Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Constraint-based AI evaluation: When exceptions matter

> 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 · 2026-10-13 · constraint-based ai evaluation · truth discovery · denial constraints · exception-aware evaluation · evidence provenance · data fusion · ai factual accuracy · constraint-based reasoning · independent holdouts

Original article: https://reliablecontext.com/articles/constraint-based-ai-evaluation-exceptions



In this article

1. Executive Summary
2. Introduction and Background
3. Definitions and evaluation targets
4. Relationship-aware inference and formal reasoning
5. Approximate constraints and business-rule exceptions
6. An independent holdout and constraint-perturbation protocol
7. Metrics, uncertainty, and stop conditions
8. Data Analysis and Evidence
9. Implications and Future Directions
10. Conclusion

Executive Summary

Constraint-based artificial intelligence (AI) evaluation should test whether cross-record relationships improve independently labeled factual accuracy while preserving evidence support, legitimate exceptions, and declared uncertainty. It should not treat rule compliance as the reference truth. Ye and colleagues' 2022 constrained truth-discovery abstract describes denial-constraint-aware optimization and reports favorable real and synthetic results, but it provides no portable effect size. (researchconnect.buffalo.edu) Livshits and colleagues' 2020 approximate-constraint work addresses a different question: how to discover useful rules that allow exceptions rather than overfit observations. (www.vldb.org)

The proposed acceptance test separates **claim evidence**, **rule provenance**, and **exception policy**. It freezes those inputs before a final independently adjudicated holdout, then compares constrained and unconstrained outputs on identical evidence. Test data must not guide model choices, including rule discovery or selection. (scikit-learn.org) The benchmark should vary incorrect rules, context mismatches, and legitimate exceptions separately, while reporting the fraction of claims exposed to each intervention. NIST's blind-evaluation program shows that independent testing can also use data that is not publicly released. (pages.nist.gov)

Three findings shape the scorecard. First, exception denominators matter: Livshits and colleagues' example has **0.95% violating ordered pairs** but requires removing **13.3% of tuples** under a different measure. (www.vldb.org) Second, source attribution and independent correctness differ: attribution evaluates whether evidence supports a statement, not whether the evidence itself supplies an independent truth label. (research.google) Third, abstention changes the denominator: selective classification trades coverage for accuracy, requiring both to be reported. (www.jmlr.org) Calibration also requires more than a lower Brier loss, which combines several aspects of predictive quality. (scikit-learn.org)

The report provides a **20-setting proposed sensitivity grid** and an explicitly hypothetical calculation. In that calculation, accuracy rises from **80% to 83%**, while **45 of 120 changed values**, or **37.5%**, are harmful overrides. The illustrative **95% Wilson interval is 29.4% to 46.4%**, assuming independent cases and calculated using the NIST formula. (www.itl.nist.gov) These are invented benchmark inputs and derived

outputs, not published system results. Approval should depend on gains surviving credible perturbations, acceptable harmful and unsupported overrides, and honest handling of unknowns. Where rule provenance or label independence cannot be defended, the claimed benefit remains conditional.

Introduction and Background

Constraint-based AI evaluation asks whether a system uses relationships among claims correctly, and whether those relationships improve independently measured factual accuracy. This report uses the term for an evaluation program, rather than for a particular algorithm. The immediate decision is whether a cross-record rule deserves influence when sources disagree. A plausible relationship is a hypothesis to test; its apparent plausibility does not supply a reference label.

Truth discovery estimates values from noisy, conflicting sources while estimating source reliability. Li and colleagues describe those quantities as interdependent; Dong and colleagues characterize fusion as resolving conflicts to identify values reflecting the real world. (arxiv.org) **Truth discovery with conflicting evidence** therefore needs more than a consistency score. The benchmark must establish whether the chosen value is correct, whether its evidence supports it, and whether a rule changed the answer.

