Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Data Extraction Error Analysis: Source or Extractor?

> 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 · data extraction error analysis · source errors · extraction errors · knowledge fusion · data provenance · entity linking · truth discovery · knowledge-based trust · calibration

Original article: https://reliablecontext.com/articles/data-extraction-error-analysis-source-or-extractor



In this article

1. Executive Summary
2. Introduction and Background
3. An attribution model with separate error layers
4. An annotation guide that makes attribution reviewable
5. Model comparisons and field-level validation
6. Analysis of key segments and identifiability
7. Data Analysis and Evidence
8. Implications and Future Directions
9. Conclusion

Executive Summary

Data extraction error analysis should attribute an apparent false triple only after separating the **observed extraction**, the **source-stated claim**, and **independent truth evidence**. A faithful extraction can reproduce a false statement; an unfaithful extraction can misrepresent a correct statement. Consistency with a passage therefore answers a different question from credibility, a distinction explicitly examined by Padia, Ferraro, and Finin (2022). (aclanthology.org) Knowledge-Based Trust likewise distinguishes extraction-process errors from factual errors in sources. (arxiv.org) Neither a changed reliability score nor disagreement with a reference graph establishes what the original page stated.

The historical scale is substantial but must retain its denominators. Dong et al. (2014) evaluated **1.6 billion unique triples**, **12 extractors**, and more than **1 billion webpages**. (www.vldb.org) Their automatically constructed reference labels covered approximately **40%**, rather than the entire corpus. (www.vldb.org) Dong et al. (2015) analyzed **2.8 billion extracted triples** and reported scores for **119 million webpages**; that scored subset met a threshold of at least **five** inferred correctly extracted triples per page. These are historical research datasets and model outputs, not current-web performance guarantees. (www.vldb.org)

This report proposes a reviewable audit: retain an authorized source capture, transcribe its exact relevant passage, adjudicate identity and predicate mappings, and assess truth using separately identified evidence. Provenance describes the entities, activities, and people producing information; it supplies the history needed for inspection, rather than a truth verdict. (www.w3.org) Version identifiers and representation validators help identify which material was examined. (support.datacite.org) (www.rfc-editor.org) Cases lacking readable passages or adequate truth evidence remain **unresolved** and are excluded only from the particular metrics they cannot support, with exclusions reported.

Compare **source-only** and **source-plus-extractor** models on the same labeled holdout, with identical priors, thresholds, and accessible evidence. Report attribution confusion matrices, field precision and recall, calibration, eligible denominators, and paired uncertainty. (scikit-learn.org) (nlp.stanford.edu) (proceedings.mlr.press) (docs.scipy.org) The worked calculations below use explicitly hypothetical fixtures, not collected benchmark results. The central decision rule is conservative: lower source reliability only for an independently established source error, lower extractor reliability for a demonstrated reading or mapping error,

and lower both when separate evidence supports both labels. Otherwise preserve uncertainty. A record system can make that decision reviewable by exposing canonical identity and source evidence; it cannot replace independent truth adjudication.

Introduction and Background

An extracted subject-predicate-object triple is an observation produced by a process. Its appearance in a graph does not establish that a page asserted it, that its identifiers refer to the intended entities, or that the assertion is true. These distinctions matter when knowledge-fusion evaluators decide which reliability estimate should change after discovering an apparent error. The Knowledge-Based Trust preprint makes the source-versus-extraction distinction explicit. (arxiv.org)

Knowledge fusion addresses evidence that has already passed through extraction and mapping. **Conventional data fusion** reconciles source-provided values; adding entity linkage and schema alignment introduces another error layer, as Dong et al. (2014) explain. (www.vldb.org) The broader survey by Zhao et al. (2020) also treats uncertainty in knowledge acquisition as a reliability problem. (link.springer.com) A disagreement between two normalized triples can therefore originate before, during, or after extraction.

The **Knowledge Vault** lineage provides the conceptual background, not a ready-made audit dataset. Its publication describes combining web extractions with prior knowledge from existing repositories. (research.google) A source passage audit must identify which evidence belongs to the page, which belongs to normalization, and which belongs to a graph prior. Otherwise a model can receive credit for predicting a familiar fact while still reading a particular page incorrectly.

