Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Negation in Information Retrieval: Recall vs. Violations

> 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 · 2026-10-30 · negation in information retrieval · semantic search · negative constraints · nevir · excluir · e-sens · retrieval evaluation · context engineering · reranking

Original article: https://reliablecontext.com/articles/negation-information-retrieval-recall-violations



In this article

1. Executive Summary
2. Introduction and Background
3. Methodology: define the objects before scoring
4. Published benchmarks and a paired example
5. Building a fixed collection and fixture ledger
6. Comparing embeddings, hybrid retrieval and rerankers
7. Analysis of key segments: scope, filters and mixed documents
8. Data Analysis and Evidence
9. Implications and future directions
10. Conclusion

Executive Summary

Negation in information retrieval requires separating the meaning of a source assertion from an instruction about which evidence may be retrieved. A passage saying an event did not occur and a question excluding a topic impose different judgments. NevIR tests contrastive document pairs, whereas ExcluIR addresses exclusionary queries; their results should not be folded into a single negation score. Weller, Lawrie and Van Durme's Negation Evaluation in Information Retrieval (NevIR, 2024) contains **2,556 pairs**, with **1,383 test pairs**, and explicitly lacks the large accompanying collection needed for collection-scale recall analysis. (arxiv.org) Zhang and colleagues' Exclusionary Neural Information Retrieval (ExcluIR, 2025) instead supplies **3,452 evaluation queries** and **70,293 training queries**. (ir.cwi.nl)

The central release question is whether a changed retriever preserves access to relevant, admissible evidence while controlling prohibited exposure. This report proposes a collection-level acceptance protocol, not a newly executed benchmark. Keep the original question fixed, freeze the collection and evidence budget, and compare a baseline with one changed component at a time. Report candidate recall, eligible-evidence recall, prohibited-unit exposure and the fraction of questions with any prohibited exposure separately. A reranker scores supplied candidates, so missing evidence needs a candidate-generation diagnosis before a ranking diagnosis. (sbert.net) Conventional document relevance can include a document containing any relevant piece; admissibility therefore needs its own span-level labels for mixed documents. (trec.nist.gov)

Published results motivate that separation without supplying production guarantees. Exclusion-Sensitive Penalization for Negative-Constraint Retrieval (E-SENS, Kim et al., 2026) evaluates **3,452 queries over 90,406 documents** and penalizes similarity to excluded intent. (arxiv.org) (aclanthology.org) Its workshop publication reports a recall versus trap-exposure trade-off, but the authors also identify incidental excluded content in useful documents and do not provide statistical uncertainty estimates. (aclanthology.org) Its reported R@5 changes from **0.9377 to 0.9366**, while V@5 changes from **0.8789 to 0.5426** between the displayed baseline and penalty setting. (aclanthology.org) As of **30 October 2026**, preprint-only wording is outdated: E-SENS appears in archival workshop proceedings. (aclanthology.org)

The proposed acceptance packet adds an independent answer-support inspection: which source spans actually support each cited claim, whether a required qualifier survives, and whether uncertainty is declared when evidence is unresolved. Retrieval coverage, exclusion compliance and answer support remain separate outcomes. A release passes only its prespecified loss policy, with unresolved labels and inaccessible evidence visible. Source versions and observation dates belong in that packet; provenance standards explicitly distinguish a revised document from its previous entity. (www.w3.org)

Introduction and Background

Semantic retrieval finds nearby representations of questions and documents. That makes it useful when wording differs, but proximity alone does not specify whether an assertion is affirmed, denied or explicitly excluded by the user. Sentence Transformers describes independently encoded query and paragraph representations and a later cross-encoder stage that scores candidates jointly with the question. Those are different ranking architectures, not definitions of factual support. (sbert.net) In **retrieval-augmented generation (RAG)**, retrieved material becomes evidence supplied to a language model; negative constraints therefore need evaluation at that evidence boundary. (aclanthology.org)