The evaluation boundary matters. A reasoning benchmark may ask what follows from supplied premises, whereas a fusion benchmark asks whether those premises and inferred values correspond to independently established facts. ProofWriter evaluates questions from supplied context and uses synthetically generated logic programs; that makes it useful for isolating reasoning, without establishing the reliability of real sources. (aclanthology.org)

For organizations combining public and private material, defensibility also includes observable evidence properties. **Reliable Context** describes a cited, verbatim, dated source behind every value. That is a first-party statement of a record property, rather than an independently measured constraint-inference result. The distinction keeps provenance guarantees separate from factual-accuracy estimates.

This report proposes an **exception-aware AI evaluation** protocol: freeze the evidence, register the constraints and exception policy, label an independent holdout, and perturb rule correctness and context. It then measures both improvements and harmful overrides. The protocol is an analyst proposal grounded in the literature below. Its suggested stress settings are experimental design choices, not published operating limits or claims about any production system.

Definitions and evaluation targets

A **denial constraint** states that a conjunction of specified predicates must never hold for the relevant records. A violation occurs when a record combination satisfies that prohibited conjunction. The formal definition concerns the supplied data and rule; it does not, by itself, identify a false source claim. (www.vldb.org) For evaluation, a violation can arise from an erroneous claim, an inappropriate rule, incompatible time scopes, or a legitimate exception.

A **hard invariant** is a rule asserted to hold throughout a precisely specified domain. A **learned heuristic** is an empirical regularity whose support depends on observations and an approximation definition. An expert-supplied rule can belong to either category. Expert provenance identifies who supplied it, while evidence and scope establish what authority it should have. Approximate-constraint research explicitly permits generally valid rules with exceptions and accepts different approximation semantics. (arxiv.org)

An **exception** here means an independently justified circumstance in which applying the general rule would be inappropriate. A **wrong constraint** makes an incorrect assertion within its stated scope. A **context-mismatched constraint** may be defensible elsewhere but misapplied to a different population, definition, or period. These are proposed benchmark labels, and annotators should retain disagreements rather than converting uncertain cases into definite rule errors.

Source support is a claim's evidential backing: identifiable passages, relevant scope and dates, and distinguishable origins. It is separate from independent truth. Rashkin and colleagues' Attributable to Identified Sources framework evaluates whether generated statements are supported by underlying sources. (research.google) **Evaluating AI evidence provenance** requires that support test alongside a correctness test, because an accurately quoted source can still disagree with an independently adjudicated label.

Table 1 compares evaluation targets. The canonical-record option is evaluated through its stated observable properties, alongside the distinct goals of inference and validation.

Approach or option	Observable target	What an evaluator still needs
Unconstrained truth discovery	Infer values from sources and source-reliability estimates. (arxiv.org)	Independent value labels and an evidence-matched baseline.
Constrained truth discovery	Infer values under cross-entity relationships, as described by Ye and colleagues. (researchconnect.buffalo.edu)	Rule provenance, scope, exceptions, and sensitivity to incorrect rules.
Constraint validation with Shapes Constraint Language (SHACL)	Validate Resource Description Framework (RDF) graphs against supplied conditions. (www.w3.org)	Evidence that the conditions are appropriate, plus separate factual labels.
Reliable Context, a canonical record system	Its stated guarantees include cited sources, declared unknowns with reasons, and reproducible past states.	Independent evaluation of those properties; these statements provide no constrained-inference benchmark.

These targets are complementary, but their measurements answer different questions. A conforming graph, a source-supported sentence, a reproducible record, and a correct inferred value each require a distinct test. The table supplies no ranking and no inference about the internal mechanism of the record system.

Relationship-aware inference and formal reasoning

Ye and colleagues' **Constrained Truth Discovery** (2022) introduces denial constraints into truth discovery as universally quantified first-order relationships. Their institutional abstract describes a constrained optimization problem, partitioning entities into disjoint groups, and minimizing an objective under corresponding arithmetic

constraints. (researchconnect.buffalo.edu) The relevant methodological idea is that relationships can contribute information when reliable sources are sparse.

