Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Conflicting Sources Fact Checking: What Explanations Must Show

> 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 · 2026-11-02 · conflicting sources fact checking · fact verification · evidence conflict fact checking · explainable fact checking · source reliability · data provenance · audit trail · knowledge fusion · system of record

Original article: https://reliablecontext.com/articles/fact-verifier-explain-when-sources-disagree



In this article

1. Executive Summary
2. Introduction and Background
3. Definitions and Taxonomy of Disagreement
4. Three Benchmarks, Three Units of Conflict
5. A Synthetic Two-Source Protocol: When Conflict Disappears and When It Stays
6. Testing Faithfulness by Removing, Swapping and Re-dating a Span
7. Scoring: Span Identification, Relation Classification, Coverage, Unsupported Resolution and Abstention
8. Data Analysis and Evidence
9. Source Reliability, Provenance and the Record Behind a Value
10. Implications and Future Directions
11. Frequently Asked Questions (FAQs)
12. Conclusion

Executive Summary

When two sources give different values for one claim, a fact verifier can return a single label with a fluent rationale, or it can show a reviewer which source spans support the claim, which contradict it, which qualify it, and which describe a different date, population or definition. This report treats the second behaviour as the object of evaluation. It offers a literature-led protocol for evidence relations, not another benchmark of verdict accuracy, and it sets aside the narrower question of whether a model can elect one value from supplied claims.

Three recent papers define three different units of conflict. Sun, Warren, Shklovski and Augenstein (2026) introduce **CLUE**, a method that identifies span-level claim to evidence and evidence to evidence relations that signal conflict or agreement, and uses them to explain model uncertainty (aclanthology.org). Yuan et al. (2026) release **ConfRAG**, with 1,814 questions, of which 57.2% show explicit contradictions (aclanthology.org); that share describes the authors' dataset, not the web. Nachshoni et al. (2025) build **NatConfQA**, where annotators label each pair of answers as conflicting or non-conflicting (arxiv.org). A span relation, a document-level contradiction and an answer pair are different units, so a protocol must name its unit before it reports a score.

The published evidence is useful but narrow. CLUE's human study had N=12 participants (arxiv.org). On ConfRAG, GPT-4.1 reaches a reason coverage of 0.157 against 0.275 for the human baseline (aclanthology.org). In AVeriTeC, the conflicting-evidence class makes up only 6.4% of training claims (arxiv.org).

The protocol has four requirements. A source-by-source relation matrix (claim, span, observation date, asserted value, scope, relation, dependence) must exist before any verdict. The explanation must change for the right reason when one decisive span is removed, swapped or re-dated. Scoring must separate span identification, relation classification, answer coverage, unsupported resolution and abstention. Results must be

reported per unit, because a generated explanation is not a causal proof of what a model used (ar5iv.labs.arxiv.org) (proceedings.neurips.cc). Two questions stay open: whether an explanation is faithful to the decisive evidence, and whether a disagreement is resolvable from the sources a reviewer may use.

Introduction and Background

Fact verification systems usually end in a label. FEVER (Thorne et al., NAACL 2018), the reference dataset for the task, contains 185,445 claims written by altering Wikipedia sentences (aclanthology.org). Its label set treats a claim as NotEnoughInfo when Wikipedia has no information that supports or refutes it (ar5iv.labs.arxiv.org). That design asks a verifier to choose among three outcomes. It does not ask the verifier to say that two good sources disagree, or that they disagree only because they describe different things.

Real evidence is messier. AVeriTeC, a dataset of 4,568 real-world claims checked by 50 organizations (proceedings.neurips.cc), adds a fourth class for conflicting evidence and cherry-picking (arxiv.org). Its guidelines require evidence documenting the conflict when that label is chosen (arxiv.org), which is the seed of this report: a verdict about disagreement is only as useful as the record behind it. Sun et al. state the explanation gap directly: existing methods express uncertainty as numbers or hedges, without revealing which evidence conflicts cause the uncertainty, which leaves users unable to resolve disagreements (aclanthology.org). The same point applies to question answering, where Nachshoni et al. argue that a response should not only incorporate several answers but also detect the conflicts and communicate them to the reader (arxiv.org).

This report is written for fact-verification and knowledge-fusion researchers who must decide whether a verdict explanation preserves the disagreement that a reviewer still has to resolve. It is a protocol, not a leaderboard. Fixtures in it are synthetic and labeled as such, and every literature figure was read on the cited page on 2 November 2026. The publisher, Reliable Context, builds a system of record for organizations that must defend every value they act on. That is an adjacent category, not a fact verifier or an explanation model, and it reappears only where the record-keeping question arises.