The practical problem arises when topical relevance and admissibility diverge. A document may be highly informative about a subject the user excluded. Another document may discuss the excluded subject only to distinguish an allowed alternative. Dropping every document containing the excluded word can remove useful evidence; passing every highly similar passage can expose prohibited material. E-SENS identifies precisely this mixed-document limitation for its document-level penalty. (arxiv.org)

For Reliable Context's context-engineering audience, the decision addressed here is a **release gate for retrieval software**: whether an embedding change, hybrid retriever or reranker supplies acceptable evidence for questions containing exclusions. Readers choosing a retrieval component should evaluate that software category directly. The protocol concerns observable evidence selection and support, with no inference about a publisher's implementation or performance.

The literature uses several operationalizations of negation. Petcu and colleagues (2025) ground a taxonomy in philosophical, linguistic and logical definitions; ExcluIR develops resources for exclusionary retrieval. (aclanthology.org) (ir.cwi.nl) Accordingly, a benchmark score is meaningful only with its task definition, annotation unit, candidate population and denominator. The report first reconstructs a published paired example, then specifies collection fixtures and metrics. Its quantitative section distinguishes **published measurements**, **arithmetic derived from them**, and **proposed measurements that have not been run**. No simulated gains, demand estimates or production failure rates are presented.

Methodology: define the objects before scoring

The protocol distinguishes four objects. This is an operational definition for the proposed evaluation, rather than a claim that any model internally represents them:

- **Query intent:** the information requested, its inclusion conditions and the scope of any exclusion.

- **Source proposition:** what a particular passage asserts, including polarity, attribution, time and qualifiers.
- **Admissible evidence span:** the exact text allowed to support a claim under that question and the evaluation policy.
- **Candidate document:** the retrieved container, potentially holding several propositions and both allowed and excluded material.

A **document-level exclusion** bans the container under a stated rule, such as membership in an explicitly excluded collection. A **span-level exclusion** bans particular evidence while allowing another passage from the same document. Neither follows automatically from a mention. This distinction extends the ordinary relevance convention: the Text REtrieval Conference (TREC) relevance documentation counts a document as relevant if any part is relevant. (trec.nist.gov)

For source-assertion negation, annotate whether the span affirms, denies, attributes or leaves unresolved the proposition being investigated. For user exclusion, annotate whether using the span breaches the instruction. Evidence contradicting an assertion can be appropriate when the task is verification. It becomes prohibited only if the question's exclusion rule makes it so. The NevIR and ExcluIR task constructions supply complementary diagnostics for these different semantics. (github.com) (arxiv.org)

Scope ambiguity must be adjudicated before scoring. Does “without references to topic X” restrict the answer's prose, its cited spans, every retrieved chunk or entire documents? These are different policies. Double negatives, exceptions and quoted denials should receive separate fixture labels and an adjudication note, not a presumed interpretation based on the presence of “not.” The linguistic taxonomy offers categories to organize such tests, rather than replacing local scope decisions. (aclanthology.org)

For each question q, define E_q as the set of known judged eligible evidence units, C_q as evidence units available in candidate text, and S_q as the units selected for the answer stage. Candidate coverage uses the same unit granularity as E_q and respects the recorded chunking and input limits. Fix unit identity and overlap rules before computing intersections. When spans overlap, deduplicate using the declared rule; repeated retrieval of the same evidence should not inflate coverage. Recall follows the ordinary positive-found over positive-total structure, but the labels here are eligible evidence units. (scikit-learn.org)

Structured filters can implement a declared property constraint. Elasticsearch documents `must_not` as excluding matches, and Weaviate documents filters that include or exclude objects by conditions. (www.elastic.co) (docs.weaviate.io) Such operations establish a result-set rule over indexed data. Whether a metadata value correctly represents the user's intended exclusion remains a separate annotation and data question. The evaluation must therefore preserve both the natural-language question and any structured filter derived from it.

