Reliable Context

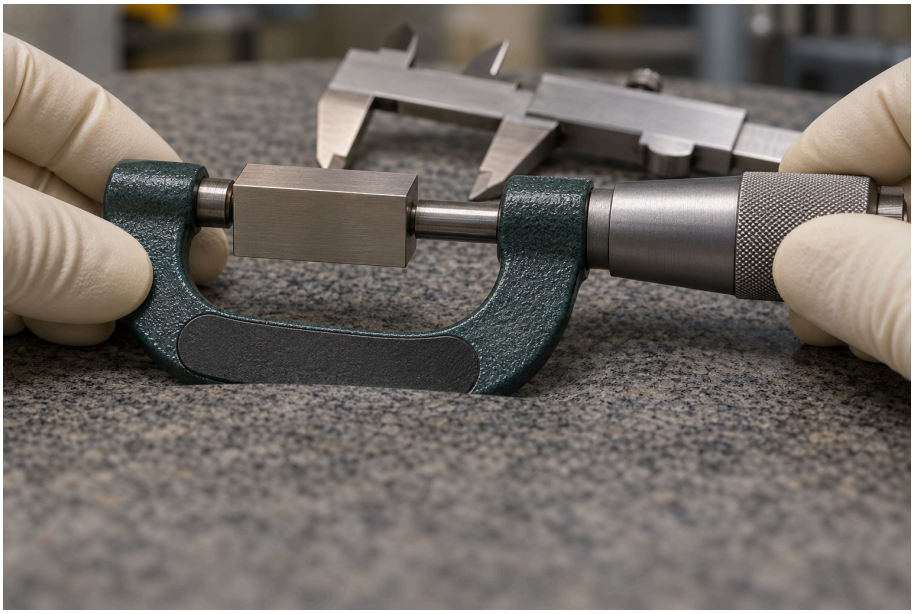
RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

How to resolve conflicting numerical claims: normalize first

> 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 · 2026-10-17 · conflicting numerical claims · truth discovery · data normalization · numerical data fusion · crh · madi-bench · data provenance · benchmark evaluation · measurement uncertainty

Original article: https://reliablecontext.com/articles/conflicting-numerical-claims-normalize-before-truth-discovery



In this article

1. Executive Summary
2. Introduction and Background
3. What constitutes a numerical conflict
4. Normalize and align before estimating reliability
5. What heterogeneous-loss truth discovery establishes
6. Separate pipeline accuracy from fusion accuracy
7. Design an auditable comparison
8. Data Analysis and Evidence
9. Implications and Future Directions
10. Conclusion

Executive Summary

To resolve conflicting numerical claims, first establish that the claims concern the **same entity, quantity kind, population, reference period, and measurement basis**. Then parse their representations, convert compatible units, preserve uncertainty, and distinguish disagreement from change over time. Only the remaining comparable claims should enter truth discovery. Equal dimensions alone do not establish equal quantity kinds, and observation time can differ from the time a result becomes available. (www.oiml.org) (docs.ogc.org)

Normalization is part of the experiment, rather than an invisible preparation step. An exact inch-to-millimetre conversion preserves equality, while rounding, missing units, and mismatched denominators require explicit annotation decisions. The National Institute of Standards and Technology (NIST) defines the standard inch as exactly **25.4 mm** and advises preserving the original rounding information when converting reported values. (www.nist.gov) The proposed evaluation therefore separates equivalent representations, genuine near-value disagreement, rounded displays, incomplete units, stale observations, and mixed-type claims.

The literature supports separating these choices from reliability estimation. **Conflict Resolution on Heterogeneous Data (CRH)** couples type-specific losses with joint truth and source-weight estimation. (cse.buffalo.edu) **Mannheim Data Integration Benchmark (MaDI-Bench)** separately evaluates normalization, matching, and fusion. Its historical first-version fusion scores of **40.20% to 84.87%** describe particular task and pipeline combinations, not a universal performance bound. (arxiv.org) The revised paper changes the evaluation context, including a pooled-cell definition that counts missing values or entities as wrong. (arxiv.org)

An explicitly **constructed fixture** below demonstrates why the scoring contract matters. With identical predictions on its comparable numerical items, median fusion wins on mean absolute error, while mean fusion wins on root mean squared error. That outcome follows from the disclosed arithmetic, consistent with the different targets of absolute and squared loss. (otexts.com) It is not evidence of general superiority or a reproduction of CRH. A defensible comparison publishes its conversions, inclusion rules, reference values, tolerance, denominator, source mix, and loss function before reporting a winner. Provenance supplies the evidence needed to inspect those decisions; it does not itself establish correctness. (www.w3.org)

