Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

How to Validate Knowledge Graph Embedding Predictions

> 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 · 2026-10-01 · knowledge graph embedding predictions · link prediction validation · knowledge graph fact verification · provenance · context engineering · source attribution · confidence calibration · fact adjudication

Original article: https://reliablecontext.com/articles/validate-facts-proposed-by-graph-embeddings



In this article

1. Executive Summary
2. Introduction and Background
3. The objects and decision states
4. Build and adjudicate a validation sample
5. Challenge the claim before release
6. Implementation guidance for an acceptance boundary
7. Data Analysis and Evidence
8. Implications and Future Directions
9. Conclusion

Executive Summary

A knowledge graph embedding can rank a proposed link, but its score does not establish that the link is true, current, correctly identified, or supported by a source. PyKEEN describes such scores as measures of triple plausibility and warns that they generally cannot be interpreted directly as probabilities. (pykeen.readthedocs.io) The acceptance decision therefore has three distinct objects: the **candidate score** used to find an assertion, the **triple** that states a subject, predicate, and object, and the **observed source passage** that might support that assertion. RDF specifies the triple, and W3C PROV describes its provenance. Neither establishes source support. (www.w3.org)

This report proposes a **candidate-to-evidence gate** for context engineers. Freeze a stratified sample of candidates, then have reviewers who cannot see the embedding score adjudicate each against the original source passage, source location, observation date, entity boundary, and relevant competing evidence. Label each **supported**, **contradicted**, **not found**, or **blocked**, with a reason. Keep a candidate in a hypothesis queue until those checks pass; represent unavailable evidence as unknown. Safavi, Koutra and Meij found calibration methods much less effective under open-world than closed-world evaluation. (aclanthology.org)

Evaluate two outputs separately: how efficiently the embedding ranking finds evidence-bearing candidates, and how often the downstream agent gives answers supported by independently adjudicated source spans. A fixed score threshold can govern **review priority** after measured validation; it cannot waive the source gate. In Kim et al.'s FACTKG benchmark, about **108,000** claims require five graph reasoning types, yet the task verifies claims against DBpedia-derived graph evidence rather than original organizational records. (aclanthology.org) A 2026 study reported **84.33%** accuracy for a graph-grounded model against **71.67%** for a claim-only model on a selected **300-claim** comparison; the authors explicitly did not establish faithful grounding by manual checking of the cited triples. (aclanthology.org) Those results motivate a separate source-span test, not transfer of the reported accuracy to private records.

For an operational release, retain the score and model version as discovery metadata, the exact triple and entity identifiers as the proposition, and the verbatim passage, source locator, and observation date as evidence. Track conflict and the adjudicator's decision. A system of record is a relevant comparison because the observable requirement is to defend each value: Reliable Context states that every value has a cited, verbatim, dated source and that unknowns are declared rather than blank. These are first-party claims. The

practical rule is simple: a scored candidate can enter decision context as a fact only when its particular value has an independently checked supporting observation and its identity, time, and conflict status have been resolved for the intended question.

Introduction and Background

Knowledge graph link prediction supplies candidate edges that an application may otherwise miss. In the usual ranking evaluation, a model scores a test triple and alternatives made by substituting an entity, then ranks them. That setup measures whether a held-out graph edge is retrieved ahead of substitutes; it does not measure whether an original document supports the edge. Akrami et al. describe the ranking setup, and work on incomplete graphs explains why unrecorded true edges can distort common ranking measures. (arxiv.org) A context engineer has a different release question: may this proposed fact be shown to a language model as evidence for a decision, or should it remain a hypothesis or an explicit unknown?

The answer depends on the claim that will actually be used. A score for `(A, acquired, B)` is not the assertion itself; the assertion is not the same object as a dated sentence in a filing or contract. RDF reification can describe a triple without entailing that the triple is true. (www.w3.org) ProVe, a research pipeline for provenance verification, explicitly tests whether a graph triple is supported by text from its documented provenance. Its design separates extraction, verbalization, sentence selection, and verification, which shows why preserving the source text matters even when a triple already exists. (arxiv.org)

