Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

When is evidence sufficient for fact checking? Omission tests

> 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 · 2026-10-23 · evidence sufficiency · fact checking · fact verification · omission test · abstention · evidence provenance · selective prediction · information fusion

Original article: https://reliablecontext.com/articles/fact-checking-evidence-omission-test



In this article

1. Executive Summary
2. Introduction and Background
3. Defining sufficiency and its alternatives
4. Gold evidence sets and adjudication
5. Constructing the deletion and restoration experiment
6. Scoring abstention and separating retrieval from inference
7. Reviewer effort, evidence order and stopping
8. Data Analysis and Evidence
9. Implications and Future Directions
10. Conclusion

Executive Summary

Evidence is sufficient for fact checking when the available, permitted evidence establishes the premises needed to support or refute a precisely specified claim at the relevant time. **Sufficiency belongs to the claim and evidence set**, under a declared decision rule. A confident model output is a different object: calibration concerns the relationship between predicted probabilities and correctness, rather than whether a particular premise is present. (proceedings.mlr.press) Editorial standards likewise require relevant supporting and challenging evidence, rather than a universal document count. (efcsn.com)

The proposed **omission test** makes that distinction experimentally observable. Start with an adjudicated sufficient evidence set, delete a necessary date, qualifier, entity link or complementary passage while preserving fluency, and test whether the verifier abstains. Restore the missing material and test whether a justified verdict returns. Include deletions of irrelevant and redundant text as controls. SufficientFacts already distinguishes essential omissions from information repeated elsewhere, providing a precedent for labeling the remaining evidence rather than assuming every deletion destroys sufficiency. (github.com) Preserve sources, transformations and observation dates so reviewers can reconstruct the evidence state; provenance describes how data was produced. (www.w3.org)

Published results motivate this experiment without supplying its outcome. **Atanasova et al. (2022)** report **21% accuracy** for omitted adverbial modifiers versus **63%** for omitted dates in their tested setting. Reported diagnostics are not deployment rates. (aclanthology.org) **Alt et al. (2026)** report **94% versus 74% verification success** for ranking versus selection. (arxiv.org) The study involved **five participants**. (arxiv.org) Average sentences read were **2.48 versus 3.85**. (arxiv.org) Sentence counts measure one aspect of effort. A new protocol should measure elapsed review time separately.

Issue a verdict only after the evidence satisfies the specified decision rule. Otherwise distinguish insufficient supplied evidence, unsuccessful retrieval, blocked access, unresolved claim scope and unresolved conflict. Evaluate sufficiency detection, unjustified acceptance, justified abstention, verdict accuracy and reviewer effort separately. **Risk and coverage must be reported together** because abstention changes which cases receive

predictions. (jmlr.org) The experiment described here is a proposed research design; no omission fixture was executed and no new accuracy estimate is asserted. Its reproducibility requirements follow established expectations to document evaluation conditions and uncertainty. (airc.nist.gov)

Introduction and Background

The question “when is evidence sufficient for fact checking?” contains two decisions. One concerns whether a statement follows from an evidence package. The other concerns whether the package is adequate to license a verdict at all. Atanasova et al. explicitly separate identifying sufficient information from predicting veracity. (direct.mit.edu) This report develops a deletion and restoration experiment for the second decision, while retaining the first as a separate outcome.

The unit of analysis is an **evidence set**, including the context needed to interpret its passages. A source URL alone does not settle whether a number refers to the intended population, period or definition. The United Kingdom's Office for Statistics Regulation makes the analogous point that publishing a source is insufficient when methods, definitions and limitations remain unclear. (osr.statisticsauthority.gov.uk) For an experiment, these contextual requirements must become inspectable fields rather than assumptions supplied by a reviewer.

Three boundaries follow. First, **source quality** and **premise completeness** are separate judgments: a reliable source can leave the relevant question unanswered. Primary sourcing remains an important editorial practice, as FactCheck.org's documented process illustrates. (www.factcheck.org) Second, **confidence** must be defined. The Intergovernmental Panel on Climate Change distinguishes its evidence-assessment confidence terminology from probability and statistical confidence; that framework is an analogy here, not a validated threshold for arbitrary claims. (www.ipcc.ch) Third, a negative search result needs a declared search space.

