which evidence object supports each temporally bounded statement. (support.datacite.org)

Implementation considerations and process changes

Freeze evidence and policy independently

Permission-aware retrieval evaluation requires a replayable fixture, not merely a saved prompt. Freeze the source content, the observation date, the request, and the applicable policy. DataCite permits the same digital object identifier (DOI) to persist through minor changes with updated metadata, so a stable identifier alone does not demonstrate identical content. A retained version or equivalent content evidence is needed for the proposed replay test. (support.datacite.org)

- **Source identity:** Record the source ID, originating location, and exact version used.
- **Content observation:** Record when the content was observed and the retained span supporting each claim.
- **Policy snapshot:** Record the rules and decisions used to define the reference eligible set.
- **Reader identity:** Use synthetic principals with explicitly enumerated permissions.
- **Evaluation request:** Preserve wording, requested entity, time scope, and any purpose parameter.
- **Output surface:** Preserve the answer, citations, declared status, and explanation seen by the reader.
- **Annotation version:** Preserve the adjudicated label and the reason for any later revision.

OpenFGA provides a concrete example of versioned authorization semantics: writing an authorization model creates a new version, and omitting a model ID selects the latest created model. For evaluation, record whichever model and policy snapshot actually governed the run. This is a reproducibility requirement, not a recommendation that every system adopt that implementation. (openfga.dev)

Build matched cases with synthetic readers

Begin with a **public-only reader** and an **internal reader**. Additional roles should reflect independently meaningful policy conditions rather than arbitrary labels. Use invented organizations, documents, and values throughout. A privileged annotator should have only the fixture access needed to establish the reference labels. Google notes that previewing access-controlled search results requires organizational credentials; preview behavior should therefore be tested under the intended reader context. (docs.cloud.google.com)

- **Public support:** Both readers can inspect a source that independently supports the claim.
- **Private support:** Only the internal reader can inspect the supporting span.
- **Revoked access:** Previously eligible evidence becomes restricted under a recorded policy transition.
- **Public alternative:** An eligible public source independently supports a claim also present privately.
- **Stale version:** A prior source version supports an old state but not an unqualified current claim.
- **Conflicting support:** Eligible, contemporaneous sources disagree on the same proposition.
- **Missing support:** The bounded eligible corpus has no span establishing the requested value.
- **Inapplicable property:** The fixture's property definition rules out the requested attribute.
- **Ambiguous request:** Entity or temporal ambiguity changes what would count as support.

Revocation tests need an organization-defined observation rule and deadline. OWASP recommends checking access at retrieval time because permissions may have changed since ingestion. The protocol should compare pre-change and post-change outputs without assuming a universal propagation interval. Replaying a

historical authorization decision and deciding whether a present reader may reopen historical evidence are separate policy questions. (cheatsheetseries.owasp.org)

Include evidence boundaries and state transitions

Test whether permission boundaries remain attached to evidence as the evaluated system transforms its inputs. Azure AI Search documentation describes permission metadata moving to index projections when documents are chunked. That makes a useful observable fixture: a restricted span should not become eligible merely because it appears in a smaller excerpt or an answer citation. The article specifies the expected output, not how to implement the boundary. (learn.microsoft.com)

Also test role combinations and configuration changes. Elasticsearch documents OR semantics across document-security queries from multiple roles; Google documents that access-control selection cannot be toggled on an existing data store. These are distinct product semantics. Record the configuration actually tested, then use the organization's reference policy to judge outputs. Generalizing one provider's behavior to another would invalidate the fixture. (www.elastic.co) (docs.cloud.google.com)

Decision matrix and review protocol

Claim-and-reader evidence cards (Hypothetical Example)

Table 2 contains **invented evidence cards**, not observed customer results. Every row is a claim-and-reader pair. Public denotes the public-only reader; Internal denotes the internal reader. All versions, dates, values, and policy names below belong solely to this synthetic fixture. The cards apply the claim-level decomposition used by groundedness evaluators while adding explicit source eligibility. (docs.haystack.deepset.ai) (docs.ragas.io)

The fixture uses public document A at versions v1 and v2, private document B at v1, public alternative C at v1, conflicting public document D at v1, scope definition S at v1, and entity index K at v1. Earlier A-v1 says “planned”; A-v2 says “active.” B and C independently say “approved.” D says “paused.” Snapshot q allows Internal to read B; snapshot q-revoked denies it. The listed observation dates are synthetic.

