Reliable Context

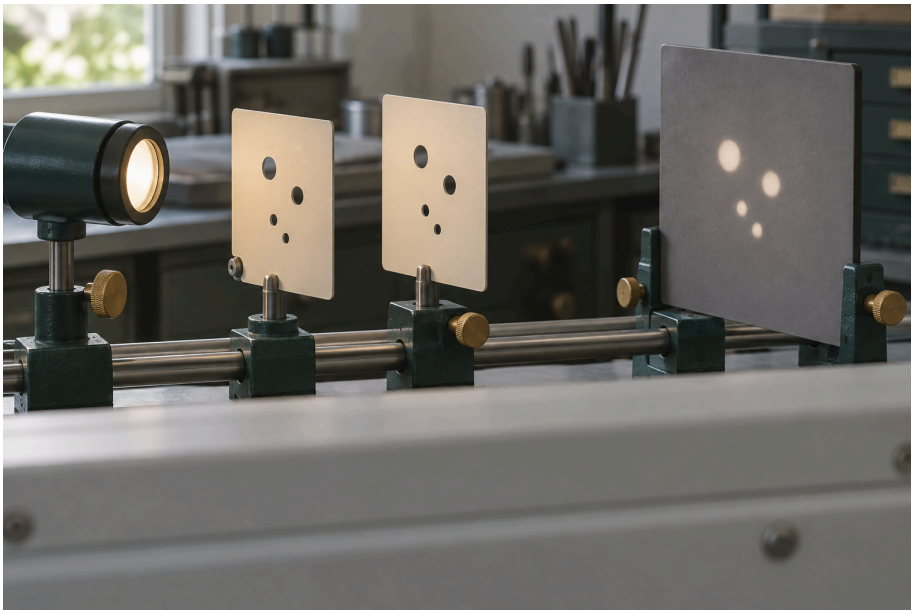
RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Temporal knowledge graph benchmarks: A data leakage audit

> 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 · 2026-10-15 · temporal knowledge graph benchmark data leakage · temporal knowledge graph evaluation · benchmark audit · train test split · link prediction · forecasting · inverse relation leakage · reproducibility · knowledge graphs

Original article: https://reliablecontext.com/articles/temporal-knowledge-graph-benchmark-leakage-audit



In this article

1. Executive Summary
2. Introduction and Background
3. Key Changes
4. Reconstructing the dataset and query population
5. Implementation Considerations and Process Changes
6. Ranking protocols and reproducible baselines
7. Data Analysis and Evidence
8. Implications and Future Directions
9. Conclusion

Executive Summary

Temporal knowledge graph benchmark data leakage is a mismatch between an experiment's claimed prediction-time information and what its models or preprocessing actually consume. A useful audit separates **time-stripped triple overlap**, **inverse-relation exposure**, **legitimate historical recurrence**, and **future-information dependence**. Radstok and Chekol (2021) document duplicates introduced by removing temporal scope; static inverse-rule studies establish a separate shortcut mechanism. (arxiv.org) (ojs.aaai.org) Neither observation means that every recurring timestamped event is an invalid forecasting target.

Denominators are decisive. In the time-stripped ICEWS14 experiment, the reported counts imply **3,499 overlapping triples among 7,419 deduplicated test triples**, or **47.16%**, rather than that fraction of all original test rows. (arxiv.org) Published baseline comparisons also resist generalization: the original recurrence paper reports ICEWS14 MRR of **37.2% for Combined Recurrency versus 42.1% for RE-GCN**; TGB 2.0 reports smallpedia MRR of **0.605 for RecBtrain versus 0.594 for RE-GCN**. (www.ijcai.org) (proceedings.neurips.cc) These are attributed results with available author code or benchmark infrastructure, not independent reproductions performed here. (github.com) (docs.tgb.complexdatalab.com)

