Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Knowledge graph completion evaluation protocol

> 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 · 2026-10-02 · knowledge graph completion evaluation protocol · knowledge graph completion · unknown triples · open-world evaluation · link prediction · negative sampling · codex · inferwiki · mean reciprocal rank

Original article: https://reliablecontext.com/articles/knowledge-graph-completion-unknown-triples-evaluation



In this article

1. Executive Summary
2. Introduction and Background
3. Key Changes
4. Implementation Considerations and Process Changes
5. Data Analysis and Evidence
6. Implications and Future Directions
7. Conclusion

Executive Summary

A knowledge graph completion (KGC) evaluation protocol should answer a narrower question than whether a model ranks held-out edges highly: does the apparent gain survive when unobserved triples are allowed to be true, false, or unresolved? Under Resource Description Framework (RDF) semantics, failure to entail a triple does not establish its negation (www.w3.org). Standard filtered link prediction removes other **known true** triples, but an undiscovered true answer remains in the candidate list and can outrank the held-out answer (pykeen.readthedocs.io) (papers.nips.cc). The resulting **mean reciprocal rank (MRR)** or Hits@K can therefore penalize a useful prediction. Yang et al. (2022) show, under specified missingness and prediction models, that even model order can reverse (papers.nips.cc).

Benchmark choice changes the question being tested. **FB15K-237** descends from a Freebase subset designed to reduce near-duplicate and inverse relations (aclanthology.org). **CoDEx** adds type-compatible, model-ranked hard candidates and manual labels for its small and medium versions; Safavi and Koutra (2020) report a **13 to 19 percentage-point** fall in triple-classification accuracy when hard negatives replace uniform ones (gemslab.github.io). **InferWiki** uses rule-guided premise and conclusion splits and a three-class evaluation, but its authors' “unknown” label identifies a candidate without a training-graph contradiction that they nevertheless judged factually incorrect (arxiv.org) (aclanthology.org). It cannot serve as a ready-made sample of genuinely unresolved real-world claims. The **WikiKG2** benchmark samples 1,000 corrupted candidates absent from its graph, a graph-membership check that is weaker than truth adjudication (ogb.stanford.edu).

The acceptance protocol in this report freezes a candidate universe and a time-stamped evidence boundary; checks entity and relation leakage; draws random, type-compatible, and model-proposed candidates; independently adjudicates a stratified sample as supported, refuted, or unresolved; and publishes both ranking and classification results with uncertainty. The **label policy** must state what evidence counts as refutation and what happens when no answer can be established. A pooled judgment set can reduce missing judgments, but its completeness remains relative to the sampled systems and candidate set (arxiv.org) (trec.nist.gov). Random negatives alone are inadequate: Cao et al. (2021) report that tested models found them easy while accuracy fell on annotated negatives (arxiv.org).

Training fixes and evaluation fixes should stay separate. Tang et al. (2022) treat corrupted training triples as unlabeled through **positive-unlabeled (PU)** risk estimation, addressing false negatives in optimization; their reported filtered MRR does not validate the test labels (www.ijcai.org). Accept a reported model gain only if its direction survives adjudicated hard-negative and unknown-aware slices, disclosed tie handling, ranking and calibration checks, and a leakage audit. Transfer to private, heterogeneous records remains an empirical question. A system of record can preserve a dated, cited value and explicitly declare an unknown, but those observable properties do not establish a KGC model's predictive accuracy.

Introduction and Background

A knowledge graph represents a proposition as a **triple**, conventionally a subject or head entity, a relation, and an object or tail entity. KGC scores candidate triples that are absent from an observed graph. The central evaluation problem is that **unobserved** describes a data state, not a truth value. A candidate can be supported

by evidence outside the graph, contradicted by evidence, or remain undecidable within the permitted evidence boundary. RDF's entailment rule makes this distinction formal: an entailed statement is true in every interpretation satisfying the source graph, while absence of entailment alone supplies no negative judgment (www.w3.org). Wikibase also distinguishes an explicit “no value” statement from a statement whose value is unknown (www.mediawiki.org).