The authors report favorable results on real and synthetic datasets, including settings with few constraints and reliable sources. The public abstract does not establish a portable numerical effect size. (researchconnect.buffalo.edu) The full paper could not be retrieved in this research session. Accordingly, this report does not assert its dataset composition, constraint provenance, metric values, reproduction settings, or sensitivity to erroneous rules. Those details remain necessary for a reproduction assessment.

Constraint-based reasoning evaluation should first distinguish logical feasibility from factual accuracy. Satisfiability modulo theories (SMT) checks satisfiability of first-order formulas relative to a logical theory. SMT-LIB describes shared standards and benchmarks as aids to evaluation and comparison. (smt-lib.org) A solver result addresses the encoded premises, not the independent authenticity of those premises.

For example, **Z3** defines a satisfiable result through an interpretation making the asserted formulas true. Its documentation also permits an unknown result when satisfiability cannot be determined. (microsoft.github.io) These provide useful separate benchmark outcomes: feasible, infeasible, and computationally undecided. Computational uncertainty should not be silently relabeled as missing real-world evidence.

Testing AI reasoning under constraints should therefore isolate several operations:

- **Premise extraction:** Does the system encode the supplied statements faithfully?
- **Scope matching:** Are compared values defined for compatible populations and periods?
- **Logical consequence:** Does the proposed answer follow from the encoded premises?
- **Feasibility handling:** Does the system report incompatible premises explicitly?
- **Evidence use:** Is the returned value supported by accessible observations?
- **World correspondence:** Does an independent label confirm the returned value?

The first and third operations can be evaluated with controlled logical tasks; the last requires evidence beyond self-consistency. ProofWriter's context-based setup illustrates the former boundary. (aclanthology.org) A system can reason correctly from a mistaken premise, so a reasoning score alone cannot authorize a cross-record override.

Hard and soft rule implementations also deserve separate reporting. Z3 documents weighted soft constraints and notes that repeating a soft constraint counts repetitions. (microsoft.github.io) For this proposed protocol, duplicated or correlated rules should consequently be recorded as an experimental factor. Otherwise, a change in rule multiplicity can be mistaken for additional independent knowledge.

Approximate constraints and business-rule exceptions

Livshits, Heidari, Ilyas and Kimelfeld (2020) examine approximate denial constraints because rules satisfied by every observed record can become overly specific in inconsistent data. Their formulation permits exceptions and seeks to avoid overfitting observations. (www.vldb.org) This supplies a reason to test approximate rules, rather than evidence that any particular mined rule is suitable for inference.

Their **ADCMiner** algorithm for mining approximate denial constraints accepts approximation semantics as input. Different definitions can yield different results, so an exception threshold is incomplete without its denominator and interpretation. (arxiv.org) The paper considers measures based on satisfying tuple pairs, tuples not participating in violations, and the largest satisfying subinstance. (www.vldb.org) These should not be collapsed into one interchangeable support percentage.

Data validation with business rule exceptions needs an explicit exception policy. SHACL, the Shapes Constraint Language, produces a validation report from data and shapes and requires input graphs to remain unchanged during validation. (www.w3.org) This illustrates a useful separation: reporting that a supplied condition is unmet need not entail changing the record to satisfy it.

For each rule, the proposed benchmark should register:

- **Statement:** Preserve both the formal predicates and the readable interpretation.
- **Provenance:** Identify the expert, mined dataset, or authoritative definition.
- **Population:** Specify which entities and record combinations are eligible.
- **Time scope:** Distinguish the period described from the observation date.
- **Approximation:** State the denominator, threshold, and estimation procedure.
- **Exceptions:** Record independently justified cases and unresolved candidates.
- **Action:** Specify whether a violation flags, penalizes, rejects, or permits abstention.

Table 2 is an illustrative rule register. Its rows are design examples, not measurements from an organization or published benchmark.

