Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Comparing Truth Discovery Methods Without Gold Labels

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

Published by Reliable Context · 2026-09-22 · truth discovery without ground truth · truth discovery evaluation · data fusion · source reliability · comptruthhyp · synthetic benchmarks · latent truth inference · confidence calibration · provenance

Original article: https://reliablecontext.com/articles/compare-truth-discovery-without-gold-labels



In this article

1. Executive Summary
2. Introduction and Background
3. Define the Four Evaluation Targets
4. Controlled Synthetic Stress Testing
5. Sparse Independent Adjudication
6. Observation-Likelihood Ranking Without Gold Labels
7. Feature Comparison
8. Performance and Benchmarks
9. Data Analysis and Evidence
10. A Go or No-Go Protocol
11. Implications and Future Directions
12. Frequently Asked Questions (FAQs)
13. Conclusion

Executive Summary

Truth discovery estimates facts and source reliability from conflicting claims, often by letting estimated facts and estimated source quality reinforce one another. That makes evaluation circular when the same source consensus supplies both the output and the answer key. The correct response is not to find one substitute score. It is to separate four targets: **fact recovery**, **source ranking**, **confidence calibration**, and **computational behavior**. Fang et al. compare **12 methods** and report that no method consistently wins across all investigated scenarios (researchers.mq.edu.au). This is an empirical literature finding, not a theorem that every new method will fail.

Three evidence tiers answer different questions. First, controlled synthetic tests expose methods to known changes in coverage, conflicts, source accuracy, dependence, participation tails, and truth cardinality. Waguih and Berti-Equille evaluated **12 algorithms**, supplied reference implementations, generated **9,120 datasets**, and used **five real-world datasets** (arxiv.org) (researchgate.net). Synthetic evidence can eliminate brittle methods, but performance depends on generator realism.

Second, a sparse, independently adjudicated probability sample is the only tier here that directly estimates real-task correctness. The reference process must not use source popularity or candidate outputs. FDA guidance gives the general statistical warning: comparison with a non-reference standard does not yield true performance, and its reference definition excludes the candidate test outcome (fda.gov). Use paired bootstrap intervals, commonly **500 to 1,000 resamples** in NIST's description, rather than selecting from point estimates (itl.nist.gov).

Third, CompTruthHyp ranks candidates by the probability they assign to observed claims (researchers.mq.edu.au). It usefully reduced a no-label candidate set in tested conditions and identified the top three methods on two real datasets (opus.lib.uts.edu.au). It does **not** prove that the highest-likelihood output is factually true, especially when source independence is false. The operational rule is therefore: reject

candidates that fail controlled stress tests, retain stable candidates supported by proxy evidence, and require independent adjudication whenever the decision needs a claim about factual accuracy or calibrated confidence.

Introduction and Background

Truth discovery is the family of methods that infers latent facts and source quality from multiple, possibly conflicting observations. A simple baseline chooses the mutually exclusive claim supported by the most sources, while iterative and probabilistic methods model trust, error, difficulty, or dependence. Pasternack and Roth describe the baseline plainly as taking a vote (aclanthology.org); Zhao et al.'s Latent Truth Model instead jointly infers records and source quality without supervision (arxiv.org).

The attraction of unsupervised fusion is also its evaluation problem. If method A calls a source reliable because it agrees with A's inferred facts, then an evaluator cannot use that same agreement to establish that A recovered reality. Independent axiomatic analysis warns that such procedures can find source consensus rather than truth (link.springer.com). Source agreement is weaker still when claims are copied, because a false value can propagate through dependent sources (vldb.org).

This report compares three evidence strategies rather than declaring one universal benchmark: controlled simulation, sparse independent adjudication, and observation-likelihood ranking. It addresses unsupervised truth-discovery method comparison, evaluation metrics, source reliability without labels, multi-source data fusion, benchmarking, circular evaluation, and latent-truth inference. The decision is whether evidence is strong enough to **reject**, **retain for a pilot**, or call a comparison **undecidable**.

Define the Four Evaluation Targets

A method can perform well on one target and fail another. Before constructing a benchmark, the evaluator should write four estimands, meaning the exact quantities to be estimated.