Most held-out link-prediction reports instead ask where a known positive ranks among entity substitutions. **Filtered evaluation** removes other triples already known to be positive, preserving the target triple for scoring (pykeen.readthedocs.io). That convention avoids one obvious penalty, but it cannot remove a true edge never entered into the known graph. The problem intensifies when the model is good at discovering missing facts: its highest-scored alternative can be a useful true answer that the benchmark calls a negative by omission (papers.nips.cc). A different protocol failure arises when many candidates have identical scores; Sun et al. (2020) traced some high scores to how ties were ranked, independently of whether missing facts were true (aclanthology.org).

This report is a **benchmark acceptance protocol** for researchers deciding whether a reported improvement should influence model selection. It compares construction, labels, splits, and metrics rather than treating “open world” as a single dataset attribute. The emphasis is on evidence one can audit: the candidate universe, dated sources behind judgments, labeling rules, rank computation, and confidence intervals. The following sections distinguish claims made by benchmark authors from results reproduced independently here. No public benchmark result below is presented as a measured result on private organizational data.

Key Changes

From absent edges to explicit label states

The first protocol change is to write down three label states before choosing a metric. **Supported** means that the frozen evidence set and adjudication rule warrant the candidate. **Refuted** means evidence warrants its negation or an explicit exclusive alternative under the declared ontology. **Unresolved** means the adjudicators cannot establish either result within the allowed sources, date, and review budget. A graph lookup can identify a known positive or an absent edge, but graph absence alone cannot assign the latter two states (www.w3.org). References, qualifiers, and explicit unknown-value statements in Wikibase illustrate why the graph's own data model may contain more epistemic structure than a binary edge table exposes (www.mediawiki.org).

The word “unknown” must be interpreted locally. InferWiki's **three-class** task uses positive, negative, and unknown labels (arxiv.org). The authors explicitly say that their unknowns are factually incorrect but lack an inferential contradiction in the training data; this is a distinction in **derivability**, not an assertion that their real-world truth is unresolved (aclanthology.org). A study that needs a genuinely unadjudicated class must create and report one. The InferWiki repository also defines its unknown label by the absence of a conflicting training triple (github.com). CoDEx's manually labeled false candidates also have an operational boundary: its annotation guideline instructs a false label when no reliable, specific supporting information can be found (gemslab.github.io). That rule is useful for a repeatable benchmark, but its false label should be read as a decision under that search procedure.

Table 1 maps benchmark labels and splits to the question each design can answer. “Unknown representation” refers to the benchmark's published evaluation construct, not to an assumed universal truth value.

Benchmark or analysis	Positive source and candidate construction	Negative validation and unknown representation	Split and principal acceptance risk
FB15K-237	Freebase-derived selected triples with near-duplicate and inverse relations reduced (www.microsoft.com) (aclanthology.org)	Corrupted candidates are not a published, source-adjudicated false set in the original split description (aclanthology.org)	Held-out facts from the selected graph; inspect residual shortcuts and unobserved truths (aclanthology.org)
CoDEx-S/M	Wikidata/Wikipedia graphs; type-compatible, top-ranked predictions supply hard candidates (aclanthology.org) (gemslab.github.io)	Candidates were manually labeled; a 100-candidate audit reported 89.5% mean agreement with original labels (gemslab.github.io)	90/5/5 train, validation, test split; test graph coverage and model-selected candidate dependence (gemslab.github.io)
CoDEx-L	Larger member of the same benchmark family (aclanthology.org)	The paper's dataset table lists no verified validation or test negatives for this variant (arxiv.org)	Do not transfer S/M hard-negative classification claims to L (arxiv.org)
InferWiki16k	Wikidata-derived, rule-mined premises in training and conclusions in test (arxiv.org)	Positive, negative, and a graph-inferential “unknown” class; the latter was still judged factually incorrect (arxiv.org) (aclanthology.org)	Check rule overlap, evidence date, and the meaning of its 1,501 unknown test triples (arxiv.org)
WikiKG2	Wikidata graph versions; later added triples form validation and test (ogb.stanford.edu)	1,000 sampled corruptions per query excluded if already in the graph; no source adjudication implied (ogb.stanford.edu)	Temporal addition is observable, but absent candidate truth remains open (ogb.stanford.edu)