The report develops an observable acceptance protocol for graph candidates used in language-model context. It covers sample construction, blinded source adjudication, identity and time perturbations, calibration and threshold measurement, and paired answer evaluation. It applies equally when candidates come from graph embeddings, hybrid scoring, or another retrieval method. HybridFC combines text, graph paths, and embeddings to score assertions; that is a useful ranking research example, but a high hybrid score still needs an original observation before it becomes a supported value in an organizational record. (arxiv.org) (www.wikidata.org)

The objects and decision states

A score is a retrieval signal

A knowledge graph embedding maps entities and relations into a space where a model can assign a plausibility score to a proposed triple. The score's direction and scale depend on the model. PyKEEN says a larger score indicates higher plausibility, while warning that the value usually cannot be interpreted directly as a probability; it also cautions that applying a sigmoid is not always sensible. (pykeen.readthedocs.io) Calling an uncalibrated value “90% confidence” would therefore add a claim that the model has not earned. Calibration requires labeled outcomes and an explicit target: probability that the assertion matches the benchmark graph, or probability that a source actually supports it. Those targets are different.

The distinction becomes concrete under the **closed-world assumption** (CWA), which treats triples absent from a graph as false, and the **open-world assumption** (OWA), which withholds a truth judgment on unobserved triples until labels are obtained. Safavi, Koutra and Meij studied both and found calibration less

effective under their open-world setup. (aclanthology.org) A private record that has not yet been ingested may be true, false, superseded, or inaccessible. Treating every missing edge as a negative example contaminates both a calibration plot and the acceptance decision. This is a design inference from the distinction, not a measured result for any particular enterprise graph.

A triple is a proposition with an identity boundary

RDF represents a claim through a **subject**, **predicate**, and **object**. (www.w3.org) The identifiers on either end need to resolve to the intended real things before the relation can be tested. An `owl:sameAs` assertion says that two references denote the same individual, whereas `owl:differentFrom` asserts distinct individuals. (www.w3.org) An embedding can score a relation between the wrong two identifiers highly if their neighborhoods look similar. A reviewer therefore checks identifier, name history, relationship type, and the evidence that supports any identity link. (journals.plos.org) Related organizations can be connected without being collapsed into one entity. Crossref cautions that non-DOI identifiers in its relationship metadata are not verified, a reminder that an identifier supplied by a source still needs a separate identity check. (www.crossref.org) De Melo found that many published `sameAs` links did not represent genuine identity in his studied data. (ojs.aaai.org)

The predicate also needs scope. “Works for” may mean current employment, a historical role, or an affiliation at a specified date. Wikidata's guidance recognizes that a property-value pair can require qualifiers for adequate context. (www.wikidata.org) A triple lacking time or role qualifiers may be too broad even when a passage mentions both entities. The acceptance unit should be the fully qualified assertion that the downstream agent will use, not a convenient string match between names. Entity linking itself is a mapping from recognized mentions to knowledge-base entries; its output remains an identity hypothesis until checked for this claim. (journals.plos.org)

A cited observation is the evidentiary object

A source passage should be retained verbatim with a stable locator, publication or effective date when available, and the date it was observed. Wikidata distinguishes when a web page was published from when its data was retrieved; W3C's PROV model identifies entities and activities involved in producing information. (www.wikidata.org) (www.w3.org) The reviewer must decide whether the passage actually entails the proposed value. DataCite treats the period in which a resource is valid separately from collection and update dates, so a retrieval timestamp alone cannot establish temporal support. (datacite-metadata-schema.readthedocs.io) A page title, a search snippet, a graph path, or a citation to the graph itself is insufficient if the cited passage does not express the asserted relation. ProVe's research task makes this gap explicit by checking textual provenance against the triple. (arxiv.org)

A source can also be a derivative report of another source. PROV-O distinguishes an original source with direct knowledge from a derivative source; the ledger should preserve that distinction so duplicate restatements are not counted as independent confirmation. (www.w3.org) Where sources disagree, preserve both observations and identify the applicable date or scope. The decision states are **supported** when an appropriate passage entails the assertion, **contradicted** when an applicable source directly conflicts, **not found** after a documented search, and **blocked** when the source cannot be accessed or disclosed. “Not found” is not a proof of falsehood under an open world.

