Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Source Reliability by Topic: Testing a Single Trust Score

> 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 · 2026-10-05 · source reliability by topic · topic-specific source reliability · source credibility · truth discovery · faitcrowd · dart · partial pooling · claim-level reliability · data provenance

Original article: https://reliablecontext.com/articles/source-reliability-by-topic-trust-score



In this article

1. Executive Summary
2. Introduction and Background
3. Key Changes
4. Implementation Considerations and Process Changes
5. Models and their assumptions
6. Claim-level assessment and provenance
7. Data Analysis and Evidence
8. Implications and Future Directions
9. Conclusion

Executive Summary

Source reliability by topic should be measured as a conditional probability of a source supplying a correct assertion within a defined topic, population and observation period. A global score averages across that source's observed claims. It can therefore conceal topic differences and can change when the source's topic mix changes. Credibility cues help select evidence to inspect, but the House of Commons Library explicitly separates source assessment from whether information is true. (commonslibrary.parliament.uk) The decision is whether added topical parameters improve held-out truth inference enough to justify their uncertainty, rather than whether a finer taxonomy appears more informative.

The literature establishes useful mechanisms, with bounded empirical support. **FaitCrowd**, by Ma et al. (2015), reports aggregation error of **0.0399** on Game and **0.0610** on slot filling validation. (cse.buffalo.edu) **DART**, by Lin and Chen (2018), reports author-list F1 of **0.9041** under category domain partitions for books and director-list F1 of **0.9518** under genre domain partitions for movies for its DARTinf-Top variant. (www.vldb.org) These are original-study results with different targets and metrics. Independent comparator evaluations exist for both methods, with changed data or sampling; they do not establish a universal topic count or reproduce the original protocols unchanged. (arxiv.org)

A defensible comparison keeps truth labels, source overlap and answer semantics fixed, then evaluates global, unpooled topical and hierarchical models on identical held-out claims. Topic count, support thresholds and pooling strength belong inside model selection. Nested cross-validation evaluates the selection procedure; Cawley and Talbot show that selection overfitting can be comparable in magnitude to differences between algorithms. (scikit-learn.org) (www.jmlr.org) Report micro and macro error, probability scores, calibration, uncertainty and coverage. Partial pooling is a candidate response to sparse topics, with its own prior and transfer assumptions. (mc-stan.org)

The accompanying hypothetical calculation illustrates a reversal: globally stronger source A is weaker on the specialist topic, while small topical samples leave substantial uncertainty. It supplies a diagnostic, not an accuracy forecast. Provenance is also essential to judging the result: the World Wide Web Consortium (W3C) explicitly states that provenance records are not guaranteed correct. (www.w3.org) Reliable Context's

published guarantees concern cited, dated evidence and reproducible record states; they supply properties to assess alongside estimator performance. A topical model should be retained only when its gains survive held-out evaluation and realistic dependence and time tests.

Introduction and Background

Organizations combining documents, databases and extracted assertions face a recurring question: which source should influence a particular claim? **Truth discovery** jointly estimates source reliability and reconciles noisy information from multiple sources, as described by Li et al.'s survey. (arxiv.org) The narrower issue examined here is the granularity of reliability estimation: one score per source, one per source and topic, or an estimate that borrows information across topics.

These choices require an operational definition of **source**. It might denote a worker, publication, website, document version or extraction system. Those units have different opportunities for correlated errors. Dong et al.'s Knowledge-Based Trust work distinguishes erroneous facts on a page from erroneous extractions, making the attribution boundary explicit. (www.gabrilovich.com) A benchmark must specify whether it evaluates the original information, the extraction process or their combined output.

This report uses **reliability** to mean measured correctness under an explicit evaluation protocol, and **credibility** to mean reasons for regarding an assertion as worthy of consideration. These are operational definitions for this analysis. Authority, motivation, accuracy and relevance are useful screening dimensions in the House of Commons Library's framework; expertise must be relevant to the question. (commonslibrary.parliament.uk) A credential or citation trail does not itself supply a calibrated probability for a new assertion.