The matrix prevents a common category error. A source-adjudicated hard negative, a graph-absent corruption, and an inferentially non-derivable claim are different objects. Their scores need separate denominators and should not be pooled as though they share a single false-label definition. A benchmark may still be valuable for its intended task while remaining insufficient for a different claim about real-world fact discovery (gemslab.github.io) (aclanthology.org) (ogb.stanford.edu).

From random corruptions to reviewed hard candidates

Uniformly replacing a head or tail is convenient because it can produce many candidates without manual review. It is also a weak adversary for triple classification. Safavi and Koutra constructed CoDEx hard candidates using type compatibility, a model's top-ranked predictions, and absence from observed graph edges, followed by manual labeling for CoDEx-S and CoDEx-M (gemslab.github.io). Their reported **13 to 19 percentage-point** decline in classification accuracy against hard rather than uniformly sampled negatives is an author result for those models and splits (gemslab.github.io). CoDEx-L should not inherit that conclusion about verified negatives because the paper lists none for its validation and test partitions (arxiv.org).

Cao et al. generated InferWiki candidates with rules and additional corruptions. Their Figure 3 groups annotated negatives together with the benchmark's inferential unknowns in its binary analysis; tested baselines classified random supplements readily, while accuracy dropped on annotated candidates (arxiv.org). These

findings support a **candidate-difficulty slice**, not a universal contamination rate for all KGC data. Without adjudicated labels for a given sample, a false-negative percentage is not estimable from graph nonmembership alone.

A practical candidate set should therefore record how each proposal arose. At minimum it needs **uniform random**, **type-compatible**, and **model-proposed hard** strata, with duplicate and known-positive filtering declared. The researcher should retain each proposal's source model or sampling rule, proposal date, entity type constraints, and inclusion probability. When the deployed workload has entity or relation frequencies unlike the benchmark, weighting must refer to the intended target population rather than to a convenient balanced test set. Candidate construction belongs beside the model score in the report.

From a single rank to a family of decision metrics

MRR averages the reciprocal rank of a target positive; **Hits@K** counts whether it appears among the first K. Both depend on which candidates are retained, which truths are known for filtering, and how equal scores are ordered. Sun et al. recommended random placement of a correct triple within a tie and reported means and standard deviations over **five runs** for their random tie protocol (aclanthology.org). Their correction addresses **score ties**. Yang et al.'s open-world analysis addresses **missing true answers**. Neither correction substitutes for the other (aclanthology.org) (papers.nips.cc).

Under Yang et al.'s setup, entity and relation sets are fixed, and the theoretical treatment begins with assumptions about independent missingness and prediction before examining correlation (papers.nips.cc). The paper reports that stronger models may receive lower conventional metric values under its model, and that increasing the number of queries does not eliminate one correlation-related inconsistency (papers.nips.cc). That is a conditional result: an acceptance report should name the assumptions and test robustness rather than announce that every empirical MRR comparison is reversed. Their suggestion to inspect less top-heavy metrics offers a sensitivity analysis, not a universal replacement for adjudication (papers.nips.cc).

Calibration asks a different question: whether a stated probability corresponds to observed correctness on a defined candidate population. Safavi, Koutra, and Meij (2020) specifically distinguished calibration under closed-world labels from open-world labels obtained for unobserved predictions and found their tested calibration methods less effective in the latter setting (arxiv.org) (aclanthology.org). Pezeshkpour et al. introduced a semi-complete triple-classification benchmark because ranking metrics alone do not reliably assess calibrated scores (www.akbc.ws). A Brier score or reliability plot can help once labels and sampling weights are credible, but scikit-learn notes that Brier loss alone mixes calibration and discrimination (scikit-learn.org).

