Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

LLM conflicting sources: a counterfactual fusion test

> 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 · 2026-10-19 · llm conflicting sources · llm source conflict resolution · data fusion · truth discovery · rag evaluation · evidence provenance · source reliability · counterfactual evaluation · abstention

Original article: https://reliablecontext.com/articles/llm-conflicting-sources-data-fusion-test



In this article

1. Executive Summary
2. Introduction and Background
3. Methodology: define the fusion decision before testing it
4. Reconstructing the published tabular results
5. A counterfactual matrix for source-based resolution
6. Analysis of key segments and transfer conditions
7. An evidence trace that preserves disagreement
8. Data Analysis and Evidence
9. Implications and Future Directions
10. Conclusion

Executive Summary

Large language models (LLMs) can reconcile conflicting supplied claims in some tested settings, but a correct answer alone does not establish source-based conflict resolution. The distinction is between agreement with reference labels and demonstrated dependence on admissible evidence. Xu et al. (2024) distinguish conflicts between context and model memory from conflicts within supplied context, a useful starting point for defining what a benchmark actually measures. (aclanthology.org) Zhou et al. (2023) explicitly connect contextual faithfulness with both knowledge conflict and abstention. (aclanthology.org)

The focal tabular-fusion study reports **Book F1 of 0.7817** for its domain-dependent one-shot prompt, compared with **0.6955** for the Latent Truth Model (LTM). (arxiv.org) Its author-maintained repository identifies **GPT-4o-mini** as the primary model and reports winning scores of **0.8172 for Movie** and **0.9119 for Flight**. (github.com) These are reported findings, not independently reproduced measurements in this report. The public implementation supplies useful inspection points, including candidate-only instructions and a scorer that aggregates values across entities. (github.com)

The Flight obfuscation result requires particular care. The best score changes from **0.9119 to 0.9048**, but the winning prompt changes too; keeping the original winning prompt fixed gives **0.8970** after masking. (arxiv.org) The implementation changes the flight identifier while retaining the associated claims, so this intervention does not remove every possible identifying cue. (github.com) A stronger evaluation should therefore separate entity masking, source-label permutation, claim order, dependent-source duplication, evidence removal and unsupported targets. Order controls have an independent motivation in Liu et al.'s context-position experiments, while explanation text requires separate scrutiny under Turpin et al.'s findings on reasoning faithfulness. (aclanthology.org) (proceedings.neurips.cc)

This report reconstructs the published comparison and proposes a **counterfactual evaluation protocol**, rather than claiming to have run that extension. It pairs set-valued correctness with exact-set accuracy, evidence support, reason-coded unknowns, calibration and repeated-run variation. Exact-set accuracy requires the complete predicted set to match. (scikit-learn.org) Paired resampling preserves matched observations, and selective prediction evaluates answering together with rejection. (docs.scipy.org) (proceedings.mlr.press) The resulting evidence package should retain literal claims, source snapshots and

observation dates, consistent with provenance and data-citation principles. (www.w3.org) (force11.org) Success would mean surviving these tests within a declared evidence regime, not proving universal conflict-resolution ability or private-domain transfer.

Introduction and Background

The query **“LLM conflicting sources”** hides several different questions. Can a model choose a value when sources disagree? Can it distinguish independent support from repeated publication of the same claim? Can it identify that the supplied evidence does not determine an answer? A benchmark needs to distinguish these capabilities before interpreting an aggregate score.

In **data fusion**, the decision concerns the value or set of values assigned to an entity's attribute after claims have been grouped. The focal implementation groups Book claims by identifier, Movie claims by title and year, and Flight claims by flight identifier before prompting. (github.com) Consequently, its fusion results do not by themselves evaluate entity resolution, the separate problem of deciding which records refer to the same thing. This boundary matters whenever an entity key can also help a model retrieve memorized facts.