The analysis is anchored to **23 October 2026** as its review date. It proposes an experiment, not a universal proof procedure or a ranking of verification products. The research question is whether a system changes its decision appropriately when the evidential basis changes. A runnable evaluation can assist inspection without establishing that its implementation is error free. (pmc.ncbi.nlm.nih.gov)

Defining sufficiency and its alternatives

A decision-relative definition

Specify a claim **c**, supplied evidence **E**, an authorized corpus **C**, a corpus snapshot **v**, a decision time **t**, and a declared rule **P**. Define sufficiency as the adjudicated judgment that E licenses a support or refutation verdict about c under those conditions. This is the protocol's operational definition. It does not claim that a finite corpus exhausts everything knowable about the world.

Before labeling E, normalize the claim's referent, predicate, quantifier, comparison class and temporal scope. These are proposed annotation fields. Their purpose is to prevent a reviewer from silently replacing an ambiguous claim with an easier one. Store the normalization alongside the evidence's provenance; W3C's framework supplies concepts for describing entities, activities and participants in data production. (www.w3.org)

The stopping rule is **premise completion plus source assessment**, not a fixed number of links. The European Fact-Checking Standards Network normally asks for at least two sources for a central claim but expressly permits cases with only one relevant source. That is an editorial policy, not a theorem that two documents are sufficient. (efcsn.com) Duplicate reports of the same underlying observation should therefore be identified as shared provenance within this protocol. The source-assessment stage should retain the underlying primary material where available, consistent with established editorial practice. (www.factcheck.org)

Distinguish the reasons a verdict is unavailable

Use separate fields for **evidence status**, **retrieval status** and **claim status**:

- **Insufficient supplied evidence:** the presented passages do not establish a necessary premise.
- **Not found in scope:** the documented search did not locate the premise in the authorized snapshot.
- **Retrieval failure:** a known relevant passage exists but the retrieval pipeline did not return it.
- **Blocked access:** a specified source could not be inspected with the available permissions or tools.
- **Underspecified claim:** the entity, period or meaning needed for adjudication remains unresolved.
- **Unresolved conflict:** applicable sources disagree and the declared source-assessment rule does not settle the issue.

If a workflow uses **unverifiable**, define it as no supported verdict under the stated claim interpretation, evidence scope and decision time. Keep the reason code separately. This scoped label should not assert falsity or permanent impossibility of verification.

These are proposed experimental categories, not interchangeable truth labels. In particular, “not found” is a scoped search result. It cannot establish universal absence. Similarly, contradiction may provide sufficient evidence for refutation when the conflicting proposition concerns the same referent and scope; incompatible sources may instead require abstention.

Do not average away such differences. Cochrane's GRADE framework treats inconsistency, indirectness and imprecision as distinct considerations in evidence certainty. Its application is evidence synthesis, so importing a GRADE category as a general fact-checking verdict would require separate validation. (www.cochrane.org)

For conflicts, the proposed adjudicator records source proximity to the underlying observation, the proposition each source actually establishes, and temporal compatibility. A later document is not automatically relevant to an earlier state. The verdict rationale must state why a competing passage changes the decision or why it does not. Providing the evidence needed to replicate verification is consistent with EFCSN's transparency requirements. (efcsn.com)

Gold evidence sets and adjudication

A **gold sufficient set** is an adjudicated reference package, not an assertion that its passages are the only possible evidence. FEVER's published format allows alternative evidence sets for a claim. (fever.ai) The proposed fixture should preserve those alternatives so that deletion of one route does not incorrectly make another route disappear from the annotation.

A set is **inclusion-minimal** when removing any remaining component breaks its sufficiency under the declared rule. This differs from having the smallest possible number of components among all sufficient sets. The experiment should retain both minimal and redundant packages. Without redundancy controls, an always-abstain response to shorter evidence could appear desirable.