- **Recovered facts:** For single-valued objects, accuracy or error rate may be appropriate. For multi-valued objects, report precision and recall because returning one correct value can still omit other true values.
- **Source ranking:** Compare estimated source order with an external source-quality criterion. Fang et al. treat value confidence and source trustworthiness as distinct elements, which is reason enough not to collapse them into one score (opus.lib.uts.edu.au).
- **Confidence calibration:** A confidence of 0.8 is calibrated only if outcomes assigned that probability occur at the claimed rate over comparable cases. Calibration is consistency between predictive distributions and observations (onlinelibrary.wiley.com). It therefore requires outcomes independent of the candidate's own inference.
- **Computational behavior:** Record runtime, peak memory, convergence failures, iterations, and scaling with objects, claims, sources, and candidate values. The 2014 benchmark treated initialization, parameter settings, convergence, and scalability as separate concerns (researchgate.net).

This decomposition prevents a common category error. A likelihood score can compare model fit without identifying factual accuracy. A synthetic accuracy score can identify recovery under the generator without proving transfer to production. A real-label accuracy estimate can establish sampled correctness without proving runtime suitability or probability calibration.

Confidence metrics also need specification. The binary **Brier score** is the mean squared difference between predicted probability and outcome; the European Centre for Medium-Range Weather Forecasts states its range as **0 for perfect to 1 for consistently wrong** (confluence.ecmwf.int). Reliability diagrams should show binning and uncertainty, because wider bins average away random noise at the cost of resolution (jmlr.org).

Controlled Synthetic Stress Testing

Capabilities

Synthetic evaluation makes the answer key known by construction and changes one data-generating factor at a time. Waguih and Berti-Equille's generator varied source coverage, conflict counts and distributions, and true-positive claims (researchgate.net). Its exponential-coverage condition represented a long tail in which few sources cover most items (researchgate.net). The framework measured algorithm quality with four metrics, not one (arxiv.org).

A stronger modern generator should cross at least six axes:

- **Coverage:** uniform, sparse, clustered, and long-tail claims per source.
- **Source count:** enough levels to reveal both statistical and computational scaling.
- **Conflict:** number of candidate values, disagreement rate, and ambiguity structure.
- **Dependence:** independent sources, copying chains, and shared upstream feeds.
- **Source accuracy:** homogeneous, bimodal, continuous, and accuracy correlated with coverage.
- **Truth cardinality:** single-valued, multi-valued, and occasionally no true value among candidates.

The dependence axis is an extension, not a feature of the original generator. Its authors explicitly excluded dependence and value similarity from scope (researchgate.net). This omission matters because ordinary voting can fail in the presence of copying (vldb.org).

Strengths and limitations

Synthetic tests provide **causal control over benchmark difficulty**. If a method's accuracy collapses only when dependence increases, that failure is interpretable. They also reveal numerical limits: the 2014 implementation reported Bayesian underflow for maximum-likelihood estimation and Latent Credibility Analysis beyond **5,000 sources** (researchgate.net). Runtime estimates were averages over **10 executions**, a better basis than one timing run (researchgate.net).

The limitation is external validity. A generator encodes assumptions about error, independence, participation, and whether the truth appears among candidates. One benchmark shows that candidate sets can exclude the true answer offered by all sources (cs.illinois.edu). A method can therefore win a synthetic world because its assumptions resemble the generator, not because it is best for the target corpus.

Use in selection

Use synthetic evidence primarily to **eliminate** candidates. Reject a method that fails a plausible operating region, produces rank reversals across seeds, or cannot finish within resource constraints. Retain a method that is robust across plausible regions, but do not call it accurate on real data until independent labels support that statement.

Sparse Independent Adjudication

Capabilities

Independent adjudication samples real objects, establishes reference outcomes without looking at source popularity or candidate results, and evaluates every candidate on the same cases. This is the strongest available evidence for factual performance on the sampled population. Probability sampling gives every population unit a known, nonzero selection chance (statcan.gc.ca), enabling design-based estimates and weights when inclusion probabilities differ.

The sampling frame should stratify by variables known before adjudication: source coverage, conflict count, object type, value cardinality, recency, and operational importance. Oversample rare but consequential strata, preserve inclusion probabilities, and report both weighted overall results and stratum results. Keep an explicit **unresolved** label instead of forcing adjudicators to manufacture truth.

Strengths and limitations

This tier directly measures sampled accuracy, precision, recall, ranking agreement, and calibration. It also exposes label uncertainty through blinded duplicate review and adjudicator agreement. Yet small labels can be misleading. In the 2014 study, the Book gold standard covered **100 objects, 7.91%** of the dataset, while the Biography gold standard covered **2,626 values, 0.069%** (researchgate.net). Coverage alone does not determine validity, but tiny or convenient subsets create wide intervals and selection risk.