This report is a **protocol and literature-based data report**, dated October 13, 2026. No new corpus was collected, and no empirical source-only versus source-plus-extractor comparison was run. The research figures retain their original study context; the worksheet, label conventions, sampling recommendations, and hypothetical calculations are the report's proposed methodology. Statistical definitions follow primary research and official documentation. (proceedings.mlr.press) (docs.scipy.org)

The practical objective is an attribution label that another reviewer can reproduce. A page that is inaccessible during review does not become a false source. A sentence that supports a different predicate does not become evidence for the emitted predicate. And a reference graph's missing edge does not, by itself, settle truth. Reading consistency must first be distinguished from credibility. (aclanthology.org)

An attribution model with separate error layers

Let **O** denote the extractor's observed triple, **S** the proposition actually stated in the captured source, **M** the proposed entity and predicate mapping, and **T** the independently adjudicated truth status of a specified proposition at a specified time. These are audit variables proposed here. They are not interchangeable probability estimates. PROV-DM supplies a vocabulary for the provenance of entities and activities; the proposed variables add the evaluation questions. (www.w3.org)

What each comparison answers

The fidelity question is whether O, interpreted through M, accurately represents S. The truth question is whether S is correct under the declared reference time and scope. The observed triple's own truth should also be checked separately. An incorrect reading can accidentally emit a true statement, while a correct reading can faithfully emit a false statement. The reading-consistency distinction supplies the conceptual basis for keeping these outcomes apart. (aclanthology.org)

Treat **entity linking** as mapping a mention to the intended real-world referent and **predicate mapping** as mapping the expressed relation to the intended schema relation. The multi-source fusion survey distinguishes instance alignment from ontology alignment. (link.springer.com) Thus an organization and a related organization can have similar names without being interchangeable subjects; a sentence expressing a proposed relationship need not assert a completed relationship. These are generic audit examples, not observations about any identifiable organization.

Use the following proposed outcome convention:

- **Correct:** The source assertion is supported as true, and the extraction and mapping faithfully represent it.
- **Source error:** The independently evaluated source assertion is false, while the extraction is faithful.
- **Extractor error:** A reading, identity, or predicate error is demonstrated, while the source assertion is supported as true.
- **Both:** The source assertion is independently false and the extraction or mapping independently misrepresents it.
- **Unresolved:** At least one necessary comparison lacks sufficient evidence; record which layer remains unresolved.

Store the underlying axis labels as well as the composite outcome. A case can have a known mapping error but unresolved source truth; it should contribute to the appropriate mapping-error metric without acquiring a complete composite label. The 2014 study already retained extraction provenance, including source and extractor context. (www.vldb.org) This protocol extends that inspection discipline with explicit eligibility rules.

Root causes and truth-label disagreement

Root-cause analysis should preserve the earliest demonstrated divergence and any subsequent divergences. Distinguish passage selection, span reading, entity linking, predicate alignment, literal normalization, temporal interpretation, and reference-label error. The extraction literature identifies predicate linkage as a task separate from earlier extraction stages. (www.vldb.org) Recording these sublabels avoids treating every failure as a generic false triple.

Keep **reference disagreement** separate from independently verified falsehood. The 2014 evaluation used a local closed-world assumption: it abstained when the subject-predicate pair was absent from its reference knowledge base. (www.vldb.org) An audit should similarly state its reference assumptions, but its proposed truth labels require adjudication appropriate to the actual claim. Absence, contradiction, outdated evidence, and unreadability are different reasons to withhold a truth decision.

An annotation guide that makes attribution reviewable

The audit unit should be a **claim-source-extractor observation**, not merely a deduplicated graph edge. The same edge can be faithfully extracted from one page and incorrectly extracted from another. Preserving the observation unit makes it possible to investigate each process rather than transferring an edge-level verdict to every supporting page. This is a proposed sampling unit, informed by provenance's separation of entities, activities, and responsible agents. (www.w3.org)

Capture, permission, and time

Retain the exact material used for reading only where permitted. Record its location, observation time, version or representation identifier, access restrictions, and applicable reuse conditions. Dublin Core's provenance term emphasizes custody changes significant to interpretation. (www.dublincore.org) DataCite recommends new identifiers for major resource versions. (support.datacite.org) Neither an identifier nor a timestamp independently certifies the content's truth.

A **source observation date** records when the evaluator saw a representation. A source publication date and a claim's effective time answer different questions and should occupy separate fields. HTTP entity tags distinguish representations, while Last-Modified reflects the origin server's belief about modification time. (www.rfc-editor.org) A capture that preserves the wrong representation cannot justify a claim about what an earlier page stated.