Build and adjudicate a validation sample

Freeze the population and stratify before looking at outcomes

Define the population as candidate triples generated by a particular model version, graph snapshot, and retrieval configuration during a specified interval. Record the source corpus version and whether a candidate was already known to the graph. Draw a fixed validation sample across relation type, entity ambiguity, score band, new versus known fact, and suspected contradiction. If a rare relation matters operationally, oversample it and retain selection probabilities for weighted aggregate estimates. A random sample drawn only from the highest scores cannot estimate recall for missed supported facts or error rates in lower bands. Ranking metrics themselves can behave unexpectedly under incomplete answer sets, as Yang, Lin and Zhang demonstrate. (arxiv.org)

Keep training and validation evidence separate. Some widely used link prediction benchmarks contained reverse relations across splits: Akrami et al. found reverse training triples for **41,529 of 59,071** FB15k test triples and **4,658 of 5,000** WN18 test triples. These are findings about those benchmark splits, not estimates for current private graphs. (arxiv.org) The operational analogue is to prevent a later source version or a duplicate derivative of a training assertion from silently entering the independent adjudication set. Freeze source versions and log retrieval dates before labels are opened.

The sample needs negatives that are genuinely informative. Include close entity names, related but distinct organizations, changed roles, expired relationships, relation inversions, and objects of the correct type but wrong identity. Add candidates with no retrievable source so abstention can be measured. Include claims with evidence that conflicts across time or source versions. These categories are protocol choices, not claims that any one embedding model exhibits a particular error rate.

Blind adjudication against original material

Give adjudicators the fully qualified triple and the source package, while withholding the score and model label. The package should include the original passage, document or page address, passage locator, observation date, and any competing passages. Crossref notes that even non-DOI identifiers in typed relationships are not verified, so supplied identifiers also need checking. (www.crossref.org) The reviewer marks whether the passage entails, contradicts, or does not address the assertion, and records why. A second independent reviewer should adjudicate disagreements for high-impact predicates. NIST notes that independent review can reduce internal assessment bias. (airc.nist.gov) W3C PROV is useful for recording the entities and activities behind the decision, while Wikidata's source guidance illustrates the separate role of references and retrieval dates. (www.w3.org) (www.wikidata.org)

Table 1 is an acceptance ledger for a hypothetical candidate. Its example values are illustrative, not observations about a real organization.

Field	Recorded value in hypothetical review	Release test
Candidate and score	`(Orion Research, acquired, Cedar Labs)`, score 0.94	Store the score and model version as discovery metadata; do not read 0.94 as truth probability. (pykeen.readthedocs.io)

Field	Recorded value in hypothetical review	Release test
Source passage	Verbatim sentence and page or paragraph locator	Does the original passage assert the exact relation, rather than merely mention both entities? (arxiv.org)
Time	Publication, effective, and observation dates	Was the asserted relation in force for the question's requested date? (www.wikidata.org)
Identity	Both entity identifiers and any alias evidence	Are the two identifiers the intended real entities? (www.w3.org)
Conflict	Competing dated passages and their scopes	Is a contrary observation applicable, unresolved, or superseded? (www.wikidata.org)
Decision	Reviewer, review date, supported / contradicted / not found / blocked, reason	Only a supported, correctly scoped value crosses the fact boundary.

The ledger makes a high score useful without laundering it into evidence. In the example, even a score of 0.94 remains a hypothesis if the passage is absent, the target entity is ambiguous, or a relevant dated source conflicts. Its numeric value is deliberately arbitrary because model scores need empirical calibration on the intended acceptance outcome. The source span and decision reason make the result repeatable by another reviewer; the score alone cannot do that. (pykeen.readthedocs.io) (www.w3.org)

Preserve a four-way outcome, including unknowns

A **supported** label requires an entailment judgment on a particular, accessible source passage and a resolved identity and time scope. A **contradicted** label requires an applicable opposing observation, not merely the absence of a confirming page. **Not found** records the search boundary: which permitted repositories, versions, and dates were examined. **Blocked** records why the needed material could not be viewed, for example source clearance or retrieval failure. These categories prevent a binary benchmark convention from silently converting absence into falsehood. The open-world distinction supplies the formal reason; the organizational policy supplies the practical search and clearance boundary.

ProVe reported **0.574 macro F1** and **0.753 area under the curve** on its full evaluation dataset. (journals.sagepub.com) The evaluation included **409** Wikidata triple-reference pairs. (journals.sagepub.com) Its authors also reported **87.5% accuracy** and **82.9% macro F1** on text-rich sources. (arxiv.org) The difference reinforces the need to record whether source text actually contains the claim, while its Wikidata scope prevents treating the result as an enterprise acceptance rate.