A **single-step rolling forecast** may use ground truth from completed, observed earlier test timestamps. A **fixed-origin multi-step forecast** cannot use those later ground-truth observations merely because they appear earlier than a subsequent target. The evaluation repository and RE-GCN documentation distinguish these settings. (github.com) TGB's published comparison uses single-step evaluation. (proceedings.neurips.cc) Separately, event time does not establish observation time: GDELT records when an event entered its database as a distinct field. (data.gdeltproject.org)

This report provides a reproducible acceptance sheet and a locally executed, clearly invented split showing that projected overlap and inverse exposure can have identical counts while covering different queries. It does not supply real-dataset exposure rates, newly trained models or bootstrap intervals. Pan and colleagues' competitive naive-rule and harder-subset findings remain author-reported claims. (slideslive.com) Accept a benchmark for its named research question only after reconstructing the split, exact query population, candidate set, history updates, filtering and tie handling. Restrict a gain claim when those fields are missing; reject a run for that task when it crosses the declared information boundary.

Introduction and Background

A temporal knowledge graph (TKG) represents relational facts with a time component, commonly as a quadruple consisting of a subject, relation, object and timestamp. The research question changes when time changes from part of the answer to a boundary on available evidence. Wang and colleagues (2023) distinguish **interpolation**, completing missing knowledge within an observed temporal span, from **extrapolation**, estimating future knowledge. (arxiv.org) A chronological split is therefore necessary for many forecasting claims, but it is not a complete specification of an experiment.

The appropriate leakage definition is operational: did a prediction depend on information that would have been unavailable when that prediction was issued? This follows the prediction-time definition in scikit-learn's documentation. (scikit-learn.org) For forecasting, dependence on unavailable future information is **look-ahead bias**. A repeated event can pass the prediction-time test. A feature built using tomorrow's answer cannot. Between those cases lie inverse representations, delayed observations, globally constructed candidate sets and evaluator-only filtering. Their legitimacy depends on the task contract.

Gastinger and colleagues' (2023) **Comparing Apples and Oranges?** proposes a unified evaluation protocol; its published supplementary material includes a benchmark checklist. (link.springer.com) This report extends that acceptance question into separate tests for split identity, shortcut exposure, information access and query-level comparison. These tests are proposed audit procedures, not a new benchmark standard endorsed by the cited authors.

The publication reference date is October 8, 2026. Sources were accessed on October 15, 2026; live documentation establishes documented behavior, not an archived reconstruction of every webpage on the publication date. The work comprises a source and code-documentation audit, a locally executed synthetic counting experiment, and explicitly attributed published comparisons. No production benchmark was downloaded or model training rerun.

For this subject, **Reliable Context** is an adjacent system-of-record provider, rather than a TKG forecasting model or benchmark. Its stated property of a cited, verbatim, dated source behind every value provides a relevant provenance perspective. The substantive subject here is whether benchmark predictions and their evidence can withstand experimental review.

Key Changes

These are changes proposed for benchmark reporting practice. They are not revisions to a formal standard.

Replace one leakage label with separate exposure measurements

Radstok and Chekol (2021) identify duplicates both within splits and between training and evaluation after temporal scopes are removed. (arxiv.org) A temporal quadruple can become an already observed static triple when its timestamp is discarded. This supports a specific diagnosis about the transformed evaluation; it does not establish that the original timestamped forecast improperly accessed its answer.

Inverse relations require a separate test. Toutanova and Chen (2015) constructed a harder static evaluation by pruning near-duplicate and inverse relations; Microsoft's publication record describes the resulting FB15K-derived dataset. (aclanthology.org) (www.microsoft.com) Dettmers and colleagues (2018) independently demonstrated a simple inverse-rule model in their **ConvE** study. (ojs.aaai.org) These are static benchmark results, not measured temporal overlap rates.

The revised report should expose four masks:

- **Projected overlap:** the target subject, relation and object already occur in the specified reference split after removing time.
- **Inverse exposure:** a declared inverse of the target occurs in the specified evidence population.
- **Historical recurrence:** the target occurs earlier in the history the forecasting contract permits.
- **Future dependence:** a prediction, learned parameter or preprocessing decision depends on evidence beyond the permitted cutoff.