Establish labels before testing systems

The proposed annotation policy has the following steps:

- **Fix the target:** record the original claim and its normalized meaning before reviewing predictions.
- **Freeze the evidence:** retain permitted source versions, observation dates and the exact displayed context.
- **Label independently:** ask reviewers to judge sufficiency before seeing system outputs or one another's labels.
- **Separate judgments:** collect sufficiency, veracity when licensed, and the indispensable premise.
- **Adjudicate explicitly:** record the competing rationales and the reason for the final label.
- **Retain uncertainty:** keep unresolved cases visible and define whether they enter the scored population.
- **Preserve alternatives:** record other known sufficient combinations rather than forcing one canonical proof path.

The rationale for independent labels is methodological: agreement concerns independently coded data, and high agreement does not establish the validity of the annotation scheme. Artstein and Poesio's survey makes both distinctions. (www.cs.brandeis.edu) Therefore report pre-adjudication agreement alongside disagreements and exclusions. A consensus label alone cannot show whether the task was well defined. The reporting package should also describe upstream sources and annotation methods, as Data Cards recommend. (research.google)

For the **oracle evidence condition**, reviewers and verifiers receive only the evidence package authorized for that trial. External knowledge is disallowed by the proposed task contract. Atanasova et al. similarly instructed annotators not to rely on world knowledge. (direct.mit.edu) A memorized correct answer must not receive credit as evidence-grounded verification when its necessary premise is absent.

Store the original claim, labels, omissions and transformations in machine-readable form. SufficientFacts' released documentation includes original labels and removed information, supporting paired analysis rather than a collection of disconnected examples. (github.com) Data Cards provide a broader documentation model covering upstream sources, annotation, training and evaluation methods. (research.google) Neither documentation format establishes that the resulting gold judgment is correct.

The proposed **fixture manifest** should link every variant to its parent claim and its evidence version. The provenance record should distinguish a source document from a passage extracted from it and from the edited passage shown to the system. That makes a mistaken extraction distinguishable from a mistaken source interpretation, using W3C's provenance vocabulary as a documentation basis. (www.w3.org)

Constructing the deletion and restoration experiment

Perturb a necessary premise, then restore it

For each adjudicated package, create **full**, **omitted** and **restored** conditions. Remove only the targeted material; preserve the rest of the trial presentation. Check grammaticality and semantic scope separately. Atanasova et al.'s omission method explicitly aims to preserve grammar and fluency. (direct.mit.edu) Fluency is a construction constraint, not evidence that the remaining package remains adequate.

Include these omission families:

- **Date or interval:** remove the period that connects an observation to the claim.
- **Qualifier:** remove a condition, exception, comparison basis or scope restriction.
- **Entity link:** remove the relationship identifying which similarly described thing the passage concerns.
- **Complementary passage:** remove a separate premise required to connect otherwise relevant passages.
- **Table context:** remove a necessary header, unit or caption while retaining the cell.
- **Redundant material:** remove a repeated premise that remains established elsewhere.
- **Irrelevant material:** remove text that the adjudicators consider unnecessary to the verdict.
- **Sham edit:** change presentation without removing a premise, to check sensitivity to formatting.

Re-adjudicate the remaining set after every edit. Some deletions remove a full sentence but leave all necessary information available elsewhere. The SufficientFacts distinction between repeated and essential information is directly relevant to this control. (github.com)

Table 1 illustrates a **hypothetical example** fixture. All entities, passages and labels in this table are synthetic design specifications, not observed annotations or benchmark results. The source/version field deliberately identifies synthetic material; a real fixture must replace it with its inspected source URL and version.

Case and claim	Source/version and observation	Removed premise	Proposed human label after deletion	Predicted label, abstention and rationale
H-date: “Entity A held role R during period T.”	Synthetic fixture H, version A; observation recorded at execution.	A passage establishes the role, but the period is deleted.	Insufficient if no other passage supplies T.	Not run; record all three outputs.
H-qualifier: “All records satisfy condition Q.”	Synthetic fixture H, version A; observation recorded at execution.	An exception limiting the source statement's scope.	Re-adjudicate whether a verdict is licensed; do not preserve the original label automatically.	Not run; rationale must address the exception.
H-link: “Entity A owns asset X.”	Synthetic fixture H, version A; observation recorded at execution.	The identity link connecting a named subsidiary to A.	Insufficient when that link is indispensable.	Not run; rationale must name the absent relationship.