The conflict taxonomy is equally consequential. A disagreement among documents is an **inter-context conflict**; disagreement between a document and learned model knowledge is a different experimental condition. Xu et al. (2024) distinguish these categories explicitly. (aclanthology.org) In retrieval-augmented generation (RAG), retrieval adds another stage whose behavior can influence the apparent quality of subsequent reconciliation. Chen, Zhang and Choi (2022) find that retrieval performance affects which knowledge sources their tested models rely on. (aclanthology.org)

This report addresses researchers designing or reviewing truth-discovery benchmarks. It reconstructs published evidence, examines what the available implementation establishes, and specifies an extension that could be preregistered. All web evidence was accessed on the stated publication date, **October 19, 2026**; study dates and versions are identified separately. It makes no new model-performance measurement. Its central question is narrower than whether an LLM knows the public answer: **does its output follow the supplied claims under interventions that preserve or deliberately change their evidential meaning?** Context-faithful prompting research provides an appropriate precedent for separating those questions. (aclanthology.org)

Methodology: define the fusion decision before testing it

Truth cardinality and evidence regimes

A **single-truth attribute** permits a single valid value within a declared scope. A **set-valued attribute** permits multiple valid values simultaneously. The authors' task descriptions identify Book authors and Movie directors as multi-truth cases, and Flight times and gates as single-truth cases. (github.com) Cardinality must be defined for the attribute and time interval, rather than inferred from the number of claims.

The evaluator should preregister separate evidence regimes:

- **Candidate-bounded fusion:** every selected value must correspond to an admissible supplied claim.

- **Evidence synthesis:** derivations from supplied claims are allowed under an explicit rule, with the derivation recorded.
- **Open-world answering:** external information may be consulted, with retrieval and its cost measured separately.
- **Insufficient-evidence decisions:** the target is an explicit unknown when admissible evidence cannot justify a selection.
- **Temporal decisions:** the target concerns a specified validity interval, with observation time recorded separately.

These are proposed experimental definitions. They prevent a model that supplies a true but unsupported answer from receiving the same interpretation as a model that reconciles the given evidence. The published candidate constraint already asks for values appearing in the sources, making enforcement and evaluation of that instruction a natural extension. (github.com)

Reliability, dependence and normalization

Source reliability should refer to an operationally defined property, such as performance on held-out labeled claims in the same domain. LTM jointly infers truth and source quality; its original formulation is unsupervised. (hanj.cs.illinois.edu) Lin and Chen's domain-aware multi-truth discovery method, DART, instead motivates domain-sensitive source expertise and partially correct multi-value answers. (researchportal.hkust.edu.hk)

Repeated claims must not automatically count as independent corroboration. Dong, Berti-Equille and Srivastava's work explicitly treats fusion with copying sources; their dependence analysis also cautions that common data alone does not establish copying. (research.google) (www.vldb.org) An evaluator should therefore distinguish experimentally known duplication from an inferred dependence relationship.

Normalization belongs inside the registered method. The focal normalizer performs transformations before person-name comparison, and its scorer matches normalized names by token-set equality. (github.com) Such equivalences can be useful, but an extension must preserve the original strings alongside normalized candidates. Otherwise a score can obscure whether the model selected the right value, whether preprocessing reconciled it, or whether a matching convention accepted a different representation.

Reconstructing the published tabular results

Kucuk, Paton, Chen and Wu (2026) compare domain-dependent/independent, zero/one-shot prompts. The setup specifies GPT-4o-mini, temperature zero and **256 output tokens**. (arxiv.org) The public configuration identifies the default model as **gpt-4o-mini**. (github.com) The scores below should be read as the authors' reported experiment, not a fresh execution or an independently established ranking.

Table 1 reconstructs the reported F1 comparison. **DD** denotes domain-dependent prompting; **C1** denotes the supplied-value constraint. The LLM row uses different winning configurations for the multi-truth tasks and Flight. (github.com)

Method or configuration	Book, multi-truth	Movie, multi-truth	Flight, single-truth	Interpretation and source
DART	0.5997	0.7768	0.8056	Reported baseline. (arxiv.org)
LTM	0.6955	0.7959	0.8214	Reported baseline. (arxiv.org)
Majority vote	0.5765	0.5826	0.7867	Reported baseline. (arxiv.org)
Source reliability vote	0.2498	0.3935	0.8214	Reported baseline. (arxiv.org)
GPT-4o-mini, selected prompt	**0.7817**, DD-1shot	**0.8172**, DD-1shot	**0.9119**, DD-C1-0shot	Author-maintained result summary. (github.com)

The table supports a bounded conclusion: the selected LLM configurations exceed the listed baseline F1 values on these tasks. It does not identify the mechanism producing that advantage. In particular, source-aware inference, normalization, public knowledge and prompt-specific behavior remain different possible contributors.