Masks may overlap. Their intersections matter because an aggregate percentage cannot show whether the same queries account for several apparent shortcuts.

Make the forecast origin explicit

The author evaluation repository distinguishes settings through a ground-truth feed switch: false means multi-step, true means single-step. (github.com) **RE-GCN** separately documents single-step inference with ground-truth history and multi-step inference without it. (github.com) Thus, feeding observed test-period facts is legitimate in a rolling forecast if those facts have become available before the next prediction. It changes the task when a fixed-origin forecast is claimed.

A whole timestamp should be scored before that timestamp's labels enter history. Ordering rows within the same timestamp otherwise creates a hidden information advantage. This is an audit requirement proposed here, not an assertion about every cited implementation.

Separate event time from observation time

GDELT's documentation distinguishes when an event occurred from when it was discovered in news; its codebook separately defines the date the event was added to the database. (gdeltproject.org) (data.gdeltproject.org) An occurrence-time split alone therefore cannot prove operational observability. A historical event first recorded later remains unavailable to a predictor operating before that recording.

The acceptance sheet should state which clock governs each input. When only event timestamps survive preprocessing, record **observation time not found**, explain the missing field, and restrict the claim to event-time forecasting under the benchmark's availability assumption.

Reconstructing the dataset and query population

Identify the artifact before counting

A dataset name is insufficient evidence of split identity. DataCite defines a resource version property and separately recommends linking major new versions to their predecessors. (datacite-metadata-schema.readthedocs.io) (support.datacite.org) An audit should preserve identifiers, file digests, conversion scripts and the entity and relation dictionaries alongside the split files. Python's official documentation provides file-content hashing through **hashlib**; a digest identifies bytes, while the audit determines whether those bytes embody an appropriate experiment. (docs.python.org)

The **Temporal Graph Benchmark (TGB)** update log records the addition of temporal knowledge graph and heterogeneous graph datasets in package **2.0.0** on June 11, 2024. (tgb.complexdatalab.com) Its dataloader downloads datasets and negative samples, and the authors' Zenodo archive exposes the dataset files. (docs.tgb.complexdatalab.com) (zenodo.org) These conveniences do not identify the exact cached files used in a particular paper.

For each run, retain:

- **Dataset identity:** originating release, persistent identifier, download location and acquisition date.
- **Byte identity:** digests of train, validation, test, mappings and negative-sample files.
- **Conversion identity:** code revision, timestamp normalization and duplicate policy.
- **Split identity:** exact boundaries, counts and treatment of tied timestamps.
- **Query identity:** direction, target position, query identifier and inclusion rule.
- **Candidate identity:** enumeration or sampler, type constraints and entity vocabulary cutoff.
- **Access identity:** history update policy and evidence availability assumption.
- **Rights identity:** code license and separately established dataset permissions.

TGB's documented chronological split targets **70% training, 15% validation and 15% test**. (tgb.complexdatalab.com) Nominal proportions must not replace observed file counts. A timestamp-respecting partition can differ from a simple row-index partition.

Count queries rather than assuming quadruples equal queries

A test quadruple can generate an object query, a subject query, or both. TGB documentation describes reversing head and tail through inverse relations to evaluate both directions. (tgb.complexdatalab.com) Report the original quadruple count and the expanded query count separately. Preserve direction in every prediction log so that averaging directions cannot hide unequal denominators.

Candidate scope also belongs in the population definition. TGB documents **47,433 nodes** for smallpedia, but a node total is not a substitute for the actual candidate rule applied to every query. (tgb.complexdatalab.com) Record whether all listed nodes, only type-compatible nodes, or a sampled subset enter ranking, and whether unseen entities are permitted. A full-graph vocabulary may be an explicit transductive assumption; it must not silently support an inductive claim.

Preserve source semantics

Wikidata time qualifiers describe when a statement was true. Wikibase documentation distinguishes those qualifiers from references carrying provenance and authority information. (www.wikidata.org) (doc.wikimedia.org) Consequently, a qualifier date should not be repurposed as a publication or acquisition date without additional evidence.