Published benchmarks and a paired example

Table 1 distinguishes the units and limitations of the core research resources. Counts describe published artifacts, not measurements made for this report.

Resource	Published unit or size	What it establishes	Boundary for a release decision
NevIR, Weller et al.	2,556 pairs: 948 training, 225 development, 1,383 test. (arxiv.org)	Both query directions must rank their corresponding document correctly. (github.com)	No accompanying large collection for recall analysis. (arxiv.org)
ExcluIR, Zhang et al.	3,452 manually annotated evaluation queries; 70,293 training queries. (ir.cwi.nl)	Exclusion expressed in the user's query; paired positive and negative documents. (arxiv.org)	Its construction does not certify the relevance or admissibility of every document for an organization's collection.
E-SENS, Kim et al.	3,452 ExcluIR queries scored over 90,406 documents. (aclanthology.org)	Gold-answer presence and designated-trap presence in the returned list. (aclanthology.org)	A designated trap is not an exhaustive inventory of prohibited spans.

The table prevents a common inference error: strong contrastive preference does not establish that all needed evidence is found, and avoiding one labeled trap does not establish that every returned span is allowed. A collection-level release decision needs additional judgments.

Reconstructing the original directions

The original NevIR dataset card provides two queries: “Which mayor did more vetoing than anticipated?” and “Which mayor did less vetoing than anticipated?” (huggingface.co) In the benchmark passages, the first document describes vetoing as unprecedented; the edited document describes expectations of unusually extensive vetoing followed by fewer vetoes than expected. These are benchmark text assertions, not independently established historical facts. (huggingface.co)

Preserve that direction: the first query prefers the first document; the second query prefers the edited document. Let the similarity function be s. The strict paired rule requires s(q_more, d_original) > s(q_more, d_edited) **and** s(q_less, d_edited) > s(q_less, d_original). The original dense evaluator uses strict comparisons; the aggregate evaluator credits fully correct pairs. (github.com)

Scoring illustration (Hypothetical Example)

If an illustrative ranking puts the original document first for both queries, only one direction is correct and the pair receives no paired credit. If the second ranking reverses appropriately, both directions are correct. These are applications of the scoring rule, not results from a tested model. **Ties remain explicit:** the strict rule gives a tied direction no credit; the proposed ledger also records the tie instead of silently resolving it by document order. (github.com)

Artifact permissions need separate checks. The original NevIR software repository includes an MIT license, while reuse of a dataset requires checking its own artifact terms. (github.com) The ExcluIR repository supplies evaluation scripts, but this research did not verify a separate benchmark license. (github.com) Neither repository was installed or executed during this report.

Building a fixed collection and fixture ledger

Choose a fixed, permitted collection, freeze the accessible document versions, and specify how questions enter the sample. Stratify by exclusion semantics and evidence layout, rather than selecting only cases where the baseline looks weak. TREC warns that judgments must match the correct document collection; reusing labels against a changed collection undermines that correspondence. (trec.nist.gov)

Use independent labels instead of one mutually exclusive “good/bad document” field:

- **Relevant:** a source contains evidence needed for the requested information.
- **Eligible:** a span meets the inclusion requirement and the declared evidence policy.
- **Trap:** a judged span or document violates the query's specified exclusion.
- **Mixed:** a document contains both eligible and prohibited spans.
- **Affirmative control:** a related query tests ordinary retrieval without the exclusion.
- **Polarity contrast:** a minimally edited fixture tests source-assertion negation separately.
- **Unjudgeable:** accessible text or scope interpretation is insufficient for a decision.
- **Unresolved exclusion:** available evidence cannot establish whether the exclusion holds.

Table 2 defines the proposed fixture ledger. It is a schema for collection, not a populated dataset or a performance claim.