What the metric and documentation establish

For multi-truth scoring, the implementation aggregates matched, missed and erroneous values across entities before calculating precision, recall and F1. (github.com) This is distinct from requiring every entity's entire answer set to be correct. An extension should retain the original scorer for comparability and add an explicitly documented exact-set evaluation.

Several reproduction details remain unresolved in the inspected material. A verified evaluated-entity denominator, the exact Book truth-file selection and a repeated-run uncertainty estimate were not established. These gaps limit what can be inferred about reproducibility. A future execution manifest should resolve the data selection and record every request setting.

Dataset totals cannot fill these gaps. The originating Book page distinguishes **1,263 available books** from a **100-book manually checked gold subset** and a broader silver standard. (lunadong.com) The Flight originator describes **over 1,200 flights**, **38 sources** and a **100-flight gold subset** collected in **December 2011**. (lunadong.com) These figures describe available datasets, not confirmed evaluation denominators. Lin's separate Book and Movie dataset page states a research-use restriction; an explicit reuse license for every processed artifact was not established. (heathersherry.github.io)

A counterfactual matrix for source-based resolution

What Flight-ID masking does and does not test

The masking intervention changes the identifier sent to the prompt while retaining that flight's claims. (github.com) It therefore tests sensitivity to that identifier under the remaining input, rather than complete removal of public-world cues. Holding the prompt fixed separates identifier sensitivity from configuration selection.

A stronger experiment should freeze model settings, prompts, parsing, normalization and evaluation labels before running transformations. Each transformed fixture should retain a reversible mapping to its original. Results should be paired at the entity and attribute level. Liu et al.'s findings on position-dependent context use motivate an order control; they do not establish the magnitude of an order effect in tabular fusion. (aclanthology.org) (cs.stanford.edu)

Table 2 specifies the proposed matrix. Invariance is appropriate only when the transformation preserves the available evidence and its declared semantics.

Intervention	What changes	What remains fixed	Diagnostic expectation
Entity masking	Replace names and public identifiers with opaque aliases.	Claims, dates, source relationships and reference labels.	Measure loss attributable to identifying cues.
Source-label permutation	Apply a global bijection to opaque source identifiers.	Source identity across records and any declared reliability metadata.	Equivalent arbitrary labels should not alter selection.
Claim-order permutation	Reorder rows, sources and candidates.	Complete claim multiset and semantic content.	Measure order sensitivity and output-set disagreement.
Dependent duplication	Add marked copies linked to an existing source family.	Number of independent observations.	Duplicated support should not count as independent confirmation.
Support removal	Remove all or selected admissible support for a target.	Unrelated claims and the original target definition.	Require a justified change or explicit unknown when support disappears.
Unsupported target	Withhold every supplied occurrence of the reference value.	Candidate-only instructions and distractor claims.	An unsupported remembered answer fails the evidence constraint.
Contradiction removal	Remove a competing claim while preserving target support.	Remaining provenance and evaluation policy.	Check whether the response changes in the policy-prescribed direction.
Semantics-preserving rewrite	Change formatting or equivalent surface forms.	Meaning, cardinality, provenance and dates.	Separate representation sensitivity from evidence sensitivity.

The matrix is an experimental proposal, not a reported result. It tests different hypotheses rather than combining every perturbation into a single robustness score. A model can legitimately change after relevant support is removed while remaining stable under arbitrary relabeling.