Case and claim	Source/version and observation	Removed premise	Proposed human label after deletion	Predicted label, abstention and rationale
H-multi: “Entity A meets both conditions P and Q.”	Synthetic complementary passages H-P and H-Q; same version.	The independent passage establishing Q.	Insufficient if no alternative establishes Q.	Not run; track the missing complementary passage.
H-repeat: “Entity A meets condition P.”	Synthetic passages H-P and H-P-copy; same version.	One repeated statement of P.	Sufficient if the retained passage independently establishes P.	Not run; abstention would be an over-rejection.

The key comparison is within the originating claim. A necessary omission should change the sufficiency judgment; an irrelevant or redundant omission should not. **Restoration** checks reversibility. If the restored package does not recover the reference judgment, inspect the edit, presentation and annotation before attributing the difference to the verifier.

Include evidence that works only in combination

The multi-passage case is essential. FEVEROUS explicitly permits evidence pieces that establish a label only when examined together. (fever.ai) It also supplies contextual information such as row and column headers for table cells. (fever.ai) A collection of individually relevant excerpts can therefore remain incomplete. The experimental package should preserve the source's interpretive context; that includes communicating definitions and limitations alongside numerical statements. (osr.statisticsauthority.gov.uk)

Design separate trials for missing content and missing access to content. In an omission trial, the necessary passage is intentionally absent from the presented package. In a retrieval trial, it remains available in the authorized corpus. This preserves the distinction between a deficient evidence set and a pipeline that failed to assemble it. Research on evidence retrieval similarly distinguishes relevance from utility to the verifier. (irlab.science.uva.nl)

Scoring abstention and separating retrieval from inference

Separate the denominators

Define the positive sufficiency label as **sufficient evidence**. Under this convention, a false positive is acceptance of an insufficient package. Report a confusion matrix for each omission type and condition, including a sufficient control population. Precision and recall use different denominators, as the standard metric definitions make explicit. (scikit-learn.org)

Table 2 specifies the proposed scoring rules. Symbols denote counts from an executed experiment; no counts or outcomes are supplied here.

Measure	Numerator / denominator	Interpretation and required qualification
Sufficiency precision	True sufficient predictions / all predicted sufficient	Penalizes acceptance of incomplete packages.

Measure	Numerator / denominator	Interpretation and required qualification
Sufficiency recall	True sufficient predictions / all gold sufficient	Detects excessive rejection of adequate packages.
Insufficiency detection	Correctly rejected gold-insufficient cases / all gold-insufficient cases	Report separately for date, qualifier, entity and passage omissions.
Coverage	Issued verdicts / all eligible trials	Include the declared eligibility and exclusion rules.
Selective risk	Incorrect issued verdicts / all issued verdicts	Undefined when no verdict is issued.
Verdict accuracy on sufficient evidence	Correct verdicts on gold-sufficient cases / all gold-sufficient cases	Count abstentions as failures for this measure; also report answered-only accuracy separately.
Unjustified verdict rate	Issued verdicts on gold-insufficient cases / all gold-insufficient cases	A factually correct guess still counts as unjustified acceptance.
Restoration recovery	Correctly accepted restored cases / restored cases adjudicated sufficient	Pair with omission detection and full-condition performance.

These denominators prevent an abstention policy from appearing successful merely because it answers less often. The coverage and selective-risk definitions follow selective-classification formalization. (proceedings.mlr.press) Under the protocol's convention, zero accepted trials yield zero coverage and **undefined risk**, because the risk denominator is zero. That is a mathematical edge case, not evidence of perfect verification. The general selective-classification objective is a trade-off between coverage and accuracy; it does not eliminate the need to inspect rejected cases. (jmlr.org)

Report the underlying confusion counts, class-specific scores and the chosen averaging rule. Macro averaging takes the unweighted mean of label-specific metrics, so it answers a different question from a pooled count. (scikit-learn.org) A pooled score must not conceal failure on a rare omission family. In particular, recall's reference population is the actual positive class, rather than only the cases accepted by the system. (scikit-learn.org)