Knowledge Vault provides another caution. It capped predicates at **10,000 instances**, described its local closed-world labels as an approximation to truth, and discarded **305 triples** adjudicators could not label (cs.ubc.ca). Excluding unresolved items changes the evaluation population and should be reported.

Use in selection

Pre-register the sampling and decision rule. Compute paired differences because every method is scored on the same objects. Bootstrap **objects**, not individual claims, to preserve within-object dependence. Report percentile or bias-corrected intervals and the number of effective sampled objects. NIST notes that uncertainty for a wider item population exceeds uncertainty for a fixed benchmark (nvlpubs.nist.gov). If the interval on a paired difference crosses the smallest operationally meaningful improvement, declare the accuracy comparison undecidable.

Observation-Likelihood Ranking Without Gold Labels

Capabilities

CompTruthHyp is the central no-gold proxy in this comparison. Fang et al. calculate the probability of observed claims under each method's output and rank that probability as a performance signal (digital.library.adelaide.edu.au). The study compared **12 methods** across single-valued and multi-valued objects (arxiv.org) (researchers.mq.edu.au). Its synthetic evaluation used **eight groups of datasets** (opus.lib.uts.edu.au).

The proxy is valuable when labels are absent because it can remove candidates that explain the observations poorly under a common likelihood model. In one Full-Optimistic synthetic setting, its ranking matched the complete-ground-truth precision ranking with Euclidean distance **0.000** and cosine similarity **1.000** (arxiv.org). On two real datasets, the authors report consistent identification of the top three methods.

Strengths and limitations

The crucial distinction is **fit to observations versus truth about the world**. CompTruthHyp interprets larger likelihood as a greater ability to capture dataset features (opus.lib.uts.edu.au). That is a legitimate model-comparison signal, but it does not observe truth. Proxy-scoring theory likewise says immediate scoring without resolved outcomes requires an appropriate proxy (ojs.aaai.org). Appropriateness is an assumption to test, not an automatic property.

The method assumes independence between sources and, in part of its formulation, independent value provision (opus.lib.uts.edu.au). The paper names neglected source-copying relations as a limitation (arxiv.org). A later review still concludes that more research is needed on no-gold comparison (link.springer.com).

Use in selection

Use observation likelihood to rank or screen candidates for a pilot, never as the sole evidence for a statement such as “method A is 90% accurate.” Check whether proxy rankings persist after de-duplicating likely source families, withholding high-coverage sources, perturbing initialization, and changing plausible hyperparameters. A rank reversal under these operations is evidence of fragility. Stability does not convert the proxy into ground truth, but instability can eliminate a candidate.

Feature Comparison

Table 1 compares what each evidence strategy identifies. The final row represents a record substrate that preserves evidence after selection, not an algorithmic substitute for the three tests.

Option	Labels required	Identifies factual accuracy?	Identifies robustness?	Main failure mode	Selection role
Controlled synthetic stress test	Complete generated labels	Only inside the simulated distribution	Yes, for varied factors	Generator assumptions favor a model family	Reject brittle candidates

Option	Labels required	Identifies factual accuracy?	Identifies robustness?	Main failure mode	Selection role
Sparse independent adjudication	Small real-world probability sample	Yes, for the sampled target population with uncertainty	Partly, when stratified	Label selection bias or reference leakage	Choose when intervals separate methods
CompTruthHyp-style observation likelihood	None	No, it measures fit to observed claims	Partly, through perturbation tests	Misspecified independence and consensus fit	Rank candidates for pilot
Provenance record practices	Not an evaluation method	No algorithmic accuracy claim	Preserves evidence for review and replay	Cannot replace benchmark labels or model comparison	Record provenance about the entities, activities, and people involved in producing data

The rows are complementary, not a league table. Synthetic tests establish controlled failure boundaries. Independent labels estimate real-task outcomes. Observation likelihood supplies a no-label screening signal. Provenance information can be used to form assessments about quality, reliability, or trustworthiness, but it does not turn retained evidence into an accuracy benchmark.

Performance and Benchmarks

Performance comparison should use a repeated matrix, not one favored configuration. For each candidate, run every combination of dataset, factor cell, initialization, seed, and pre-declared hyperparameter setting. Retain failures and timeouts in the denominator.

