Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Knowledge graph fact verification: a controlled evidence test

> Reliable Context is a kernel of truth: a system of record for organizations that fuse private and public information and must defend every value they act on. One canonical record per real thing, a cited source and an observation date behind every value, unknowns declared rather than left blank, and any past state of the record reproducible.

Published by Reliable Context · 2026-10-29 · knowledge graph fact verification · knowledge graph fact checking · graph evidence · factkg · evidence provenance · counterfactual evaluation · knowledge fusion · llm evaluation · fact verification benchmark

Original article: https://reliablecontext.com/articles/graph-evidence-fact-verification-controlled-test



In this article

1. Executive Summary
2. Introduction and Background
3. Methodology and verification targets
4. Published studies and their comparability
5. Paired graph controls and an auditable fixture
6. Analysis of key reasoning segments
7. Data Analysis and Evidence
8. Implications and Future Directions
9. Conclusion

Executive Summary

Knowledge graph fact verification requires two separate tests: whether a verdict is correct and whether it depends on the supplied evidence. A graph path can support an answer without establishing that the model used that path. This distinction motivates the controlled evaluation proposed here. Prior work on rationale evaluation separates evidence sufficiency from evidence removal, while intervention studies show that reliance on an explanation can vary across tasks. (www.eraserbenchmark.com) (www.anthropic.com)

FactKG supplies approximately **108,000 claims** spanning one-hop, conjunction, existence, multi-hop and negation reasoning. (aclanthology.org) Upadhyay et al. (2026) report **92.68% accuracy** for a neural model over linearized subgraphs versus **66.54%** for their symbolic baseline. (aclanthology.org) Their separate large language model (LLM) experiment reports gains of **12.67 and 9.33 percentage points** for GPT-4o-mini and GPT-4.1-mini, respectively, with graph grounding. (aclanthology.org) These are different experiments, and neither establishes that every correct answer causally depends on its cited triples. The institutional project page reports paired accuracies rising from **71.67% to 84.33%** and **74.67% to 84.00%**, respectively. (tw.rpi.edu)

Shami, Marchesin and Silvello (2026) evaluate different graph collections and web retrieval. Their results vary across datasets, labels and models; their retrieval intervention therefore cannot reproduce or contradict the FactKG experiment by comparing headline scores. (arxiv.org) The broader literature also distinguishes checking existing assertions from predicting missing links. (dl.acm.org) Evidence-grounded validation must specify its reference graph, time and entailment rules before interpreting any improvement.

This report proposes **five paired conditions**: claim-only, graph-only, matched graph evidence, irrelevant evidence and contradicted evidence. A second intervention removes a necessary edge from matched evidence, with an irrelevant-edge perturbation as a control. The primary outcomes are verdict accuracy, directional response to evidence changes, audited evidence sufficiency, valid paths, calibration, abstention and measured cost. No new model experiment was run for this report; the fixture and acceptance rules below are a proposed protocol.

The practical threshold is an auditable dependency claim: the cited evidence supports the verdict, changing necessary evidence changes the answer appropriately, and irrelevant changes leave it stable. A missing edge should ordinarily lead to insufficient evidence rather than automatic refutation under open-world semantics. (www.w3.org) Provenance must also connect the graph assertion to its documentary source: references establish origin, while even referenced information can be incorrect. (www.wikidata.org)

Introduction and Background

A verifier may return the correct answer because it remembers a familiar association, recognizes a linguistic pattern, or reasons over the supplied graph. Ordinary accuracy does not distinguish these explanations. **Evidence-based fact validation** therefore needs an explicit estimand: the change in a fixed verifier's behavior when evidence changes while the assertion remains fixed. Behavioral testing provides a useful precedent. Ribeiro et al.'s **CheckList** organizes evaluation around capabilities and test types rather than aggregate scores alone. (www.microsoft.com)

The representation matters. A **knowledge graph** expresses entities and their relationships; in the Resource Description Framework (**RDF**), graphs consist of subject-predicate-object triples. (www.w3.org) A natural-language assertion must be aligned with this representation before a path can count as evidence. RDF datasets can separate a default graph from named graphs, a useful distinction when retaining evidence from different sources. (www.w3.org) Entity identity, predicate direction, time and scope belong to the verification problem. A path connecting two names is insufficient if it expresses the wrong relationship. Knowledge graph evidence validation must check the witness, its entailment and its source separately.