Preserve the information needed for a fair test

A source-label test needs special care. Reassigning a different alias on every record would destroy the cross-record identity needed to estimate source reliability. The proposed transformation instead preserves a consistent alias mapping. Likewise, deleting a source's declared expertise changes evidence, whereas replacing an arbitrary identifier need not. DART's domain-sensitive treatment of source expertise shows why these are distinct interventions. (researchportal.hkust.edu.hk)

The independent-source baseline and dependent-duplication condition should also be reported separately. DART's original formulation assumes source independence. (www.vldb.org) A deliberately dependent fixture changes that assumption and should be labeled accordingly.

A useful execution sequence is:

- **Freeze:** register evidence regimes, equivalence rules and selection criteria.
- **Snapshot:** preserve the original claims before transformations.
- **Pair:** link every transformed entity to its original evaluation unit.
- **Isolate:** test individual interventions before combined stress conditions.
- **Audit:** check that masking has not accidentally changed claim meaning.
- **Inspect:** classify failures as selection, support, parsing or abstention errors.
- **Release:** publish permitted fixtures, mappings and evaluation code.
- **Restrict:** retain nonpublic evidence only under its applicable permissions.

These controls ask whether the answer follows evidence. They do not make explanation text a causal measurement. Turpin et al. (2023) show that generated chains of thought can omit factors affecting predictions, while Anthropic's original faithfulness study intervenes on reasoning text itself. (proceedings.neurips.cc) (www.anthropic.com) Claim-removal experiments therefore complement explanation inspection, though redundant support and genuinely irrelevant claims must be distinguished before interpreting an unchanged answer.

Analysis of key segments and transfer conditions

Single truths, multiple truths and unknowns

Single-truth evaluation should report attribute accuracy alongside unsupported selections and justified abstentions. Set-valued evaluation should additionally distinguish missing members, extra members and exact-set errors. DART's treatment of partially correct source answers is a reminder that a source can contribute useful evidence without supplying a complete correct set. (researchportal.hkust.edu.hk)

The focal Flight prompt already permits **N/A** for unknown attributes. (github.com) An extension should preserve that capability while separately evaluating whether abstention is appropriate. A string representing an unknown is not evidence that the system recognizes insufficient support.

The protocol should define:

- **Not found:** admissible evidence does not establish a value.
- **Blocked:** required evidence exists or is expected, but access is unavailable.
- **Not applicable:** the attribute does not apply under the fixture's schema.
- **Unresolved conflict:** admissible alternatives remain without a registered tie-breaking basis.
- **Known empty set:** evidence positively supports no members for a set-valued attribute.
- **Parser failure:** the output cannot be interpreted under the required schema.

These are proposed output states, not interchangeable truth labels. A known empty set differs from not knowing the set, and a blocked source differs from a source that was inspected and contained no relevant claim. Selective prediction provides the formal framework for measuring decisions that include rejection. (proceedings.mlr.press)

Public knowledge, private fixtures and retrieval

A public benchmark can mix evidence reconciliation with learned knowledge. Kortukov et al. (2024) find that knowledge-update conclusions depend on whether conflicts use realistic documents or synthetic counterfactuals. (openreview.net) Neither construction should be treated as a complete substitute for the other.

The proposed extension should compare public fixtures with permissioned nonpublic fixtures where reuse and access are authorized. Synthetic aliases should preserve relational structure without inadvertently making the problem easier. Held-out entities, source families and time periods should be separated deliberately; otherwise information needed for a source-reliability estimate can cross the evaluation boundary.

For RAG, record both the evidence that was available and the evidence actually retrieved. Cattan et al. (2025) use a conflict taxonomy and evaluate responses with explicit conflict-category information, while Chen, Zhang and Choi find limited confidence response to contradictions in their setup. (research.google) (aclanthology.org) These findings motivate separate retrieval, conflict-detection and selection measurements.

Finally, **MaDI-Bench is not the same experimental object**. Its current paper is version 2, dated **October 5, 2026**. (arxiv.org) Its **P3** workflow uses an LLM to configure PyDI integration methods, rather than directly asking it to elect every fused value. (arxiv.org) Stage-specific evaluation can help locate an integration gap, but an end-to-end score cannot isolate the effect of direct LLM value fusion. The benchmark's official project page explicitly supports scoring individual stages. (wbsg-uni-mannheim.github.io)

An evidence trace that preserves disagreement

A provenance trace should make the decision inspectable without replacing the original claims. The World Wide Web Consortium's **PROV** model distinguishes entities, activities and responsible agents; its primer treats revisions as new entities. (www.w3.org) A benchmark can use these concepts to distinguish a source snapshot, an extracted claim, a transformation and a scored output.

