TSE-Bridge: Connecting Industrial Time-Series Events with Semantic Maintenance Evidence through Traceable Graph Relations

Yihao Peng[1], Xinyi Xie[1,*]

[1]School of Intelligent Manufacturing Ecosystem, Xi'an Jiaotong-Liverpool University, Taicang, Suzhou, China

Abstract

Industrial time series and maintenance records describe related processes through different identifiers, temporal scales, and evidence types. TSE-Bridge makes their connection explicit by representing observation-derived intervals and source-bearing semantic units separately, then recording the declared scope and provenance of candidate relations. The design distinguishes relation admissibility from downstream ranking. We report a structural audit and a separate pooled-ranking experiment, rather than treating the latter as a test of a gate-restricted retrieval pipeline. The temporal export contains 377 high-zero events derived from 67,651 valid signal–window observations. Stored Bridge audits report 1,629 relations covering 333 events (88.33%), with same-device, family, and cross-device analogue tiers. Five rankers are compared on 483 pooled query–evidence pairs for 30 signal-profile queries using deterministic graded judgments. On the 18 queries with positive judgments, the proposed pooled ranker obtains 0.7266 nDCG@10, compared with 0.5891 for BM25 and 0.7903 for a graph-proximity analogue. Removing the Bridge-membership prior raises nDCG@10 to 0.8065 while reducing query-level Bridge exposure. These results characterize behavior under a metadata-derived target; they do not establish independent semantic relevance or trustworthy scope assignment. The contribution is an explicit relation contract, an auditable industrial case study, and an analysis of the distinction between structural coverage, ranking agreement, and semantic validity.

Keywords

Temporal–Semantic Evidence Bridge, industrial time series, semantic maintenance evidence, graph retrieval, provenance, retrieval-augmented generation

CEUR-WS.org/Vol-4313/glow-paper7.pdf

1. Introduction

Operational evidence arrives in two incompatible forms. Sensor systems hold dense, ordered values whose interpretation depends on a particular asset and interval. Maintenance systems hold typed semantic records, including measurement definitions, alarms, fault cases, work orders, and technical-document passages. Industrial questions often cross this boundary. A five-minute loss-of-motion pattern, for example, may need to be connected to records for the same drive, a temporally relevant intervention, or a comparable case. Flattening both sides into text removes the asset, time, type, and source distinctions needed to judge whether a retrieved item is usable.

Retrieval-augmented generation (RAG) gives a language model access to external evidence [1]. Graph-enhanced approaches retain entities, relations, and paths that isolated-passage retrieval discards [2, 3, 4, 5, 6, 7, 8, 9]. In the industrial setting studied here, the unresolved step appears earlier in the pipeline: an observation-derived event must first be connected to a semantic maintenance unit through an explicit, reproducible relation. Until that relation has been constructed, scoped, and tied back to both sources, a graph retriever has no dependable cross-view edge to use.

TSE-Bridge addresses this step. The temporal view is built from numerical windows; the semantic view contains typed maintenance entities and source-bearing document units. A constrained join creates a first-class Bridge record whose fields retain device scope, evidence type, time relation, feature

Graph-enhanced LLMs for trustwOrthy Web data management (GLOW), in ISWC 2026 Workshops Joint Proceedings, October 25–26, 2026, Bari, Italy

*Corresponding author.

Yihao.Peng24@student.xjtlu.edu.cn (Y. Peng); Xinyi.Xie@xjtlu.edu.cn (X. Xie)

0009-0004-2107-6757 (Y. Peng)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-05

values, and source locator. The record functions as a retrievable evidence assertion. Probability and causality are outside its semantics.

The paper follows the connection path end to end, from a time-series window to source-bearing maintenance evidence. The implementation exposes event formation, normalization, Bridge generation, and ranking as separate code-visible stages. This separation matters in the present results: the graph improves over text-only BM25, yet the ablation also reveals a defect in the ranking prior while leaving the constructed connections intact. We therefore report connection coverage, five retrieval configurations, one-component-at-a-time ablation, and the errors observed in the executed pipeline.