The survey by Qudus et al. (2025) covers **27 approaches and 35 benchmarks**, illustrating the breadth of assertion-verification research. (dl.acm.org) Breadth does not create a common experimental population. A benchmark of verbalized claims, a collection of candidate triples and a corpus of web evidence can require different retrieval and reasoning operations. ERASER likewise warns that even nominally shared datasets can have different splits. (www.eraserbenchmark.com)

Reliable Context publishes from a system-of-record perspective. The analysis here concerns the evidence needed to justify a verification claim, rather than a comparison of commercial systems. No publisher performance result is inferred from the literature.

The report distinguishes published measurements from a proposed controlled experiment. Its decision question is narrow: **which controls must be present before a researcher claims that supplied graph evidence determined a verdict?** The answer requires both semantic auditing and behavioral intervention. The World Wide Web Consortium's (W3C) provenance model describes information about the entities, activities and people involved in producing data, a useful foundation for preserving that audit trail. (www.w3.org)

Methodology and verification targets

Define the object being checked

Claim validation is the assignment of a verdict to a specified assertion under declared evidence and interpretation rules. **Graph evidence** is the exact set of triples or source records admitted to that decision. **Model memory** denotes information available through model parameters rather than supplied records. A **grounded verdict**, as operationalized here, must be justified by admissible evidence and respond appropriately to controlled changes in that evidence. This is a proposed definition of the evaluation target, not an assertion that internal model processes have been identified.

Graph-based fact verification should separate three questions:

- **Structural validity:** Do the cited identifiers, predicates and directed edges actually occur in the evaluated graph?
- **Entailment:** Does that evidence establish the particular assertion under the declared rules?
- **Documentary validity:** Does the graph's source support the value at the relevant time and scope?

OWL, the Web Ontology Language, can check consistency and make implicit knowledge explicit; those functions do not independently certify documentary accuracy. (www.w3.org) Wikidata's documentation similarly distinguishes references, which identify origin, from ranks used to prioritize values, and says normal rank makes no judgment of accuracy or currency. (www.wikidata.org)

The target must also distinguish **graph-relative truth** from **world truth**. A graph-relative label asks what follows from a snapshot and its rules. World truth additionally requires suitable documentary evidence. Under OWL's open-world assumption, absence may reflect missing information. (www.w3.org) An evaluation that labels every absent triple false is making an additional benchmark convention, not deriving a universal rule of knowledge.

Freeze an auditable evaluation unit

The proposed evaluation unit is an assertion together with its immutable evidence package and all paired variants. RDF describes graphs as static snapshots; the identity of that snapshot must therefore accompany its claims. (www.w3.org) DBpedia's Databus explicitly distinguishes a dataset's abstract identity from individual versions. (www.dbpedia.org)

Before running inference, register:

- **Assertion identity:** Preserve the original wording, structured assertion, entity identifiers and reasoning tags.
- **Evidence identity:** Save the exact triples, their ordering, retrieval scores and any truncation.
- **Temporal scope:** Record the graph version, source publication date, observation date and claim-validity interval separately.
- **Label policy:** Specify what constitutes support, explicit contradiction and insufficient information.
- **Unknowns:** Distinguish not found, blocked and not applicable, retaining the reason for each.

- **Model conditions:** Freeze model identifier, prompt, demonstrations, output vocabulary and sampling settings.
- **Eligibility:** State why an assertion qualifies for each intervention and which cases are excluded.

The package should include annotation disagreements and unresolved source access. A digital object identifier (**DOI**) for a collection does not necessarily identify the version used: DataCite distinguishes collective identifiers from version-specific identifiers, and Zenodo recommends citing the exact research artifact. (support.datacite.org) (zenodo.org)

Published studies and their comparability

Table 1 distinguishes the published populations and evidence interventions. Its purpose is methodological comparison, not ranking.