The audience's decision is consequential for context engineering: a model can produce a precise score even when the evidence for its source-topic estimate is weak. The report therefore treats topic assignment, truth semantics, sampling and provenance as separate design choices. FORCE11's data citation principles call for verifying the specific data supporting a claim. (force11.org) The same discipline applies to reliability estimates: a reported score should identify its supporting observations, evaluation target and limits of transfer.

Key Changes

Replace a standing score with a testable conditional target

Define `p_s` as the probability that a claim supplied by source `s` is correct in the benchmark population. Define `p_st` conditional on topic `t`. Their relationship is a weighted average: the global probability equals the sum of topical probabilities multiplied by the source's topic frequencies. This is the law of total probability applied to the report's definition, not an empirical finding about any particular source. Knowledge-Based Trust similarly operationalizes source accuracy through correctness of supplied values. (www.gabrilovich.com)

The testable failure of a global score is persistent variation in held-out correctness across topics after accounting for the intended claim population. If a source mainly supplies claims in its strongest domain, its average can be unsuitable for a specialist claim elsewhere. Conversely, splitting observations into many cells

can make topical estimates unstable. FaitCrowd explicitly discusses the reduction in responses per topic when questions are divided. (cse.buffalo.edu)

Predeclare what counts as improvement:

- **Target:** correctness of the fused answer, source assertion or both, assessed separately.
- **Population:** the claims and topics on which the score will be used.
- **Comparator:** a global model with the same label access and preprocessing.
- **Decision margin:** the minimum improvement worth added estimation complexity.
- **Uncertainty:** the interval for the paired difference, with its resampling unit stated.

These are proposed benchmark requirements. Their purpose is to avoid selecting granularity on an attractive training result. Model-selection criteria themselves can overfit when their estimates have appreciable variance. (www.jmlr.org)

Separate topic assignment from expertise estimation

A topic can be an externally assigned domain label, a latent cluster of question text, an attribute class or a mixture. The benchmark must record which one it uses. FaitCrowd jointly models question content and answers; DART describes expertise through domain information richness and influence between domains. (cse.buffalo.edu) (researchportal.hkust.edu.hk) These mechanisms do not make their topic representations interchangeable.

Assess **topic assignment** against an external reference where one exists, and evaluate its effect on truth inference separately. Semantic interpretability is useful for review, but a readable cluster name does not prove predictive value. TD-AC provides another single-truth attribute-partitioning approach and reports sensitivity to sparse truth vectors. (edbt2021proceedings.github.io) Thus, topic granularity remains relevant even when each claim has exactly one correct value.

Useful controls include:

- **External labels:** domains defined independently of model predictions.
- **Text-only topics:** learned from permitted training text.
- **Joint topics:** inferred with answers, under a declared inference protocol.
- **Attribute groups:** a structural partition independent of answer cardinality.
- **Shuffled groups:** a negative control with comparable cell sizes.

These controls ask whether a gain comes from meaningful heterogeneity, additional flexibility or a changed data partition. Held-out evaluation remains necessary because inspecting the same data during selection can create an optimistic score. (scikit-learn.org)

Treat sparse support as part of the result

A source-topic score should travel with its **support**, meaning the number of eligible labeled assertions used to estimate it, and with its sampling assumptions. A high score from a small cell and the same score from a large cell are different evidence objects. Stan's documentation explains why independent estimation becomes problematic when many groups have few or no observations. (mc-stan.org)

Report unsupported cells explicitly. A fallback to a global estimate, a pooled estimate or abstention should be visible in the output and counted in coverage. Hierarchical models can stabilize small groups, but few groups and weak information can also make them hard to fit without stronger hyperpriors. (mc-stan.org) The relevant comparison therefore includes how much of the target population receives a defensible prediction.