The 2014 evidence shows why. Under random truth allocation, tested methods differed by no more than **2 percentage points** in precision (researchgate.net). Parameters tuned for Book precision did not maximize precision elsewhere (researchgate.net). Across five real datasets, the study found no clear aggregate benefit over majority voting after considering precision and runtime (researchgate.net). These are results of one experimental study, not permanent rankings.

Stability belongs beside accuracy. The benchmark found repeatability issues for Latent Truth Model and 3-Estimates (researchgate.net). Dawid-Skene-style latent label models add related concerns: the ordinary estimator maximizes a non-convex likelihood (jmlr.org), and high-probability recovery of all labels may be impossible under sparse sampling (jmlr.org).

Table 2 defines the stability report to publish for every candidate.

Axis	Minimum report	Evidence of concern	Decision consequence
Seeds	Distribution across repeated seeds	Wide metric or rank spread	Retain uncertainty, do not report best seed
Initialization	All documented initial states	Different fixed points or failures	Reject if production initialization is uncontrolled

Axis	Minimum report	Evidence of concern	Decision consequence
Hyperparameters	Pre-declared plausible grid	Winner changes across nearby settings	Require nested tuning or call undecidable
Convergence	Iterations, tolerance, non-convergence rate	Results depend on stop rule	Compare only converged runs and report failures
Scale	Runtime and peak memory against claims, sources, objects	Superlinear growth or numerical failure	Restrict operating envelope or reject
Source perturbation	Remove high-coverage and related-source clusters	Proxy or truth rank reverses	Investigate dependence before selection

The table turns “stable” into inspectable evidence. No universal cutoff is defensible across tasks. Teams should set an operational tolerance before viewing results, then publish full distributions and rank reversals rather than inventing a generic threshold.

Data Analysis and Evidence

The strongest quantitative lesson is that a single point score is insufficient. Suppose 400 independently adjudicated objects are evaluated by methods A and B. A is correct on 332 objects, **83.0%**; B is correct on 324, **81.0%**. The apparent lead is **2.0 percentage points**, but the relevant statistic is the paired per-object difference, not two independent proportions. Resample the 400 objects with replacement **1,000 times**, recompute the paired difference, and report its interval. NIST describes bootstrap resampling as typically repeated **500 to 1,000 times** (itl.nist.gov). If the interval includes zero or the pre-declared practical threshold, the data do not select a winner.

For source rankings, use rank correlation against independently estimated source outcomes, but include uncertainty because sources with few sampled claims have noisy rates. For confidence, report Brier score, reliability bins, and sample counts per bin. For multi-valued facts, report micro and macro precision, recall, and F1 so high-volume objects do not conceal low-volume failures. For runtime, show medians and upper quantiles across identical hardware and record failed runs.

Table 3 is a worked example of circular evaluation (Hypothetical Example). There are five sources and one object. Three related sources report X, while two independent sources report Y. Independent adjudication later establishes Y.

Evaluator	Answer key	Method choosing X	Method choosing Y	Valid inference
Consensus as gold	X, because 3 of 5 sources report it	100%	0%	Rewards reproduction of consensus
Observation fit	Claims themselves	Likely favored if dependence is ignored	Likely penalized	Compares fit under assumptions, not truth
Independent adjudication	Y, established without candidate outputs	0%	100%	Directly evaluates this object's correctness

The calculation is intentionally simple: consensus scores the consensus-reproducing method as perfect even though the independent answer disagrees. Copying is the mechanism that makes three observations less than three independent pieces of evidence. Dong et al. describe source-dependence and truth estimation as interdependent (vldb.org). The example does not prove Y without adjudication; it proves only that consensus cannot serve simultaneously as candidate and referee.

Existing studies also show how rankings move with evaluation design. Fang et al.'s two displayed sparse-label experiments did not produce a completely correct ranking until label coverage reached **60%** (arxiv.org). In a separate field, changing the aggregation mechanism altered the state-of-the-art conclusion in about **30% of 296 setups** (aclanthology.org). These numbers do not supply universal thresholds. They quantify sensitivity in specific experiments and justify reporting uncertainty and analysis choices.

A Go or No-Go Protocol

The following protocol converts the evidence tiers into a falsifiable selection procedure.