Introduction and Background

Sources report different numbers for reasons that belong to different analytical categories. Some disagreements disappear after conversion. Others survive conversion because the sources measured different things, used different populations, referred to different times, or revised an estimate. The U.S. Census Bureau, for example, explicitly treats its latest population-estimate vintage as superseding earlier estimates for the same dates. That is a versioning relationship, not necessarily a new observation of population change. (www.census.gov)

A numerical comparison must therefore begin with a **claim specification**. A literal value becomes interpretable only through its unit, quantity, scope, time, and provenance. Statistical rates require a denominator and period, while a reference period identifies the point or interval to which the collected data relate. (www.oecd.org) An algorithm can reconcile numbers only relative to a specification of what counts as the same question.

The evidence snapshot is October 17, 2026. This report examines how to resolve conflicting numerical claims when evaluating **truth discovery**, the joint estimation of candidate truths and source reliability represented by methods such as CRH. (cse.buffalo.edu) Its audience is researchers deciding whether an apparent improvement comes from reliability estimation or from earlier choices about representation and scoring. Reliable Context is the publisher; the subject here is evaluation methodology, with no product performance inference.

The central recommendation is to make **conflict annotation** an explicit experimental stage. Researchers should preserve the original claims, publish each transformation, and report separate scores for each comparability stratum. MaDI-Bench provides a useful precedent for separating normalization accuracy from fusion and releasing the associated evaluation resources. (wbsg-uni-mannheim.github.io) (github.com) The constructed experiment below extends that separation to tolerance and loss sensitivity. It deliberately holds source weights fixed so that changes caused by the scoring contract remain visible.

What constitutes a numerical conflict

A **numeric claim** is an assertion that a specified quantity has a value for a specified entity and context. The **quantity kind** identifies what is measured; the **unit** supplies the reference used to express its magnitude. Matching physical dimensions is insufficient to establish that the quantities have the same meaning, as the International Vocabulary of Metrology explicitly notes. (www.oiml.org)

For evaluation, define **genuine disagreement** as non-equivalence between otherwise comparable assertions under a published comparison rule. This is an operational definition, not a claim that every measurement difference establishes an error. **Measurement uncertainty** characterizes dispersion in values attributed to the measurand, whereas a **tolerance interval** specifies permissible property values in conformity assessment. These concepts answer different questions. (www.oiml.org) An evaluation tolerance is a further analytical choice: it states which prediction errors the benchmark will count as acceptable.

Use an annotation vocabulary that keeps those questions separate:

- **Equivalent representation:** The values become equal after an authorized conversion.
- **Comparable disagreement:** The context matches, but the values differ under the declared rule.

- **Rounded display:** The apparent difference may reflect reporting precision.
- **Uncertain comparison:** Available measurement information does not justify a decisive label.
- **Temporal difference:** The assertions refer to different observation periods.
- **Scope difference:** The population, denominator, or aggregation differs.
- **Insufficient context:** A unit, date, or quantity definition is missing.
- **Identity uncertainty:** The records have not been established to concern the same entity.

These are proposed evaluation labels. Their purpose is to prevent a single accuracy statistic from mixing parsing success, measurement compatibility, identity decisions, and truth selection. The distinction matters because a matching error can exclude the source record containing the correct value before fusion begins. (arxiv.org)

Closeness also needs a precise operational rule. In a **(Hypothetical Example)**, values of 0, 0.4, and 0.8 share the same unit and context. Under a pairwise tolerance of 0.5, the middle value is close to each endpoint, but the endpoints are not close to each other. These are constructed values. Consequently, pairwise tolerance does not automatically define transitive equivalence classes. A benchmark must specify whether it compares predictions with a reference, forms clusters, or applies another rule.

Finally, **unknown** is an annotation outcome. An absent unit does not authorize inference from numerical magnitude alone. The recommended policy is to preserve the literal claim and record why it was excluded. This keeps incomplete evidence from silently becoming an incorrect-value label, a distinction that pooled accuracy can otherwise obscure when missing outputs count as wrong. (arxiv.org)

Normalize and align before estimating reliability

Parse and convert without adding meaning

Parsing should preserve the source literal alongside the interpreted number. Record decimal conventions, grouping separators, scale words, unit symbols, and any range or inequality. MaDI-Bench illustrates why representation has its own evaluation target: its current normalization scorer compares numerical forms as text, rather than treating every numerically equivalent rendering as interchangeable. (github.com) That is a benchmark convention, not a universal definition of numerical truth.