Two scope limits apply. The report does not test whether a language model can elect one value from a set of claims. It also does not treat unresolved conflict as evidence that a claim is false: an unresolved disagreement is a state of the evidence, and the correct output for that state is a visible disagreement.

Definitions and Taxonomy of Disagreement

Precise terms prevent a common evaluation error, which is scoring one kind of disagreement with a metric built for another. The following are working definitions; where a published definition exists, it is cited.

- **Claim.** A proposition with an explicit scope: subject, attribute, value, and the time and population it covers. A claim without scope cannot be checked against a source that has one.
- **Source assertion.** A statement by a named source at an observation date, reproduced verbatim. The W3C PROV model defines provenance as information about the entities, activities and people involved in producing a piece of data or thing (www.w3.org). An entity is a thing with some fixed aspects

(www.w3.org), an activity occurs over a period of time and acts upon entities (www.w3.org), and an agent bears some form of responsibility for an activity taking place (www.w3.org).

- **Evidence span.** The smallest verbatim stretch of a source assertion that carries the value and its qualifiers. SciFact formalizes a similar idea for abstracts: a rationale is a minimal collection of sentences that supports or refutes a claim (ar5iv.labs.arxiv.org).
- **Support.** The span asserts the claim's value for the claim's scope.
- **Contradiction.** Two assertions that cannot both hold for the same scope. NatConfQA's annotators apply this test to answers that cannot be simultaneously true (arxiv.org).
- **Qualification.** The span supports part of the claim while narrowing its scope or adding a condition, and the verdict must keep the condition.
- **Temporal mismatch.** Two assertions are both correct at different dates. Fact-check markup already treats verdicts as time-bound: schema.org's ClaimReview type inherits a property for a publisher who wants to indicate that a review may no longer be relevant (schema.org).
- **Definitional or population mismatch.** Two assertions use different counting rules or populations. Schema.org also lets a review or rating be scoped to a part or facet of the item reviewed (schema.org).
- **Genuine multi-answer case.** Several answers are valid. Wikipedia's neutral point of view policy, a community rather than research source, says that when reputable sources contradict each other and are about equal in prominence, editors should describe both points of view and work for balance (en.wikipedia.org). The same policy warns against giving a false impression of parity (en.wikipedia.org).
- **Dependent source.** A source that copies another. Dong, Berti-Equille and Srivastava observe that some sources copy others, often without proper attribution on the web (lunadong.com), and that voting falls short in the presence of copying (lunadong.com).

Existing label sets draw these lines differently. MultiFC keeps each fact-checking site's own labels, so the number of labels varies by domain from 2 to 27 (ar5iv.labs.arxiv.org). SciFact assigns one label per claim, and its authors note that a claim being both supported and refuted by different abstracts never occurs in the dataset (ar5iv.labs.arxiv.org). The protocol therefore requires an explicit mapping from each dataset's labels to the relations above before any pooled score is reported.

Three Benchmarks, Three Units of Conflict

CLUE, ConfRAG and NatConfQA all concern conflicting evidence and are often cited together. They differ in the unit they score, in where ground truth comes from, in what the experimenter varies and in what the system must output. *Table 1* below summarizes those differences from the papers' own descriptions, with AVeriTeC added as a contrast at the verdict level.

Resource	Unit of conflict	Ground truth	Intervention or task	Output scored
CLUE (Sun et al., 2026)	Span relations between claim and evidence, and between evidence passages (aclanthology.org)	No gold explanations required; 600 instances per dataset (arxiv.org)	Relation-guided prompting against a span-agnostic baseline; needs white-box access (arxiv.org)	Generated uncertainty explanation, scored by an entropy-based faithfulness test and by human ranking

Resource	Unit of conflict	Ground truth	Intervention or task	Output scored
ConfRAG (Yuan et al., 2026)	A question whose retrieved documents form clusters with mutually exclusive claims (aclanthology.org)	Annotators with a third adjudicator (aclanthology.org); a boolean contradiction field per question (huggingface.co)	Answer clustering, answer coverage and reason coverage (aclanthology.org)	Cluster agreement, matched answers and matched reasons
NatConfQA (Nachshoni et al., 2025)	A pair of answers, labeled conflicting or non-conflicting (arxiv.org)	Built from fact-checking data (aclanthology.org); 269 conflicting and 408 non-conflicting instances (arxiv.org)	Include all valid answers and indicate their conflicts	Answer and conflict identification scores
AVeriTeC (Schlichtkrull et al., 2023)	Whole claim, with a fourth verdict class	4,568 claims from 50 fact-checking organizations (proceedings.neurips.cc)	Verdict plus evidence	Verdict score, with a free-text justification (proceedings.neurips.cc)

Table 1 shows why a single "conflict detection" score would be uninterpretable. CLUE evaluates an explanation of a model's own uncertainty, ConfRAG evaluates how a model organizes a document set, and NatConfQA evaluates whether each answer pair is flagged. The protocol scores each unit separately and names the unit in every table header.