1. **Define outputs.** State whether the method must recover one value, several values, abstain, rank sources, or emit probabilities.
2. **Freeze estimands.** Specify fact metrics, source-ranking metrics, calibration metrics, runtime limits, and the smallest practically important difference.
3. **Build the claim graph.** Identify likely shared feeds and copying families before scoring. Source count is not evidence count when origins overlap.
4. **Run synthetic factor cells.** Cross coverage, source count, conflict, dependence, long-tail participation, source accuracy, and truth cardinality. Include cases in which no offered candidate is true.
5. **Repeat every stochastic run.** Vary seed, initialization, source order, and plausible hyperparameters. Record convergence and all failures.
6. **Apply the no-gold proxy.** Compute observation-likelihood ranking on the unlabelled corpus, then repeat after source-family and high-coverage-source perturbations.
7. **Draw an independent holdout.** Use probability sampling or record inclusion probabilities. Blind adjudicators to source popularity, candidate names, scores, and outputs.
8. **Quantify uncertainty.** Use paired object-level bootstrap intervals, show sample sizes by stratum, and retain unresolved adjudications as a reported outcome.
9. **Check calibration.** Score probabilities only against independently adjudicated outcomes and disclose binning. Do not calibrate against inferred truths.
10. **Make one of three decisions.** Reject, retain for pilot, or undecidable. A pilot is not evidence of factual superiority unless its outcomes are independently measured.

The operational rules are strict. **Reject** a candidate that fails a plausible stress cell, exceeds resources, fails convergence, or changes materially under benign perturbations. **Retain for pilot** candidates that survive stress tests and rank consistently under proxies, while labeling factual accuracy unknown. **Select** among

retained methods only when independent labels and uncertainty support the pre-declared practical difference. Otherwise mark the comparison **undecidable** and collect targeted labels.

Evidence custody should be designed with the evaluation. W3C PROV was designed to enable interoperable provenance exchange on the Web (w3.org). Provenance is information about entities, activities, and people involved in producing a piece of data or thing. That supports review and replay of selected values, while benchmark validity remains the evaluator's responsibility.

Implications and Future Directions

Four unresolved problems should shape new work.

- **Distribution realism:** Synthetic generators need empirical audits showing that coverage, conflict, dependence, and error structures resemble the target corpus. Knowledge Vault combined text, tables, page structure, and human annotations, illustrating how heterogeneous real evidence can be (research.google).
- **Hidden dependence:** Copying, shared feeds, and common extraction pipelines violate naive independence. Dawid-Skene identification likewise assumes annotator responses are conditionally independent given the latent label (proceedings.neurips.cc).
- **Label selection bias:** Adaptive allocation can preferentially send workers to easy tasks, biasing the observed sample (nature.com). Probability sampling, recorded inclusion weights, and unresolved labels reduce but do not erase this problem.
- **Non-identifiability:** Latent truths and source reliabilities may not be uniquely recoverable without structural assumptions. Pairwise co-occurrence results identify parameters only under explicit conditions (proceedings.neurips.cc). Unlabelled agreement cannot validate those conditions by itself.

Future benchmarks should publish generators, seeds, configurations, failure logs, adjudication protocols, inclusion probabilities, and retained evidence. Temporal metadata also matters because sources and facts change; the TDGT benchmark, for example, annotates ground-truth records with temporal information (webdatacommons.org). Better artifacts will not abolish the need for labels. They will make the boundary between measured fact, model assumption, and unresolved uncertainty visible.

Frequently Asked Questions (FAQs)

How can truth discovery be evaluated without ground truth?

It can be evaluated partially. Synthetic labels measure behavior under controlled assumptions. Observation likelihood measures fit to observed claims. Stability tests measure sensitivity. None directly establishes real-world factual accuracy. A sparse, independent holdout is required for that claim.

What is the best unsupervised truth-discovery method?

The literature does not support a universal winner. Fang et al. report no consistent leader across all investigated scenarios (researchers.mq.edu.au). Method choice should follow the target's cardinality, dependence, coverage, abstention needs, calibration requirements, and resource envelope.

Can source reliability be estimated without labels?

Yes, conditionally. Dawid and Skene estimate observer error rates even when the true response is unavailable (econpapers.repec.org). Later consistency results remain conditional on independence assumptions (proceedings.mlr.press). The estimate is model-dependent, not independently verified reliability.

Which metrics should a benchmark report?

Report fact accuracy or precision and recall, source-rank agreement, Brier score and reliability plots for probabilities, convergence rate, runtime, memory, and stability distributions. Keep the targets separate and include paired uncertainty intervals.