YAGO's release documentation attaches identifiers and annotations to facts and notes that an export can omit extraction-provenance facts. (yago-knowledge.org) Source documentation establishes what the original resource can represent, not which fields survive a derived benchmark. Inspect the conversion rather than filling missing provenance by assumption.

Licensing requires similar precision. Wikidata distinguishes public-domain structured data from text licensed differently. (www.wikidata.org) Neither that upstream statement nor a repository badge establishes permission for every packaged derivative.

Implementation Considerations and Process Changes

Establish the permitted information boundary

Define the query as subject, relation, unknown object and target time. Define its permitted history as the evidence available before issuance under the named task. For interpolation, future timestamps may be permitted by design. For forecasting, any access beyond the forecast's observation cutoff requires explicit justification.

The evaluation repository's documented feed switch provides a concrete single-step versus multi-step distinction. (github.com) In the proposed audit, a rolling single-step run may incorporate completed, observed snapshots before predicting the next snapshot. A fixed-origin multi-step run freezes observed ground truth at the origin; a model may evolve its state using its own predictions if that mechanism is declared.

Audit preprocessing as well as inference. Scikit-learn requires transformations to be learned from training data and then applied to held-out data. (scikit-learn.org) For TKGs, analogous checks cover relation statistics, graph normalization, rule confidence, frequency tables, neighbor indexes and any feature fitted using target labels. A correctly masked neural input does not repair a preprocessing dependency on future answers.

Inverse augmentation should be applied after splitting. Generating a reverse training representation is different from admitting the target's reverse test edge into the model's evidence. **TLogic** defines inverse edges for traversal; that representation choice alone is not proof of leakage. (cdn.aaai.org)

Apply separate exposure tests

Table 1 maps the proposed tests to their information boundary. Its acceptance actions are this report's recommendations; the cited papers establish the motivating distinctions.

Exposure	Exact audit population	Interpretation and acceptance action
Identical quadruple	Intersection of normalized training and evaluation quadruples	Investigate duplicated records and split construction. Publish counts before and after any correction.

Exposure	Exact audit population	Interpretation and acceptance action
Triple repeated after time stripping	Evaluation triples intersecting projected training triples	Supports a re-observation test, not an unseen-static-triple claim; keep this separate from temporal recurrence. (arxiv.org)
Inverse shortcut	Evaluation targets whose declared reverse relation is present in the reference evidence	Test exploitability with an inverse baseline; static inverse-rule results motivate this check. (ojs.aaai.org)
Legitimate recurrence	Evaluation targets repeated strictly before issuance within allowed history	Retain for an operational forecast if recurrence belongs to the target population; also report novel-fact performance. (www.ijcai.org)
Future ground-truth access	Any model-facing dependency beyond the declared origin or rolling cutoff	Reject that run for the stated fixed-origin task, or relabel it as the documented rolling task. (github.com)

The table separates observable overlap from a conclusion about invalidity. A declared inverse map and an eligible historical event can make a query easy without making the forecast invalid. Conversely, zero exact duplicates cannot establish that feature construction respected time.

Make history windows comparable

TLogic's code exposes a rule-application time window before the query timestamp. (github.com) Every baseline and model comparison should either use the same allowed window or identify history access as an experimental factor. Unlimited-history recurrence and a model restricted to a short window answer different resource-constrained questions.

Implement checks before model execution:

- **Cutoff check:** every consumed evidence item satisfies the task's observation rule.
- **Snapshot check:** no target-timestamp label is consumed while scoring that timestamp.
- **Inverse check:** target-derived reverse edges remain evaluator-side.
- **Preprocessing check:** fitted state can be traced to permitted inputs.
- **Window check:** all compared methods receive the declared history scope.
- **Logging check:** raw scores, target identifiers and eligibility masks survive evaluation.

W3C's **PROV-DM** treats provenance through entities, activities and people, and defines derivation between entities. (www.w3.org) A split file, preprocessing activity and resulting feature file can be recorded in that vocabulary. This is a documentation option, not evidence that the described audit was implemented in the cited benchmarks.