Table 3 presents a synthetic fixture **(Hypothetical Example)**. Its observation dates indicate when each statement was seen, not when its asserted status became valid. No row represents an actual organization or source.

Claim and source	Literal synthetic statement	Observed date	Relationship and alternatives	Permitted decision under the example's policy
c-alpha, s-alpha	“Entity Q has status active.”	2026-09-01	Independent origin; supports active and contradicts inactive for the same target interval.	Preserve active as a supported candidate.
c-beta, s-beta	“Entity Q has status inactive.”	2026-09-03	Independent origin; supports inactive and contradicts active.	Preserve inactive as a supported candidate.

Claim and source	Literal synthetic statement	Observed date	Relationship and alternatives	Permitted decision under the example's policy
c-copy, s-beta-copy	“Entity Q has status inactive.”	2026-09-03	Explicit copy of c-beta; adds a publication, not an independent origin.	Retain the claim and dependency edge.
Evaluated output	“Unknown: unresolved conflict.”	Fixture evaluation	No declared quality distinction or validity-time evidence resolves the alternatives.	Preserve both alternatives and identify the missing basis for selection.

A publication majority would favor inactive in this example, while a count of independent origins would be tied. The stipulated policy therefore returns an unresolved conflict. A later observation alone does not resolve the status at the target time. The example illustrates the policy's consequences, not a universal requirement to abstain whenever claims disagree.

The proposed run manifest

The release should retain:

- **Identity:** fixture identifier, entity mapping and attribute definition.
- **Claims:** literal text, normalized value, source identifier and claim identifier.
- **Time:** observation date, asserted validity interval and snapshot identifier.
- **Dependence:** known duplication edges and separately inferred relationships.
- **Execution:** model identifier, request settings, prompt bytes and tool permissions.
- **Artifacts:** repository revision, parser, normalizer and reference-label hashes.
- **Outputs:** raw response, parsed selection, cited claims and unknown reason.
- **Evaluation:** scorer version, exclusions, run identifier and paired results.

FORCE11's citation principles call for identification of the specific timeslice, version or granular data supporting a claim. (force11.org) Its access principle includes the associated documentation and code needed for informed reuse. (force11.org) This supports releasing an evidence package rather than only an answer table.

Persistent identifiers need additional detail. **DataCite** permits some minor updates under the same digital object identifier (DOI), while providing relations between versions. (support.datacite.org) **Zenodo** distinguishes separate version records, each with its own files and identifier. (help.zenodo.org) An evaluator should consequently record version-specific identifiers and content hashes rather than assuming that a landing-page URL identifies immutable bytes.

The **RO-Crate 1.2** specification offers a concrete packaging option, describing contextual metadata and a machine-readable JSON-LD document. JSON-LD means JavaScript Object Notation for Linked Data. (www.researchobject.org) Using a standard representation can aid inspection; its presence does not certify that a selected value is correct.

Data Analysis and Evidence

The Book difference between the selected LLM score and LTM is **0.0862 F1**, calculated as **0.7817 minus 0.6955**. (arxiv.org) For Movie, the corresponding difference is **0.0213**, using **0.8172 minus 0.7959**. (github.com) (arxiv.org) These are arithmetic differences on the reported scale, not percentage-point improvements in entity accuracy.

For Flight, the best-to-best masking difference is **0.0071**, whereas the fixed original winning prompt declines by **0.0149**. (arxiv.org) The former includes selecting a new winner; the latter holds the prompt fixed. Neither establishes statistical significance without paired outcomes and an appropriate uncertainty analysis.

The proposed exact-set denominator is every eligible labeled entity-attribute pair, with a numerator counting complete matches. The support denominator is every selected value, with a numerator counting selections justified by identified admissible claims. Report these separately: restricting support checks to correct answers would hide unsupported selections.