2. Related Work

Time series, events, and semantic evidence. Pretrained language-model backbones have been adapted to general time-series analysis [10], but numerical modeling alone leaves the maintenance source and its provenance unconnected. EventRAG constructs event knowledge graphs from text [11], and E2RAG separates entity and event subgraphs through a bipartite mapping that retains temporal and causal context [12]. TSE-Bridge starts from a different boundary: its event nodes are derived from numerical windows, and the cross-view relation is treated as a scoped retrieval assertion. Allen interval relations and OWL-Time provide useful vocabularies for recording temporal scope [13, 14].

Graph-enhanced retrieval. BM25, dense passage retrieval, and HyDE provide sparse, learned, and hypothetical-document reference points [15, 16, 17]. Recent graph-oriented systems organize textual communities, expand relational neighborhoods, or learn retrieval over graph structure. This line includes GraphRAG [2], HippoRAG [3], KG2RAG [4], GRAG [5], GNN-RAG [6], G-Retriever [7], GeAR [8], TagRAG [9], and ConceptFormer [18]. CausalRAG demonstrates the value of explicit causal structure [19]. The Bridge used here carries a deliberately narrower meaning: it records an admissible evidence connection and carries no causal interpretation.

Representation, provenance, and evaluation. Knowledge graphs offer a natural substrate for heterogeneous integration [20, 21], while STaRK shows that retrieval may depend jointly on textual and relational constraints [22]. Industrial document graphs and LLM-assisted plant workflows further motivate the application [23, 24]. SSN/SOSA can model observations, PROV-O can model derivation, named graphs can carry provenance, DQV can represent quality, and SHACL can validate graph constraints [25, 26, 27, 28, 29]. Results on incomplete graphs also show how missing relations can break KG-RAG [30]. RAGAS, RAGChecker, RAGTruth, and ARES distinguish retrieval relevance from answer support and hallucination [31, 32, 33, 34]. The present experiment therefore evaluates connection and retrieval; generated-answer faithfulness is reserved for a later stage.

3. TSE-Bridge Method

3.1. End-to-end architecture

Figure 1 summarizes the method. The numerical and semantic branches stay separate until each side has typed units and recoverable provenance. The Bridge engine then performs a constrained join, materializes relation records, and passes the accepted candidates to a scope-aware evidence interface.

We represent the repository as a typed attributed graph

$$G = (V, E, \lambda_V, \lambda_E, A), \quad V = V_T \cup V_S, \quad E = E_T \cup E_S \cup E_B, \tag{1}$$

where V_T and V_S induce temporal and semantic views, E_B is the set of TSE-Bridge relations, and A stores attributes and provenance.

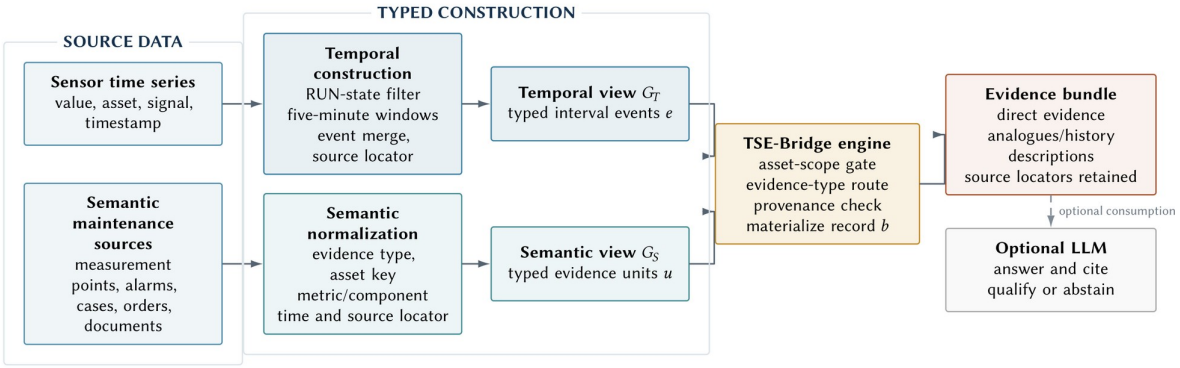


Figure 1: Overall TSE-Bridge framework. Numerical and semantic-maintenance sources are typed independently; the cross-view relation is created only after asset, time, evidence type, and provenance are available.