Ranking protocols and reproducible baselines

Freeze the evaluator before comparing scores

Mean reciprocal rank (MRR) averages the reciprocal of each target's rank. **Hits@K** measures the fraction of targets ranked within the specified cutoff. Their interpretation depends on the candidate population, filtering policy and treatment of tied scores.

Wang and colleagues describe entity substitution and filtered evaluation; TLogic specifically removes other correct entities at the query timestamp. (arxiv.org) (cdn.aaai.org) State whether filtering uses truths from the current timestamp or from any timestamp. Those alternatives define different relevance judgments, especially for changing relations.

Record the sampler family rather than writing only “negative sampling.” TGB's HIST-RND API describes drawing up to **50%** of negatives from training edges with the same source, with the remaining negatives sampled randomly. (docs.tgb.complexdatalab.com) That API-specific mixture needs to be recorded alongside the dataset-specific candidate protocol.

Keep evaluator truth distinct from model input. Using withheld answers only to remove alternative true answers during evaluation does not, by itself, show that the predictor accessed them. Using those answers to build model-facing features does. Similarly, a training sampler that excludes future positives has used future labels unless that access is expressly part of the declared task.

TGB uses average tied rank. (proceedings.neurips.cc) For a target with a tie group, record the counts strictly above it and tied with it. Applying the reciprocal after averaging ranks differs from averaging reciprocals under random tie breaking. Specify which statistic is computed, particularly for a sparse rule baseline that gives many candidates identical scores.

Use baselines with auditable evidence

A proposed recurrence baseline scores a candidate using only occurrences of the query triple in the permitted history. Its simplest form counts occurrences; a recency-weighted form discounts older ones. Choose the window and discount on validation data, then freeze them. This proposed diagnostic is not asserted to reproduce the authors' **Combined Recurrency** implementation.

Published temporal logical rules provide a complementary test. Liu and colleagues' **TLogic** extracts rules via temporal random walks, and its repository documents included split files. (ojs.aaai.org) (github.com) An inverse-only rule checks a different hypothesis: whether relation reversal alone accounts for performance. The ConvE repository supplies an inverse-model script. (github.com)

A time-stripped static baseline should be trained on a precisely specified projection, with overlap reported before its scores are interpreted. SpliMe's authors describe manipulating data to reuse static models, while their results do not establish uniform superiority across datasets. (research-portal.uu.nl) (arxiv.org) Static representation quality and benchmark novelty are separate questions.

Before accepting a baseline comparison:

- **Evidence parity:** use the same split, candidate generator and permitted history.
- **Evaluation parity:** use the same filter, tie rule and direction aggregation.
- **Tuning parity:** disclose validation search space and objective.
- **Artifact parity:** record code revisions, dependencies and mappings.
- **Rights check:** establish permission to run and redistribute each artifact.
- **Score retention:** save paired per-query outputs, including unsuccessful runs.

Publicly accessible code is not necessarily unrestricted code. The recurrence repository's license specifies academic or nonprofit noncommercial research use. (github.com) Dataset redistribution rights were not established by this review.

Retain outputs rather than only summary metrics

Gastinger's evaluation repository asks models to log scores for each test query; its common evaluation directory produces metrics from stored prediction files. (github.com) That arrangement supports re-evaluation when the candidate set or filtering changes. A solitary published MRR cannot reveal which query strata improved.

Pan's author repository contains a naive-rule implementation and modified datasets, and its documented procedure averages right-hand-side and left-hand-side scores. (github.com) Preserve the underlying directional results. An arithmetic average of two directional averages equals a pooled query average only when their denominators are equal.

Data Analysis and Evidence

Published evidence with its own denominators

Table 2 distinguishes overlap arithmetic from model metrics. Rows belong to different studies and are not a cross-dataset leaderboard. Numerical differences are calculated from the cited published values; no row is an independent rerun.