Implementation Considerations and Process Changes

Freeze the evaluation target and evidence boundary

An acceptance test begins with a written estimand: “For which query distribution, entity universe, source set, time window, and review budget should this model rank missing facts?” The **candidate universe** can be all in-scope entities, a typed subset, or a fixed retrieved shortlist. These choices change denominator and difficulty.

The **evidence boundary** fixes permitted public and private sources and their observation dates; adjudicators should receive the same snapshot. The time boundary is especially important when a benchmark uses later graph additions as test positives, as WikiKG2 does (ogb.stanford.edu). A later-added edge is evidence of later graph inclusion, not automatic proof of when its underlying fact first became true.

The record for each judged candidate should make its chain of custody inspectable. W3C's provenance ontology represents entities, activities, and agents and supports chains connecting used and generated evidence (www.w3.org). For an evaluation item, the practical fields are the triple, normalized entities, relation definition, original proposal stratum, source locator, quoted supporting or contradicting passage, observation date, adjudicator, decision rule, and uncertainty reason. The Reliable Context site describes dated, verbatim citations behind values and explicit reasons for not-found or blocked values; that illustrates the kind of **observable record property** an evaluation set can require, without asserting model accuracy.

The target and label policy can be committed in a short protocol:

- **Population:** Declare entity and relation eligibility, type constraints, and the intended query distribution.
- **Time:** Freeze graph, evidence sources, and adjudication cutoffs before test scoring.
- **Positive rule:** State what evidence supports a triple and whether temporal qualifiers are required.
- **Negative rule:** State what explicit contradiction, exclusivity axiom, or exhaustive source warrants refutation.
- **Unresolved rule:** Record not found, inaccessible or blocked, and not applicable separately, with reasons.
- **Coverage:** Report how many candidates lack a conclusive judgment in every sampling stratum.
- **Identity:** Preserve the mapping from mentions to canonical entities and audit entity merges or splits.

Adjudicate a stratified sample without manufacturing negatives

Judgment collection should blind reviewers to model identity and score where feasible, while preserving the evidence needed to decide the triple. A useful design samples from each candidate stratum and score band, including the highest-ranked predictions and low-scored controls. **Two independent reviewers** can label a subset, resolve disagreements under a written rule, and report agreement by relation and label. The CoDEx paper's separate audit of **100** candidate negatives reported **89.5%** mean agreement with its original labels; InferWiki's audit of **300** test triples, balanced by label, reported **84.3%** agreement (gemslab.github.io) (aclanthology.org). Those are dataset-specific checks, not expected agreement rates for a new graph.

Pooling candidates from several systems can reveal relevant answers that one system's output would miss. Zhou et al. (2022) used an information-retrieval-style pool on sampled FB15K-237 queries and reported changed ordering among **13** tested KGC models under expanded judgments (arxiv.org). TREC's documentation is precise about the limit: pooled relevance judgments are considered complete for the particular document set assembled, not for all possible documents (trec.nist.gov). The same limit applies to triples. A held-out model can propose valid candidates outside the pool, so the report should quantify unjudged top-ranked items and test how conclusions change if some are true.

The adjudication workflow should report these denominators:

- **Proposed:** Count unique triples per stratum before known-positive filtering.

- **Eligible:** Count candidates after type, temporal, and identity checks.
- **Reviewed:** Count those actually seen by adjudicators, with sampling probabilities.
- **Supported:** Count source-backed positives under the written evidence rule.
- **Refuted:** Count explicit negatives under the written refutation rule.
- **Unresolved:** Count evidence-limited cases, separated by reason.
- **Disputed:** Count reviewer disagreements before and after resolution.
- **Coverage:** Give the share judged conclusively in each score band, with uncertainty intervals.

A **false-negative contamination estimate** is permissible only for a known sampling frame and reviewed sample. If (n) graph-absent candidates are sampled with equal probability and (x) are later supported, the sample estimate is (x/n); its uncertainty should reflect the sampling design and reviewer error. For unequal inclusion probabilities, use design weights.