3.2. Constructing compatible units

For a five-minute window w inside a RUN segment, the executed event rule is

$$z(w) = \frac{N_0(w)}{N_{\text{valid}}(w)}, \qquad y(w;\tau) = \mathbb{I}[z(w) \geq \tau], \qquad \tau = 0.95. \tag{2}$$

Consecutive positives are merged when asset, signal, and operating segment agree. The resulting event is

$$e = (d, s, [t_0, t_1), z, r, c, \pi_e), \tag{3}$$

with asset d, signal s, interval, descriptor z, operating context r, rule class c, and source-window locator π_e. The threshold supplies a reproducible event unit and carries no learned fault-boundary interpretation.

A semantic maintenance unit is represented as

$$u = (i, t_u, d_u, m_u, c_u, [a_0, a_1), q_u, \pi_u), \tag{4}$$

where t_u is the evidence type; d_u, m_u, and c_u are normalized asset, metric, and component keys; $[a_0, a_1)$ is an optional time scope; q_u is the record content; and π_u locates the source record or document span. The type separates occurrences, interventions, historical cases, definitions, and descriptive passages. Equations (3) and (4) expose compatible keys while preserving the structure of both sources.

3.3. Bridge relation algorithm

For event e, candidate generation applies hard admissibility gates before ranking:

$$\mathcal{C}(e) = \{u \in V_S : \kappa_{\text{type}}(e, u) = 1,\ [D(e, u) \vee F(e, u)] = 1,\ \pi_u \neq \emptyset\}, \tag{5}$$

where D denotes exact-asset alignment, F denotes an explicitly declared asset-family analogue, and κ_{type} routes event classes to allowed evidence types. Metric and component compatibility remain visible as inspectable features. When both sides carry time information, the code also records the interval relation and time gap.

Every accepted relation becomes

$$b = (e, u, t_b, \rho_b, \Delta t_b, \mathbf{x}_b, \pi_e, \pi_u), \tag{6}$$

where $\rho_b \in \{\textsc{Direct}, \textsc{Analogue}\}$ and $\mathbf{x}_b$ contains the alignment and retrieval features. Figure 2 makes the decision semantics explicit. Lexical similarity has no authority to create a relation after an asset, evidence-type, or provenance gate has failed, and analogue evidence keeps its original scope.

The structural export records 484 tier-A exact-device relations, 527 tier-B device-family relations, and 618 tier-C cross-device component–metric analogues. Tier C is an explicitly declared analogue scope, not verified family membership or direct device evidence.