Ledger field	Required content	Audit purpose
Question and version	Original wording, fixture family, inclusion condition, exclusion and its scope.	Prevents unrecorded rewriting or a changed definition of success.
Source identity	URL or permitted locator, content hash, document version, observation date and accessibility status.	Connects judgments to the text actually inspected.
Evidence text	Verbatim supporting and trap spans, offsets, necessary surrounding context.	Separates container relevance from span admissibility.
Labels and review	Relevant, eligible, trap and mixed flags; annotators, agreement counts, adjudication reason, unresolved status.	Preserves disagreement rather than hiding it.
Retrieval trace	Candidate membership and rank, score, reranked position, ties and filter decisions.	Distinguishes omission from ordering.
Answer trace	Selected evidence, cited claim, used offsets, missing qualifiers and uncertainty outcome.	Connects the answer inspection to the retrieval run.

The ledger should make an include/exclude decision reproducible without trusting the original annotator's memory. The World Wide Web Consortium (W3C) provenance primer treats each revision as a new entity; this supports recording a changed document as a new observation instead of overwriting the earlier evidence. (www.w3.org)

Before tuning, apply these additional sampling safeguards:

- **Group shared material:** keep related source documents and minimally edited questions in the same split.

- **Check duplicates:** compare normalized question text, passage text and source identifiers across splits.
- **Reserve a test partition:** choose penalty settings without examining held-out outcomes.
- **Retain difficult cases:** preserve mixed documents, ambiguous scope and unresolved exclusions.
- **Record sampling losses:** disclose discarded questions and inaccessible evidence with reasons.

This review did not establish document-group independence for the original benchmark splits; the proposed checks require a local audit.

For labeling quality, have reviewers independently judge both relevance and exclusion scope, then retain disagreements and adjudications. Report agreement with its unit and denominator. An inaccessible page is a recorded access outcome, not a negative relevance label. When pooled judgments are incomplete, call the denominator **known judged eligible evidence**, not all relevant evidence in the collection. NIST's research on pooled judgments explains why sparse judgments can be insufficiently complete for reuse. (www.nist.gov) (trec.nist.gov)

Comparing embeddings, hybrid retrieval and rerankers

Preserve the original question as the reference input. If a method derives a trap query, metadata filter or instruction, retain that derived input alongside it. E-SENS, for example, preserves the original-query positive signal and subtracts a penalty based on excluded intent. That is an evaluated selection transformation, not permission to alter the user's requested information. (arxiv.org)

The comparison should isolate a component change wherever feasible:

- **Freeze the collection:** identical source versions, access permissions and document identifiers.
- **Freeze segmentation:** record chunk boundaries, overlap, deduplication and parent-document links.
- **Freeze candidate depth:** preserve the candidate budget when testing only a reranker.
- **Freeze evidence capacity:** specify selected units and the text budget actually passed onward.
- **Change one component:** embedding, hybrid combination, filter policy or reranker in a named comparison.
- **Record effective inputs:** query text, prefixes, derived exclusions and truncation settings.
- **Retain every position:** capture candidate ranks, post-ranking positions and ties before answer generation.

A reranker cannot select an eligible passage absent from its supplied candidates. Sentence Transformers explicitly places cross-encoding after candidate retrieval. (sbert.net) If an embedding change alters candidate generation, evaluate that end-to-end comparison separately from reranking a common pool. OpenSearch documents hybrid combination across subqueries and a candidate-depth bound on the results normalized and combined, which makes that bound part of the experimental configuration. (docs.opensearch.org)

Application programming interface (API) acceptance is not an exclusion guarantee. Cohere's rerank endpoint accepts a query and texts, but its current guide explicitly limits instructions that change relevance judgment. (docs.cohere.com) Voyage's embedding endpoint documents query/document input types and a

prepended retrieval prompt. That is an input convention, not evidence that arbitrary exclusion instructions will be obeyed. (docs.voyageai.com) Record the exact endpoint, model identifier and effective input; do not infer instruction-following from a parameter name.