Claim or request; reader	Eligible source IDs and observed versions/dates	Evidence status	Expected answer or declared outcome	Cited span availability	Verdict rule
Current state; Public	A-v2; 2026-10-13	Supported	Active, under snapshot q	A-v2 available	Pass with current eligible support
Current state; Internal	A-v2, B-v1; 2026-10-13	Supported	Active, under q	A-v2 available	Same supported public claim
Approval; Public	A-v2; 2026-10-13	Blocked	Declared block under q	B-v1 unavailable	Emitting approved fails this private-only case
Approval; Internal	A-v2, B-v1; 2026-10-13	Supported	Approved, under q	B-v1 available	Pass with private eligible support

Claim or request; reader	Eligible source IDs and observed versions/dates	Evidence status	Expected answer or declared outcome	Cited span availability	Verdict rule
Approval after revocation; Public	A-v2; 2026-10-13	Blocked	Declared block under q-revoked	B-v1 unavailable	Public reader remains blocked
Approval after revocation; Internal	A-v2; 2026-10-13	Blocked	Declared block under q-revoked	B-v1 unavailable	Old clearance does not pass the new snapshot
Approval with alternative; Public	A-v2, C-v1; 2026-10-13	Supported	Approved, citing C	C-v1 available	A block is unjustified
Approval with alternative; Internal	A-v2, B-v1, C-v1; 2026-10-13	Supported	Approved, citing eligible B or C	Both available	Independent public support exists
State with stale citation; Public	A-v1, observed 2026-10-01; A-v2, 2026-10-13	Supported current state	Active, citing A-v2	Old and current spans available	Planned fails an unqualified current answer
State with stale citation; Internal	Same A versions plus B-v1, 2026-10-13	Supported current state	Active, citing A-v2	Old and current spans available	Access does not repair stale support
State with conflict; Public	A-v2, D-v1; 2026-10-13	Conflicting	State disagreement; clarify requested authority or scope	Both spans available	Unsupported certainty fails
State with conflict; Internal	A-v2, D-v1, B-v1; 2026-10-13	Conflicting	Same disagreement; clarify	Both state spans available	Extra unrelated access does not resolve conflict
Unspecified deadline; Public	A-v2; 2026-10-13	Not found	Not found within declared corpus	No supporting span	No inferred date
Unspecified deadline; Internal	A-v2, B-v1; 2026-10-13	Not found	Not found within declared corpus	No supporting span	Broad access still lacks support
Human birthdate of project; Public	S-v1; 2026-10-13	Not applicable	Property applies only to persons	Scope-definition span available	Missing data is not the reason
Human birthdate of project; Internal	S-v1; 2026-10-13	Not applicable	Same scope reason	Scope-definition span available	Same applicability verdict
Which Alder project?; Public	K-v1; 2026-10-13	Clarification needed	Ask which project the reader means	Ambiguity span available	Do not select an entity silently
Which Alder project?; Internal	K-v1; 2026-10-13	Clarification needed	Same entity clarification	Ambiguity span available	Clearance does not disambiguate identity

The alternative-source rows isolate the central decision: identical claims can become answerable without changing the reader's access to the private source. The conflict rows isolate a different problem: greater clearance does not necessarily resolve competing support. For actual fixtures, retain the evidence entities and

their derivations so a reviewer can reconstruct these distinctions; PROV supplies the vocabulary for that provenance relationship. (www.w3.org)