Permission to read, retain, and redistribute must be checked separately for each fixture. Creative Commons Attribution licenses impose attribution conditions and may leave other necessary permissions outside the license. (creativecommons.org) No permission to redistribute arbitrary research PDFs or web captures is inferred here. The worked passages below are original, synthetic text, so they contain no redistributed real-source passages. An actual release should document rights, minimize retained material, and respect access and retention conditions.

Required annotation fields

The proposed worksheet should include:

- **Observation identifier:** A stable reference to this claim-source-extractor event.
- **Observed triple:** The emitted identifiers, predicate, value, and qualifiers.
- **Source locator:** The exact address, document version, page, or section.
- **Observation date:** When that representation was captured or inspected.
- **Verbatim passage:** Exact supporting text with enough surrounding context to interpret it.
- **Source-stated proposition:** A separate interpretation of what the passage asserts.
- **Entity mapping:** Mention, candidate referent, selected referent, and justification.
- **Predicate mapping:** Source relation, target relation, direction, and scope.
- **Independent truth evidence:** Its identity, effective time, authority, and relation to the source.
- **Eligibility status:** Which fidelity and truth comparisons can actually be completed.
- **Error layer and reason:** Composite outcome plus specific demonstrated suberrors.
- **Reviewer record:** Reviewer, disagreement, adjudication, and confidence with reasons.

This extends a provenance record with adjudication fields rather than treating custody metadata as a truth label. (www.dublincore.org) Human agreement should be measured separately from model accuracy; official evaluation documentation describes Cohen's kappa as a comparison between human annotators' labelings. (scikit-learn.org) Reviewer confidence describes the quality of available evidence, not the extractor's confidence score.

Table 1 illustrates the worksheet using wholly original synthetic passages. Each row is a **Hypothetical Example**, with observation and truth assessments supplied solely for demonstration. The source and extraction axes follow the reading-consistency distinction, not a claim about a real system. (aclanthology.org)

Claim and observation	Synthetic verbatim source passage	Observed extraction	Mapping check	Independent truth fixture	Error layer	Reviewer confidence
Hypothetical Example: affiliation, observed at audit time	"Aster joined Meridian."	Aster joined Meridian	Identity and relation match	Separate fixture confirms the affiliation	Correct	High: both checks available
Hypothetical Example: affiliation, observed at audit time	"Aster joined Meridian."	Aster joined Meridian	Identity and relation match	Separate fixture explicitly contradicts the affiliation	Source error	High: faithful reading, contradicted assertion
Hypothetical Example: project ownership, observed at audit time	"Aster supports Meridian's project."	Aster owns the project	Support mapped incorrectly to ownership	Separate fixture confirms support only	Extractor error	High: predicate mismatch demonstrated
Hypothetical Example: affiliation, observed at audit time	"Aster joined Meridian."	Aster joined Solstice	Object identity changed	Separate fixture contradicts the source affiliation	Both	High: separate evidence for both layers
Hypothetical Example: affiliation, observed at audit time	Not available; access blocked	Aster joined Meridian	Cannot inspect passage or mapping	Truth evidence unavailable	Unresolved	Low: no attribution permitted

The table's confidence values are categorical judgments within invented fixtures, not measured reviewer reliability. Notice that the last row preserves the extraction and the access failure without implying a negative source statement. In a real audit, the independent fixture must itself be inspected and linked to the same entities and time. PROV-DM provides the background for retaining those evidence relationships. (www.w3.org)

Independent review and abstention

Reviewers should read the passage before seeing a model's reliability score. A second reviewer should assess disputed entity, predicate, and truth labels using the same captured material. Preserve both initial decisions, the adjudicated outcome, and the reason for any change. Kappa can describe annotation agreement, but it does not substitute for claim-level adjudication. (scikit-learn.org)

Use explicit **not found**, **blocked**, and **not applicable** states as proposed audit conventions. A not-found passage is not a passage stating a negation. A blocked source is not a demonstrated extractor mistake. If only extraction fidelity is decidable, retain that result and declare source truth unresolved. The separation mirrors the literature's distinction between consistency and credibility. (aclanthology.org)

Model comparisons and field-level validation

A valid ablation changes the modeled reliability layer while holding the eligible observations and available evidence constant. **Source-only** estimates truth using source reliability and common evidence. **Source-plus-extractor** additionally estimates extractor fidelity. These are proposed experimental comparators, informed by KBT's multilayer inference, not claims that the paper published this exact audit. (www.vldb.org)