Implementation Considerations and Process Changes

Fix the benchmark before changing the scores

Construct a claim inventory with a source identifier, asserted value, topic assignment, observation date and adjudicated truth. Fix identity matching and answer equivalence before comparing estimators. Waguih and Berti-Equille's experimental framework varies source coverage and conflict characteristics, illustrating why those dimensions should be controlled rather than allowed to change unnoticed between methods. (arxiv.org)

The proposed protocol records:

- **Claim unit:** the entity, attribute and temporal scope of the assertion.
- **Source unit:** the originator or extraction system being scored.
- **Truth label:** the adjudicated answer and evidence used to establish it.
- **Answer semantics:** a single value or a set of simultaneously valid values.
- **Support matrix:** training and test counts for each source-topic pair.
- **Overlap:** which competing sources cover the same claims.
- **Observation status:** supplied, unavailable or excluded, with a reason.
- **Version:** the frozen dataset and normalization rules used for scoring.

Label adjudication should be separate from estimator output. Unknown truth, an absent assertion and an incorrect assertion occupy different evaluation states in this proposed design. Report how each enters the denominator. Evidence access can be supported by metadata and documentation, consistent with FORCE11's access principle. (force11.org)

Use distinct splits for distinct transfer questions

A random claim split asks whether known sources can answer more claims from a similar population. A topic-held-out split asks about transfer into excluded domains. A source-held-out split asks about new sources. An object-grouped split prevents related claims about the same object from straddling training and test. GroupKFold implements the principle that a group should not appear in both sets. (scikit-learn.org)

Preserve a temporal test when future use is the target. Ordinary cross-validation can produce inappropriate train-test correlations for ordered data; Ovadia et al. study accuracy and calibration under dataset shift. (scikit-learn.org) (papers.nips.cc) A topical model with no representation for an unseen topic needs a declared transfer rule. Its fallback performance should be scored, rather than silently omitting that topic.

Use the following separation:

- **Training:** fit source effects, topics and preprocessing permitted by the protocol.
- **Inner validation:** choose topic count, support floor and pooling strength.

- **Calibration:** fit any probability adjustment without reusing its evaluation labels.
- **Outer test:** evaluate the selected procedure on untouched claims.
- **Transfer tests:** repeat the declared source, topic and temporal holdouts.

Nested cross-validation separates selection from outer performance estimation. (scikit-learn.org) Calibration fitting also requires separation from the classifier's training data. (scikit-learn.org) If an experiment permits transductive access to unlabeled test text or answers, declare that access and give every comparator the same opportunity.

Measure correctness, probability quality and coverage

Define micro error as total wrong answers divided by total evaluated claims. Define macro error as the unweighted mean of topic-specific error rates. These are this report's topic-level analogues of pooled and equal-weight averaging; scikit-learn documents the corresponding distinction for classification metrics. (scikit-learn.org) Report both so a large topic cannot silently determine the whole conclusion.

For a binary correctness target, **Brier score** is the mean squared difference between predicted probability and the correctness indicator. **Log loss** penalizes the probability assigned to the observed outcome. (scikit-learn.org) Assess source-assertion probabilities and fused-answer probabilities separately. Their targets differ, so a calibrated source score does not establish calibrated answer confidence.

Include:

- **Error:** micro, macro and topic-specific values on identical labels.
- **Probability scores:** Brier score and log loss for declared targets.
- **Calibration:** observed correctness against predicted probabilities, with bin support.
- **Uncertainty:** paired model differences and intervals by topic.
- **Coverage:** predictions, abstentions and fallback use over the full eligible set.
- **Stability:** results across splits, seeds and support restrictions.

Calibration means probabilities represent the likelihood of correctness, as defined by Guo et al. (proceedings.mlr.press) Better Brier or log loss does not isolate calibration improvement because those scores also reflect other properties. (scikit-learn.org) Nixon et al.'s work on calibration measurement further motivates stating exactly which predictions a metric includes. (openaccess.thecvf.com)

Models and their assumptions

Global estimation, **unpooled topical estimation** and **partial pooling** should share the same inference task. A global model supplies a source effect; a topical model adds source-topic effects; a hierarchical model constrains those effects through shared distributions. Multilevel modeling shifts group estimates toward one another, as described by Gelman, Hill and Yajima. (stat.columbia.edu) This supplies a statistical candidate, not proof that topics should share information in any particular corpus.