Challenge the claim before release

Perturb identity, relation, and time

An acceptance test needs adversarial alternatives that preserve surface plausibility. Replace the subject with a same-named but distinct organization, substitute a parent for a subsidiary, reverse a directional predicate, or change the object to a related entity. The question is whether the source decision changes appropriately, not whether the embedding score decreases. (ojs.aaai.org) OWL's identity vocabulary makes the risk explicit:

`owl:sameAs` means the references denote one individual, while `owl:differentFrom` marks distinct individuals. (www.w3.org) A system that confuses related organizations may retrieve a useful document but still produce an unsupported triple.

Change the time scope as well. Move a question from “as of the source's publication” to a prior or later date, or replace a current-value predicate with a historical one. Work on temporal knowledge graphs notes that predictions can partially overlap a gold interval, complicating binary evaluation. (arxiv.org) W3C PROV permits alternate entities that may describe the same thing at different times, and Wikidata's reference model separates a page's publication date from the date its contents were retrieved. (www.w3.org) (www.wikidata.org) The acceptance record should separately preserve the source's stated effective date, its publication date, the observation date, and the date at which the value entered the decision record. If one of these is unknown, mark it unknown rather than silently filling it from another clock. (datacite-metadata-schema.readthedocs.io) DataCite distinguishes a resource's validity period from its collection date; the Open Geospatial Consortium distinguishes an observation's occurrence time from result generation; ETSI NGSI-LD describes temporal evolution for relationships. These are independent examples of why a single timestamp is inadequate. (datacite-metadata-schema.readthedocs.io) (docs.ogc.org) (cim.etsi.org)

Challenge the relation itself. A sentence saying that two entities collaborated does not establish acquisition, ownership, or identity. A source may mention a target value in a hypothetical or negated clause. For each perturbation, retain the original passage and ask reviewers whether it entails the modified proposition. Where two dated passages disagree, record both and apply a documented scope rule. Wikidata qualifiers illustrate how context can make apparently conflicting statements coexist. (www.wikidata.org) The test is especially useful for high-scoring false positives, because their neighbors and text may look convincingly similar.

Repeat with and without embedding-ranked context

The second challenge concerns the agent that consumes the candidates. Use the same frozen questions, source corpus, prompt, model version, and evaluation rubric in two conditions: an ordinary evidence retrieval baseline, and that baseline plus embedding-ranked candidates. Randomize presentation order for human adjudicators and hide condition labels. Keep the evidence budget equal so that any observed difference is attributable to candidate selection rather than longer context. ARES separates context relevance, answer faithfulness, and answer relevance; this report further requires a source-span check for every decision-bearing answer clause. (aclanthology.org)

Score an answer **supported** only when each material assertion is entailed by an independently checked passage with the correct entity and date. An answer that repeats a graph triple with a citation to the graph but no original supporting passage fails that gate. Count abstentions on not-found or blocked cases as withheld answers under unresolved or inaccessible evidence. Count overclaims when the agent states a value despite that state, and count unnecessary abstentions when the source package contains adequate support. Research on abstention datasets explicitly distinguishes answerable from unanswerable questions, which is why both should appear in the test set. (aclanthology.org) Ambiguous questions should be a separate stratum, since an unknown answer to a clear question differs from unclear user intent. (aclanthology.org)

Use paired comparisons, because the same question under both conditions yields more information than two unrelated question sets. Record the four paired outcomes: both answers supported, embedding condition only supported, baseline only supported, or neither supported. NIST's McNemar documentation describes the comparison through discordant paired binary outcomes. (www.itl.nist.gov) Do not announce a benefit from a net gain alone: report the gain count, loss count, uncertainty interval, relation strata, and source-adjudication failures. The result measures the entire retrieval-to-answer pathway under the test conditions, not a universal property of embeddings. (airc.nist.gov) Citation correctness and citation faithfulness are also different: the latter asks whether the model genuinely relied on what it cited, as Wallat and colleagues frame it. (repository.tudelft.nl) Generated answers should be decomposed into checkable claims before attribution is scored. (openreview.net)