What the CLUE evidence supports

CLUE is a single-paper result with a specific scope. It runs on three open-weight models, Qwen2.5-14B-Instruct, OLMo-2-1124-13B-Instruct and Gemma-2-9B-it (arxiv.org), and it samples 600 instances from each of two datasets, HealthVer and DRUID (arxiv.org). Its headline claim is that explanations track model uncertainty more faithfully and align better with decisions than span-agnostic prompting, and that human raters judged them more helpful, more informative, less redundant and more logically consistent (aclanthology.org).

Three details temper that claim. First, the baseline's faithfulness correlations are negative on both datasets and all three backbones (arxiv.org), so on that metric the comparison is against a weak reference. Second, the human study is a ranking exercise: N=12 participants ranked three explanations per item (arxiv.org), and agreement was moderate to low (arxiv.org). Third, the relation labeler is fallible: in a manual check, GPT-4o labeled 129 of 150 span pairs correctly (arxiv.org). The paper also limits itself to conflict-driven uncertainty, noting that real-world uncertainty may stem from other sources, including insufficient evidence (arxiv.org).

What ConfRAG and NatConfQA add

The construction pipeline includes clustering documents into coherent viewpoints (huggingface.co). ConfRAG pairs each of 1,814 real-world questions with an average of 9.58 retrieved paragraphs (aclanthology.org), and 1,037 questions, 57.2% of the total, are labeled contradictory (aclanthology.org). The annotation agreement is a Cohen's kappa of 0.642, which the authors call moderate consistency (aclanthology.org). The three tasks quantify how well a model organizes and explains contradictory content (aclanthology.org). The dataset is hosted on Hugging Face under a card that lists the CC BY 4.0 license (huggingface.co), while the GitHub

repository states that the dataset is distributed for research use only (github.com). The same README says the code is released under the CC-by-4.0 License (github.com). The dataset statements differ, so a team planning reuse should confirm terms with the authors.

NatConfQA derives its conflicting set from fact-checking triplets that contain at least one supporting and one refuting item, which yielded 188 candidate groups before manual annotation (arxiv.org). The annotation process then produced 269 conflicting instances and 408 non-conflicting ones. Within each group, annotators wrote questions and answers and labeled every answer pair as conflicting or non-conflicting.

A Synthetic Two-Source Protocol: When Conflict Disappears and When It Stays

The central test is whether an explanation separates apparent conflict from genuine conflict. Two fixtures illustrate the difference. Both are **(Hypothetical Example)** material built for this report: the companies, plants, numbers and dates are invented, and no result is claimed for any system.

Fixture A: apparent conflict that disappears. The claim is "Fixture Co. had 12,400 employees." Source S1 is an annual report observed on 2026-03-02 stating 12,400 employees at 31 December 2024. Source S2 is a registry profile observed on 2026-09-27 stating 14,900 people at 30 June 2026, of whom 2,300 are contractors. A naive comparison sees 12,400 against 14,900. Aligned for population (14,900 minus 2,300 contractors is 12,600 employees) and for date (18 months apart), the assertions are compatible, and S2 qualifies the claim instead of contradicting it.

Fixture B: genuine conflict that remains. The claim is "Plant P has a rated capacity of 500 units per day." Source S3 is the operator's page observed on 2026-03-02, stating 500 units per day for 2026. Source S4 is a regulator's register observed on 2026-09-20, stating 350 units per day effective 1 January 2026. Source S5 is a trade directory observed on 2026-09-27, stating 500 units per day "as reported by the operator". S3 and S4 use the same definition and overlapping dates, so they contradict. S5 repeats S3 and is a dependent source, so it must not count as a third vote. Nothing in the fixture says whether the operator or the register prevails, so the disagreement is unresolved from the permitted sources.

Table 2 below gives the relation matrix that the protocol requires before any verdict. Its columns follow the minimum a reviewer needs: each cited source with its observation date, the verbatim span, the asserted value, the time and population scope, and the relation to the claim.