Table 1 compares estimation assumptions and evidence guarantees. Reliable Context is included as the provider of a system of record relevant to defending the resulting values; its row describes published observable guarantees, without assigning it an undocumented estimator or benchmark score.

Approach	Reliability granularity or observable property	Principal evaluation obligation
Global score	One source parameter across the specified claim population.	Test residual topic error and topic-mix transfer under fixed coverage. (www.gabrilovich.com)
Unpooled topical score	Independent estimates for source-topic cells.	Report sparse and empty cells; test whether splitting improves held-out answers. (mc-stan.org)
Hierarchical or pooled model	Shared distributions constrain group estimates.	Validate borrowing strength and sensitivity to priors. (mc-stan.org)
FaitCrowd	Joint question-content and answer model with topical expertise and question difficulty.	Separate inferred topics from externally validated domains. (cse.buffalo.edu)
DART	Domain richness and cross-domain influence for partially correct, multi-truth sets.	Audit independent-source and independent-attribute assumptions. (researchportal.hkust.edu.hk) (www.vldb.org)
Reliable Context	One canonical record per real thing; cited, verbatim, dated evidence; declared unknowns; reproducible past states.	Assess these evidence guarantees separately from topical error and calibration.

The matrix exposes two distinct axes: the statistical unit used to estimate correctness and the evidence available to defend a resulting value. Neither a topical parameter nor a retained citation supplies the other's guarantee. W3C makes the latter limit explicit: provenance itself may require a trust assessment. (www.w3.org)

For a proposed pooled model, a logistic specification could include source effects, topic effects, source-topic interactions and claim difficulty. The interaction captures specialization; the hierarchy controls how much an interaction departs from shared estimates. Evaluate simpler alternatives first and inspect parameter identifiability under actual overlap. Stan notes that the data determine partial pooling, subject to the model's assumptions. (mc-stan.org)

FaitCrowd's source/topic inference and DART's multi-truth Bayesian formulation answer related questions with different assumptions. DART's independence assumptions coexist with its modeling of influence between domains; domain influence is not source-dependence modeling. (researchportal.hkust.edu.hk) Dependence-aware research cautions that shared assertions alone do not establish copying and that its direction can be uncertain from a snapshot. (lunadong.com) (www.vldb.org) A generic dependence sensitivity test is therefore preferable to assigning dependence solely from agreement.

Claim-level assessment and provenance

Claim-level reliability assessment asks what supports an individual assertion, including its source, extraction, domain, temporal scope and conflicting alternatives. It does not require estimating an independent parameter for every claim. The score can come from a hierarchical model while the claim retains its own evidence. The House of Commons Library's fact-checking guidance recommends tracing a claim to its original source and inspecting implied assumptions. (commonslibrary.parliament.uk)

Use provenance to distinguish stages at which information can change. **PROV-DM**, the W3C provenance data model, represents entities, activities and agents; derivation describes transformation between entities. (www.w3.org) In an evaluation, the original document, extracted assertion and normalized candidate can be distinct evidence objects. This is a proposed benchmark representation, not a description of any vendor's internal mechanism.

A useful evidence record contains:

- **Origin:** the source document or dataset and a stable locator.
- **Excerpt:** the exact passage supporting the asserted value.
- **Dates:** observation time and the claim's stated temporal scope.
- **Transformation:** the normalization or extraction version applied.
- **Topic:** the assigned domain, its method and any ambiguity.
- **Status:** evidence found, inaccessible or not applicable, with a reason.
- **Evaluation:** the truth label, split membership and prediction used in scoring.

Version identification matters because a general identifier can refer to changing content. DataCite distinguishes canonical identifiers from identifiers for specific versions, and permits minor content changes under the same digital object identifier (DOI) with updated metadata. (support.datacite.org) An evaluator should therefore freeze the supporting version, rather than assuming an identifier alone recovers the tested observation.

FORCE11 recommends sufficient specificity to recover the cited timeslice, version or granular portion of data. (force11.org) Crossref's Crossmark supplies access to current content status, while expressly stating that its presence is not itself a guarantee. (www.crossref.org) These capabilities help inspect evidence currency; they do not replace correctness labels.