Illustrative rule	Provenance and scope	Exception evidence	Violation denominator and evaluation
End is no earlier than start	Definition supplied by a domain expert; same event, clock and timezone basis.	Evidence that fields describe different clocks or events makes scope uncertain.	Eligible event records; test normalization separately from inference.
Subset count does not exceed total count	Expert-supplied set definition; same membership basis and observation period.	Evidence that the alleged subset includes members excluded from the total.	Eligible linked record pairs; adjudicate set membership before enforcing.
A mined relationship usually holds	Training-only discovery; specified population and definitions.	Independent labels establish legitimate counterexamples.	Explicit pair, tuple-participation, or repair denominator. (arxiv.org)
A source-supported value conflicts with a rule	Rule origin and claim origins documented independently.	Preserve evidence for both the claim and applicability dispute.	Count affected claims and overrides separately from rule violations.

The register turns exceptions into inspectable evidence. It also prevents a rule from gaining artificial authority merely because its source was an expert or its training support was high. Later work on fast approximate denial-constraint discovery likewise motivates exceptions as a way to avoid overfitting dirty data, without establishing a factual-fusion tolerance threshold. (mashuai-ms.github.io)

For an **illustrative counterexample (Hypothetical Example)**, suppose a record reports an active total of **12**, while a related record reports **30** historical members. The rule “subset count cannot exceed total count” is inapplicable if “historical members” is not a subset of “active members.” Raising the active total would improve compliance with the incorrectly scoped rule while creating a factual error. The invented values demonstrate a category mismatch, not a published performance result.

An independent holdout and constraint-perturbation protocol

The proposed benchmark separates **claim evidence**, **constraints**, and **exception policy** into independently versioned inputs. It freezes them before final test labels are exposed. This follows the general evaluation principle that information unavailable at prediction time can cause leakage, and that test data must not guide model choices. (scikit-learn.org) Rule discovery is a model choice for this purpose, even if its output looks like a declarative business statement.

The holdout should be independently adjudicated from retained evidence, not labeled by the same rule whose utility is being tested. Adjudicators should see relevant context but remain blinded to system predictions during initial labeling. The independence claim must identify who labeled the records, which sources they used, and whether their evidence overlaps the rule-discovery population.

A workable sequence is:

- **Define the unit:** Choose entity-attribute-time claims and eligible cross-record links.
- **Split by dependence:** Group related entities, shared source origins, or repeated observations.
- **Freeze evidence:** Retain the exact source material available at the prediction cutoff.
- **Register rules:** Fix mined and expert rules, weights, scope tests, and exception policy.
- **Adjudicate labels:** Establish correct, indeterminate, and genuinely inapplicable targets.
- **Compare methods:** Run unconstrained and constrained inference on identical evidence.
- **Perturb inputs:** Inject documented wrong rules, context shifts, and legitimate exceptions.
- **Repeat interventions:** Record random seeds and separate repeated perturbations from new labels.
- **Release an audit record:** Preserve predictions, changes, seeds, exclusions, and denominators.

Group-aware validation addresses dependencies within subjects or other groups; time-oriented evaluation should test future observations when that matches deployment. (scikit-learn.org) A graph of rule-linked records may contain a large connected component, making complete separation impractical. In that case, disclose residual connectivity and report separate source-family and time splits, rather than declaring independence without qualification.

AI evaluation using private and public data adds an access boundary. NIST's AI Technology Evaluation describes blind evaluation and supports data not publicly released. (pages.nist.gov) This shows that independent testing need not imply public disclosure of the evidence. It does not supply a ready-made cross-record constraint benchmark.

For this proposal, evaluate each prediction against the evidence its intended reader may access. Freeze permissions along with the evidence snapshot, and distinguish absent public evidence from inaccessible private evidence. The Internet Engineering Task Force's OAuth rich authorization standard supports fine-

grained authorization data and resource-specific actions; these are examples of explicit access semantics, not a requirement to implement OAuth in the benchmark. (datatracker.ietf.org)

The perturbation should vary **rule error rate** and **legitimate exception frequency** independently. Report both the proportion of rules corrupted and the proportion of claims exposed to those rules. A rarely applicable wrong rule and a globally applied wrong rule have different opportunities to cause harm.

Use structurally plausible mistakes as well as random corruption. Examples include an incorrect inequality, a stale validity interval, a category mapping that omits a legitimate group, and several rules derived from one biased source family. These are proposed stress cases. Keep an untouched control set so that an apparent improvement cannot be attributed solely to changing the test distribution.