The recommended scorecard has separate components:

- **Set F1:** report matched, missed and extra values, with the aggregation rule stated.
- **Exact-set accuracy:** require complete agreement for each evaluated entity-attribute set.
- **Evidence support:** measure whether each selected value is licensed by identified supplied claims.
- **Unknown quality:** evaluate justified abstention and inappropriate abstention separately.
- **Calibration:** assess probability predictions and correctness frequencies on held-out data.
- **Stability:** report variation across repeated runs and paired transformations.

With true positives TP, false positives FP and false negatives FN, **F1 = 2TP / (2TP + FP + FN)**. (scikit-learn.org) Exact-set accuracy requires full-set agreement. (scikit-learn.org) In a second calculation **(Hypothetical Example)**, predicting a set containing a single correct member when the reference contains that member plus another yields F1 **2/3** and exact-set accuracy **0**, directly from those definitions. (scikit-learn.org) This explains why a stronger set score need not imply more fully resolved records.

For confidence, use a proper score such as the **Brier score**, but do not equate a lower Brier loss with better calibration alone. (scikit-learn.org) For abstention, plot selective accuracy against coverage, the proportion of eligible cases receiving answers, with unsupported selections shown separately. (proceedings.mlr.press)

Uncertainty analysis should resample paired entities and retain their claims and perturbations together. This is the proposed fusion-specific adaptation of paired bootstrap resampling; SciPy's implementation preserves shared indices across paired arrays. (docs.scipy.org) Confidence intervals should be computed for differences between methods, with the confidence level and resampling procedure disclosed. Repeated model calls address another uncertainty source: Blackwell, Barry and Cohn document nondeterminism even with temperature zero and a fixed seed in their studied setting. (www.turing.ac.uk)

Finally, contextual benchmarks provide perspective rather than substitute evidence. **ConfRAG** contains **1,814 questions**, averaging **9.58 retrieved paragraphs**, with explicit contradictions in **57.2%** of its questions. (aclanthology.org) Those figures describe its constructed benchmark, not general conflict prevalence. Its answer-clustering and coverage metrics examine a different output objective from selecting fused tabular values. (aclanthology.org)

Implications and Future Directions

An LLM source-conflict benchmark should yield an inspectable explanation of its score. The proposed acceptance question is whether performance survives transformations that remove irrelevant cues, responds appropriately when evidence changes, and remains useful when abstention is allowed. These are experimental criteria to measure, rather than guarantees inferred from fluent explanations.

Baseline comparisons need identical records, candidate normalization, label access and evidence budgets. An open-world LLM should not be described as a better closed-world fusion algorithm merely because it can use information unavailable to its comparator. Conversely, a source-aware baseline should receive the same permitted reliability metadata. LTM's joint quality inference and DART's domain-aware expertise illustrate distinct assumptions worth preserving in the comparison. (hanj.cs.illinois.edu) (researchportal.hkust.edu.hk)

The next research priorities are:

- **Contamination controls:** compare identifiable public records with masked and authorized nonpublic fixtures.
- **Dependency controls:** distinguish repeated publication from additional independent observations.
- **Temporal controls:** evaluate a declared target time while retaining observation history.
- **Selection controls:** reserve held-out records for prompt and threshold evaluation.
- **Failure inspection:** publish unsupported selections, unresolved cases and parsing exclusions.
- **Transfer evidence:** test conditions resembling intended use before generalizing.

The National Institute of Standards and Technology's **AI Risk Management Framework** calls for documented test sets, metrics and tools, uncertainty measures, and assessment under conditions similar to deployment. (airc.nist.gov) These are relevant documentation principles, not certification of this proposed protocol.

Reliable Context's position as a system-of-record publisher gives this discussion an emphasis on defensible values and retained evidence. The present report evaluates research methods and proposed tests; it supplies no performance result for the publisher or account of its internal implementation.

A useful follow-up publication would release a versioned fixture and paired execution manifest, including negative and unknown results. Zenodo's distinction between file versions and metadata edits provides one practical model for maintaining that record. (help.zenodo.org) The research claim should then match the evidence: a measured gain under the released conditions, with its remaining failure modes visible.