Score the observable response

The proposed scoring unit is a substantive claim-and-reader pair, with answer-level summaries retained for practical decisions. Claim decomposition follows the grounding approach documented by Haystack; sufficiency and support remain separate judgments in MLflow. Before scoring, declare whether partially answered requests count as useful outputs or require every requested claim to pass. A supported introductory sentence must not rescue an unsupported operative value. (docs.haystack.deepset.ai) (mlflow.org)

- **Authorized evidence support:** Supported emitted claims divided by all substantive emitted claims. Empty responses receive an undefined claim-support rate, alongside their outcome label.
- **Improper disclosure:** Required-block cases exposing prohibited content, including disallowed explanation metadata, divided by all required-block cases. Also report the count of affected answers.
- **Justified block:** A block receives credit only when no sufficient eligible alternative exists and the declared explanation is permitted. Mere failure to retrieve support is insufficient.
- **False block:** Supported, answerable cases receiving an access block divided by all supported, answerable cases. Keep mistaken not-found declarations in a separate error count.
- **Citation accuracy:** Supporting eligible citations divided by all emitted citations, plus supported cited claims divided by claims requiring citations. Report reader-access resolution failures separately.
- **Correct abstention status:** Correctly labeled not-found, not-applicable, conflict, and clarification outcomes divided by the reference cases for each status, without pooling them into block accuracy.

RAG citation accuracy evaluation needs citation precision and claim coverage. They address different omissions: irrelevant citations reduce the former, while uncited substantive claims reduce the latter. ALCE explicitly separates correctness and citation quality. No metric is a substitute for a declared reference corpus and policy. If a role has no required-block cases, its disclosure rate is undefined rather than zero. (arxiv.org)

Adjudicate policy and support separately

OpenFGA recommends testing authorization models and supports assertions about authorization API results. Such tests can establish an eligibility reference, but answer support still needs its own annotation. LangSmith distinguishes reference-based offline evaluation from deterministic code checks. Use deterministic checks for declared statuses, source IDs, and citation structure; use calibrated reviewers for entailment, ambiguity, and whether an explanation reveals prohibited metadata. (openfga.dev) (docs.langchain.com)

- **Policy dispute:** Have the designated policy owner resolve the eligibility rule without changing the model answer.
- **Support dispute:** Have an evidence annotator examine the exact span and the claim's qualifiers.
- **Temporal dispute:** Determine which source state answers the requested time scope.
- **Disclosure dispute:** Review the answer, citation, and explanation against the same reader policy.
- **Annotation correction:** Publish the changed label and reason, then rescore all affected runs.
- **Unresolved disagreement:** Preserve competing labels and report sensitivity to the disputed cases.

Automated judges should not become the undisclosed authority for policy decisions. ARES uses human annotations with prediction-powered inference to mitigate judge prediction errors. That finding supports retaining human calibration for support assessment, but it does not prescribe a universal annotation count for permission tests. Pre-register sampling and adjudication rules so difficult cases cannot disappear from the reported denominator. (arxiv.org)

Data Analysis and Evidence

What the published results establish

Jeong and Lee's large-scale experiment sampled **1,000 questions** (www.researchgate.net), used **eight user profiles** (www.researchgate.net), and drew on **13,783 supporting-fact documents** corresponding to **7,405 HotpotQA questions**. (www.researchgate.net) Google Cloud Platform (GCP) supplies role-based access control (RBAC); Amazon Web Services (AWS) supplies attribute-based access control (ABAC). (www.researchgate.net) These are experimental conditions, not customer deployment measurements. The author paper is accessible through ResearchGate; the institutional record independently establishes the study's purpose. (snu.elsevierpure.com)

The authors report perfect document-permission **Match** across the profiles. (www.researchgate.net) They also report **token-overlap F1 of 0.31** for the user with no document access, suggesting internal knowledge or correct guessing as possible explanations. (www.researchgate.net) The methodological implication is limited but important: document filtering agreement and answer correctness do not establish reader-eligible evidence support. The no-access result does not demonstrate that a private document leaked.