Study or artifact	Population and partition	Evidence and graph identity	Reported outcome and interpretation
Kim et al. (2023), FactKG	Approximately 108,000 natural-language claims; five reasoning types. (aclanthology.org)	DBpedia 2015-10 through WebNLG; graph-relative verification. (aclanthology.org)	A claim benchmark, not an intervention test proving that each correct verdict used its cited path.
Upadhyay et al. (2026)	Official test set of 9,041; separate 300-claim LLM sample restricted to examples with at least ten triples. (aclanthology.org)	Linearized subgraphs for neural models; full one-hop subgraphs in the LLM prompting comparison.	Full-test neural/symbolic accuracy and sampled LLM gains are distinct results.
Shami et al. (2026), FactCheck	Table 2: FactBench 2,800, YAGO 1,386, DBpedia 9,344. (arxiv.org)	Candidate triples plus retrieved web text, rather than the same supplied-subgraph intervention. (arxiv.org)	Class-specific F1, the harmonic mean of precision and recall, and configuration-dependent retrieval effects. (arxiv.org)

The table establishes why raw scores should remain within their own designs. The independent variables include assertion representation, evidence retrieval, class balance and prompt format. A stronger result on one combination cannot identify the contribution of another combination's graph reasoning.

What FactKG exposes, and what it leaves to the evaluator

Kim et al. report an **8:1:1** split of claims into training, development and test sets, with disjoint triple sets across the splits. (aclanthology.org) This is a construction description, not permission to regenerate a new random claim split. The later paper reports **86,367 training, 13,267 validation and 9,041 testing claims**. (aclanthology.org) Researchers should record the released files actually used, rather than infer final claim counts from the construction ratio. The discrepancy between the original paper's stated claim-split ratio and the later reported release counts remains unresolved here. The historical DBpedia documentation dates its input Wikipedia dumps to October 2015, separately from later access to those files. (downloads.dbpedia.org)

The original dataset documentation says its released test dictionary omits the **Evidence** field. (github.com) Consequently, a benchmark label is not automatically a complete gold evidence annotation. ERASER, by comparison, explicitly distributes labels and annotator-marked supporting text, illustrating a different evidence

annotation contract. (www.eraserbenchmark.com) A study claiming evidence sufficiency must supply or audit the necessary evidence separately.

FactKG's negatives include entity or relation substitutions. (aclanthology.org) Such construction makes matched controls important: an evaluator should ask whether lexical or entity-type cues predict the label without evidence. The official representation also adds reverse-relation triples. (github.com) An audited path must preserve the distinction between an inverse traversal and an assertion with the original predicate direction.

What the newer comparisons establish

The Upadhyay prompt templates differ beyond graph availability: claim-only responses use True/False and pretrained-knowledge explanations, while grounded responses use SUPPORTED/REFUTED and triple identifiers. (aclanthology.org) The printed appendix also omits demonstrations. (aclanthology.org) The reported difference therefore estimates a combined prompting intervention. That does not invalidate the result; it narrows the causal interpretation and the artifacts needed for reproduction.

Shami et al. freeze collected search evidence and filter documents directly originating from source knowledge bases. (arxiv.org) Source filtering is useful, but indirect copies and model exposure remain separate questions. Table 2 lists 9,344 DBpedia facts, whereas the prose says 9,934; this report preserves the table value and leaves the discrepancy unresolved. (arxiv.org) Their dataset card exposes an aggregated “train” partition, which should not be mistaken for a documented held-out experimental split. (huggingface.co) A reproducible comparison needs the actual evaluated identifiers, not an inference from a hosting platform's partition name.

Paired graph controls and an auditable fixture

The proposed controlled evaluation keeps the same held-out assertions across all eligible conditions. The primary contrast fixes the system instruction, label vocabulary, demonstrations, response schema and output budget. Only the evidence slot changes. Require separately scored fields for a best-judgment label, which may use prior knowledge, and an evidence-only label. The claim-only arm measures prior knowledge through the first field and appropriate abstention through the second. Keep those instructions identical across conditions. Never compare the fields as though they were the same task: a correct best judgment without evidence is compatible with memory, but does not identify its internal origin. CheckList's distinction between invariance tests and directional expectation tests is directly useful: some perturbations should preserve an answer, while others should change it in a specified direction. (ijcai.org)

Graph-only needs an explicit definition. A graph without a target assertion cannot be evaluated for assertion truth. Here, graph-only retains a structured candidate assertion and the graph, while withholding the natural-language claim. Use stable identifiers and separately mask familiar aliases where feasible. This arm tests the contribution of natural-language wording and lexical familiarity. It does not make every relation name semantically opaque.