Fixture	Source and observation date	Span (synthetic)	Asserted value and scope	Relation to claim	Reviewer note
A	S1, annual report, observed 2026-03-02	"12,400 employees at 31 December 2024"	12,400; payroll employees; 2024-12-31	**Supports**	Scope matches the claim as stated
A	S2, registry profile, observed 2026-09-27	"14,900 people ... including 2,300 contractors, at 30 June 2026"	14,900 total; 12,600 payroll employees after removing contractors; 2026-06-30	**Qualifies** (temporal and population mismatch)	Not a contradiction once aligned; verdict must state both dates

Fixture	Source and observation date	Span (synthetic)	Asserted value and scope	Relation to claim	Reviewer note
B	S3, operator page, observed 2026-03-02	"rated capacity 500 units per day"	500; rated capacity; 2026	**Supports**	Self-reported by the operator
B	S4, regulator register, observed 2026-09-20	"rated capacity 350 units per day, effective 1 January 2026"	350; rated capacity; from 2026-01-01	**Contradicts** S3	Same definition, overlapping dates
B	S5, trade directory, observed 2026-09-27	"500 units per day (as reported by the operator)"	500; derived from S3	**Supports, but dependent on S3**	Adds no independent weight; a 2 to 1 majority is an artifact

Table 2 makes three expectations testable. For Fixture A, a verdict of "claim false, sources disagree" fails because the relation is a qualification. For Fixture B, a verdict of "supported, 2 of 3 sources agree" fails because S5 is not independent. A verdict of "unresolved: 500 per S3, 350 per S4, S5 derives from S3" passes. Provenance standards already supply vocabulary for the S5 relation: PROV defines a derivation as a transformation of an entity into another, or an update of an entity resulting in a new one (www.w3.org), and notes that it is important for secondary sources to reference the primary sources from which they were derived (www.w3.org).

The protocol that follows has five requirements:

- **Matrix first.** The relation matrix is produced and stored before a verdict is generated, and the verdict is a function of the matrix.
- **Verbatim spans.** Every row carries the span text exactly as published and the observation date, so a reviewer can replicate the check. The International Fact-Checking Network Code of Principles asks signatories to provide all sources in enough detail that readers can replicate their work (www.ifcncodeofprinciples.poynter.org).
- **Both directions.** The matrix lists evidence that appears to support the claim and evidence that appears to undermine it (www.ifcncodeofprinciples.poynter.org).
- **Independence flag.** Each row states whether its source derives from another row, because checking key elements against more than one named source only helps if the sources are distinct (www.ifcncodeofprinciples.poynter.org).
- **Declared unresolved state.** When the matrix holds an unresolved contradiction, the output says so and lists the competing values.

Testing Faithfulness by Removing, Swapping and Re-dating a Span

Whether an explanation is faithful differs from whether it reads well. Jacovi and Goldberg define a faithful interpretation as one that accurately represents the reasoning process behind the model (ar5iv.labs.arxiv.org). They argue that faithfulness evaluation should not involve human-provided gold labels (ar5iv.labs.arxiv.org),

that conflating plausibility with faithfulness is harmful (ar5iv.labs.arxiv.org), and that the binary notion should give way to a graded one (aclanthology.org). Human preference for an explanation therefore measures plausibility, and only intervention on the input measures whether the explanation tracks the decision.

Published tests supply the building blocks. ERASER defines comprehensiveness by the idea that a model ought to be less confident once rationales are removed (ar5iv.labs.arxiv.org), and sufficiency by whether the rationale contains enough signal to come to a disposition (ar5iv.labs.arxiv.org). VitaminC supplies the contrastive design for fact verification, with evidence pairs nearly identical in language and content but different in whether they support the claim (ar5iv.labs.arxiv.org).

The protocol applies these ideas to the relation matrix with five interventions, each with an expected change in the explanation and in stated uncertainty:

- **Remove the decisive span.** Delete the S4 span in Fixture B. The conflict should leave the explanation, the verdict should become supported on a single self-reported source, and the explanation should say that independent corroboration is missing. If it still mentions a disagreement, it is unfaithful.
- **Swap the value.** Replace S4's 350 with 500. The relation should flip from contradiction to support, and uncertainty should fall. VitaminC reports that a classifier trained on FEVER flips its verdict on only 56% of contrastive cases (ar5iv.labs.arxiv.org), against 86% after training on contrastive data (ar5iv.labs.arxiv.org), so flip rate is a measurable property.
- **Re-date the span.** Change S2's date in Fixture A to 31 December 2024. The temporal explanation should disappear. S2 then gives 12,600 employees (14,900 minus 2,300 contractors) against S1's 12,400 on the same date, so a 200-person difference should remain once contractors are excluded. The temporal part of the qualification should disappear while the population part still applies, and the explanation should name the removed date gap as the reason. Report the 200-person residual as an unresolved difference, and call it a contradiction only if the fixture states that both figures use the same definition, since a counting-rule difference could also explain it.
- **Add a copy.** Add a second directory that repeats S3. The verdict should not gain confidence, and the explanation should mark the new row as dependent.
- **Paraphrase neutrally.** Reword a non-decisive sentence. Nothing in the verdict or explanation should move.