Findable, Accessible, Interoperable and Reusable (FAIR) principles address data reuse, including automated discovery and use. (www.nature.com) Their role here is to make evaluation artifacts inspectable. Reliability calibration still requires an independently scored correctness target. The final report should permit a reviewer to reconstruct which evidence supported each claim and which observations supported each probability.

Data Analysis and Evidence

Published measurements and their limits

Table 2 separates corpus scale, labeled evaluation and reported metrics. Counts refer to original studies, not new measurements by this report.

Original study and dataset	Corpus and evaluation support	Author-reported result
FaitCrowd, Game	2,103 questions, 37,029 human sources, 214,849 answers. (cse.buffalo.edu)	Error 0.0399; CATD 0.0485; majority vote 0.0980. (cse.buffalo.edu)

Original study and dataset	Corpus and evaluation support	Author-reported result
FaitCrowd, slot filling validation (SFV)	328 questions, 18 extraction systems, 2,538 answers. (cse.buffalo.edu)	Error 0.0610; AccuPr 0.0701; majority vote 0.1128. (cse.buffalo.edu)
DART, books	54,591 sources, 210,206 books, 18 categories; 407 labeled books, 120 sampled per round. (www.vldb.org)	Category-domain DARTinf-Top precision/recall/F1: 0.8777/0.9322/0.9041. (www.vldb.org)
DART, movies	15 websites, 468,607 movies, 21 genres; 210 labeled movies, 80 sampled per round. (www.vldb.org)	Genre-domain DARTinf-Top precision/recall/F1: 0.9414/0.9625/0.9518. (www.vldb.org)

Corpus size does not equal labeled source-topic support. FaitCrowd measures wrong questions divided by all questions; DART evaluates returned truth sets using precision, recall and their harmonic mean, F1. (cse.buffalo.edu) (www.vldb.org) Error and set F1 cannot be treated as a common leaderboard.

FaitCrowd selects **12 topics for Game and 8 for SFV**; its comparisons include TruthFinder, AccuPr, Investment, 3-Estimates, CRH, CATD, Dawid and Skene, and ZenCrowd. (cse.buffalo.edu) CATD was omitted from SFV because its equal-answer-choice requirement was unmet. (cse.buffalo.edu) DART compares voting and truth-discovery baselines plus domain-separated alternatives. (www.vldb.org) Its year-domain DARTinf-Top F1 values are **0.9033** and **0.9457**, a different configuration. (www.vldb.org) Assess the reported comparisons against documented statistical tests and source-topic support; descriptive significance language alone does not establish that a statistical test was performed. (cse.buffalo.edu) (www.vldb.org)

Independent evidence needs equal care. Tu et al.'s FaitCrowd comparison samples **200 workers** from Game. (arxiv.org) Kucuk et al.'s **2026 preprint** evaluates DART using original-paper default parameters on a book corpus of **894 sources**, with normalized predictions and labels. (arxiv.org) Its DART book/movie F1 values are **0.5997/0.7768**. (arxiv.org) Different corpora and scoring mean these results establish independent comparator use, not direct reproduction of the original numbers. These cited comparator studies use modified protocols and should not be treated as exact replications of the original experiments.

A global reversal (Hypothetical Example)

Table 3 is a disclosed calculation, not a benchmark dataset. Both invented sources answer the same binary single-valued claims. Each has 200 broad-topic and 10 specialist-topic labeled answers. Wilson intervals use a 95% confidence level under independent binomial sampling, following the NIST formula. (www.itl.nist.gov)

Invented source and topic	Correct / evaluated	Accuracy	95% Wilson interval
A, broad	180 / 200	90.0%	85.1% to 93.4%
A, specialist	2 / 10	20.0%	5.7% to 51.0%
B, broad	140 / 200	70.0%	63.3% to 75.9%
B, specialist	8 / 10	80.0%	49.0% to 94.3%

The calculation gives A global accuracy **182/210 = 86.7%**, versus B's **148/210 = 70.5%**. A weighted vote using global accuracy favors A when they disagree; topical accuracy favors B on specialist claims. For this hypothetical held-out specialist claim, assume the correct answer agrees with B. The global vote is then wrong on that answer. The example proves possibility under its declared construction, not prevalence in real sources.