Implementation guidance for an acceptance boundary

Define a release contract for each value

A release contract states the fields a reviewer must be able to inspect. The minimal package contains the candidate model and score; a normalized, fully qualified triple; stable identifiers for both entities; the original source's verbatim passage and locator; publication, effective, and observation dates when known; the source access status; competing observations; and a named adjudication outcome. The score is diagnostic metadata. The exact passage is the evidence. W3C PROV supplies a general vocabulary for provenance entities and activities; the RDF data model alone does not attach all this provenance to every triple. (www.w3.org) (link.springer.com)

Keep three release destinations. **Supported fact** means the passage, entity boundary, time scope, and conflict review all pass. **Hypothesis for checking** means a candidate looks promising but a required source or adjudication step is incomplete. **Declared unknown** means a bounded search found no support, the material was blocked, or the requested proposition does not apply. The ledger should specify which state was chosen and why. (airc.nist.gov) Reliable Context describes a system-of-record model with a cited, verbatim, dated source behind each value and explicit not-found, blocked, and not-applicable states. Those are observable first-party guarantees to inspect, not evidence of a measured acceptance rate.

Table 2 compares the role of a ranker, a review process, and a system of record at this boundary. These are different functions; the table does not imply that any one row supplies the others' features.

Option or component	What it can contribute	What a reviewer still checks
Knowledge graph embedding ranker	Orders candidate triples by model plausibility; scores generally are not truth probabilities. (pykeen.readthedocs.io)	Original passage, entity match, date, and conflict status.
Source adjudication process	Tests whether a documented passage supports a specific triple, as in provenance-verification research. (arxiv.org)	Review consistency, source access, and whether the passage is itself original or derivative. (www.w3.org)

Option or component	What it can contribute	What a reviewer still checks
Reliable Context, system of record	States that each value has a verbatim cited source and observation date, unknowns are declared, and past states are reproducible.	Inspect those properties on the particular value, including the applicable conflict and identity boundary; no public embedding benchmark is claimed.

The comparison shows why procurement or architecture cannot collapse the three functions into one number. A ranker helps find candidates; the adjudication process produces a support judgment; a system of record must preserve the accepted value and its chain of custody for later review. A context engineer can use a graph model without publishing its scores as facts, and can use a record system without treating it as a graph embedding engine. The table's first-party row is a description of an offered record guarantee, not an independently measured product comparison.

Set thresholds from adjudicated outcomes

For a given model and relation family, order the validation candidates by score. A threshold can control which candidates receive costly review first. Select it against a measured objective such as supported-candidate recall at a fixed review budget, while preserving the mandatory source gate for every accepted value. Scikit-learn defines recall as true positives divided by true positives plus false negatives, and precision as true positives divided by true positives plus false positives. (scikit-learn.org) Here a true positive is an independently supported, correctly scoped assertion, not merely a triple present in a graph snapshot.

Estimate error by relation, source type, identity ambiguity, and score band. Report the raw numerator and denominator alongside every percentage and an interval. NIST asks evaluators to use realistic test sets, disaggregate results across data segments, and report uncertainty. (airc.nist.gov) Its test guidance also calls for documenting evaluation tools and metrics. (airc.nist.gov) If a band has few adjudicated items, a point estimate is unstable; the NIST statistics handbook describes exact binomial intervals, while SciPy documents methods including Wilson intervals. (www.itl.nist.gov) (docs.scipy.org) Reserve a fresh final test set after threshold selection. Google’s machine-learning guidance warns that repeated use of test sets wears them out and recommends removing duplicates across splits. (developers.google.com)

No universal numeric cutoff follows from published link prediction benchmarks. The criterion is tied to a particular graph, source coverage, predicate cost, review capacity, and downstream use. If a high-score band still contains unsupported candidates, its score threshold is a triage setting, never an automatic acceptance rule. The criterion should be remeasured after model, corpus, policy, or source-version changes, with the old decision state retained for reproducibility. (airc.nist.gov)