Table 2 gives a minimal **hypothetical example**, not model output or a statement about real organizations. The candidate claim is “Aster is a member of Harbor through a partner.” The stipulated rule is: a partnership edge to an intermediary plus that intermediary's membership edge supports the claim. No additional edges exist in

this synthetic fixture.

Condition	Exact supplied records or target	Expected evidence-only label	Audit purpose
Claim-only	Natural-language claim; empty evidence slot	Insufficient evidence under the evidence-only instruction	Establish the fixed baseline; a guessed answer is not demonstrated grounding.
Graph-only	Candidate `memberViaPartner(Aster, Harbor)`; E1: `(Aster, partner, Birch)`; E2: `(Birch, member, Harbor)`	Supported under the stipulated rule	Test the structured assertion without its prose wording.
Matched graph evidence	Claim plus E1 and E2	Supported	Require the complete witness E1, E2.
Irrelevant evidence	Claim plus I1: `(Cedar, partner, Delta)`; I2: `(Delta, member, Estuary)`	Insufficient evidence	Test response to a graph with matched shape but unrelated identifiers.
Contradicted evidence	Claim plus E1 and C2: `(Birch, notMember, Harbor)`	Refuted only under the fixture's explicit exclusion rule	Test directional response to an explicit contradiction, with no alternative partner route.
Necessary-edge removal, derived from matched	Claim plus E1; E2 removed	Insufficient evidence	Test dependence on a certified necessary edge.
Irrelevant-edge perturbation, derived from matched	Claim plus E1, E2 and I1; remove I1	Supported	Estimate false changes under a label-preserving intervention.

The fixture separates the five main conditions from two interventions derived from the matched condition. Its negative predicate and exclusion rule are explicitly stipulated. Ordinary graph absence would not justify that refuted label. Real-world assertions also need independent documentary sources; a synthetic fixture has none and must remain labeled accordingly.

For a real assertion, save the minimal witness alongside the full supplied graph. Check for alternative witnesses before certifying an edge as necessary. If a second path still establishes the claim, removing the first path should not force a verdict change. ERASER's removal and rationale-only evaluations offer a methodological precedent, although its original objects are text tokens rather than graph edges. (www.eraserbenchmark.com)

Execute the interventions with these controls:

- **Necessity audit:** Certify the edge against the full admitted graph and rule set.
- **Budget matching:** Match irrelevant evidence on triple count, token budget and graph shape.
- **Lexical matching:** Preserve relevant name overlap where the test targets semantics rather than entity recognition.
- **Ordering control:** Randomize or balance triple order without changing the graph.
- **Contradiction audit:** Insert a genuinely incompatible assertion, not merely a different relation.

- **Residual-evidence check:** Remove duplicated statements and source text that restate the deleted edge.
- **Run pairing:** Retain an assertion-level key across all arms and repeated model calls.
- **Label separation:** Keep original binary benchmark scoring distinct from newly audited insufficient-evidence labels.

A cited path remains an object to inspect. Intervention on that path provides behavioral evidence of dependence, not complete identification of a model's internal computation. Anthropic's faithfulness study makes the related distinction operational by changing reasoning content and observing predictions. (www.anthropic.com)

Analysis of key reasoning segments

FactKG's reasoning types should define evidence requirements before subgroup scores are inspected. They are not interchangeable difficulty bins. The following proposed audits translate those categories into observable proof obligations; the category vocabulary comes from the original benchmark. (aclanthology.org)

- **One-hop:** Require the precise predicate, subject and object, with original direction and qualifiers.
- **Conjunction:** Require evidence for every conjunct; test removal of each independently.
- **Existence:** Require a witness for support; require an explicit closure or exclusion rule for refutation.
- **Multi-hop:** Require every edge in a valid composed path and the rule licensing that composition.
- **Negation:** Record the scope of negation, explicit negative evidence and any closed-world convention.
- **Overlapping tags:** Report each subgroup denominator and preserve overlaps rather than summing them as independent populations.