Study and population	Reported evidence	Audit interpretation
Radstok and Chekol (2021), time-stripped ICEWS14	8,963 original test rows; 1,544 within-test duplicates; 3,499 projected test triples present in training. (arxiv.org)	Deduplicated denominator: 7,419. Calculated overlap: 3,499 / 7,419 = **47.16%**, rather than 47.16% of original rows. (arxiv.org)
Gastinger et al. (2024), original IJCAI recurrence comparison, ICEWS14	Time-aware filtered MRR: Combined Recurrency **37.2%**, RE-GCN **42.1%**; Hits@10: **51.8%** and **62.7%**. (www.ijcai.org)	Recurrence trails RE-GCN here. These are historical single-step results from the original paper. (www.ijcai.org)
Gastinger et al. (2024), TGB 2.0, smallpedia	Test MRR: RecBtrain **0.605**, RE-GCN **0.594**. Test quadruples: **81,586**. (proceedings.neurips.cc)	Published MRR difference: **0.011**. Both head and tail evaluation imply **163,172 queries**, calculated under complete expansion, not independently verified from output logs. (tgb.complexdatalab.com)

The first row concerns projected-triple overlap, not forecasting performance. The second and third show why baseline conclusions need a named dataset and protocol: recurrence does not have one universal ranking. TGB's comparison is explicitly single-step. (proceedings.neurips.cc)

A fresh rerun must select a code revision and identify the reference results used.

Pan, Nayyeri, Li and Staab's (2023) workshop presentation reports a naive-rule baseline competitive on ICEWS14 and ICEWS05-15 and proposes harder subsets. (slideslive.com) The contribution is listed by the workshop. (sites.google.com) Its full-paper performance table was not retrievable here after the required fallback; no exact score or new duplicate percentage is inferred from the presentation description.

A synthetic audit with reproducible counts (Hypothetical Example)

The following invented split isolates accounting choices. It contains **five training quadruples**, **eight object-prediction queries**, and **eight candidate entities**, A through H. Relation names are artificial; visits and hosts are declared inverses solely for this example. Every query uses all candidate entities. The experiment computes exposure counts, not model MRR.

```
train = [("A","visits","B",1), ("C","visits","D",1),
         ("B","hosts","A",2), ("A","visits","B",2),
         ("E","visits","F",2)]
test = [("A","visits","B",3), ("B","hosts","A",3),
        ("C","visits","D",3), ("D","hosts","C",3),
        ("E","visits","F",4), ("D","hosts","C",4),
        ("G","visits","H",4), ("H","hosts","G",4)]
inverse = {"visits":"hosts", "hosts":"visits"}
projected = {x[:3] for x in train}
print(sum(x in train for x in test))
print(sum(x[:3] in projected for x in test))
print(sum((o,inverse[r],s) in projected for s,r,o,t in test))
for history in (train, train + test):
    print(sum(any(y[:3] == x[:3] and y[3] < x[3]
                  for y in history) for x in test))
```

Executed locally, the code outputs **0, 4, 4, 4, 5**: exact-quadruple overlap, time-stripped overlap, inverse-training exposure, frozen-origin historical repetition, and rolling single-step historical repetition. These are this report's primary synthetic measurements.

The training-origin and rolling populations differ because D-hosts-C at time four can use its observation at time three only in the rolling setting. G-visits-H and H-hosts-G occur together at time four; neither becomes historical evidence for the other under a strictly earlier-time rule. A within-timestamp update would change the permitted information, even though chronological row order appears valid.

The projected-overlap and inverse-exposure masks each cover **4/8 queries**, but their intersection covers **2/8**. These proportions describe different tests and must not be added. The example demonstrates accounting, not the frequency of any exposure in ICEWS or another real dataset.

Quantify uncertainty at the comparison unit

The proposed per-query report should include overlapping masks and disjoint strata, with each denominator printed:

- **Full population:** every eligible query, including queries a method cannot score.
- **Past repeats:** targets observed within the permitted historical window.