Unit conversion requires a known quantity and conversion rule. NIST's exact conversion gives **1 inch = 25.4 mm**. (www.nist.gov) Temperature also illustrates why a single generic multiplication rule is inadequate: NIST specifies subtracting **32** and dividing by **1.8** for Fahrenheit-to-Celsius conversion. (www.nist.gov) The International System of Units (SI) Brochure distinguishes absolute Celsius temperature, which involves the **273.15 K** offset, from temperature differences, whose Celsius and kelvin numerical magnitudes agree. (www.bipm.org)

Conversion should not manufacture measurement precision. NIST's rounding guidance ties the converted representation to the original rounding error. (www.nist.gov) A source displayed to a coarse increment may be compatible with a more detailed reference without supplying that reference's precision. Preserve the displayed value, the conversion, and any explicitly assumed rounding interval separately.

A reproducible preprocessing record should include:

- **Literal:** The source's unchanged value and surrounding unit text.
- **Parser:** The locale and syntax rule applied.
- **Quantity:** The operational definition of the measured attribute.
- **Conversion:** The formula, constants, and target unit.
- **Precision:** Reported digits and any stated rounding convention.
- **Uncertainty:** The source's interval and its stated interpretation.
- **Status:** Converted, unchanged, ambiguous, or excluded, with a reason.

Align semantics, denominators, and time

Common units do not align different denominators. The OECD's statistical handbook defines rates using events and a population over a period; it also distinguishes percentage-point differences from percentage changes. (www.oecd.org) Comparing rates requires checking eligibility, coverage, geographic boundaries, and the numerator's event definition. Converting a count into a rate requires the corresponding population, rather than a convenient population from another period.

Price basis is another semantic condition. The U.S. Bureau of Labor Statistics (BLS) explains that constant-dollar estimates adjust nominal costs for consumer-price changes and requires consistent index base periods in its documented calculation. (www.bls.gov) Two amounts with the same currency symbol can therefore represent different measurement bases. A conversion to a common price basis is an additional documented transformation.

Keep several time fields. Open Geospatial Consortium (OGC) Observations distinguishes **phenomenonTime**, when a result applies to the observed characteristic, from **resultTime**, when it becomes available, and **validTime**, the interval during which it is assumed applicable for use. (docs.ogc.org) The Dublin Core Metadata Initiative (DCMI) separately defines formal issuance metadata, while Request for Comments (RFC) 3339 provides explicit relationships between timestamp offsets and Coordinated Universal Time (UTC). (www.dublincore.org) (www.rfc-editor.org)

The recommended final stages are **duplicate detection**, followed by conflict resolution. Preserve a derivation chain for repeated or transformed evidence instead of treating each appearance as independent support. The PROV Ontology (PROV-O) provides derivation and alternate-entity relations suitable for representing such lineage, although using them at claim level remains an application modeling choice. (www.w3.org) Neither the newest publication nor the most frequently repeated value should win by an unstated rule.

What heterogeneous-loss truth discovery establishes

CRH's objective and update rules

Li and colleagues' **2014 CRH** paper uses categorical disagreement and numerical absolute or squared deviation divided by each item's across-source standard deviation. Fixed weights yield weighted voting, medians, or means. (cse.buffalo.edu)

Writing source k's aggregate loss as D_k, CRH minimizes sum w_k D_k subject to **sum exp(−w_k) = 1**. For positive losses, its fixed-truth update is **w_k = −log(D_k / sum D_j)**. Thus negative exponentials, rather than weights, sum to unity. (cse.buffalo.edu) The specified convergence result reaches a stationary point; it does

not establish a global joint optimum. (cse.buffalo.edu)

Its squared numerical loss divides by **standard deviation**, rather than variance. (cse.buffalo.edu) By direct algebra, replacing an item's values and truth with a times their original values, for positive a, multiplies that item's squared-over-standard-deviation loss by a. Absolute-over-standard-deviation loss remains unchanged. This is a derivation from the published formula, not an empirical result about the authors' code.