For uncertainty, report **numerators, denominators and confidence intervals**, with the sampling unit declared. Wilson intervals are available for appropriately modeled binomial proportions. (docs.scipy.org) They should not be attached mechanically to dependent variants of the same claim. The proposed paired analysis resamples originating claims while keeping all their variants together. SciPy's paired bootstrap preserves correspondence across arrays; its paired option alone does not construct arbitrary claim-level clusters. (docs.scipy.org) General cluster-resampling guidance likewise warns that observation-level resampling can ignore grouped structure. (www.stata.com)

Run oracle and retrieved evidence conditions

In the oracle condition, supply the adjudicated packages directly. The foundational omission study largely excluded document and sentence retrieval, making that boundary particularly relevant. (direct.mit.edu) In the retrieved condition, run the complete pipeline against the same frozen corpus.

Track the following stages separately:

- **Corpus availability:** is a necessary sufficient route present in the authorized snapshot?
- **Retrieval completeness:** did the pipeline return a whole sufficient route?
- **Sufficiency judgment:** did the verifier recognize whether its actual input was adequate?
- **Verdict inference:** when the input was sufficient, did it reach the correct verdict?
- **Operational abstention:** did the interface represent the reason for withholding a verdict accurately?

A document retrieval score cannot substitute for the second stage. FEVEROUS's evidence criterion requires a complete annotated set, not an arbitrary relevant fragment. (fever.ai) Conversely, retrieving a gold set does not guarantee correct interpretation. FEVER's annotators recorded necessary evidence sentences for supported and refuted claims, making the evidence annotation a separate object from the verdict label. (fever.ai)

Keep all variants of a parent claim together when allocating training, validation and test data. The proposed split applies the established requirement to account for dependencies between training and test samples. (pmc.ncbi.nlm.nih.gov) Choose thresholds on development data, then freeze them before evaluating the test set. Document test sets, metrics and tools, following NIST's evaluation-documentation expectations. (airc.nist.gov)

Reviewer effort, evidence order and stopping

Evidence ordering adds a sequential decision: how much material must a reviewer inspect before a verdict becomes licensed? Alt et al. define a sufficient prefix as one containing an entire gold evidence set. (arxiv.org) Their framing preserves access to the remaining evidence while bringing sufficient material forward. (cris.biu.ac.il)

The proposed omission experiment should therefore evaluate **reviewer effort** alongside verdict quality:

- **Passages inspected:** count evidence units opened and actually considered.
- **Time to justified decision:** measure elapsed review time, with pauses and censoring rules declared.
- **Missing-premise identification:** record whether the reviewer names the absent information.
- **Redundancy burden:** measure material inspected beyond an adjudicated sufficient prefix.
- **Unresolved reviews:** retain abstentions and unfinished trials in the reported population.

The stopping policy has separate sufficient and insufficient branches. For sufficient cases, stop after the reviewer can identify a complete applicable evidence route and has applied the declared source-assessment rule. For insufficient cases, stop after the predefined search or review budget is exhausted, and report the remaining gap. **Budget exhaustion establishes a stopping condition, not truth or falsity.**

Do not calculate “time to sufficient evidence” as if an unreachable event occurred in a deliberately insufficient trial. Alt et al.'s minimal-sufficient-rank definition itself assumes that a sufficient set exists. (arxiv.org) In omitted cases, report time to justified abstention and whether the missing premise was identified.

Randomize presentation order and counterbalance conditions in the proposed human study. Avoid showing a reviewer multiple variants of the same claim where remembering a removed premise would defeat the evidence boundary. Record whether a source remained accessible outside the trial interface. Preserve the

claim-to-source trail so another reviewer can reconstruct the process, as transparent fact-checking standards require. (efcsn.com) Archive the inspected evidence where permitted; EFCSN recommends preserving original claims and evidence for traceability. (efcsn.com)

The retained source must also include enough interpretive context. A reviewer who finds a number quickly but lacks its definition has not necessarily reached an adequate stopping point. OSR's guidance explicitly joins figures with their context, strengths and limitations. (osr.statisticsauthority.gov.uk) This is why the experiment should score missing-premise rationales, rather than treating speed alone as the objective.