Long-input handling also belongs in the manifest. Cohere's endpoint reference describes truncation, while its guide describes chunking behavior; those descriptions should be resolved against the chosen endpoint and settings rather than generalized into one universal behavior. (docs.cohere.com) The evaluation must inspect the text that survived preprocessing. A qualifier absent from the representation or answer context cannot be credited merely because it exists elsewhere in the parent source.

Analysis of key segments: scope, filters and mixed documents

Report segments by failure mechanism rather than just a global mean. The proposed segments are source-assertion polarity, explicit topic exclusion, entity exclusion, structured-property exclusion, mixed documents and unresolved evidence. These are evaluation strata, not claims about the distribution of real queries. Petcu and colleagues' taxonomy provides additional linguistic distinctions; stratification should preserve any locally important scope classes. (aclanthology.org)

Filters test indexed predicates. Pinecone documents metadata filters that limit results to matching records or documents. (docs.pinecone.io) Elasticsearch's Boolean exclusion removes documents matching its specified queries. (www.elastic.co) Neither documented operation establishes that missing metadata means a property is absent. If eligibility depends on an unknown property, keep an unresolved label and assess the effect of withholding that evidence under the prespecified policy.

Array scope needs inspection. Qdrant notes that separate conditions may match different array elements, while its nested filter requires an element satisfying the whole nested condition. (qdrant.tech) Pinecone separately documents array exclusion when any stored value matches. (docs.pinecone.io) These are operationally consequential differences in predicate scope. A fixture should record whether the intended constraint concerns an entity, one nested assertion, a source tag or an entire document.

Mixed documents need span judgments. A relevant paragraph can coexist with an excluded comparison. Blanket rejection may remove needed evidence; blanket acceptance may expose a prohibited passage. E-SENS explicitly identifies relevant documents containing incidental trap evidence as a limitation. (arxiv.org) Define two boundaries: retrieval exposure at the selected-context boundary and actual use at the answer-support boundary. Candidate exposure may be reported separately, without pretending it is identical to evidence shown to an answer model.

Unresolved exposure needs bounds. The query-exposure denominator includes empty results. If selected evidence is unjudged, the observed violation rate is a lower bound: report an upper bound that additionally counts every question with unresolved exposure. Questions whose exclusion policy itself remains unresolved stay in a separate status count. Do not silently score either form of uncertainty as compliance.

Control questions need independent reporting. A method can improve exclusion compliance by becoming overly restrictive. Affirmative controls reveal whether ordinary evidence access deteriorates; unresolved-exclusion fixtures reveal whether the system declares uncertainty. The independent NevIR reproduction by van den Elsen and colleagues (2025) reports unreliable transfer from fine-tuning on one negation dataset to the

other, supporting evaluation across task families rather than extrapolating from one success. (www.dare.uva.nl) Its repository supplies reproduction resources, not evidence that this report executed them. (github.com)

Data Analysis and Evidence

Metrics and their denominators

Table 3 specifies the proposed measurements. None of its entries is a measured result for a deployed system.

Metric	Numerator and denominator	Required companion disclosure
Candidate recall	For each q, eligible units in C_q / all units in E_q.	Candidate budget; questions with no known eligible evidence; judgment coverage.
Eligible-evidence recall@k	For each q, eligible units in S_q / all units in E_q, with at most k selected units.	Macro average over defined questions; optional micro average reported separately.
Prohibited-unit exposure	Judged prohibited selected units / all judged selected units, aggregated across the declared question set.	Unjudged selected units and document-versus-span policy.
Queries with any exposure	Questions with at least one judged prohibited selected unit / all questions with an adjudicated exclusion policy.	Unresolved exposure, empty result sets and judgment coverage; policy-unresolved questions counted separately.
Strict paired accuracy	Pairs with both directional rankings correct / all eligible labeled pairs.	Tied directions and missing candidates, separately counted.
Unsupported-use rate	Answers containing any unsupported or inadmissible cited use / all evaluated answers.	Atomic claim judgments and answers declining to answer, separately labeled.
Qualifier and uncertainty outcomes	Counts by prespecified qualifier-loss and uncertainty labels / their applicable inspected answer sets.	Overlapping labels retained; no combination into retrieval recall.