Thus rescaling only some attributes can change their relative contribution to source totals and potentially change inferred source weights. Uniformly multiplying every contribution by the same factor would cancel in the weight-update ratios. This distinction makes a unit-sensitivity check necessary for a heterogeneous evaluation; it does not prove a ranking reversal on any published dataset. (cse.buffalo.edu)

Evidence and limits of attribution

The paper tests weather, stock, flight, and simulated data. It reports **218 errors among 580 categorical weather entries**, against **266** for its best baseline. These are study-specific error counts, not a unit-normalization experiment or a transferable accuracy guarantee. (cse.buffalo.edu) The paper should be credited for its tested loss-based framework without attributing untested parsing or unit-disambiguation capabilities to it.

An author's publication page identifies an original code-and-data archive. The archive's contents were not inspected in this review, so the equations above are paper-verified rather than code-verified. (cse.buffalo.edu) A reproduction should publish policies for zero dispersion, zero total source loss, missing claims, initialization, and stopping. Those policies must be stated as implementation choices unless confirmed in the selected implementation.

Temporal methods address a related but distinct problem. Li and colleagues' **2015 DynaTD** uses smoothing and decay, while Zhi and colleagues' **2018 EvolvT** models truth transition, source quality, and dependency, using Kalman filtering for numerical evolution. (pmc.ncbi.nlm.nih.gov) (experts.illinois.edu) These models motivate time-aware evaluation. They do not make claims from different observation dates automatically comparable. Even a reference set deserves scrutiny: EvolvT's stock experiment describes reference values obtained by majority agreement among selected providers. (hanj.cs.illinois.edu) A label's construction method is part of the evidence.

Separate pipeline accuracy from fusion accuracy

MaDI-Bench, by Aaron Steiner, Ralph Peeters, and Christian Bizer, evaluates an integration pipeline with separate resources for normalization, matching, and fusion across its base tasks. (wbsg-uni-mannheim.github.io) Its normalization tests include values already in the required form, so unnecessary transformations can reduce accuracy. (wbsg-uni-mannheim.github.io) This gives normalization a measurable failure mode beyond failure to change a malformed value.

Version identification is essential. The arXiv record dates the first version to **June 29, 2026**, and the revision to **October 5, 2026**. (arxiv.org) The first version evaluates human-designed P1, best-of-breed P2, and large language model (LLM)-based P3. (arxiv.org) Its fusion table reports **40.20%** for Products/P1 and **84.87%** for

Games/P3. (arxiv.org) These endpoints cross tasks and pipelines. They are neither a within-task treatment effect nor a performance interval for arbitrary fusion systems.

The current website describes **four pipelines**, with revised evaluation material. (wbsg-uni-mannheim.github.io) The historical percentages must therefore remain attached to their version. This review did not independently reconstruct the first version's effective scored-cell denominators from its historical evaluator. It does not multiply those percentages by the paper's inventory of test values to invent correct-cell counts.

The revised paper defines fusion accuracy by pooling ground-truth cells across attributes under attribute-specific comparators and counting missing values or entities as wrong. (arxiv.org) The current repository reports **793 graded Games test cells** and **836 Products test cells**. (github.com) Its second-version gold excludes cells whose verified value cannot be obtained from any source record of the entity even after normalization; those blanked cells are not graded. (github.com) These are explicit task-design decisions about attainability and denominator membership.

Stepwise and end-to-end results answer different questions. A fusion method can fail because it selects poorly from a correct cluster, or because an upstream matcher withheld the record containing the answer. (arxiv.org) The revised paper also separates reference-free structural metrics, a verified gold slice, and broader silver-reference evaluation. (arxiv.org) Human verification uses web research, but the existence of that process does not remove the need to report label provenance. (arxiv.org)

A suitable experiment therefore runs an **oracle-input fusion condition**, using correctly aligned claims, alongside an **end-to-end condition**, using actual preprocessing outputs. It reports normalization and matching outcomes separately, then attributes only the remaining difference to fusion. This is a proposed experimental design informed by the benchmark's stage separation, not a result already established by its tables.

Design an auditable comparison

A constructed fixture with independent reference definitions

Table 1 specifies a **(Hypothetical Example)**, constructed for this article. A, B, and C are fictional sources, not measurements collected from organizations. Reference values are declared independently of their votes. Every numerical value in the fixture is an experimental input; the only external physical conversion used is NIST's exact inch conversion. (www.nist.gov) All comparable lengths are normalized to millimetres (mm).