A pass requires the explanation and the uncertainty to move in the expected direction for the stated reason. CLUE's faithfulness metric follows related logic: its authors adapt a counterfactual test so that inserted tokens that shift predictive entropy should be mentioned in the explanation (arxiv.org). The protocol extends that logic from tokens to spans and relations.

Two cautions apply. A generated explanation is not a causal proof of what a model used: Turpin and colleagues find that chain-of-thought explanations can systematically misrepresent the true reason for a model's prediction (proceedings.neurips.cc). And a model's prior can override span evidence: Xie and colleagues find that models can be highly receptive to external evidence (proceedings.iclr.cc), yet show a strong confirmation bias when external evidence contains information consistent with their parametric memory

(proceedings.iclr.cc), which is one reason a decisive-span test should use fictional or post-training fixtures. White-box measures are also unavailable for closed systems, since CLUE itself needs model internals that closed APIs do not provide (arxiv.org); input-level interventions are the portable part of the protocol.

Scoring: Span Identification, Relation Classification, Coverage, Unsupported Resolution and Abstention

The protocol reports six scores, each with a defined denominator.

- **Span identification.** Precision and recall of the spans an explanation cites against gold decisive spans, per relation. An explanation that cites the right source but the wrong sentence earns no credit.
- **Relation classification.** A per-relation confusion matrix over support, contradiction, qualification, temporal mismatch, population or definition mismatch, unrelated and dependent. Pooling relations hides the failures that matter: on AVeriTeC the conflicting class is rare, and a baseline reached precision of about 10% on it even with gold evidence (arxiv.org). Because human annotators also disagree for legitimate reasons, gold labels should be distributions where annotators diverge, in line with the argument that models should capture the full distribution of plausible human judgments (aclanthology.org).
- **Answer coverage.** The share of distinct defensible answers an output names, penalized for invented ones. A normalization such as matched answers divided by the square root of the product of predicted and gold answer counts is one choice; in a hypothetical case with 3 gold answers, 4 predicted answers and 2 matches, the score is 2 divided by the square root of 12, or 0.577.
- **Unsupported resolution rate.** The share of fixtures with an unresolved gold conflict where the output names one value as correct without any matrix row that licenses the choice. NatConfQA's error analysis is a starting taxonomy: in the authors' manual analysis of 160 sampled responses, 80 contained an error, and in 42% of the cases the authors analysed, which appear to be the erroneous responses, the model simply chose one answer to avoid dealing with the conflict (arxiv.org). The paper does not state the base of that 42% explicitly, and the likeliest reading is a share of the errors found, not a rate across all responses, so it is a taxonomy to borrow, not a baseline for this metric.
- **Abstention and selective accuracy.** Geifman and El-Yaniv motivate selective prediction as reducing the error rate by abstaining when in doubt while keeping coverage high (ar5iv.labs.arxiv.org), and define the risk-coverage curve as risk as a function of coverage (ar5iv.labs.arxiv.org). Kamath and colleagues propose selective question answering under domain shift (ar5iv.labs.arxiv.org), evaluate the area under the risk-coverage curve (ar5iv.labs.arxiv.org), and note that softmax-based abstention fails because models are overconfident on out-of-domain inputs (ar5iv.labs.arxiv.org). In a hypothetical audit of 100 verdicts, a system that answers all 100 and is right on 78 has accuracy 0.78; a system that answers 70, is right on 63 and abstains on 30 has selective accuracy 0.90 at coverage 0.70. Those abstentions count as good only if most land on fixtures whose gold state is unresolved.
- **Counterfactual sensitivity.** The share of interventions in the previous section where explanation and uncertainty both move as expected.

For relation classification, take a hypothetical audit of 200 source pairs with 40 gold qualifications. A system labels 22 correctly, 10 as contradictions and 8 as support, so qualification recall is 22 divided by 40, or 0.55. If it assigns the qualification label to 30 pairs in total, precision is 22 divided by 30, or 0.73. The ten qualification-to-contradiction errors produce false "sources disagree" verdicts, as in Fixture A, and should be reported by name instead of folded into overall accuracy.

Data Analysis and Evidence

The published numbers cannot be pooled, because each belongs to a different unit and construction. This section lists them with their methods and flags what each does not show.

Prevalence figures are dataset properties.