Klisura and colleagues' policy-length ablation concerns **text-to-SQL**, not RAG. For GPT-to-GPT zero-shot verification, their Table 8 reports false-permit rates of **5.5%, 17.3%, and 33%** for short, medium, and long policies. (aclanthology.org) The reported policy lengths average approximately **2,700, 8,000, and 22,700 characters**. (aclanthology.org) These results motivate complexity-stratified testing; they are not expected disclosure rates for an enterprise document system.

Citation completeness requires separate attention. Gao and colleagues report that even their best evaluated models lack complete citation support **50% of the time on ELI5**. (arxiv.org) That benchmark result concerns the models and tasks studied, not a current product ranking. It supplies evidence for retaining citation-support checks alongside correctness and permissions. ARES, the Automated RAG Evaluation System, likewise separates context relevance, answer faithfulness, and answer relevance. (arxiv.org)

Utility and disclosure denominators (Hypothetical Example)

Table 3 is an **invented arithmetic example** derived entirely from the following stated counts, not an empirical benchmark. It concerns resolvable claims classified as answerable or required-block cases. Not-found, not-applicable, conflict, and clarification cases require separate status reporting. This separation implements the answer/refuse distinction studied by Klisura and colleagues while adding the article's eligible-evidence condition. (aclanthology.org)

Reference eligibility class	System answers	System declares a block
Sufficient eligible support: 40 cases	30 correctly supported answers	10 unjustified blocks
Required block, no eligible support: 20 cases	2 improper disclosures	18 justified blocks

The table yields **eligible answer utility = 30/40 = 75%**, **false-block rate = 10/40 = 25%**, and **improper disclosure rate = 2/20 = 10%**. An overall correct-outcome calculation gives **48/60 = 80%**, combining supported answers and justified blocks. That aggregate hides the disclosure denominator. Every value here is a transparent calculation on synthetic counts, with no inferred production performance.

For a role breakdown consistent with the table, assign the public reader 20 eligible cases, 14 supported answers, 6 false blocks, 12 required-block cases, and 2 disclosures. Its utility is **14/20 = 70%** and disclosure rate **2/12 = 16.7%**, rounded. Assign the internal reader 20 eligible cases, 16 supported answers, 4 false blocks, and 8 required-block cases with no disclosures. Its utility is **16/20 = 80%** and observed disclosure rate **0/8 = 0%**. These are invented counts, not evidence that either role is safe.

Report status strata separately: private-only restrictions, revocation, and public alternatives should each have their own case counts and outcomes. For this proposed protocol, define citation resolvability as the number of emitted citations whose exact supporting spans are accessible to the specified reader divided by all emitted citations checked under that reader and policy snapshot. Keep broken and inaccessible citations in the denominator. Report technical fetch failures separately from policy denials, and report missing citations separately. If no citations are emitted, label the rate undefined. OWASP recommends independent verification of cited-document existence and hashes; a successful locator check is evidence of availability, not proof of entailment. (cheatsheetseries.owasp.org)

Implications and Future Directions

Evaluating grounded AI answers requires the conjunction of relevance, authorized evidence support, correct temporal scope, and inspectable citations. Tooling can assess parts of that conjunction. Ragas distinguishes context ranking from faithfulness, while MLflow distinguishes retrieval sufficiency from generated-answer support. The article's reader-specific protocol connects those dimensions through a common eligibility reference rather than asserting that the existing metrics already enforce permissions. (docs.ragas.io) (mlflow.org)

Auditable AI answers need enough retained context for a permitted reviewer to explain why a verdict changed. A change can arise from a revised source, revoked access, a different reader, or corrected annotation. DataCite recommends connecting versions through related identifiers; PROV is extensible across domains. Together these standards provide vocabulary for an evidence trail, without determining the organization's disclosure policy or dispute-resolution rules. (support.datacite.org) (www.w3.org)

Several limitations remain. Relevant ranking can depend on information outside the visible result set: Elasticsearch notes that document-security queries do not alter the global statistics used for relevance scoring. An output-only evaluation should therefore limit its claims to observed source eligibility, answer support, and disclosure surfaces. It cannot establish the absence of every possible inference channel or internal exposure. (www.elastic.co)