Data Analysis and Evidence

Table 3 separates published measurements from calculations and their transfer limits. None is a result of the proposed deletion and restoration protocol.

Source and setting	Verified measurement or calculation	Interpretive boundary
Atanasova et al., reported omission diagnostics	Adverbial-modifier detection accuracy **21%**; date-modifier detection accuracy **63%**. (aclanthology.org)	Reported figures; abstract-level aggregation was not reconstructed.
SufficientFacts released diagnostic sets	FEVER **1,000**, HoVer **1,000**, VitaminC **600** instances. FEVER and HoVer each have **600 constituent** and **400 sentence** omissions; VitaminC uses constituent omissions. (huggingface.co)	These are released test-set compositions, not estimates of deployment prevalence.
Atanasova et al., Table 4, HoVer BERT sufficiency	Macro F1 **58.15 to 74.98**; **16.83 points** calculated gain; **three seed runs**. (aclanthology.org)	Same-model baseline versus augmented data.
Atanasova et al., Table 4, HoVer ALBERT veracity	Macro F1 **76.88 to 79.44**; **2.56 points** calculated gain. (aclanthology.org)	Different model; gains are not jointly achieved.
Alt et al., controlled reviewer study	**Five participants**. (arxiv.org) Success **94% ranking versus 74% selection**. (arxiv.org) Average sentences read **2.48 versus 3.85**. (arxiv.org)	Small, curated study; sentences read are an effort proxy, not elapsed-time savings.

The sufficiency and veracity differences above are direct subtraction of the cited table cells. The abstract's “up to” gains are not used as a single jointly achieved result. The table comparison gives a reproducible denominator and model identity for each calculated change. **Macro F1 points**, percentages and relative percentage improvements must remain distinct.

SufficientFacts selected cases the tested models initially answered correctly, limiting deployment interpretation. (aclanthology.org) Its author-maintained card reports **three annotators** and **0.81 Fleiss' kappa**. (huggingface.co) That describes annotation agreement, not proof that each gold label is valid.

A new report should publish per-type confusion matrices before aggregate scores. For each system, plot **selective risk against coverage** at frozen operating thresholds, with claim-level uncertainty intervals and the accepted-error count visible. The underlying rationale is the risk-coverage trade-off of selective classification. (jmlr.org) Add paired full-to-omitted and omitted-to-restored transitions, including cases where the sufficiency judgment changes but the factual answer stays the same.

No new confidence interval can be responsibly computed from the headline percentages alone without the relevant denominator and sampling structure. With repeated variants or reviewer observations, the independent unit cannot simply be assumed to be the displayed row. Resampling must respect that structure. (www.stata.com) Report interval failures too: bootstrap intervals can be undefined for degenerate bootstrap distributions. (docs.scipy.org)

Implications and Future Directions

The main deliverable should be a **documented omission fixture and scoring rule**, not a broad claim that one verifier knows when evidence is enough. Data Cards emphasize dataset origins, development, intended use and evolution; those fields help describe what the fixture can reasonably test. (research.google) For Reliable Context's audience, the practical output is an auditable evidence package with an explicit reason when the evidence does not license a verdict.

Several transfer questions remain. Artificial deletions may differ from naturally missing information. Fluency checks do not cover every semantic artifact. A system may learn the editing pattern rather than recognize a missing premise. Authorized corpora, source quality and temporal requirements can vary between deployments. NIST calls for documenting generalizability limits beyond development conditions. (airc.nist.gov) These are research questions for follow-up experiments, not measured failures of the protocol.

Release decisions also need separate attention. FAIR principles call for detailed provenance and a clear data-use license; documenting a source does not by itself permit redistribution. (www.gofair.foundation) Pin the original fixture revision, retain the applicable license text, and distinguish the license of code from the rights governing source passages. Record that history using an explicit provenance representation so another team can inspect the transformations. (www.w3.org) This report does not redistribute either study's evidence corpus or certify a current tagged release.