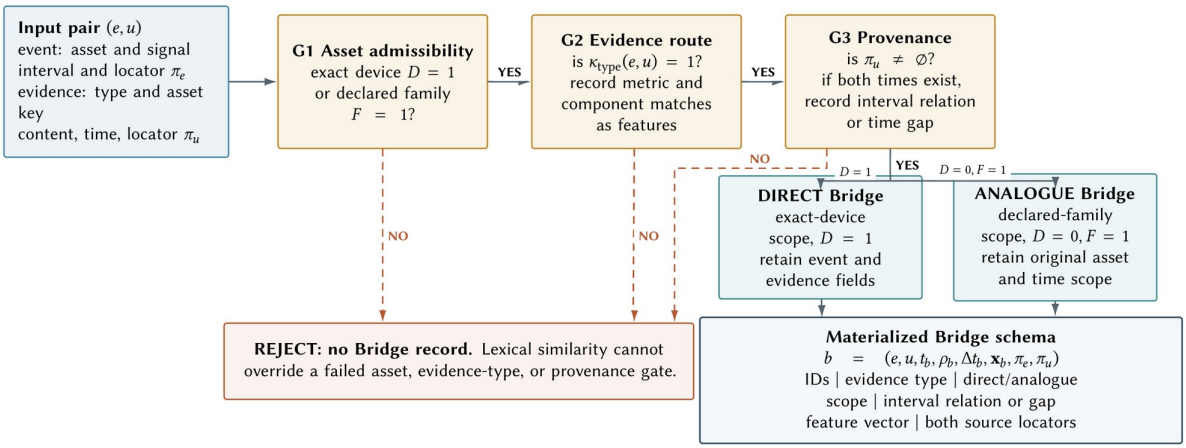


Figure 2: Bridge relation algorithm as one complete decision trace. Hard gates decide whether a relation exists and how it is scoped; ranking occurs only after the Bridge record has been materialized.

3.4. Ranking and evidence-bundle interface

The implemented TSE-Bridge ranker orders accepted candidates with a transparent feature-fusion policy:

$$S(e,u) = \operatorname{norm}\{0.82[0.25D + 0.23M + 0.17C + 0.12T + 0.08P + 0.05Q + 0.20B - 0.20F] + 0.18H\}, \tag{7}$$

where M and C are metric and component alignment; T, P, and Q encode evidence type, provenance, and quality; B is the current Bridge prior; and H is an LSA fallback. The coefficients should be read as fixed ranking heuristics. Because each term is stored by the code, Equation (7) can be rerun with any single component removed.

The top-k candidates are packaged into direct, analogue/historical, and descriptive channels instead of a single undifferentiated list (Figure 3). Every item retains the event interval, evidence scope, and source locator. An optional generator consequently receives the distinctions needed to cite evidence, qualify an answer, or abstain.

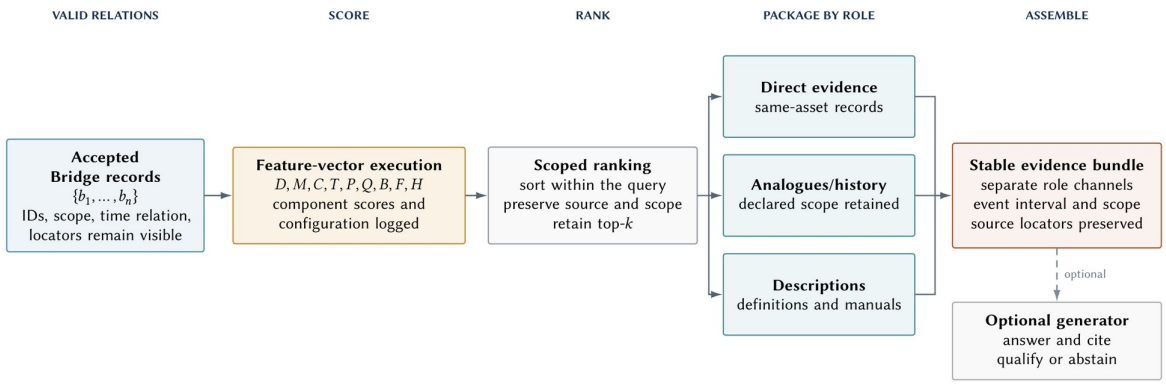


Figure 3: Bridge ranking and evidence packaging. The ranker selects among already valid relations; the bundle preserves direct, analogous, and descriptive roles for downstream grounding.

4. Experimental Evaluation

4.1. Setup and questions

The main temporal evidence is from Period B, beginning 6 July 2026. The upstream state analysis scans 46 non-empty device exports and identifies 20 devices eligible for RUN-conditioned analysis; 19 provide complete five-minute RUN windows. The event-building input contains 4,738 device-level windows and 67,651 valid signal–window observations across 437 represented continuous signals. At $\tau = 0.95$, 2,784 candidate signal–windows are merged into 377 events across eight devices and 30 device–signal profiles.