Audit split leakage, ranking code, and training labels separately

A split audit checks whether a test target is inferable by trivial duplication, inverse relation, alias, or pre-existing evidence that was intended to be held out. Toutanova and Chen removed near-duplicate and inverse-duplicate relations in constructing FB15KSelected, a precursor to FB15K-237 (aclanthology.org). CoDEx removed relations with more than **50%** entity-pair overlap and used a **90/5/5** split while keeping validation and test entities and relations present in training (arxiv.org) (gemslab.github.io). InferWiki deliberately put grounded rule premises in training and conclusions in test, making rule availability part of its measured task (arxiv.org). Another inferential benchmark notes that random splits can miss whether a model learns target inference patterns (proceedings.kr.org). These are different split hypotheses; none is a generic guarantee against all leakage. A separate open-world KGC line studies unseen entities, which is a different generalization problem from missing facts among fixed entities (ojs.aaai.org).

Before comparing model scores, reproduce the **ranking code** on a small disclosed set. Record whether head and tail queries are both scored, how known positives from train, validation, and test are filtered, whether test facts other than the target are removed, how ties are handled, and whether candidate counts vary by relation. PyKEEN's documentation describes filtered ranking as suppressing other known true triples (pykeen.readthedocs.io). Sun et al. show that tie placement can change reported performance enough to warrant a declared random policy and repeated runs (aclanthology.org). Compare implementations using the same fixed score vector, not merely the same model checkpoint.

Training-time negative sampling needs a separate audit. Tang et al. argue that corrupted triples used as negatives can be potential true facts and propose PUDA's **positive-unlabeled risk estimator** for KGC training (www.ijcai.org). Kamigaito and Hayashi show that negative-sampling loss behavior depends on choices such as margin and number of samples (proceedings.mlr.press). Those findings inform optimization. The PyKEEN 1.0 paper likewise documents distinct training approaches and loss functions as experimental choices (www.jmlr.org). They do not certify the evaluation set: a model can be trained with positive-unlabeled risk and still be scored against unadjudicated random corruptions. The training report should disclose loss, sampling policy, observed-positive coverage, and relation constraints independently from the test-label report.

The final comparison should be predeclared:

- **Ranking:** Give filtered MRR, Hits@1, Hits@10, and query-level intervals on the frozen pool.
- **Judged ranking:** Repeat on adjudicated candidates and show unjudged rates at each K.
- **Classification:** Give a three-state confusion matrix and precision-recall measures for supported and refuted labels.
- **Calibration:** Give reliability by score band and prevalence-adjusted Brier or proper-score analysis on reviewed items.
- **Slices:** Separate uniform, type-compatible, hard, relation, time, and source-availability groups.
- **Robustness:** Recompute under alternate tie placement, plausible unknown labels, and removal of suspected leakage.
- **Decision:** State whether the effect direction and practical threshold survive every required slice.

An acceptance record should also identify conditions that stop a score from being treated as decisive:

- **Label gap:** Too many top-ranked candidates remain unresolved for the proposed use.
- **Slice reversal:** A reported gain changes sign on reviewed hard candidates.
- **Source drift:** Adjudicators used evidence outside the frozen date or permission boundary.
- **Identity drift:** Two versions of an entity were compared as one, or one entity was split across identifiers.
- **Metric drift:** Tie handling, filtering, or candidate counts differ between compared runs.
- **Selection drift:** The reviewed sample cannot be related back to its proposal population.
- **Uncertainty overlap:** The effect is too small for the declared decision threshold under the measured intervals.
- **Transfer gap:** The private target population has not been tested under the same label policy.

Data Analysis and Evidence

The published numbers below are **author-reported** and refer to different tasks, samples, and metrics. They cannot be ranked against each other as if they were a common leaderboard. *Table 2* records the claims with their denominators or baseline and the limits of the available verification.