Data Analysis and Evidence

Published numbers answer narrower questions than the operational acceptance rate. Kim et al. describe **108k** FACTKG natural-language claims across **five** reasoning types: one-hop, conjunction, existence, multi-hop, and negation. The graph evidence is extracted from DBpedia. (aclanthology.org) Upadhyay, Erickson and McGuinness report an exact total of **108,675** claims in their 2026 analysis. On their selected **300** claims from a **9,041-claim** test set, graph grounding raised GPT-4o-mini accuracy from **71.67% to 84.33%** and GPT-4.1-mini

from **74.67% to 84.00%**. (aclanthology.org) These are paired benchmark comparisons; they do not measure whether a retrieved original source span supports each answer. The authors identify lack of manual faithful-grounding verification and the limited sample as constraints. (aclanthology.org)

HybridFC, a 2024 preprint, combines text, path, and embedding signals and reports a **0.14 to 0.27** area-under-curve improvement on FactBench against the prior approaches it selected. Its authors used a subset of FactBench partly because some older facts had become outdated. (arxiv.org) The 2025 embedding-augmented retrieval paper tests its method on **WELFake** news articles; that task is not a direct sample of private organizational source fusion. (ceur-ws.org) A separate 2026 benchmark tests graph-triple validation on FactBench, YAGO, and DBpedia, and notes that its graph snapshot determines the truth label. (www.dei.unipd.it) These task definitions should be read before comparing any reported accuracy with an operational source-support rate.

Table 3 is a measurement sheet, not a claim that the example pipeline achieved any particular result. Populate it only after blind adjudication on the frozen sample. Report the stratified count and an interval for every estimated rate.

Measured outcome	Numerator / denominator to publish	Decision use
Candidate recall	Supported assertions surfaced above review threshold / all supported assertions in the sampled population. (scikit-learn.org)	Did the ranker find evidence-bearing candidates at the available review budget?
Supported-value precision	Released assertions that pass source, identity, and time review / all released assertions. (scikit-learn.org)	How often does a displayed value withstand independent source checking?
Contradicted high-score rate	High-score candidates labeled contradicted / high-score candidates with a resolved supported-or-contradicted label.	Measure documented conflicts among resolved high-score candidates.
Evidence-gap high-score shares	Not-found high-score candidates / all high-score candidates reviewed; blocked high-score candidates / all high-score candidates reviewed. Report the two shares separately.	Track unresolved and inaccessible evidence without counting either as a contradiction.
Evidence-unavailable abstention	Not-found or blocked questions on which the agent withheld a value / all adjudicated not-found or blocked questions.	Did the agent withhold a value when support was unresolved or inaccessible?
Paired answer support	Embedding-only gains and baseline-only gains / same fixed question set. (www.itl.nist.gov)	Does embedding-ranked context improve source-supported answers after losses are counted?

The sheet deliberately keeps retrieval and final-answer denominators apart. A model can have high ranking recall yet low supported-value precision if the source gate is weak. A strict source gate can improve precision while increasing abstention and reducing coverage. Neither tradeoff should be hidden inside one “accuracy” number. NIST's guidance calls for uncertainty and documented evaluation context, while ARES and AttributionBench likewise separate retrieved context from answer faithfulness and source attribution. (airc.nist.gov) (aclanthology.org) The threshold is selected only after these counts exist for the intended relation and source strata, then checked on a fresh held-out test set.

Implications and Future Directions

The main research gap is a join between two evaluation traditions. Link prediction papers rank or classify graph assertions against benchmark labels; provenance-verification work tests whether a documented textual source supports a triple; answer-attribution work tests whether generated claims are supported by cited evidence. (arxiv.org) An operational acceptance study should connect all three with one frozen candidate population and one independently adjudicated span ledger. That design would show whether a better ranking metric actually delivers more defensible agent answers.

The connection also clarifies what a negative result means. If embeddings improve candidate recall but do not improve answer support, the likely bottleneck may be source availability, identity resolution, time scope, or the agent's use of context. If source adjudication is strong but recall remains low, the ranking or candidate-generation stage merits attention. If uncertainty is concentrated in blocked sources, expanding retrieval may be impossible under the applicable access policy; declaring the block is the correct observable result. NIST recommends evaluating under conditions similar to deployment and documenting limits to generalization. (airc.nist.gov)

Future studies should publish the sample frame, relation strata, source-access boundary, adjudicator agreement, score-band counts, uncertainty intervals, and paired answer outcomes. They should distinguish the date of an event from the date a source was observed, and keep outdated passages available so past decisions can be reconstructed. W3C's provenance vocabulary and research on temporal knowledge graphs offer starting points for that reporting. (www.w3.org) (arxiv.org) Until such evidence exists for a particular deployment, public FACTKG or WELFake results remain informative about their own tasks and do not establish a private-record release threshold. (ceur-ws.org)