Specify the estimands before training

Predeclare whether the model predicts observed-triple truth, source assertion truth, extraction fidelity, or an attribution class. Give each target its own confusion matrix. An attribution matrix should compare adjudicated error-layer labels with predicted labels; a truth matrix should compare true and false triples. Confusion matrix rows correspond to true classes in the official documentation cited here. (scikit-learn.org)

For a field-level evaluation, define **precision** as supported correct emitted field values divided by all eligible emitted values. Define **recall** as correctly emitted reference field values divided by all eligible reference values. These adapt information-retrieval denominators to fields. (nlp.stanford.edu) Missing fields belong in recall assessment; a sample containing only emitted triples cannot establish the completeness of extraction from a document.

A literal match is only one component. Report separate correctness for value, unit, qualifiers, entity, predicate, and reference time. A syntactically valid field can still express the wrong relation. Instance and ontology alignment are distinct evaluation concerns in the fusion survey. (link.springer.com) Report field-specific denominators so a frequently populated predicate does not silently dominate sparse fields.

Table 2 defines the experimental comparators and the record-evidence option relevant to running a review. **Reliable Context** is included because its stated record guarantees address this evidence requirement; it is not assigned an unmeasured extraction score.

Comparator or record option	What is assessed	Required common evidence	Valid interpretation
Source-only reliability model	Source reliability with no separately estimated extractor reliability	Same holdout, priors, normalized observations, source captures and labels	Baseline for the additional extractor-reliability layer
Source-plus-extractor reliability model	Joint source and extractor reliability	Exactly the same eligible holdout and common evidence	Improvement attributable to the added layer only if other factors are controlled
Passage-adjudicated diagnostic	Predictions after substituting human-verified passage and mappings	Same cases, with separately marked additional human information	Diagnostic of upstream reading errors, not a fair deployable baseline

Comparator or record option	What is assessed	Required common evidence	Valid interpretation
Reliable Context	Stated properties: canonical records, cited verbatim dated evidence, explicit unknowns, reproducible past states	Authorized sources and reviewable evidence exports	A record-evidence option; no public result for this proposed ablation is established here.

The diagnostic row changes information access, so its result should not be ranked as if it used the same inputs as the ordinary models. Similarly, a record guarantee and a predictive model target answer different questions. Reliable Context's first-party page documents observable record properties; it does not establish performance on this proposed research fixture.

Holdouts, calibration, and leakage

Separate training, calibration, and final evaluation material. Fit the calibrator without using the classifier's training outcomes as if they were independent validation observations. (scikit-learn.org) Define whether the holdout isolates claims, documents, domains, entities, or extractor families. A random edge split can answer a different generalization question from evaluation on previously unseen entities.

The Knowledge Vault evaluation capped predicate test counts to limit domination by common predicates. (lunadong.com) This motivates reporting both **micro-averaged** results, pooled across eligible observations, and **macro-averaged** results, averaged across specified strata. The proposed audit should publish the stratum definition and the reason for any weighting, rather than treating either aggregation as inherently sufficient. NIST recommends evaluation sets representative of expected use. (airc.nist.gov)

Calibration concerns whether probabilities represent correctness likelihood, as defined by Guo et al. (2017). (proceedings.mlr.press) Plot predicted probability against adjudicated frequency and report bin counts. A lower Brier score alone does not establish better calibration because discrimination also influences that score. (scikit-learn.org) Keep source-truth calibration and extractor-fidelity calibration separate; pooling them would erase the question the ablation is meant to answer.

A prespecified comparison should make these controls explicit:

- **Training access:** The same allowed observations and prior information.
- **Calibration access:** Separate fitting data and an unchanged evaluation holdout.
- **Decision thresholds:** Fixed thresholds or a disclosed common selection procedure.
- **Eligibility:** Identical case membership for every paired comparison.
- **Information changes:** Mark added human adjudication as a distinct diagnostic.

Analysis of key segments and identifiability

The primary analysis should stratify by **extractor**, **predicate**, **topic**, and **source coverage**. These are proposed audit strata, not an assertion that the available studies benchmarked every requested segment. NIST's framework allows disaggregating accuracy measurements across data segments. (airc.nist.gov) Coverage must include excluded observations as well as the labeled cases, since eligibility can vary with page accessibility and evidence availability.