- ConfRAG: 1,037 of 1,814 questions are labeled contradictory, which is 57.2% (aclanthology.org). The questions were selected for the benchmark and documents were retrieved for them, so the share is not an estimate of how often web sources conflict. The Hugging Face viewer lists 1.81k rows (huggingface.co).
- NatConfQA: 269 conflicting and 408 non-conflicting instances, a conflicting share of 39.7% (269 of 677) by construction, because conflicting instances were selected from fact-checking groups.
- AVeriTeC: in the training split, supported, refuted, conflicting and not-enough-evidence claims are 27.6, 56.8, 6.4 and 9.2 percent (arxiv.org).

The spread from 6.4% to 57.2% reflects different definitions and sampling. None of these numbers is a prevalence of conflict on the web, and none should be quoted as one.

Model performance shows a gap to human performance.

- ConfRAG: GPT-4.1 scores 0.466 on cluster agreement, 0.448 on answer coverage and 0.157 on reason coverage (aclanthology.org), against 0.691, 0.510 and 0.275 for the human baseline (aclanthology.org). Computed from those rows, GPT-4.1 reaches 67% of human cluster agreement (0.466 of 0.691), 88% of human answer coverage (0.448 of 0.510) and 57% of human reason coverage (0.157 of 0.275). The weakest component is the explanation of why answers differ, which is the part a reviewer needs most.
- VitaminC: built from over 100,000 Wikipedia revisions that modify an underlying fact (ar5iv.labs.arxiv.org); the 56% to 86% flip rates above come from the same paper.
- Turpin and colleagues report accuracy dropping by as much as 36% on 13 BIG-Bench Hard tasks when models were biased toward wrong answers (proceedings.neurips.cc), a reminder that fluent explanations can sit beside shifted behaviour.

Source-quality figures show that popularity and trust differ. Dong and colleagues (PVLDB 2015) estimate the trustworthiness of 119M webpages (www.vldb.org), applying it to a database of 2.8B facts extracted from the web (www.vldb.org), using a method published in the Proceedings of the VLDB Endowment, Vol. 8, No. 9 (www.vldb.org). The method scores web sources from the facts they provide and not from link popularity, which is one reason a majority of links is not a majority of independent witnesses.

Source Reliability, Provenance and the Record Behind a Value

The secondary questions in this topic, how to evaluate source reliability, what provenance and lineage are, how to resolve conflicting data and what an audit trail for a value should contain, all reduce to what a reviewer can inspect after the verdict.

Evaluating source reliability and credibility

Fact-checking practice gives concrete criteria. The IFCN Code of Principles asks fact-checkers to identify the source of all significant evidence (www.ifcncodeofprinciples.poynter.org), to use the best available primary, not secondary, sources (www.ifcncodeofprinciples.poynter.org), to identify the relevant interests of the sources they use (www.ifcncodeofprinciples.poynter.org) and to apply the same high standards to evidence on equivalent claims regardless of who made the claim (www.ifcncodeofprinciples.poynter.org). Wineburg and McGrew's study compared how professional fact checkers, historians and first-year college students evaluated online information (purl.stanford.edu), which shows that source evaluation is a practice that can be studied; the protocol therefore records evaluation steps explicitly. In their Knowledge-Based Trust paper, Dong and colleagues treat agreement among extractors as a signal of reliability (www.vldb.org), and Dong and colleagues also show that voting falls short in the presence of copying (lunadong.com). That is why the independence flag matters: reliability estimated from agreement among copies is circular.

Provenance and lineage

Provenance supports this check. The W3C PROV Data Model says provenance can be used to form assessments about quality, reliability or trustworthiness (www.w3.org). Cheney, Chiticariu and Tan describe why-provenance as the source tuples that witness the existence of an output tuple (homepages.inf.ed.ac.uk), and note that provenance adds value to data by explaining how it was obtained (homepages.inf.ed.ac.uk). When sources give conflicting accounts of provenance itself, PROV observes that users need to determine the provenance of provenance (www.w3.org). For a verdict, lineage means the matrix rows plus the chain from each span to its original publisher.

Fact-check markup shows what a single verdict can carry

Schema.org defines ClaimReview as a fact-checking review of claims made in some creative work (schema.org). Google's documentation notes that fact-checking projects use a variety of rating schemes with subtle differences (developers.google.com), and that the textual rating is the only value shown in search results (developers.google.com). Its required rating carries a short truthfulness label (developers.google.com), the claim is summarized in a short text (developers.google.com), and the structured data must not mismatch the page content (developers.google.com). Schema.org adds that reviews can themselves be rated (schema.org). The same documentation requires the claim to be attributed to a distinct origin (developers.google.com) and the analysis to be traceable and transparent about sources and methods (developers.google.com). These are sound requirements for a published verdict. The documentation reviewed here does not describe how two verdicts that disagree are represented side by side, which illustrates the gap that a relation matrix fills.