Stratum and reference	Literal claims from A; B; C	Alignment and adjudication
Equivalent units, reference 25.4 mm	1 inch; 25.4 mm; 2.54 cm	Normalize to 25.4; 25.4; 25.4 mm. Inch conversion is exact; centimetres use the SI prefix relation. (www.nist.gov)
Near-value disagreement, reference 100 mm	100 mm; 101 mm; 101.2 mm	Keep all values. They concern the same constructed item and observation time.
Rounded display, reference 100.04 mm	100.0 mm; 100.04 mm; 100.1 mm	Assume A rounds to the nearest 0.1 mm using round-half-up, with ties rounded away from zero, giving an interval from 99.95 inclusive to 100.05 exclusive. Keep point-score and interval interpretation separate.

Stratum and reference	Literal claims from A; B; C	Alignment and adjudication
Source-mix stress, reference 10 mm	12 mm; 10 mm; 10 mm	Keep all values. The high-weight source deliberately disagrees with the reference.
Missing unit, reference withheld	12, unit unstated; no claim; no claim	Exclude from comparable numeric scoring; record the missing unit.
Stale observation, current reference 10 mm	12 mm at an earlier observation; 10 mm currently; 10 mm currently	Exclude the earlier claim from current-time fusion. Preserve it in history.
Mixed-type item, reference alpha	beta; alpha; alpha	Treat as categorical. No numerical distance between the labels is defined.

The table separates equivalence, uncertainty, and eligibility rather than declaring every unequal string an error. Its rounded display is compatible with the reference under the stated reporting assumption, although its point value is unequal. This follows the distinction between numerical representation and uncertainty; the assumption is synthetic, not attributed to a real source. (www.oiml.org) (www.nist.gov) The missing-unit and stale strata test abstention and routing, rather than numerical estimation.

Freeze the evaluation contract

Use source weights **A = 0.6, B = 0.3, C = 0.1**, declared synthetic constants. They are fixed across items and deliberately not learned from their reference labels. Compare median plus categorical mode, arithmetic mean plus mode, weighted numeric mean plus mode, and heterogeneous combinations of numerical loss with weighted categorical disagreement. CRH supplies the loss-based motivation, while the fixed-weight comparison isolates the aggregation stage. (cse.buffalo.edu)

Before running the comparison, publish:

- **Eligibility:** Which claims enter each item and why others are excluded.
- **Reference:** The exact target and how it was established.
- **Loss:** Absolute, squared, categorical, or another specified function.
- **Weights:** Their values, training evidence, and any coverage adjustment.
- **Tolerance:** Absolute or relative limits, including boundary equality.
- **Denominator:** Every eligible item and missing-output policy.
- **Duplicates:** The treatment of common upstream evidence.
- **Version:** Frozen input, annotation, and evaluator identifiers.
- **Alternatives:** Rejected candidates and adjudication reasons.
- **Provenance:** Verbatim evidence, source location, and relevant dates.

A provenance graph can link a normalized value to its literal input through derivation. PROV-O also distinguishes entity generation from the subject's observation time. (www.w3.org) Store the reference adjudication as a separate evidence object so that a later change to the gold value can trigger rescoring without rewriting the original claims.

Data Analysis and Evidence

Compare predictions before comparing winners

Table 2 reports calculations from the constructed fixture. The numerical denominator contains only the **four comparable length items**. Mean absolute error (**MAE**) averages absolute deviations; root mean squared error (**RMSE**) takes the square root of average squared deviations, retaining the target unit. (scikit-learn.org) Values below are article calculations rounded for display; equality tests use unrounded decimal arithmetic.

Method and fixed losses	Predictions in fixture order, mm	Categorical output	MAE, mm	RMSE, mm
Median plus mode	25.4; 101; 100.04; 10	alpha	0.250000	0.500000
Mean plus mode	25.4; 100.733333; 100.046667; 10.666667	alpha	0.351667	0.495547
Weighted mean plus mode	25.4; 100.42; 100.022; 11.2	alpha	0.409500	0.635752
Heterogeneous squared loss plus weighted vote	25.4; 100.42; 100.022; 11.2	beta	0.409500	0.635752
Heterogeneous absolute loss plus weighted vote	25.4; 100; 100; 12	beta	0.510000	1.000200

The median and mean reverse order between MAE and RMSE. Their predictions do not change. Absolute loss and squared loss reward different summaries, consistent with the textbook distinction between median-targeting and mean-targeting prediction. (otexts.com) The squared heterogeneous row shares its numerical predictions with the weighted mean; only categorical aggregation changes. The absolute heterogeneous row is the weighted median, which selects A because its weight exceeds half the total. These are fixed-weight estimators, not complete CRH runs.