The two exposure measures answer different questions: the proportion of selected material that breaches the policy, and the frequency with which a question encounters any breach. Empty result sets can reduce exposure while destroying recall. Undefined recall must remain visible when E_q is empty; scikit-learn's explicit zero-division policy illustrates why this case requires a stated convention. (scikit-learn.org) Here the proposed policy reports it as undefined and excludes it from the defined-recall average, while retaining the question in the status counts.

Published quantities and a trade-off calculation

E-SENS evaluates four embedding models. (aclanthology.org) Its archival table uses shared displayed penalty checkpoints and reports, for text-embedding-3-small, **R@5 = 0.9377** and **V@5 = 0.8789** at beta zero; at **beta 0.30**, it reports **R@5 = 0.9366** and **V@5 = 0.5426**. These are the paper's gold-document and designated-trap hit rates, not this protocol's prohibited-unit exposure measure. (aclanthology.org) Subtracting those published

rates gives a recall change of **−0.11 percentage points** and a trap-hit change of **−33.63 percentage points**. (arxiv.org) This arithmetic does not establish statistical significance, transfer to another collection or a production setting. The authors report no statistical uncertainty estimates. (aclanthology.org)

A separate fixed-pool preprint by Bian and colleagues (2026) reports strict preference increasing from **52.30% to 56.62%** while preferred-passage mean reciprocal rank falls from **0.5764 to 0.3055**. Its comparisons are conditional on pair coverage in one fixed candidate configuration. (arxiv.org) Thus preference improvement can coexist with worse position among other candidates; it is not a substitute for collection recall.

For the proposed evaluation, plot eligible-evidence recall against query-level prohibited exposure across prespecified thresholds or penalties, with a companion plot for unit-level exposure. Keep the evidence budget fixed, show uncertainty by resampling query families or shared-document groups, and identify dominated settings. Resample baseline and changed-method outcomes together within each group; preserve ties and missing-candidate statuses in every replicate. Report group counts and the interval construction rather than treating edited pairs as independent observations. Select a setting on development data under the loss policy, then evaluate it on held-out groups. **No curve or numeric gain has been measured for this proposed protocol.**

Implications and future directions

The release decision needs an explicit loss policy before results are examined. It should state the consequences of losing eligible evidence, exposing prohibited evidence and returning uncertainty. Avoid collapsing these outcomes into a weighted score unless the weights and acceptable trade-offs are documented.

A reviewable acceptance packet should contain:

- **Scope policy:** the distinction between assertion polarity, entity exclusion, topic exclusion and document bans.
- **Frozen manifest:** source hashes, versions, observed text, question families and effective model inputs.
- **Coverage accounting:** candidate omissions, judgment coverage, unjudgeable cases and inaccessible sources.
- **Separate outcomes:** recall, unit exposure, query exposure, paired preference and answer support.
- **Prespecified decision:** acceptance limits, uncertainty method and treatment of unresolved labels.
- **Failure ledger:** exact supporting and trap spans, rankings and adjudication notes for each rejected setting.

For the secondary answer-stage test, keep the answer model, instructions and evidence budget fixed. Compare baseline and changed retrieval outputs for the same questions. Inspect cited claims against the exact supplied spans, marking inadmissible use, unsupported inference, lost qualifiers and appropriate uncertainty independently. This analysis tests the downstream consequence of evidence selection; it does not retroactively turn a ranking score into factual support. Candidate-only controls can help distinguish absence of evidence from failure to use evidence already supplied.