Study and test set	Reported result and comparator	Metric and interpretation	Uncertainty and reproduction status
Safavi and Koutra (2020), CoDEx hard negatives (aclanthology.org)	Accuracy fell **13 to 19 percentage points** versus uniform negatives for tested models (gemslab.github.io)	Triple classification, not a cross-dataset MRR improvement	Paper-level claim; no independent rerun established here
Tang et al. (2022), FB15k-237 ablation (www.ijcai.org)	PUDA **0.369** versus ordinary positive-negative DistMult **0.313** (www.ijcai.org)	Filtered MRR after a training intervention	Paper-level claim; test-label completeness not established
Zhou et al. (2022), sampled FB15k-237 (arxiv.org)	Expanded judgments changed ranking among **13** models (arxiv.org)	System-order sensitivity to judgment completeness	Paper-level claim; no operational private-graph replication established
Yang et al. (2022), theory and artificial family graph (papers.nips.cc)	Stronger modeled systems may receive lower MRR or Hits@K (papers.nips.cc)	Conditional order reversal under modeled missingness	Assumption-bound analysis, not an empirical prevalence estimate

The table gives no pooled effect size because its rows mix accuracy, MRR, model order, and theoretical behavior. The absence of a confidence interval in this summary means one was not extracted and verified from the cited source; it does not mean the original study had no variability analysis. Sun et al. did report means and standard deviations over five random-tie runs, illustrating one reproducible way to expose protocol sensitivity (aclanthology.org). In a new benchmark, **uncertainty belongs at the query level and the judgment-sampling level**, because more random seeds do not repair systematic missing labels.

A transparent toy query shows the direction of the missing-truth penalty. There are **three** candidates: (t), an observed positive; (u), a true but unrecorded positive; and (f), a genuinely false candidate. Model A ranks (u,t,f); Model B ranks (t,f,u). Under the observed graph, (u) is not filtered. Under complete truth labels, the other positive is filtered when each positive is evaluated. These are arithmetic assumptions of this example, not benchmark measurements. *Table 3* gives the resulting scores.

Model and ordering	Observed-label MRR	Complete-label filtered MRR	Observed Hits@1	Complete Hits@1
A: (u,t,f)	(1/2 = 0.50)	((1+1)/2 = 1.00)	**0**	**1.00**
B: (t,f,u)	(1/1 = 1.00)	((1+1/2)/2 = 0.75)	**1**	**0.50**

The calculation is explicit: for A, the unknown-but-true (u) pushes (t) to observed rank two, while complete judgments allow each true answer to be evaluated with the other true answer filtered. For B, (f) stays ahead of (u), so the newly judged positive receives rank two. **Observed labels prefer B; complete labels prefer A.** Yang et al.'s paper analyzes why this kind of order reversal can occur under formal missingness assumptions, although this toy list is constructed here and is not their experiment (papers.nips.cc). The example does not show a typical reversal frequency. It shows why the acceptance test must inspect newly judged top predictions.

Implications and Future Directions

Acceptance should be conditional on the proposed use. If the workload is **ranking for human review**, precision among top proposals and the fraction of unresolved top proposals may matter more than a full-entity MRR. Meilicke et al. describe the practical setting in which an expert reviews only the top of a ranking, while standard KGC evaluation is an artificial controlled approximation (link.springer.com). If the workload needs **probability-like scores**, per-query ranking alone is insufficient; a model can rank perfectly within queries while producing scores that are not comparable across queries (proceedings.mlr.press) (www.akbc.ws). Calibration must be evaluated on a labeled, representative candidate population, with the unknown class preserved rather than silently converted to false (arxiv.org) (scikit-learn.org).

A public benchmark can test methodology, but **private heterogeneous records** add source permissions, time-varying evidence, entity ambiguity, and records that cannot be exposed for annotation. A transfer study should preregister the same label policy and candidate strata, then report how many proposed facts were supported, refuted, unresolved, or blocked by access policy. The chain of custody for each label makes disagreement traceable; W3C provenance chains offer a general representation, while explicit unknown-value distinctions in

Wikibase show that unknown need not be an empty field (www.w3.org) (www.mediawiki.org). Reliable Context's stated guarantee of reproducible past record states is relevant as an evidence-retention property, but no public KGC benchmark here verifies its predictive performance or any private-graph transfer claim.