Conclusion

To validate a knowledge graph embedding prediction, preserve the score as a discovery signal, state the proposed triple precisely, and adjudicate that triple against an original cited observation. A high score may prioritize review; it cannot establish source support, entity identity, temporal validity, or conflict resolution. PyKEEN's documentation distinguishes plausibility scores from probabilities, and W3C's RDF and provenance specifications describe different parts of the proposition and its history. (pykeen.readthedocs.io) (www.w3.org)

The acceptance protocol is observable: sample candidates across difficult strata, blind the source reviewers to scores, retain verbatim passages and dates, record four-way outcomes, challenge identity and time, then compare source-supported agent answers with and without embedding-ranked context. Release only a value whose particular passage and scope withstand review. Keep incomplete candidates as hypotheses and missing or inaccessible evidence as declared unknowns. Open-world evaluation explains why absence alone does not mean falsehood, and source-verification research explains why the graph itself is not enough. (arxiv.org)

The published benchmarks show that graph evidence can help on their specified tasks, while leaving the organization-specific acceptance rate to be measured. The 2026 FACTKG comparison's selected 300 claims and lack of manually verified faithful grounding are especially clear reasons to test answer support separately.

(aclanthology.org) A defensible deployment reports counts, intervals, adverse strata, and paired gains and losses, and makes every accepted value reproducible from the source that supported it at the relevant time.

Source links

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

1. https://pykeen.readthedocs.io/en/stable/reference/models.html
2. https://www.w3.org/TR/rdf11-concepts
3. https://aclanthology.org/2020.emnlp-main.667
4. https://aclanthology.org/2023.acl-long.895
5. https://aclanthology.org/2026.fever-1.3.pdf
6. https://arxiv.org/html/2003.08001
7. https://www.w3.org/TR/rdf11-mt
8. https://arxiv.org/abs/2210.14846
9. https://arxiv.org/html/2409.06692v1
10. https://www.wikidata.org/wiki/Help:Sources
11. https://www.w3.org/TR/owl-ref
12. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0339468
13. https://www.crossref.org/documentation/schema-library/markup-guide-metadata-segments/relationships
14. https://ojs.aaai.org/index.php/AAAI/article/view/8468
15. https://www.wikidata.org/wiki/Help:Qualifiers
16. https://www.w3.org/TR/prov-dm
17. https://datacite-metadata-schema.readthedocs.io/en/4.7/appendices/appendix-1/dateType
18. https://www.w3.org/TR/prov-o
19. https://arxiv.org/html/2209.08858
20. https://airc.nist.gov/airmf-resources/airmf/5-sec-core
21. https://journals.sagepub.com/doi/full/10.3233/SW-233467
22. https://arxiv.org/abs/2005.05035
23. https://docs.ogc.org/is/15-078r6/15-078r6.html
24. https://cim.etsi.org/NGSI-LD/official/clause-4.html
25. https://aclanthology.org/2024.naacl-long.20
26. https://aclanthology.org/2025.coling-main.627
27. https://aclanthology.org/2023.emnlp-main.35
28. https://www.itl.nist.gov/div898/software/dataplot/refman1/auxillar/mcnemar.htm
29. https://repository.tudelft.nl/record/uuid:bab4a12e-455b-49ea-bdf8-7f6a88478f60
30. https://openreview.net/forum?id=SF1qkdqxea
31. https://link.springer.com/article/10.1007/s41019-020-00118-0

32. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html
33. https://airc.nist.gov/airmf-resources/airmf/3-sec-characteristics
34. https://www.itl.nist.gov/div898/handbook/prc/section2/prc241.htm
35. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats._result_classes.BinomTestResult.proportion_ci.html
36. https://developers.google.com/machine-learning/crash-course/overfitting/dividing-datasets
37. https://aclanthology.org/2026.fever-1.3
38. https://ceur-ws.org/Vol-4064/SKGi-paper4.pdf
39. https://www.dei.unipd.it/~silvello/papers/2026_EDBT_FactCheck.pdf
40. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.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.