Audit trail for data values

An audit trail should let a reviewer reproduce what was believed at a past date and what changed since. In the United States, 21 CFR 11.10(e) is one concrete statement of the property for electronic records in a regulated scope: record changes shall not obscure previously recorded information (www.ecfr.gov), and audit trails are secure, computer-generated and time-stamped (www.ecfr.gov). This is a US federal rule cited only as an example of the engineering property, not as guidance on any compliance question. NIST's AI Risk Management Framework, released on 26 January 2023 (www.nist.gov), is intended for voluntary use (www.nist.gov) and states more generally that accountability presupposes transparency (nvlpubs.nist.gov); as of the NIST page read on 2 November 2026, version 1.0 is being revised (www.nist.gov).

What a system of record for data conflicts must preserve

Table 3 below compares ways of representing a disagreement to a reviewer, including a system of record, which is the category the publisher operates in. The rows describe what each approach lets a reviewer see, not which is better; they answer different questions.

Approach	What a reviewer can see	What it hides or cannot show	Evidence base
Single verdict label	One rating from a rating scheme (developers.google.com), with a short truthfulness label (developers.google.com)	Which sources disagree, and why	Label sets such as FEVER's NotEnoughInfo class (ar5iv.labs.arxiv.org)
Numeric confidence or hedge	A scalar or hedging phrase	Which evidence conflicts cause the uncertainty (aclanthology.org)	Limits stated by the CLUE authors
Span-relation explanation (CLUE-style)	Spans linked by agree, disagree or unrelated relations	Causal faithfulness; requires white-box access	Single-paper result on two datasets
Enumerated answers with conflict flags (NatConfQA style)	All valid answers and which pairs conflict	Source-level provenance and dates, unless added	Benchmark with pair-level labels
System of record (Reliable Context)	One canonical record per real thing; where sources disagree, the ambiguity is shown rather than hidden; a verbatim cited source and an observation date behind every value; unknowns declared with a reason; past states reproducible	It is not a fact verifier, an explanation model or an evaluation harness, and it does not by itself show that a downstream explanation is faithful. It is also not a vector database, RAG framework, data catalog, MDM or data-quality suite, CRM or chatbot; readers who need one of those should use one	Vendor description by the publisher; no independent benchmark or evaluation is cited in this report

Table 3 places the system of record where it belongs: it addresses the record, not the explanation. A verifier can be tested with the protocol above, and a record can be checked for whether each value carries its cited source and observation date, whether unknowns are declared, and whether a past state can be reproduced. The publisher describes Reliable Context as running inside the customer's perimeter with one database per

organization; this is a vendor description that the report has not independently verified. The two checks are complementary, since a faithful explanation of an incomplete record is still an explanation of an incomplete record.

A resolution ladder for conflicting data

For the practical question of how to resolve conflicting data from different sources, the matrix suggests a fixed order of operations:

- **Align scope.** Convert dates, populations and definitions to a common basis before comparing values, as in Fixture A.
- **Check independence.** Collapse copies and derivations into their origin, as with S5.
- **Check authority by attribute.** Decide in advance which source type is authoritative for which attribute; a rule written after the conflict is a preference, not a method.
- **Check recency.** Compare observation dates and effective dates separately.
- **Escalate or abstain.** If the conflict survives, return it with both values and the reason it is unresolved. Wikipedia's policy for contested assertions is a useful analogy: attribute them as opinions rather than state them as facts (en.wikipedia.org), and present conflicting findings in a disinterested tone (en.wikipedia.org).

Implications and Future Directions

Failure cases the protocol should expose. An explanation can cite the correct sources but assign the wrong relation, for example reading a qualification as a contradiction. It can assign the correct relation yet be unfaithful, as the faithfulness literature warns (proceedings.neurips.cc). It can inherit labeler errors, since CLUE's pipeline relies on a relation labeler that was right on 129 of 150 sampled pairs, as reported above. It can reflect rater limits: a study with N=12 non-specialist raters and moderate to low agreement cannot settle expert utility. And it can be unresolvable by design, because the permitted sources lack the rule that settles the case.

Unanswered transfer questions. CLUE was evaluated on two datasets and three models of comparable size. Whether span-relation explanations transfer to long enterprise documents, closed models, non-English sources or larger evidence sets is not established by the papers. ConfRAG isolates the generator, so it says little about retrieval-stage disagreement. A faithful explanation of relations in one corpus does not show that a verifier handles another definition of conflict.

What a research group can do next. Build a small, openly licensed fixture set in which the relation, the decisive span and the permitted-source resolution are labeled by at least two annotators, report agreement per relation, and publish intervention outputs and not only aggregate scores. Treat "unresolved from permitted sources" as a first-class gold label, because systems that always resolve cannot be tested without it.