For conjunctions, removing a required conjunct should normally eliminate support while leaving the other conjuncts intact. For multi-hop claims, every transition must retain its entity identity and direction. A visually plausible route is insufficient if a predicate's semantics do not compose with the next one. The benchmark's original description distinguishes all-triple conjunction support from connecting-path support. (aclanthology.org)

Existence and negation require particular care. A found witness can establish an existential statement in the admitted graph. Failure to retrieve a witness cannot establish that none exists. A counterfactual edit that only removes one witness may leave another. The distinction between absent and false follows directly from open-world semantics; any closed-world test needs an explicit completeness boundary. (www.w3.org)

Temporal claims need another axis. A record may support an assertion at an earlier time without supporting it now. Original FactBench models include the timespan in which a fact holds. (github.com) The proposed evaluation should therefore distinguish source publication time, observation time and assertion-validity time. A date qualifier is part of the claim's meaning, not optional provenance decoration.

Lexical cue tests should preserve the target relation. Alias changes, syntactic paraphrases and reordered triples can be label preserving. A swapped entity or altered negation can instead change the assertion itself. These tests need separate expected outcomes, as behavioral testing distinguishes invariance from directional change. (www.microsoft.com) Combining them into a single “robustness” average hides which capability failed.

Finally, report evidence retrieval and evidence reasoning separately. Give the verifier a certified sufficient subgraph to measure its conditional reasoning ability, then compare this with the retrieved subgraph. The FactCheck artifact combines structured facts with extracted questions, text and metadata, illustrating how many stages can intervene before a verifier sees evidence. (huggingface.co) A retrieval miss should not silently become a reasoning error or an unsupported world-truth judgment.

Data Analysis and Evidence

Read the published numbers with their denominators

The Upadhyay language-model experiment uses a **300-claim** sample restricted to evidence-rich cases. (aclanthology.org) Its project page reports graph-grounded accuracies of **84.33%** and **84.00%** for the two mini-models. (tw.rpi.edu) Those values apply to that sample and prompt comparison; they are not an estimate for arbitrary private graphs.

The paired GPT-4.1-mini table is more informative than the accuracy difference alone: **224/300** claim-only answers and **252/300** grounded answers were correct; **49** cases improved and **21** deteriorated, giving a net improvement of **28/300**, or **9.33 percentage points**, calculated as 49 minus 21. (aclanthology.org) This calculation does not identify whether the improvements came from necessary-edge use, because those interventions were not the contrast being tested.

Shami et al.'s Table 5 reports GPT-4o mini's FactBench true/false F1 changing from **0.52/0.72 to 0.91/0.90** with retrieval; Gemma2's DBpedia values change from **0.85/0.36 to 0.79/0.38**. (arxiv.org) Their Table 2 reports a **0.99** positive-label proportion for YAGO. (arxiv.org) These class-specific results illustrate why accuracy, F1 and a heavily imbalanced population should not be placed on a shared scale.

Report evidence use independently of accuracy

Table 3 specifies the proposed reporting schema. Every entry requires its raw denominator, and an unrun experiment should display “not measured” rather than a synthetic result. This follows the broader requirement to document test sets, evaluation methods and uncertainty. (airc.nist.gov)

Measure	Operational numerator and denominator	Interpretation and required companion
Verdict accuracy	Correct verdicts / all eligible cases, with parse failures retained	Score best judgment and evidence-only outputs separately by arm and subgroup; state how abstention is scored.
Evidence-use sensitivity	Changed verdicts / all certified necessary-edge removal cases	Also report directionally correct changes; an arbitrary flip is not success.
False-change rate	Changed verdicts / all label-preserving irrelevant-edge perturbations	Separate edge addition, removal and order changes.
Evidence sufficiency	Verdicts whose cited subset independently entails the label / all audited verdicts	Include empty citations and inadequate witnesses in failures.
Path validity	Valid cited paths / all cited paths, plus fully valid cases / all cases	Validity includes edge existence, direction, identity and qualifiers.