When is a small adjudicated sample still required?

It is required whenever the decision must support a claim about factual accuracy, real-world error, or confidence calibration. It is also required when proxy rankings disagree, source dependence is substantial, or the cost difference between candidates is material enough to require a defensible winner.

Conclusion

Truth-discovery evaluation without comprehensive gold labels is possible, but only as a layered argument. Controlled synthetic data can reveal failure boundaries. Sparse independent adjudication can estimate correctness for a defined population. Observation likelihood can rank how well candidates explain observed claims. Stability and scaling tests can determine whether a method behaves reproducibly within an operating envelope.

The layers are not interchangeable. A no-gold proxy does not establish factual truth, a synthetic win does not establish transfer, and a small convenience holdout does not establish population performance. The defensible comparison pre-defines four targets, stresses plausible data regimes, repeats stochastic runs, records failures, checks dependence, samples real items independently, and reports paired uncertainty.

The final decision vocabulary should remain narrow: **reject**, **retain for pilot**, or **undecidable**. Select a factual winner only when an independent sample separates candidates by a pre-declared meaningful amount. Where that evidence is absent, the honest result is not that the highest-consensus or highest-likelihood method is true. It is that the available evidence supports a pilot, identifies a failure, or leaves the choice unresolved.

Source links

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

1. https://researchers.mq.edu.au/en/publications/from-appearance-to-essence-comparing-truth-discovery-methods-with

2. https://arxiv.org/html/1409.6428
3. https://www.researchgate.net/publication/265972324_Truth_Discovery_Algorithms_An_Experimental_Evaluation_TechReport_May_2014_httparxivorgabs14096428
4. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/statistical-guidance-reporting-results-studies-evaluating-diagnostic-tests-guidance-industry-and-fda
5. https://itl.nist.gov/div898/handbook/eda/section3/bootplot.htm
6. https://opus.lib.uts.edu.au/bitstream/10453/142029/2/From%20Appearance%20to%20Essence%20Comparing%20Truth%20Discovery%20Methods%20without%20Using%20Ground%20Truth.pdf
7. https://aclanthology.org/C10-1099.pdf
8. https://arxiv.org/abs/1203.0058
9. https://link.springer.com/article/10.1007/s10458-022-09569-3
10. https://www.vldb.org/pvldb/vol2/vldb09-pvldb47.pdf
11. https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2007.00587.x
12. https://confluence.ecmwf.int/spaces/FUG/pages/673551584/Section%2B8.3.5%2BUsing%2Bverification%2Bmetrics%2Bwith%2Bthe%2Boutput
13. https://www.jmlr.org/papers/v23/22-0658.html
14. https://www.vldb.org/pvldb/vol2/vldb09-335.pdf
15. https://blender.cs.illinois.edu/paper/truthkdd1.pdf
16. https://www.statcan.gc.ca/sites/default/files/media/14252-eng.pdf
17. https://www.cs.ubc.ca/~murphyk/papers/kv-kdd14.pdf
18. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-3.pdf
19. https://digital.library.adelaide.edu.au/server/api/core/items/3c5768d4-f9b9-4176-b00b-9e818a1e0630
20. https://arxiv.org/html/1708.02029v1
21. https://ojs.aaai.org/index.php/AAAI/article/view/10590
22. https://link.springer.com/content/pdf/10.1007/978-3-032-18765-9_2.pdf
23. https://www.jmlr.org/papers/v17/14-511.html
24. https://www.jmlr.org/papers/volume17/14-511/14-511.pdf
25. https://aclanthology.org/2021.acl-long.179
26. https://www.w3.org/TR/prov-implementations
27. https://research.google/pubs/knowledge-vault-a-web-scale-approach-to-probabilistic-knowledge-fusion
28. https://proceedings.neurips.cc/paper_files/paper/2019/file/c0e19ce0dbabbc0d17a4f8d4324cc8e3-Paper.pdf
29. https://www.nature.com/articles/s41598-022-10794-9
30. https://proceedings.neurips.cc/paper/2019/file/c0e19ce0dbabbc0d17a4f8d4324cc8e3-Paper.pdf
31. https://webdatacommons.org/TDGT
32. https://econpapers.repec.org/article/blajorssc/v_3a28_3ay_3a1979_3ai_3a1_3ap_3a20-28.htm
33. https://proceedings.mlr.press/v38/jaffe15.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.