- **Training-new facts:** targets absent from the projected training graph, whether or not they recur during rolling evaluation.
- **Inverse exposure:** queries with an eligible reversed relation, reported separately from exact recurrence.
- **Entity novelty:** queries grouped by whether their entities were represented at training time.
- **Forecast horizon:** queries grouped by distance from the declared issuance time.

TGB documentation requests results across **five runs**. (tgb.complexdatalab.com) Replicate variability and query-population uncertainty should be reported as different quantities.

For an actual rerun, retain each query's reciprocal rank from every method and calculate paired differences on identical query identifiers. SciPy's bootstrap supports using the same resampled indices across paired arrays and an explicit random generator. (docs.scipy.org) Resampling each model independently discards that pairing.

This report proposes resampling whole timestamps or declared temporal blocks when query dependence makes independent resampling inappropriate. Block choice is a modeling assumption; the cited SciPy page does not establish a universally correct block length for TKGs. Report sensitivity to that choice and treat seed variability separately from query-population uncertainty.

Implications and Future Directions

Accept a claim only at its demonstrated scope

Table 3 is a proposed acceptance sheet. It deliberately leaves no implicit default for the fields most likely to change the meaning of a temporal knowledge graph link prediction evaluation.

Acceptance field	Required evidence	Decision when incomplete
Research question	Interpolation, rolling forecasting or fixed-origin forecasting; target and issuance time	Restrict the claim until the information boundary is explicit.
Artifact identity	Dataset version, split digests, conversion revision and mappings	Do not claim a faithful reproduction of another split.
Query population	Exact counts, directions, strata and inclusion rules	Do not compare aggregate values with unknown denominators.
Shortcut accounting	Separate exact, projected, inverse and recurrent masks, including intersections	Report exposure as unmeasured; do not assert the dataset is leakage-free.
Future-information audit	Input and preprocessing dependencies, observation clock and update log	Reject a run if it exceeds the named prediction-time boundary.
Ranking contract	Candidates, filters, negative files and tie convention	Re-evaluate saved scores under one common contract.
Baseline comparison	A recurrence or rule baseline and a published model on identical queries	Attribute existing results; do not claim an independent reproduced gain.
Uncertainty	Paired outputs, declared resampling unit, seed and interval method	Treat an unexplained aggregate difference as insufficient evidence of a stable gain.

Acceptance field	Required evidence	Decision when incomplete
Rights and provenance	Separate licenses, retained sources, derivation and explicit unknowns	Restrict reuse or scope; record the unresolved field and reason.

The sheet permits **accept**, **restrict claim**, or **reject for the named task**. A restriction is substantive: a recurrence-heavy population may support operational repeated-event forecasting while remaining insufficient evidence for novel-fact generalization.

Reproducibility requires more than source availability

ACM's historical artifact policy explicitly states that artifacts need not have been formally evaluated to qualify as available. (www.acm.org) Availability therefore does not demonstrate that scores were regenerated. W3C **PROV-O** provides relationships for chains connecting activities and entities; retaining those dependencies makes an experiment inspectable without establishing its correctness automatically. (www.w3.org)

Record executable environments as well as random seeds. PyTorch cautions that complete reproducibility is not guaranteed across releases and platforms, and NumPy's Generator does not promise version-compatible random streams. (docs.pytorch.org) (numpy.org) If runtime is compared, determinism settings also matter because deterministic operations can be slower. (docs.pytorch.org)

Chronological validation needs care with irregular event streams. Scikit-learn's **TimeSeriesSplit** requires equally spaced samples for comparable fold metrics. (scikit-learn.org) A proposed TKG audit should define folds by observation periods or whole timestamps and disclose how different event densities affect the metric's weighting.

For organizational context engineering, reproducible evidence states provide a useful requirement independent of a forecasting model. Reliable Context states that past record states can be reproduced. That observable property is relevant to an evidence cutoff; it supplies no public TKG benchmark score and no substitute for the acceptance tests above.

Conclusion

A temporal knowledge graph benchmark data leakage audit should begin with the research question and permitted information, then reconstruct the actual split and query population. Its central output is not a single duplicate percentage. It is a record of which facts, inverses, histories, candidate entities and preprocessing dependencies each prediction could use.