Tolerance, relative loss, and source mix

Table 3 scores those same numerical predictions against the declared references. Accuracy means the count satisfying absolute error less than or equal to the stated tolerance, divided by **four**. All entries are calculations from Table 1. No missing-unit claim, stale claim, or categorical item enters this denominator.

Numeric estimator	Exact equality	Tolerance 0.05 mm	Tolerance 0.5 mm	Tolerance 1 mm
Median	3/4	3/4	3/4	4/4
Mean	1/4	2/4	2/4	4/4
Weighted squared-loss estimate	1/4	2/4	3/4	3/4
Weighted absolute-loss estimate	2/4	3/4	3/4	3/4

Changing tolerance changes ties and relative standing without changing any source-reliability estimate. At the broader threshold, mean fusion catches median fusion, while weighted fusion still misses the stressed item. The table is a counterexample to treating a single accuracy threshold as a property of the method itself. It does not estimate how often such reversals occur outside this fixture.

Relative loss changes the item weighting again. For this fixture, mean absolute percentage errors are **0.2500%**, **1.8517%**, **3.1095%**, and **5.0100%** for median, mean, weighted squared, and weighted absolute estimates respectively, calculated with nonzero references. Compared with absolute MAE, the shorter stressed item receives greater relative influence. Percentage loss also requires a zero-reference policy: scikit-learn documents an epsilon convention and warns that percentage errors can become arbitrarily large. (scikit-learn.org)

A source-mix intervention illustrates a separate mechanism. Removing synthetic source A from the stressed item leaves B and C both at the reference value, so every listed numeric estimator returns it. This is a designed sensitivity check, not evidence that removing high-weight sources usually helps.

For a substantive benchmark, report:

- **Stratum scores:** Error and eligible count for each annotation category.
- **Exact accuracy:** Correct count and denominator before tolerance is applied.
- **Tolerance accuracy:** The threshold, unit, and inclusive boundary rule.
- **Numeric loss:** MAE and RMSE separately within comparable units.
- **Coverage:** Exclusions, abstentions, and missing outputs with reasons.
- **Source sensitivity:** Fixed, learned, and perturbed weights separately.
- **Label sensitivity:** Results under plausible alternative references.
- **Sampling uncertainty:** The resampling unit and interval interpretation.

The fixture is too small and deliberately constructed to support population inference. Bootstrap confidence intervals concern repeated-sampling coverage, and degenerate distributions can produce undefined interval endpoints in documented implementations. (docs.scipy.org) Prediction-interval coverage is yet another quantity: it measures the proportion of observations inside prediction intervals, rather than confidence in the estimated score. (scikit-learn.org)

Implications and Future Directions

A defensible benchmark distinguishes **truth uncertainty**, **measurement uncertainty**, and **evaluation uncertainty**. The first concerns the reference assertion, the second concerns values attributed to the measurand, and the third concerns what a finite test establishes about method performance. The metrological definition of measurement uncertainty and the repeated-sampling interpretation of confidence intervals do not make those quantities interchangeable. (www.oiml.org) (docs.scipy.org)

The immediate research priority is a factorial evaluation of preprocessing and fusion. Hold the aligned claims fixed while varying loss, then hold the loss fixed while varying normalization, tolerance, time alignment, and source mix. Repeat the experiment across naturally occurring data and controlled perturbations. MaDI-Bench's

released task and evaluation artifacts provide a concrete basis for stage-aware comparisons, but its changing gold conventions also show why the release identifier must accompany a score. (github.com)

Particular attention should go to **uncertain units**, **context-dependent tolerances**, and **ambiguous valid time**. A timestamp can be syntactically explicit while leaving the observation period unclear. OGC's separate time concepts and RFC 3339's offset convention address different parts of that problem. (docs.ogc.org) (www.rfc-editor.org) Researchers should test abstention when semantics remain unresolved, rather than force every input into a scalar.

Public and private evidence also require a common evaluation contract. This report does not establish the availability or absence of a comprehensive private/public fusion benchmark. It identifies the unresolved requirement: independent evaluators need access to permitted evidence, annotation provenance, and reproducible snapshots before they can distinguish a reliability gain from an unavailable-source effect.