Before making stronger claims, the proposed evaluation should add:

- **Natural missingness:** compare controlled deletions with independently collected incomplete cases.
- **Domain transfer:** test whether labels and thresholds remain meaningful under new evidence conventions.
- **Temporal transfer:** ensure the evaluation snapshot matches the period the scientific claim addresses.
- **Adversarial adaptation:** test targeted omissions without presenting a guessed robustness rate.
- **External adjudication:** ask whether independent reviewers reconstruct the same necessary premises.

Distribution matters: leakage research requires the test population to match the distribution about which a scientific claim is made. (pmc.ncbi.nlm.nih.gov) Historical snapshots and dependent claim variants need deliberate splitting rather than a random allocation that silently changes the target task. (pmc.ncbi.nlm.nih.gov)

Finally, an evidence package can be inspectable without making its verdict correct. Conversely, a correct answer may rest on information unavailable under the trial's rules. The protocol's value is to preserve these distinctions in observable outputs, then test them under conditions where sufficiency has been independently adjudicated.

Conclusion

Evidence is enough when the available package licenses the proposed verdict under a stated claim interpretation, corpus boundary, snapshot and decision rule. The operational question is whether the necessary premises are present, applicable and adequately supported. A confidence score or a larger collection of links cannot answer that question by itself.

The omission test turns this requirement into a reproducible comparison. Establish a sufficient package, remove a necessary premise while preserving fluency, and ask whether the system recognizes the gap. Restore the premise and check recovery. Keep redundant and irrelevant deletions as controls so that blanket abstention is not rewarded. Include complementary passages because the missing step may be a relationship between sources rather than a word inside one passage.

An informative evaluation reports sufficiency judgments, unjustified verdicts, abstention, correctness on sufficient evidence, retrieval completeness and reviewer effort separately. Its denominators must show what was answered, rejected and excluded. Its uncertainty analysis must preserve the dependence between variants of an originating claim.

When evidence remains inadequate, report the actual reason: an absent premise in the supplied package, a documented retrieval gap, blocked access, ambiguous claim scope or unresolved conflict. Stop searching under a declared policy, and preserve the gap rather than converting budget exhaustion into a verdict. The resulting record supports a narrow, defensible conclusion about the evidence state that was actually examined.

Source links

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

1. https://proceedings.mlr.press/v70/guo17a.html
2. https://efcsn.com/code-of-standards
3. https://github.com/copenlu/sufficient_facts/blob/master/data/sufficient_facts/README.md
4. https://www.w3.org/TR/prov-overview
5. https://aclanthology.org/2022.tacl-1.43
6. https://arxiv.org/html/2601.21387v1
7. https://jmlr.org/papers/v11/el-yaniv10a.html
8. https://airc.nist.gov/airmf-resources/airmf/5-sec-core
9. https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00486/112498/Fact-Checking-with-Insufficient-Evidence
10. https://osr.statisticsauthority.gov.uk/guidance/regulatory-guidance-on-intelligent-transparency/pages/2
11. https://www.factcheck.org/our-process
12. https://www.ipcc.ch/site/assets/uploads/2010/07/uncertainty-guidance-note.pdf
13. https://pmc.ncbi.nlm.nih.gov/articles/PMC10499856
14. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-14
15. https://fever.ai/dataset/fever.html
16. https://www.cs.brandeis.edu/%7Ecs140b/CS140b_docs/Artsein_Poesio_2008_Inter-coder_agreement.pdf

17. https://research.google/pubs/data-cards-purposeful-and-transparent-dataset-documentation-for-responsible-ai
18. https://fever.ai/2021/task.html
19. https://irlab.science.uva.nl/wp-content/papercite-data/pdf/zhang-2023-from.pdf
20. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html
21. https://proceedings.mlr.press/v97/geifman19a/geifman19a.pdf
22. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats._result_classes.BinomTestResult.proportion_ci.html
23. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bootstrap.html
24. https://www.stata.com/support/faqs/statistics/bootstrap-with-panel-data
25. https://cris.biu.ac.il/en/publications/user-centric-evidence-ranking-for-attribution-and-fact-verificati
26. https://huggingface.co/datasets/copenlu/sufficient_facts
27. https://aclanthology.org/2022.tacl-1.43.pdf
28. https://www.gofair.foundation/fair-principles

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.