Segment-specific denominators

For each extractor, count inspected emitted observations, readable passages, decidable mappings, truth-labeled source assertions, and unresolved cases. For each predicate, distinguish precision from completeness and record the number of eligible reference fields. Precision and recall use different populations, so a single unlabeled accuracy percentage cannot answer both questions. (nlp.stanford.edu)

For each topic, define relevance before inspection. A source can be accurate on a narrow topic while providing little evidence about another. For each coverage stratum, distinguish pages with dense evidence, sparse evidence, and unavailable evidence. The KBT paper reports that more than a billion webpages yielded only a single extracted triple, an explicit limitation for page-level estimation. (www.vldb.org) Sparse coverage should therefore remain visible in the audit rather than being converted into assumed reliability.

The proposed conditional source-error rate is the number of independently false source assertions divided by the number of source assertions with decidable truth in the stated stratum. The conditional extractor-error rate is the number of independently demonstrated fidelity failures divided by observations with decidable fidelity. Binomial-rate reporting should retain both successes or errors and the eligible number of trials. (docs.scipy.org)

Do not require identical denominators for these separate rates: some readable passages lack independent truth evidence. Do require identical eligibility for the paired model comparison. Otherwise one model may appear stronger because it is tested on easier cases. NIST's emphasis on representative evaluation supports documenting this distinction between an audit's inspected population and its scored subset. (airc.nist.gov)

Reconciling conflicting data sources

For conflicting sources, the proposed protocol first checks whether their claims share the same entity, predicate, qualifiers, and effective time. Preserve both verbatim statements before deciding whether they actually contradict each other. Alignment and conflict detection are distinct components of the multi-source fusion literature. (link.springer.com)

A newer capture is not automatically a more current assertion. Compare claim scope, source version, and independent evidence before selecting a canonical value. Keep the alternative, the decision reason, and any unresolved mapping visible. Representation validators identify changed content rather than certify factual correctness. (www.rfc-editor.org) A source-only reliability score cannot resolve an ambiguity caused by attaching one source's statement to the wrong entity; the mapping must be adjudicated first.

Where attribution is not identifiable

Multiple agreeing extractors do not provide independent confirmation if they share the same entity linker or source-processing assumption. The 2014 fusion study specifically identifies common linkage mistakes across extractors. (www.vldb.org) The proposed audit should record shared components where known and retain a dependence warning where they are unknown.

A single source and a single extractor can leave competing explanations observationally indistinguishable. If the source cannot be read, a false observed triple may reflect a source error, an extraction error, or both. Extra model confidence does not supply the missing passage. Reading consistency requires comparing the output

with its provenance text. (aclanthology.org)

Similarly, independent truth evidence copied from the same underlying assertion is not an independent witness for this protocol. The evaluator should document evidence lineage rather than count repeated addresses as separate confirmations. Provenance supplies the entities and activities through which such dependence can be inspected. (www.w3.org) Where that lineage remains unknown, report sensitivity to alternative dependence assumptions instead of assigning certainty.

Data Analysis and Evidence

The historical studies establish why the proposed audit is necessary, while the hypothetical fixtures demonstrate how its denominators and comparisons work. They should not be combined into a single performance claim.

What the published evaluations establish

The 2014 reference set labeled **650 million** triples, including **200 million** labeled correct, under an automatically constructed knowledge-base comparison. (www.vldb.org) Its inspection of **25** false triples assigned **44%** to triple-identification errors, **44%** to entity-linkage errors, **20%** to predicate-linkage errors, and **4%** to source-provided errors. Categories overlapped, and the sample is too small to establish current population rates. (www.vldb.org)

The 2015 manual check sampled **100** websites with KBT above **0.9**; it selected **10** triples per website from the leading **three** predicates with inferred extraction correctness above **0.8**. This selective inspection is not an unrestricted test of all extracted facts. (www.vldb.org) Neither study provides a newly source-adjudicated comparison of the exact ablations proposed here.

Hypothetical attribution counts

Consider a **Hypothetical Example** containing **100** inspected observations: **50** correct, **10** source errors, **15** extractor errors, **5** both, and **20** unresolved. These counts are invented inputs, not observed data. The **80** fully adjudicated cases imply a conditional source-error rate of **15/80 = 18.75%** and extractor-error rate of **20/80 = 25%**. Both cases contribute to both numerators; unresolved cases contribute to neither eligible denominator.