Useful next experiments include:

- **Unit rescaling:** Change only one attribute's unit and inspect learned weights.
- **Rounding perturbation:** Vary display precision while preserving latent references.
- **Temporal shifts:** Move observation times without changing publication dates.
- **Gold alternatives:** Rescore items with defensible competing adjudications.
- **Source dependence:** Group shared upstream claims before estimating reliability.
- **Coverage changes:** Add withheld sources without changing the scoring population.

These are proposed tests, not reported benchmark results. They can expose assumptions omitted from a leaderboard. The record needed for them is an evidence history: literal claim, transformation, reference decision, and every excluded alternative. PROV-O's derivation and alternative relations supply vocabulary for that history; they do not select the correct number. (www.w3.org)

Conclusion

Resolving conflicting numerical claims begins with deciding whether the claims conflict at all. The evaluation must establish common quantity, unit, scope, observation period, and measurement basis before it asks which source to trust. Normalization rules, rounding assumptions, time alignment, and missing-context policies belong in the published method.

The constructed fixture demonstrates a narrow but decisive result: the ranking of fixed predictions can change when the evaluator changes absolute loss to squared loss, and tolerance can change which methods tie. That result establishes a possibility, not a prevalence estimate. Its source weights, references, exclusions, and denominators are visible so that the conclusion can be recomputed.

CRH contributes a framework for heterogeneous losses and reliability estimation. MaDI-Bench contributes explicit separation of integration stages and versioned evaluation resources. Their findings should retain their own experimental scope. A published fusion score cannot establish the benefit of reliability estimation when the comparator used different conversions, reference labels, or input clusters.

The recommended reporting standard is therefore an **auditable comparison contract**: preserve source literals and dates, declare comparability, publish conversions and tolerances, separate numeric from categorical scoring, and vary the assumptions that could determine the winner. Where a reference remains uncertain, disclose the uncertainty and show the alternative outcomes. A canonical value becomes defensible through that evidence and its reproducible adjudication, rather than through agreement alone.

Source links

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

1. https://www.oiml.org/en/files/pdf_v/v002-200-e10.pdf/@@download/file/V002-200-e10.pdf
2. https://docs.ogc.org/as/20-082r4/20-082r4.html
3. https://www.nist.gov/pml/owm/si-units-length
4. https://cse.buffalo.edu/~lusu/cse726/papers/Resolving%20Conflicts%20in%20Heterogeneous%20Data%20by%20Truth%20Discovery%20and%20Source%20Reliability%20Estimation.pdf
5. https://arxiv.org/html/2606.30371v1
6. https://arxiv.org/abs/2606.30371
7. https://otexts.com/fpp3/toolbox.html
8. https://www.w3.org/TR/prov-o
9. https://www.census.gov/programs-surveys/popest/about/faq.html
10. https://www.oecd.org/content/dam/oecd/en/publications/reports/2007/06/data-and-metadata-reporting-and-presentation-handbook_g1gh7b15/9789264030336-en.pdf
11. https://wbsg-uni-mannheim.github.io/MaDI-Bench
12. https://github.com/wbsg-uni-mannheim/MaDI-Bench
13. https://arxiv.org/html/2606.30371
14. https://www.nist.gov/pml/owm/metric-si/unit-conversion/approximate-conversions-us-customary-measures-metric
15. https://www.bipm.org/documents/20126/41483022/SI-Brochure-9-EN.pdf/2d2b50bf-f2b4-9661-f402-5f9d66e4b507?download=true&t=1671101192839&version=1.11
16. https://www.nist.gov/pml/special-publication-811/nist-guide-si-appendix-b-conversion-factors
17. https://www.bls.gov/eci/factsheets/eci-constant-dollar-factsheet.htm
18. https://www.dublincore.org/specifications/dublin-core/dcmi-terms
19. https://www.rfc-editor.org/rfc/rfc3339.html
20. https://cse.buffalo.edu/~jing/publications.htm
21. https://pmc.ncbi.nlm.nih.gov/articles/PMC4688022
22. https://experts.illinois.edu/en/publications/dynamic-truth-discovery-on-numerical-data
23. https://hanj.cs.illinois.edu/pdf/icdm18_szhi.pdf
24. https://docs.python.org/3/library/decimal.html
25. https://scikit-learn.org/stable/modules/model_evaluation.html
26. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bootstrap.html
27. https://scikit-learn.org/stable/auto_examples/ensemble/plot_gradient_boosting_quantile.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.