Future benchmark releases would be more informative if they published **judgment coverage by rank and stratum**, relation-level uncertainty, source and observation dates, and an explicit procedure for revisiting disputed labels. Macro metrics can reveal whether performance concentrates in frequent query types; Zhou et al. report greater sampling stability for several macro measures in their FB15k-237 study (arxiv.org). A supplementary global ranking can test whether scores are comparable across queries, as Brunink et al. propose, while retaining per-query results for the standard task (proceedings.mlr.press). Neither aggregation replaces source-grounded adjudication. The most useful negative result would be a model gain that disappears on reviewed hard candidates: it would identify the exact assumption under which the gain mattered.

Conclusion

A credible **knowledge graph completion evaluation protocol** treats graph absence as an observation, not a false label. It states the evaluation population and evidence date, samples candidates from both easy and hard distributions, adjudicates enough cases to expose missing truths, and reports supported, refuted, and unresolved outcomes separately. It also audits split leakage and score ties before interpreting MRR or Hits@K. CoDEx, InferWiki, FB15K-237, and WikiKG2 each provide useful but different tests; their labels and candidate sets should be read literally rather than combined into a single claim of open-world validity (gemslab.github.io) (aclanthology.org) (ogb.stanford.edu).

The acceptance decision is straightforward to state even when it is laborious to measure: a reported gain is persuasive only when it persists under the declared candidate population, reviewed hard and random slices, a reproducible rank implementation, and uncertainty from both queries and human judgments. Training with positive-unlabeled methods may improve robustness to sampled false negatives, but it does not turn unknown evaluation triples into verified false ones (www.ijcai.org). When an important slice remains largely unjudged, the honest result is **undetermined under that slice**, with its coverage and evidence boundary published. That conclusion is more useful for model selection than an apparently precise score whose denominator silently changes with what the graph happened to know.

Source links

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

1. https://www.w3.org/TR/rdf11-mt
2. https://pykeen.readthedocs.io/en/stable/tutorial/understanding_evaluation.html
3. https://papers.nips.cc/paper_files/paper/2022/hash/378226e5df7eded3e401de5c9493143c-Abstract-Conference.html
4. https://aclanthology.org/W15-4007.pdf
5. https://gemslab.github.io/papers/safavi-2020-codex.pdf
6. https://arxiv.org/html/2108.01387

7. https://aclanthology.org/2021.acl-long.534.pdf
8. https://ogb.stanford.edu/docs/linkprop
9. https://arxiv.org/abs/2205.04105
10. https://trec.nist.gov/data/reljudge_eng.html
11. https://www.ijcai.org/proceedings/2022/312
12. https://www.mediawiki.org/wiki/Wikibase/DataModel
13. https://aclanthology.org/2020.acl-main.489.pdf
14. https://github.com/TaoMiner/inferwiki/blob/main/readme.md
15. https://www.microsoft.com/en-gb/download/details.aspx?id=52312
16. https://aclanthology.org/2020.emnlp-main.669
17. https://arxiv.org/html/2009.07810
18. https://papers.nips.cc/paper/2022/file/378226e5df7eded3e401de5c9493143c-Paper-Conference.pdf
19. https://arxiv.org/abs/2004.01168
20. https://aclanthology.org/2020.emnlp-main.667
21. https://www.akbc.ws/2020/papers/1uufzxsxfL
22. https://scikit-learn.org/stable/modules/calibration.html
23. https://www.w3.org/TR/prov-o
24. https://proceedings.kr.org/2023/45
25. https://ojs.aaai.org/index.php/AAAI/article/view/11535
26. https://www.ijcai.org/proceedings/2022/0312.pdf
27. https://proceedings.mlr.press/v162/kamigaito22a.html
28. https://www.jmlr.org/papers/v22/20-825.html
29. https://link.springer.com/article/10.1007/s00778-023-00800-5
30. https://proceedings.mlr.press/v284/brunink25a.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.