Metrics, uncertainty, and stop conditions

AI factual accuracy evaluation should compare predictions with independently adjudicated labels. The truth-discovery survey distinguishes categorical accuracy or error rates from continuous-value measures such as mean absolute error and root mean square error. (arxiv.org) Report the metric that matches the target, along with label exclusions, units, and answerability.

The minimum scorecard should include:

- **Paired accuracy change:** Constrained accuracy minus unconstrained accuracy on the same labeled claims.
- **Override rate:** Claims whose selected value changes, divided by eligible claims.
- **Beneficial overrides:** Baseline-wrong claims changed to independently correct values.
- **Harmful overrides:** Baseline-correct claims changed to independently wrong values.
- **Unsupported overrides:** Changes lacking independent evidential support under a frozen audit rubric.
- **Rule-exposed accuracy:** Accuracy among claims actually reached by the perturbed rules.
- **Coverage and selective risk:** Fraction answered and error rate among answered claims.
- **Calibration and uncertainty:** Probability quality and intervals at the appropriate dependence unit.

An **unsupported override** is not automatically a wrong answer. It measures an evidential deficiency. A **harmful override** requires a correctness label. Likewise, “strongly supported” should be defined independently of the constrained model, using source passages, applicability, dates, and source independence. Otherwise, the system can redefine support to excuse its own changes. Attribution frameworks provide the source-support target, while independent truth labels provide the correctness target. (research.google)

For probabilistic predictions, calibration concerns whether stated probabilities match correctness frequencies. Guo and colleagues (2017) define this target explicitly. (proceedings.mlr.press) Scikit-learn's calibration guidance warns that Brier loss combines calibration, discrimination, and uncertainty, so a lower Brier score alone does not demonstrate improved calibration. (scikit-learn.org) Use reliability diagrams and stratify by exception status and rule exposure; fit any recalibration procedure on data independent of model training and final testing. (scikit-learn.org)

Abstention should remain visible. Selective classification trades coverage for accuracy, so an apparent gain produced by answering fewer difficult claims must be accompanied by the coverage change. (www.jmlr.org) The Web Ontology Language (OWL) open-world assumption also makes clear that an absent statement may still be true. (www.w3.org) A declared unknown is therefore not interchangeable with a false value, a nonexistent target, or a solver's computational unknown.

Report paired uncertainty estimates. SciPy's bootstrap documentation describes resampling with replacement and supports using the same resampled indices across arrays. (docs.scipy.org) For dependent claims, the proposal is to resample independent source or entity clusters and recompute the paired statistic; simply applying an observation-level bootstrap would not establish cluster independence. Report degenerate or undefined intervals, rather than replacing them with a precise-looking result. (docs.scipy.org)

Stop conditions should be specified before testing. Withhold approval if gains vanish under credible wrong-rule settings, harmful overrides exceed the decision owner's prespecified budget, exception cases lose accuracy, or unknown cases become unsupported assertions. Also withhold approval if rule provenance or holdout independence cannot be defended. These are proposed acceptance principles, not literature-derived universal numerical thresholds.

Data Analysis and Evidence

The quantitative evidence answers different questions. Livshits and colleagues' example has **2 violating ordered pairs out of 210**, approximately **0.95%**, yet requires deleting **2 of 15 tuples**, approximately **13.3%**, to obtain the relevant satisfying subinstance. (www.vldb.org) The same dataset therefore supports very different apparent exception rates depending on the definition. Their sampling F1 score compares discovered constraint sets with those found using the complete dataset, not factual correctness of fused values. (www.vldb.org)

Adjacent benchmarks reinforce the need for separate scores. Gao and colleagues (2023) report that the best models evaluated on their ELI5 citation task lacked complete citation support **50% of the time**. This is a citation-support result, not a fusion error rate. (aclanthology.org) WILDS comprises **10 datasets** studying distribution shifts, and its authors report that evaluated robustness methods did not consistently improve over standard models. (wilds.stanford.edu) (ai.stanford.edu) Neither result establishes a safe constraint-error threshold.