Wilson **95%** intervals calculated from those hypothetical counts are **11.7% to 28.7%** for source errors and **16.8% to 35.5%** for extractor errors. These are original binomial calculations, using the interval approach documented by Statsmodels. (www.statsmodels.org) They express sampling uncertainty under an independent-trial assumption, not uncertainty caused by unresolved truth or dependence between copied pages. The unresolved share must still be reported separately.

Hypothetical paired ablation

Table 3 uses a separate **Hypothetical Example** of **200** truth-labeled observations, evenly divided between true and false claims. Both models score the exact same cases. Positive means a true claim; TP, FP, FN, and TN denote true positive, false positive, false negative, and true negative. The confusion-matrix convention is documented by scikit-learn. (scikit-learn.org) All entries are invented inputs or original arithmetic.

Model or paired statistic	TP / FP / FN / TN	Accuracy	Precision / recall	Uncertainty on the common holdout
Source-only	72 / 18 / 28 / 82	154/200 = 77%	72/90 = 80% / 72/100 = 72%	Wilson 95% accuracy interval: 70.7% to 82.3%. (www.statsmodels.org)
Source-plus-extractor	84 / 16 / 16 / 84	168/200 = 84%	84/100 = 84% / 84/100 = 84%	Wilson 95% accuracy interval: 78.3% to 88.4%. (www.statsmodels.org)
Paired difference	25 improved; 11 regressed; 164 unchanged	7 percentage points	Compare each model's predictions on the same case	Paired percentile-bootstrap 95% interval, rounded outward to whole points: 1 to 13 percentage points. (docs.scipy.org)

The marginal intervals describe each hypothetical model's accuracy; the paired interval addresses the improvement directly. The following runnable calculation specifies the procedure for this **Hypothetical Example**. It uses **NumPy 2.5.2**, an explicit **Generator with PCG64**, and seed **20261013**. The ordered paired-outcome vector contains **25** improved cases (+1), followed by **11** regressed cases (-1), then **164** unchanged cases (0). Each entry is the source-plus-extractor correctness indicator minus the source-only correctness indicator for the same case. Resample **200** indices with replacement for each of **10,000** replicates in one call, then take the **2.5th and 97.5th percentiles** using `method="linear"`. For the table display, round the lower endpoint down and the upper endpoint up to whole percentage points; the code also prints the endpoints before that rounding.

```
# Run with NumPy 2.5.2.
import numpy as np
from math import floor, ceil

assert np.__version__ == "2.5.2", "Use NumPy 2.5.2 for this fixture."
outcomes = np.array([1] * 25 + [-1] * 11 + [0] * 164, dtype=np.int64)
rng = np.random.Generator(np.random.PCG64(20261013))
indices = rng.integers(
    0, outcomes.size, size=(10000, outcomes.size), dtype=np.int64
)
replicates = 100 * outcomes[indices].mean(axis=1)
low, high = np.percentile(replicates, [2.5, 97.5], method="linear")
print(f"Before whole-point rounding: {low:.1f} to {high:.1f} percentage points")
print(f"Table display: {floor(low)} to {ceil(high)} percentage points")
# Before whole-point rounding: 1.5 to 13.0 percentage points
# Table display: 1 to 13 percentage points
```

The unchanged outcomes comprise **143** cases both models get correct and **21** both get wrong; both groups contribute zero to the paired difference. Paired resampling preserves corresponding indices. (docs.scipy.org) An actual audit should resample at the dependence unit, such as source family, when observations within that unit are correlated.

The positive interval in this invented fixture demonstrates the calculation only. It is not evidence that adding an extractor layer improves a real system. A measured result would require released labels, predictions, exclusions, and the prespecified resampling procedure. Confidence intervals for proportions depend on

method and assumptions; exact alternatives can be conservative. (www.statsmodels.org)

Implications and Future Directions

For knowledge-fusion research, the most useful deliverable is a **source-adjudicated evaluation fixture**, with permissions and evidence lineage sufficient to reproduce its judgments. Its value comes from separating the questions a model answers, not from adding another aggregate leaderboard. Padia and colleagues' consistency-versus-credibility distinction is directly relevant to that separation. (aclanthology.org)

A practical release should provide:

- **Authorized captures:** Enough retained context to reproduce each source reading.
- **Versioned evidence:** Stable references to the exact inspected material.
- **Axis labels:** Source truth, extraction fidelity, mapping correctness, and observed-triple truth.
- **Declared unknowns:** Reasons and eligibility effects for every incomplete assessment.
- **Adjudication history:** Initial reviews, disagreements, and final rationales.
- **Fixed comparisons:** Common holdout membership and controlled information access.
- **Paired predictions:** Per-case outputs supporting reproducible uncertainty calculations.
- **Segment results:** Coverage and denominators alongside conditional error rates.