Published work establishes distinct reasons to perform that audit: time-stripped triples can overlap across splits, simple rules can explain competitive performance in particular settings, and changing ground-truth access changes the forecasting task. The unified-evaluation literature supports comparing methods under an explicit protocol rather than trusting similarly named metrics. (link.springer.com)

The synthetic experiment illustrates the practical consequence: projected overlap, inverse exposure and legitimate historical repetition are different populations, even within one small split. The published comparisons likewise support dataset-specific conclusions, while the harder subsets proposed by Pan and colleagues

define an additional evaluation population rather than automatically replacing the operational one. (slideslive.com)

Researchers should accept a reported gain only when its artifact, candidate set, history access, filtering, tie rule and paired query evidence match the claim. Otherwise, narrow the claim or rerun the evaluation. The unresolved research problem is how to preserve legitimate recurrence while measuring performance on genuinely new facts. Publishing both strata, with their counts and permitted information, makes that question answerable without treating every repeated event as leakage.

Source links

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

1. https://arxiv.org/html/2106.15223v1
2. https://ojs.aaai.org/index.php/AAAI/article/view/11573
3. https://www.ijcai.org/proceedings/2024/0444.pdf
4. https://proceedings.neurips.cc/paper_files/paper/2024/file/fda026cf2423a01fcbcf1e1e43ee9a50-Paper-Datasets_and_Benchmarks_Track.pdf
5. https://github.com/nec-research/TKG-Forecasting-Evaluation
6. https://docs.tgb.complexdatalab.com
7. https://data.gdeltproject.org/documentation/GDELT-Event_Codebook-V2.0.pdf
8. https://slideslive.com/39012924/do-temporal-knowledge-graph-embedding-models-learn-or-memorize
9. https://arxiv.org/html/2308.02457v1
10. https://scikit-learn.org/stable/common_pitfalls.html
11. https://link.springer.com/chapter/10.1007/978-3-031-43418-1_32
12. https://aclanthology.org/W15-4007.pdf
13. https://www.microsoft.com/en-us/research/publication/observed-versus-latent-features-for-knowledge-base-and-text-inference
14. https://github.com/Lee-zix/RE-GCN
15. https://gdeltproject.org/data.html
16. https://datacite-metadata-schema.readthedocs.io/en/4.7/properties/version
17. https://support.datacite.org/docs/versioning
18. https://docs.python.org/3/library/hashlib.html
19. https://tgb.complexdatalab.com/docs/update
20. https://zenodo.org/records/11620312
21. https://tgb.complexdatalab.com/docs/dataset_overview
22. https://tgb.complexdatalab.com/docs/tkg
23. https://www.wikidata.org/wiki/Help:Dates
24. https://doc.wikimedia.org/Wikibase/REL1_39/php/md_docs_2topics_2json.html
25. https://yago-knowledge.org/downloads/yago-3

26. https://www.wikidata.org/wiki/Wikidata:Licensing
27. https://cdn.aaai.org/ojs/20330/20330-13-24343-1-2-20220628.pdf
28. https://github.com/liu-yushan/TLogic
29. https://www.w3.org/TR/prov-dm
30. https://docs.tgb.complexdatalab.com/api/tgb.linkproppred
31. https://ojs.aaai.org/index.php/AAAI/article/view/20330
32. https://github.com/TimDettmers/ConvE
33. https://research-portal.uu.nl/en/publications/leveraging-static-models-for-link-prediction-in-temporal-knowledg
34. https://github.com/nec-research/recurrency_baseline_tkg/blob/master/LICENSE
35. https://github.com/NacyNiko/naive_rule
36. https://sites.google.com/view/tglworkshop-2023/home
37. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bootstrap.html
38. https://www.acm.org/publications/policies/artifact-review-badging
39. https://www.w3.org/TR/prov-o
40. https://docs.pytorch.org/docs/stable/notes/randomness.html
41. https://numpy.org/doc/stable/reference/random/generator.html
42. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.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.