Table 3 specifies a **proposed sensitivity grid (Hypothetical Example)**. Percentages are chosen stress settings, not observed error rates. Each cell would hold paired accuracy change, calibration measures, coverage, unsupported-override rate, and confidence intervals.

Injected wrong-rule rate	0% legitimate exceptions	1% legitimate exceptions	5% legitimate exceptions	10% legitimate exceptions
0%	Clean control	Exception stress	Exception stress	Exception stress
1%	Wrong-rule stress	Joint stress	Joint stress	Joint stress
5%	Wrong-rule stress	Joint stress	Joint stress	Joint stress
10%	Wrong-rule stress	Joint stress	Joint stress	Joint stress

Injected wrong-rule rate	0% legitimate exceptions	1% legitimate exceptions	5% legitimate exceptions	10% legitimate exceptions
20%	Severe wrong-rule stress	Joint stress	Joint stress	Joint stress

The grid contains **20 proposed settings**. Its exception denominator is eligible rule applications confirmed by independent adjudication; the injected-rule denominator is distinct registered rules. Small rule sets may not realize every percentage exactly; report actual integer counts and realized exposure. Hold the claims fixed when possible and change only the rule set or preidentified exception strata. Record how any synthetic exceptions alter the evidence, so the experiment does not confuse a rule intervention with an unexplained label change. Approximation semantics must remain fixed across comparable cells. (arxiv.org)

A **worked calculation (Hypothetical Example)** shows why aggregate gains are insufficient. Assume **1,000 independently labeled claims**, an unconstrained baseline with **800 correct**, and a constrained output with **830 correct**. The accuracy change is **80% to 83%**, or **3 percentage points**. Suppose **120 values change**: **75** changes correct baseline errors and **45** introduce errors. The net gain is **75 minus 45 = 30** correct claims, yet the harmful share among overrides is **45/120 = 37.5%**. All these values are invented for the calculation.

Under an independent-binomial assumption, the NIST Handbook's Wilson formula gives a **95% interval of approximately 29.4% to 46.4%** for that illustrative harmful-override proportion. (www.itl.nist.gov) The endpoints were calculated from **45/120**, not reported by NIST. They describe conditional harm among overrides, not uncertainty in the paired accuracy gain. Clustered records would require a dependence-aware analysis.

If **18** of those overrides lack support under the frozen evidence rubric, the unsupported-override rate is **18/120 = 15%**. That figure still does not establish how many unsupported changes are factually wrong. NIST's statistical evaluation work distinguishes performance conditional on a fixed benchmark from generalized accuracy, which is why one favorable test set cannot establish robustness across plausible rule environments. (www.nist.gov)

Implications and Future Directions

A constraint should influence inferred values only when its benefit survives independently labeled testing under credible rule and exception uncertainty. This is an evaluation conclusion, not a claim that constraints must always be soft. A genuinely scoped invariant and a learned high-support heuristic can warrant different treatments, provided the benchmark tests each treatment separately.

Distribution shift makes that distinction operational. The WILDS authors distinguish domain generalization from subpopulation shift and choose splits and metrics with domain experts. (ai.stanford.edu) In the proposed constraint test, those ideas translate into separate evaluations for changed contexts and uncommon legitimate exceptions. A globally favorable average may conceal a concentrated loss in one applicable subgroup.

Probability quality also needs testing after perturbation. Ovadia and colleagues (2019) examine accuracy and calibration under dataset shift and find traditional post-hoc calibration insufficient in their experiments. (papers.neurips.cc) This supports testing confidence under changed conditions; it does not establish that any

specific calibration method will correct errors from a bad rule.

The major unresolved issue is **shared bias between evidence and rules**. A constraint mined from one source population can agree with labels drawn from the same population without receiving an independent test. Holding out records does not remove that shared origin. D'Amour and colleagues' 2020 publication-page account describes predictors with comparably strong held-out performance that behave differently in deployment domains. (research.google) For rule-driven fusion, this motivates holding out source families and applicability contexts, alongside conventional entity splits.