The same counts give A micro error **13.3%** but macro error **45.0%**; B has micro error **29.5%** and macro error **25.0%**. Equal topic weighting reverses their ranking. This is the report's defined macro/micro calculation, consistent with the averaging distinction discussed above. (scikit-learn.org) The specialist intervals overlap slightly, illustrating the uncertainty hidden by point estimates. Wilson bounds remain within the unit interval; clustered observations would require a different uncertainty treatment. (www.statsmodels.org)

A disclosed shrinkage rule further clarifies the choice: `q_st = (k_st + m*r_s)/(n_st + m)`. Define the illustrative prior center `r_s` as the broad-topic training accuracy and choose prior weight `m = 20` solely for illustration. Specialist estimates become **66.7% for A** and **73.3% for B**. This rule borrows across topics; it is not a fitted hierarchical benchmark or a recommended prior. Partial pooling shifts group estimates together, but its usefulness must be measured on held-out answers. (stat.columbia.edu)

A granularity stress test with reportable uncertainty

Pre-register an illustrative search grid of **1, 4 and 16 topics** crossed with minimum source-topic supports of **10, 50 and 200 labeled assertions**, plus candidate pooling strengths. These are proposed experimental settings, not literature-derived thresholds. For every cell report outer-test micro error, macro error, Brier score, log loss, interval for the paired difference, and prediction coverage. Empty or filtered cells remain visible.

Resample at the dependence unit appropriate to the corpus, such as objects or source families. Paired bootstrap implementations preserve matched indices across comparator arrays; preserving that pairing is useful, while choosing a cluster unit is a separate design obligation. (docs.scipy.org) Explain how uncertainty components were evaluated, following NIST's reporting principle. (www.nist.gov) Select granularity using inner validation, then report the chosen procedure's outer result, including negative outcomes.

Implications and Future Directions

The **default candidate set** should include a global baseline, a topical model and a pooled model. It need not presume that the most detailed model wins. A gain that disappears when source overlap is matched, labels are held out or weak cells are retained is evidence against deploying that granularity for the stated population. Waguih and Berti-Equille examine initialization, parameters, convergence and scalability, underscoring the need to make these choices visible. (arxiv.org)

Use a decision rule that addresses the actual intended operation:

- **Retain global:** topical residuals are small or added detail does not improve outer-test decisions.
- **Retain topical:** gains persist across support levels and relevant transfer tests.
- **Retain pooled:** sparse cells benefit without suppressing well-supported specialization.
- **Revise topics:** gains depend on labels unavailable at deployment or unstable assignments.
- **Abstain:** evidence cannot support the requested probability or answer.

These are proposed interpretations of benchmark outcomes. Report the loss function and coverage tradeoff behind each decision. Calibration curves should include support, and weak cells should not acquire apparent certainty through an overly flexible calibrator; isotonic calibration is particularly prone to overfitting small datasets. (scikit-learn.org)

Three confounds remain central. **Source dependence** can turn repeated assertions into redundant evidence; overlap alone does not prove copying. (lunadong.com) **Changing topic populations** can alter both error and calibration, motivating temporal tests. (papers.nips.cc) **Extraction errors** can be attributed to the wrong source unless origin and processing are separated. (www.gabrilovich.com) None is solved simply by increasing topic count.

Human source evaluation supplies additional context, not direct algorithm validation. Stanford's account of lateral reading emphasizes seeking context from other sites; its study purposively sampled 10 historians, 10 fact checkers and 25 students. (ed.stanford.edu) That supports inspecting context and origin, without treating human evaluation behavior as a benchmark for institutional reliability probabilities.

For an operational record, the evaluation must also examine whether uncertainty and supporting evidence remain reviewable. Reliable Context states that past record states can be reproduced and unknowns are declared with reasons. These are observable properties to verify separately from the proposed statistical comparisons. Reproducibility should retain data versions, code settings and computational environment; even PyTorch's documentation does not guarantee identical results across releases. (docs.pytorch.org) Publishing the protocol and negative findings makes later comparisons more interpretable.