Measure	Operational numerator and denominator	Interpretation and required companion
Calibration	Probability error and reliability curves on a fixed labeled population	Preserve label probabilities, bin counts and the probability source.
Abstention and coverage	Abstentions / all cases; answered cases / all cases	Pair coverage with error among answered cases and reasons for abstention.
Cost	Total retrieval, inference and auditing resources / all attempted cases	Preserve failed attempts, hardware, elapsed time and repeated-call settings.

The table prevents a common interpretive error: a high verdict-change rate can reward instability unless the changes follow the expected labels. Conversely, a stable answer can be correct when alternative sufficient evidence survives. Neither measure should be interpreted without the necessity audit.

Calibration means probability estimates represent the likelihood of correctness, rather than confidence merely sounding persuasive. Guo et al. define this target in their calibration research; their experiments concern image and document classification, not direct measurements of graph verification. (proceedings.mlr.press) Use held-out calibration data and publish subgroup reliability, especially for insufficient-evidence cases.

For binary rates, a Wilson interval is an available method; SciPy documents an implementation without continuity correction. (docs.scipy.org) For paired differences, preserve the assertion as the resampling unit, including all its conditions and repeated calls. Cluster related assertions when they share a documentary source or generated template. The choice of resampling unit is part of the proposed protocol and must be declared before results are interpreted.

Cost also belongs in the comparison. Shami et al. report retrieval often exceeding direct assessment's computational cost by a factor of six or more. (arxiv.org) Such measurements need their original workload and infrastructure, not extrapolation to another deployment. NIST's framework likewise calls for uncertainty, benchmark comparisons and formalized reporting. (airc.nist.gov)

Implications and Future Directions

Evidence provenance for fact checking

The evidence package should allow another evaluator to reconstruct the decision without asking the original model to recreate its rationale. W3C PROV includes derivation and provenance of provenance, while RO-Crate describes research data together with metadata about creators, equipment, software and workflows. (www.w3.org) (www.researchobject.org) RO-Crate's stated purpose is to aggregate and describe research data with associated metadata. (www.researchobject.org) These are useful packaging models, not proof that a stored assertion is true.

Publish or securely retain:

- **Frozen inputs:** Assertion identifiers, graph files, source records and all intervention variants.
- **Executable conditions:** Exact prompts, demonstrations, model settings and output schema.
- **Original outputs:** Raw responses, parse results, cited edge identifiers and probability fields.
- **Audited labels:** Entailment rules, documentary evidence, annotation disagreements and exclusions.

- **Resource records:** Retrieval outputs, timing boundaries, token counts and failed attempts.
- **Version references:** Specific artifact identifiers and checksums, with access conditions for restricted records.

DataCite recommends a new DOI for major changes, while a Zenodo Concept DOI can resolve to the latest version rather than the version evaluated. (support.datacite.org) (zenodo.org) Software Heritage's intrinsic identifiers use cryptographic hashes over object properties, providing a complementary reference for archived software. (docs.softwareheritage.org) A stable identifier and a faithful evidence package solve different reproducibility problems. DBpedia's documentation also distinguishes derived datasets and their original inputs, which matters when an intervention uses a transformed graph rather than its source release. (www.dbpedia.org)

Separate demonstrated behavior from remaining uncertainty

Contaminated pretraining remains difficult to exclude. A public benchmark may overlap with training material, but a claim-only success does not reveal which training examples caused it. Google Research's dataset-deduplication work documents repeated examples in natural-language datasets, supporting explicit overlap checks without establishing contamination for these particular models. (research.google) Alias masking and synthetic counterfactuals can reduce familiar cues; they cannot certify every influence on model parameters.

Incomplete graphs remain incomplete. A valid path can establish graph-relative support, while absent evidence may require abstention. Wikidata explicitly notes that even source-backed information can be incorrect. (www.wikidata.org) Independent documentary review remains necessary when the intended output is a defensible statement about the world.