For practitioners. Do not infer that an unresolved conflict means a claim is false, and do not infer that agreement among several links means independent confirmation. Both inferences substitute a count for a relation.

Frequently Asked Questions (FAQs)

How should a verifier handle conflicting sources in fact checking? Produce the relation matrix first, align scope, flag dependent sources and return a verdict that states the relations. The IFCN Code asks for both supporting and undermining evidence (www.ifcncodeofprinciples.poynter.org).

How can conflicting data from different sources be resolved? Follow the ladder: scope, independence, authority by attribute, recency, then escalation or abstention. Reliability estimated from agreement alone is vulnerable to copying (lunadong.com).

Is a conflicting-evidence verdict the same as a false claim? No. A conflicting verdict states an unresolved state of the evidence, and AVeriTeC's fourth class exists for conflicting evidence and cherry-picking.

Does a generated explanation prove what the model used? No. Faithfulness requires intervention tests, and explanations can misrepresent the true reason for a prediction (proceedings.neurips.cc).

Is 57.2% the share of web questions with conflicting sources? No. It is the share of questions in the ConfRAG authors' dataset that exhibit explicit contradictions (aclanthology.org).

Conclusion

A single verdict can be correct and still hide what a reviewer has to resolve. The literature offers three useful units, span relations in CLUE, document-level contradiction in ConfRAG and answer pairs in NatConfQA, and none alone tells a reviewer whether an explanation preserved the disagreement. The protocol here fills that gap with a relation matrix that precedes the verdict, interventions that remove, swap, re-date or duplicate a decisive span, and scores that separate span identification, relation classification, answer coverage, unsupported resolution and abstention.

The evidence base is small. CLUE is a single-paper result with a small human ranking study, ConfRAG's contradiction share describes its own dataset, and NatConfQA's conflicting share is set by construction. A generated explanation is not a causal proof, and a majority of sources is not a majority of independent witnesses. The two open questions, whether an explanation is faithful to the decisive evidence and whether a disagreement is resolvable from permitted sources, can be tested with fixtures of the kind described here.

Source links

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

1. https://aclanthology.org/2026.acl-long.2110
2. https://aclanthology.org/2026.acl-long.11
3. https://arxiv.org/html/2508.12355v1
4. https://arxiv.org/html/2505.17855v2
5. https://aclanthology.org/2026.acl-long.11.pdf
6. https://arxiv.org/html/2305.13117
7. https://ar5iv.labs.arxiv.org/html/2004.03685

8. https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html
9. https://aclanthology.org/N18-1074
10. https://ar5iv.labs.arxiv.org/html/1803.05355
11. https://proceedings.neurips.cc/paper_files/paper/2023/hash/cd86a30526cd1aff61d6f89f107634e4-Abstract.html
12. https://www.w3.org/TR/prov-overview
13. https://www.w3.org/TR/prov-dm
14. https://ar5iv.labs.arxiv.org/html/2004.14974
15. https://schema.org/ClaimReview
16. https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view
17. https://lunadong.com/publication/lifespan_vldb.pdf
18. https://ar5iv.labs.arxiv.org/html/1909.03242
19. https://huggingface.co/datasets/OracleY/ConfRAG
20. https://aclanthology.org/2025.uncertainlp-main.13
21. https://arxiv.org/html/2505.17855v1
22. https://github.com/XaiverYuan/ConfRAG
23. https://www.ifcncodeofprinciples.poynter.org/know-more/the-commitments-of-the-code-of-principles
24. https://aclanthology.org/2020.acl-main.386
25. https://ar5iv.labs.arxiv.org/html/1911.03429
26. https://ar5iv.labs.arxiv.org/html/2103.08541
27. https://proceedings.iclr.cc/paper_files/paper/2024/hash/99261adc8a6356b38bcf999bba9a26dc-Abstract-Conference.html
28. https://aclanthology.org/Q19-1043
29. https://ar5iv.labs.arxiv.org/html/1705.08500
30. https://ar5iv.labs.arxiv.org/html/2006.09462
31. https://www.vldb.org/pvldb/vol8/p938-dong.pdf
32. https://purl.stanford.edu/yk133ht8603
33. https://homepages.inf.ed.ac.uk/jcheney/publications/provdbsurvey.pdf
34. https://developers.google.com/search/docs/appearance/structured-data/factcheck
35. https://www.ecfr.gov/current/title-21/chapter-I/subchapter-A/part-11/subpart-B/section-11.10
36. https://www.nist.gov/itl/ai-risk-management-framework
37. https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

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.