Conclusion

LLMs can resolve some conflicting-information tasks in tested settings, but source-based resolution requires more evidence than a high aggregate F1. The focal results make direct tabular fusion worth investigating. They do not, by themselves, establish independence from memorized knowledge, robustness to dependent sources, or transfer to private evidence.

The proposed counterfactual evaluation keeps claims and their provenance explicit while changing identifiers, labels, order, duplication and admissible support. Its purpose is to separate useful invariances from justified sensitivity. Stability under arbitrary relabeling is desirable; stability after the removal of essential support may

indicate a different problem.

Evaluation should preserve the original scoring convention for comparison while adding exact-set correctness, source support, unknown quality, calibration and repeat-run variation. A model should receive distinct credit for selecting a supported value, identifying a genuinely unresolved conflict and declaring that relevant evidence is unavailable. Those outcomes answer different operational questions.

The durable research product is therefore a reproducible evidence package: inputs, transformations, model configuration, raw outputs, scoring rules and uncertainty estimates. A generated explanation can assist inspection, but the strongest test changes the evidence and measures the consequence. Researchers can then state precisely what improved, which information was available, and where the method still cannot justify a fused answer.

Source links

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

1. https://aclanthology.org/2024.emnlp-main.486
2. https://aclanthology.org/2023.findings-emnlp.968
3. https://arxiv.org/html/2606.28062
4. https://github.com/hiraberil/LLM-DataFusion
5. https://github.com/hiraberil/LLM-DataFusion/blob/main/llm/prompt_builder.py
6. https://github.com/hiraberil/LLM-DataFusion/blob/main/run.py
7. https://aclanthology.org/2024.tacl-1.9
8. https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html
9. https://scikit-learn.org/stable/modules/model_evaluation.html
10. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bootstrap.html
11. https://proceedings.mlr.press/v97/geifman19a.html
12. https://www.w3.org/TR/prov-primer
13. https://force11.org/group/joint-declaration-of-data-citation-principles-final
14. https://github.com/hiraberil/LLM-DataFusion/blob/main/data/loader.py
15. https://aclanthology.org/2022.emnlp-main.146
16. https://hanj.cs.illinois.edu/pdf/vldb12_bzhao.pdf
17. https://researchportal.hkust.edu.hk/en/publications/domain-aware-multi-truth-discovery-from-conflicting-sources
18. https://research.google/pubs/data-fusion-resolving-conflicts-from-multiple-sources
19. https://www.vldb.org/pvldb/vol2/vldb09-pvldb47.pdf
20. https://github.com/hiraberil/LLM-DataFusion/blob/main/data/normalizer.py
21. https://github.com/hiraberil/LLM-DataFusion/blob/main/config.py
22. https://github.com/hiraberil/LLM-DataFusion/blob/main/evaluation/metrics.py
23. https://lunadong.com/fusionDataSets.htm
24. https://heathersherry.github.io

25. https://cs.stanford.edu/~nfliu/papers/lost-in-the-middle.tacl2023.pdf
26. https://www.vldb.org/pvldb/vol11/p635-lin.pdf
27. https://www.anthropic.com/news/measuring-faithfulness-in-chain-of-thought-reasoning
28. https://openreview.net/forum?id=xm8zYRfrqE
29. https://research.google/pubs/dragged-into-a-conflict-detecting-and-addressing-conflicting-sources-in-retrieval-augmented-llms
30. https://arxiv.org/html/2606.30371
31. https://wbsg-uni-mannheim.github.io/MaDI-Bench
32. https://support.datacite.org/docs/versioning
33. https://help.zenodo.org/docs/deposit/manage-versions
34. https://www.researchobject.org/ro-crate/specification/1.2/introduction.html
35. https://proceedings.mlr.press/v97/geifman19a/geifman19a.pdf
36. https://www.turing.ac.uk/news/publications/towards-reproducible-llm-evaluation-quantifying-uncertainty-llm-benchmark-scores
37. https://aclanthology.org/2026.acl-long.11
38. https://airc.nist.gov/airmf-resources/airmf/5-sec-core

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.