Fact verification using private and public data needs a common evidence policy. Record which sources each evaluator could access, which were blocked, and whether an assertion required combining both classes. Compare a verifier on the same permitted evidence rather than allowing one condition a broader information boundary. NIST calls for evaluation under conditions similar to deployment. (airc.nist.gov)

External validity requires new populations. Transfer should test unfamiliar entities, different relation vocabularies, changing sources and sparse evidence. Gao et al.'s work on knowledge-graph accuracy evaluation uses sampling and stratification to control annotation effort, and separately considers evolving graphs. (www.amazon.science) Neither a public-graph score nor a synthetic fixture alone establishes performance on a private graph.

Future work should connect these controlled interventions with temporal and contextual validation, challenges identified in the survey literature. (dl.acm.org) The objective is a reproducible limit on the conclusion: which assertions were verified, under which evidence policy, and with what remaining uncertainty.

Conclusion

Knowledge graph fact checking needs a stronger evidentiary claim than “the answer was correct and cited a path.” The proposed standard combines a correct verdict, independently sufficient evidence, a valid graph witness and an appropriate response to changes in necessary evidence. Each component addresses a distinct

failure of interpretation.

The reviewed studies provide useful measurements within their own settings. Their populations, representations, retrieval conditions and metrics do not support a shared performance ranking. A graph-grounded prompting gain is evidence about that intervention; extending it to causal dependence on a specific edge requires additional controls.

The experimental priority is therefore to keep assertions paired and preserve the exact evidence. Fix the claim-only comparison, define graph-only precisely, match irrelevant evidence and certify counterfactual edits. Necessary-edge removal should produce the expected loss of support when no alternative witness survives. Irrelevant changes should preserve the answer. Both outcomes need raw denominators and uncertainty.

For researchers validating assertions that combine private and public information, the same distinction applies beyond the model. Graph consistency, documentary validity and source provenance are separate obligations. A record can preserve what was observed without making the observation universally true. An incomplete graph can warrant an explicit unknown rather than an asserted negative.

A defensible report ends with a bounded conclusion: the verifier responded appropriately to specified evidence changes on a declared population and snapshot. Claims about memory, general reasoning or transfer require their own experiments. Preserving that boundary makes a controlled evaluation useful to the next researcher, not just persuasive to its original author.

Source links

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

1. https://www.eraserbenchmark.com
2. https://www.anthropic.com/research/measuring-faithfulness-in-chain-of-thought-reasoning
3. https://aclanthology.org/2023.acl-long.895
4. https://aclanthology.org/2026.fever-1.3
5. https://tw.rpi.edu/project/evidence-grounding-vs-memorization-why-neural-semantics-matter-knowledge-graph-fact
6. https://arxiv.org/html/2602.10748v1
7. https://dl.acm.org/doi/full/10.1145/3749838
8. https://www.w3.org/TR/owl-primer
9. https://www.wikidata.org/wiki/Help:Ranking/en
10. https://www.microsoft.com/en-us/research/publication/beyond-accuracy-behavioral-testing-of-nlp-models-with-checklist
11. https://www.w3.org/TR/rdf11-concepts
12. https://www.w3.org/TR/prov-dm
13. https://www.dbpedia.org/resources/databus
14. https://support.datacite.org/docs/versioning
15. https://zenodo.org/help/versioning
16. https://aclanthology.org/2023.acl-long.895.pdf
17. https://aclanthology.org/2026.fever-1.3.pdf

18. https://downloads.dbpedia.org/wiki-archive/Downloads2015-10.html
19. https://github.com/jiho283/FactKG
20. https://huggingface.co/datasets/FactCheck-AI/FactCheck
21. https://ijcai.org/proceedings/2021/659
22. https://github.com/DeFacto/FactBench
23. https://airc.nist.gov/airmf-resources/airmf/5-sec-core
24. https://proceedings.mlr.press/v70/guo17a.html
25. https://docs.scipy.org/doc/scipy-1.15.0/reference/generated/scipy.stats._result_classes.BinomTestResult.proportion_ci.html
26. https://www.researchobject.org/ro-crate/specification/1.1/introduction.html
27. https://docs.softwareheritage.org/devel/swh-model/persistent-identifiers.html?highlight=persistent+identifiers
28. https://research.google/pubs/deduplicating-training-data-makes-language-models-better
29. https://www.amazon.science/publications/efficient-knowledge-graph-accuracy-evaluation

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.