A useful research release would preserve a **constraint-perturbation manifest**, the rule register, independent label provenance, and paired prediction changes. It should expose which claims were reached by each rule, which exceptions were recognized, and which judgments remained unresolved. These are proposed reproducibility artifacts. SHACL's allowance for provenance metadata in validation reports illustrates that rule-checking outputs can carry contextual evidence as well as a conformance result. (www.w3.org)

Future work should test whether independent adjudicators can reliably distinguish bad rules, legitimate exceptions, and insufficient evidence. Where they cannot, the benchmark should report sensitivity to alternative adjudications. A defensible unknown is a substantive result. The truth-discovery survey already discusses an unknown output for contradictory constraints and cases lacking a truth. (arxiv.org) Treating that possibility explicitly is more informative than forcing every conflict into a single asserted value.

Conclusion

Constraint-based AI evaluation should decide whether rules improve independently established values, how often they cause harmful or unsupported overrides, and whether the system handles exceptions and unknowns appropriately. Constraint satisfaction is one diagnostic within that evaluation. It cannot supply the reference truth used to judge its own benefit.

The proposed acceptance test separates claim evidence, rule provenance, and exception policy; freezes them before final testing; and compares constrained and unconstrained inference on identical accessible evidence. It then varies incorrect rules, context mismatches, legitimate exceptions, and correlated rule origins. Accuracy, calibration, selective coverage, and override outcomes remain separately inspectable.

The literature supports these distinctions without supplying a universal allowable rule-error rate. Constrained truth discovery motivates relationship-aware inference, approximate constraint discovery makes exception semantics explicit, and attribution and robustness research show why provenance and aggregate predictive scores answer different questions. The appropriate acceptance boundary follows the decision's consequences and independently measured uncertainty.

For the researcher choosing whether a cross-record relationship should influence a disputed value, the practical requirement is a reproducible paired comparison with credible labels and visible failure conditions. When those labels or rule origins are unavailable, the result should remain conditional. When a benefit survives realistic perturbations and its harmful overrides remain within a prespecified budget, the evidence supports using the constraint within the tested scope. Extending it to a new context requires another applicability test.

Source links

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

1. https://researchconnect.buffalo.edu/en/publications/constrained-truth-discovery
2. https://www.vldb.org/pvldb/vol13/p1682-livshits.pdf
3. https://scikit-learn.org/stable/common_pitfalls.html
4. https://pages.nist.gov/ai-technology-evaluation
5. https://research.google/pubs/measuring-attribution-in-natural-language-generation-models
6. https://www.jmlr.org/beta/papers/v11/el-yaniv10a.html
7. https://scikit-learn.org/stable/modules/calibration.html
8. https://www.itl.nist.gov/div898/handbook/prc/section2/prc241.htm
9. https://arxiv.org/abs/1505.02463
10. https://aclanthology.org/2021.findings-acl.317.pdf
11. https://arxiv.org/abs/2005.08540
12. https://www.w3.org/TR/shacl
13. https://smt-lib.org/about.shtml
14. https://microsoft.github.io/z3guide/docs/logic/basiccommands
15. https://microsoft.github.io/z3guide/docs/optimization/softconstraints
16. https://mashuai-ms.github.io/pubs/vldb2023a.pdf
17. https://scikit-learn.org/stable/modules/cross_validation.html
18. https://datatracker.ietf.org/doc/html/rfc9396
19. https://arxiv.org/html/1505.02463
20. https://proceedings.mlr.press/v70/guo17a.html
21. https://www.w3.org/TR/owl2-primer
22. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bootstrap.html
23. https://aclanthology.org/2023.emnlp-main.398
24. https://wilds.stanford.edu
25. https://ai.stanford.edu/blog/wilds
26. https://www.nist.gov/publications/expanding-ai-evaluation-toolbox-statistical-models
27. https://papers.neurips.cc/paper_files/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html
28. https://research.google/pubs/underspecification-presents-challenges-for-credibility-in-modern-machine-learning

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.