The research is developing beyond the original paired task. CoverageBench-Negate, described in an October 2026 preprint by Abrishamchian and colleagues, spans **six datasets and roughly 63,000 original/edited evidence pairs** and tests retrieval when corpus evidence conflicts with pretrained priors. That is complementary to user exclusion, not an interchangeable label set. (arxiv.org) E-SENS is now archival workshop work, with submission rules specifying double-blind review; archival status does not establish independent replication of its numerical results. (aclanthology.org) (groundlm.github.io)

Future collection studies should publish group-aware splits, source-version manifests and adjudication instructions alongside scores. ExcluIR's artifact and the independent NevIR reproduction already provide evaluation resources that can inform such work. (github.com) Their existence supports inspectability, while successful reproduction and transfer still require execution and measurement.

Conclusion

Negation in information retrieval should be evaluated as a set of distinct evidence-selection problems. Source-assertion polarity asks what the evidence says. User exclusion asks which evidence is permissible. Document relevance identifies a potentially useful container, while span admissibility identifies the text that may support the answer. Keeping these objects separate makes a release decision interpretable.

The proposed protocol holds the original question and collection fixed, labels relevance and exclusion independently, and compares changed components under a declared evidence budget. It reports whether needed evidence reaches the candidate set, whether selection retains it, how much prohibited material is exposed and how many questions encounter any exposure. Ties, unresolved labels and inaccessible sources remain visible. A later answer-support inspection records what the generated claim actually used.

Published benchmarks provide diagnostic evidence, not a universal exclusion guarantee. Paired ranking, collection coverage, designated-trap avoidance and cited-claim support have different units and denominators. The acceptance policy should preserve those differences rather than using a single negation score to obscure them.

A defensible release packet therefore consists of the frozen manifest, the fixture ledger, separate recall and exposure measurements, uncertainty estimates and source-span judgments. If a changed method preserves the evidence required by the question while meeting the prespecified exclusion policy, it can be accepted within that tested scope. If not, the ledger should show whether the next intervention belongs in collection access, candidate generation, filtering, ranking or answer support.

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/2305.07614v2
2. https://ir.cwi.nl/pub/35863
3. https://sbert.net/examples/sentence_transformer/applications/retrieve_rerank/README.html
4. https://trec.nist.gov/data/reljudge_eng.html
5. https://arxiv.org/html/2608.30130v3

6. https://aclanthology.org/2026.groundlm-1.28
7. https://aclanthology.org/2026.groundlm-1.28.pdf
8. https://www.w3.org/TR/prov-primer
9. https://aclanthology.org/2025.findings-emnlp.839
10. https://github.com/orionw/NevIR/blob/main/src/negations/dense.py
11. https://arxiv.org/html/2404.17288v1
12. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html
13. https://www.elastic.co/docs/reference/query-languages/query-dsl/query-dsl-bool-query
14. https://docs.weaviate.io/weaviate/search/filters
15. https://github.com/orionw/NevIR/blob/main/src/negations/evaluate.py
16. https://huggingface.co/datasets/orionweller/NevIR
17. https://github.com/orionw/NevIR/blob/main/LICENSE
18. https://github.com/zwh-sdu/ExcluIR
19. https://www.nist.gov/publications/comparison-pooled-and-sampled-relevance-judgments
20. https://docs.opensearch.org/latest/query-dsl/compound/hybrid
21. https://docs.cohere.com/reference/rerank
22. https://docs.voyageai.com/reference/embeddings-api
23. https://docs.pinecone.io/guides/search/filter-by-metadata
24. https://qdrant.tech/documentation/search/filtering
25. https://www.dare.uva.nl/id/3a7eb311-2976-4635-96cc-66af9f782520
26. https://github.com/thijmennijdam/NevIR-reproducibility
27. https://arxiv.org/html/2610.26260v1
28. https://arxiv.org/abs/2610.37162
29. https://groundlm.github.io/grouplm_emnlp2026/cfp.html

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.