Policy complexity deserves independent stratification. Klisura and colleagues report declining reliability with longer and more complex policies, but their text-to-SQL setting differs from prose answers, citations, and mixed-source retrieval. Future studies should vary policy interactions separately from evidence completeness and question difficulty. Replication should preserve these distinctions instead of treating refusal performance as a transferable security guarantee. (aclanthology.org)

Reliable Context's guarantees state that client-visible values have an excerpt, locator, and observation date, while explicitly distinguishing what a document says from whether it is correct. Those are vendor-stated observable properties suitable for acceptance testing. They do not replace the reader-specific support verdict or establish independent accuracy, certification, or compliance. Teams needing a search engine or RAG framework should evaluate those categories for their retrieval requirements.

Conclusion

An answer is defensible for a specified reader when its substantive claims have sufficient eligible support at the requested time, its citations expose the supporting spans to that reader, and its explanations respect the same policy. A correct claim supported only by an inaccessible source fails that standard. An independently sufficient public source can change the verdict without changing access to the private source. Those paired cases should anchor the evaluation.

The proposed protocol makes this decision observable through frozen evidence versions, policy snapshots, synthetic readers, and claim-and-reader cards. It treats supported answers, declared blocks, bounded not-found results, inapplicable properties, and clarification requests as distinct outcomes. It also preserves conflicts rather than converting a disputed claim into unsupported certainty. These are protocol definitions, not claims about any vendor's internal implementation.

The reporting requirement follows directly: retain the eligibility reference, disclose denominators by role and evidence status, and report utility separately from improper disclosure. A test with no observed disclosures remains bounded by its cases, roles, source versions, and policy conditions. Citation accuracy and reproducibility must accompany the scores so a reviewer can defend the evaluation itself.

The decisive question is therefore specific: given this request, this reader, these evidence versions, and this policy snapshot, should the system answer, declare a block, report a bounded absence, or ask for clarification? A permission-aware RAG evaluation succeeds when it can justify that decision and reproduce the evidence behind it.

Source links

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

1. https://docs.ragas.io/en/latest/concepts/metrics/available_metrics/faithfulness
2. https://docs.haystack.deepset.ai/docs/faithfulnessevaluator
3. https://mlflow.org/docs/latest/genai/eval-monitor/scorers/llm-judge/rag
4. https://www.researchgate.net/publication/397278007_Permission-Aware_RAG_Identity_and_Access_Management_IAM-Based_Access_Filtering_in_Multi-Resource_Environments

5. https://aclanthology.org/2026.findings-eacl.316
6. https://aclanthology.org/2026.findings-eacl.316.pdf
7. https://cheatsheetseries.owasp.org/cheatsheets/RAG_Security_Cheat_Sheet.html
8. https://www.crossref.org/documentation/retrieve-metadata/text-and-data-mining
9. https://arxiv.org/abs/2305.14627
10. https://snu.elsevierpure.com/en/publications/permission-aware-rag-identity-and-access-management-iam-based-acc
11. https://www.w3.org/TR/prov-dm
12. https://arxiv.org/abs/2309.15217
13. https://csrc.nist.gov/pubs/sp/800/205/final
14. https://www.elastic.co/docs/deploy-manage/users-roles/cluster-or-deployment-auth/controlling-access-at-document-field-level
15. https://learn.microsoft.com/en-us/azure/search/search-document-level-access-overview
16. https://docs.aws.amazon.com/kendra/latest/dg/user-context-filter.html
17. https://docs.cloud.google.com/generative-ai-app-builder/docs/data-source-access-control
18. https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/context_precision
19. https://support.datacite.org/docs/versioning
20. https://openfga.dev/docs/getting-started/immutable-models
21. https://openfga.dev/docs/modeling/testing
22. https://docs.langchain.com/langsmith/evaluation-concepts
23. https://arxiv.org/abs/2311.09476

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.