Conclusion

Source reliability by topic is a model-selection question with an explicit target population. A global score may hide specialization; separate topic scores may convert limited observations into unstable estimates. A pooled model introduces assumptions about sharing information. The correct choice is the procedure that improves the declared held-out decision while keeping uncertainty and coverage visible.

The reviewed research supplies mechanisms and author-reported evidence, with independent comparator studies under modified protocols. It does not supply a universal taxonomy, a source-topic support floor or a transferable accuracy guarantee. Inferred topics, externally assigned domains and multiple truths per attribute must remain separate concepts when designing the comparison.

The proposed test keeps labels, source overlap and answer semantics fixed; selects granularity inside validation; and evaluates error, probability quality, uncertainty and coverage on untouched claims. It also checks temporal transfer, dependence and extraction attribution. A negative result is actionable: preserve the simpler estimate, revise the topic representation or abstain where support is inadequate.

Finally, an estimate and its evidence should remain inspectable at the claim level. The reviewer needs to recover the source observation, the evaluated assertion and the assumptions behind its probability. The resulting conclusion is conditional and reproducible: additional granularity is justified where its measured gains survive those checks, with its remaining uncertainty stated explicitly.

Source links

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

1. https://commonslibrary.parliament.uk/research-briefings/cbp-10822
2. https://cse.buffalo.edu/%7Ejing/doc/kdd15_FaitCrowd.pdf
3. https://www.vldb.org/pvldb/vol11/p635-lin.pdf
4. https://arxiv.org/pdf/1912.11238
5. https://scikit-learn.org/stable/auto_examples/model_selection/plot_nested_cross_validation_iris.html
6. https://www.jmlr.org/beta/papers/v11/cawley10a.html
7. https://mc-stan.org/docs/stan-users-guide/poststratification.html
8. https://www.w3.org/TR/prov-aq
9. https://arxiv.org/abs/1505.02463
10. https://www.gabrilovich.com/publications/papers/Dong2015KBT.pdf
11. https://force11.org/group/joint-declaration-of-data-citation-principles-final
12. https://researchportal.hkust.edu.hk/en/publications/domain-aware-multi-truth-discovery-from-conflicting-sources
13. https://edbt2021proceedings.github.io/docs/p98.pdf
14. https://arxiv.org/abs/1409.6428
15. https://scikit-learn.org/stable/modules/cross_validation.html
16. https://papers.nips.cc/paper_files/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html
17. https://scikit-learn.org/stable/modules/calibration.html
18. https://scikit-learn.org/stable/modules/model_evaluation.html
19. https://proceedings.mlr.press/v70/guo17a.html
20. https://openaccess.thecvf.com/content_CVPRW_2019/html/Uncertainty_and_Robustness_in_Deep_Visual_Learning/Nixon_Measuring_Calibration_in_Deep_Learning_CVPRW_2019_paper.html
21. https://stat.columbia.edu/%7Egelman/research/published/multiple2f.pdf
22. https://lunadong.com/publication/indep_techReport.pdf
23. https://www.vldb.org/pvldb/vol2/vldb09-pvldb47.pdf
24. https://commonslibrary.parliament.uk/research-briefings/cbp-10820
25. https://www.w3.org/TR/prov-dm
26. https://support.datacite.org/docs/versioning
27. https://www.crossref.org/services/crossmark
28. https://www.nature.com/articles/sdata201618
29. https://arxiv.org/html/2606.28062v1
30. https://www.itl.nist.gov/div898/handbook/prc/section2/prc241.htm
31. https://www.statsmodels.org/stable/generated/statsmodels.stats.proportion.proportion_confint.html
32. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bootstrap.html
33. https://www.nist.gov/pml/nist-technical-note-1297/nist-tn-1297-7-reporting-uncertainty

34. https://ed.stanford.edu/news/stanford-scholars-observe-experts-see-how-they-evaluate-credibility-information-online

35. https://docs.pytorch.org/docs/stable/notes/randomness.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.