These are proposed release requirements. Version-specific identifiers and custody descriptions provide established building blocks for implementing them. (support.datacite.org) (www.dublincore.org) They do not eliminate the need to inspect permission terms for each retained source. (creativecommons.org)

For organizational context engineering, the same reasoning becomes an acceptance test: can a reviewer recover the passage, identity, dates, alternatives, and reason for an unknown? **Reliable Context** states that a value points to its verbatim source sentence and that past states are reproducible. Those observable guarantees are relevant evidence properties, not an independent measurement of extraction accuracy or a description of its internal mechanism.

Future studies should test whether extractor reliability remains identifiable under shared mapping components, sparse source coverage, and changing evidence availability. A useful experiment would vary these conditions while preserving known source assertions and independent labels. Calibration should be assessed on the probability target actually predicted, rather than imported from an unrelated classification benchmark. Guo et al.'s calibration experiments used image and document classification datasets. (proceedings.mlr.press)

The key unresolved measurement problem is the combination of inaccessible sources and dependent evidence. More transparent provenance helps expose that problem, but transparency and accuracy remain distinct characteristics in NIST's framework. (airc.nist.gov) Report the non-identifiable cases as a result, then state which additional evidence would resolve them.

Conclusion

An apparent false triple does not contain its own explanation. Defensible **data extraction error analysis** begins with the captured source passage, preserves the distinction between source assertions and extracted observations, and examines entity and predicate mappings before assigning blame to a reliability layer. The literature's separation of reading consistency and credibility supplies the central conceptual rule. (aclanthology.org)

The proposed audit turns that rule into a reproducible annotation process. Each observation retains its passage, mapping, independent truth evidence, observation date, eligibility status, and reviewer rationale. A source error requires independent evidence that the source-stated proposition is false. An extractor error requires evidence that the output misrepresents the passage or its mapping. Both labels require both demonstrations; incomplete evidence remains unresolved.

A valid model comparison uses the same labeled holdout, controlled priors and evidence, and explicitly declared probability targets. Confusion matrices reveal where labels change; field precision and recall retain their distinct denominators; calibration and paired intervals address different performance questions. (scikit-learn.org) (nlp.stanford.edu) (proceedings.mlr.press) (docs.scipy.org) The hypothetical calculations illustrate these reporting requirements without asserting measured improvement.

The operational threshold is therefore a reviewable attribution, not a ranking change. Preserve what was read, identify what remains unknown, and report which conditional rates can legitimately be estimated. Provenance makes the evidence chain inspectable. (www.w3.org) Independent source inspection and truth adjudication determine whether that chain supports a source error, an extractor error, both, or no defensible attribution.

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/2022.deelio-1.5
2. https://arxiv.org/abs/1502.03519
3. https://www.vldb.org/pvldb/vol7/p881-dong.pdf
4. https://www.vldb.org/pvldb/vol8/p938-dong.pdf
5. https://www.w3.org/TR/prov-dm
6. https://support.datacite.org/docs/versioning
7. https://www.rfc-editor.org/rfc/rfc9110.html
8. https://scikit-learn.org/stable/modules/model_evaluation.html
9. https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-unranked-retrieval-sets-1.html
10. https://proceedings.mlr.press/v70/guo17a.html
11. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bootstrap.html
12. https://link.springer.com/article/10.1007/s11280-020-00811-0
13. https://research.google/pubs/knowledge-vault-a-web-scale-approach-to-probabilistic-knowledge-fusion
14. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.binomtest.html

15. https://www.dublincore.org/specifications/dublin-core/dcmi-terms
16. https://creativecommons.org/licenses/by/4.0
17. https://scikit-learn.org/stable/modules/calibration.html
18. https://lunadong.com/publication/kv_kdd.pdf
19. https://airc.nist.gov/airmf-resources/airmf/3-sec-characteristics
20. https://www.statsmodels.org/stable/generated/statsmodels.stats.proportion.proportion_confint.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.