The retrieval benchmark contains 30 fixed signal-profile queries and 483 query–evidence pairs. Twenty profiles have at least one Proposed-Bridge candidate and ten do not. The inspectable pool contains 212 unique evidence IDs, with 11–22 candidates per query: 288 measurement points, 103 fault cases, 60 maintenance orders, 31 documents, and one alarm. Deterministic feature rules assign 25 pairs grade 2, 129 grade 1, and 329 grade 0. The pool is a deduplicated union of five generators, so the experiment is pooled reranking rather than full-corpus retrieval. The labels share device, metric, component, evidence-type, and family-mismatch features with the scorer; metrics therefore describe agreement with a frozen target, not diagnostic truth or independent semantic relevance.

4.2. RQ1: Executed temporal–semantic connections

TSE-Bridge connects 333 of 377 events, giving 88.33% coverage, and every retained relation has a source locator. Connected events receive 4.89 candidates on average. Within-type alternatives occur for 264 events (70.03%), and 237 events expose at least one cross-device analogue. Figure 4 shows the complete reduction and structural profile. In practical terms, the run reaches source-bearing evidence for most numerical events, yet it also reproduces a familiar maintenance-data problem: a valid connection rarely settles which record should appear first. The 70.03% ambiguity rate is therefore part of the retrieval problem rather than a pipeline failure.

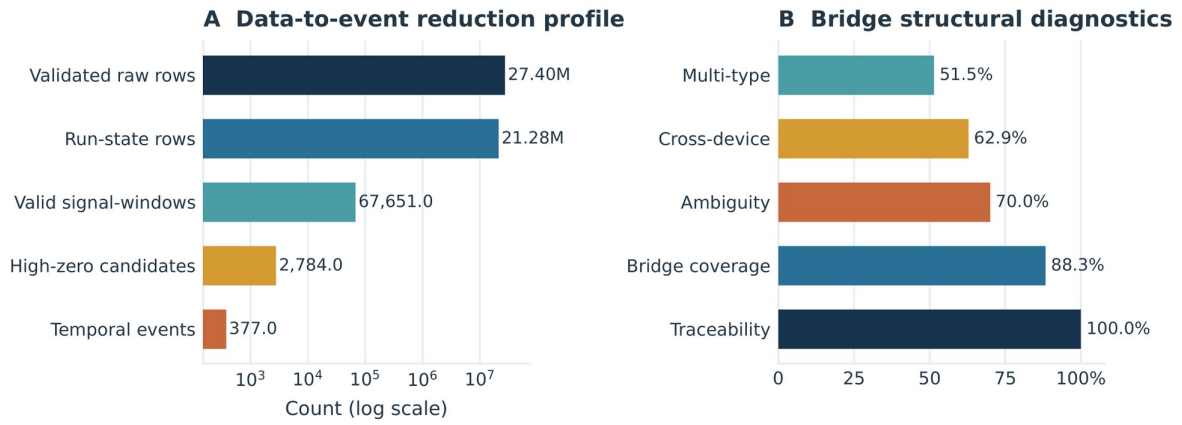


Figure 4: Code-generated data-to-event and Bridge profile. Coverage and traceability measure executable connectivity; multi-type, cross-device, and ambiguity rates identify the subsequent selection problem.

4.3. RQ2: Retrieval effectiveness

Figure 5 compares five configurations on the fixed rule-qrel target. TSE-Bridge reaches 0.4733 P@5, 0.8692 Recall@10, and 0.7266 nDCG@10, compared with 0.3800, 0.7058, and 0.5891 for BM25. The corresponding gains are 0.0933, 0.1634, and 0.1375. Across the 30 queries, the temporal–semantic features recover more relevant evidence and place it higher than sparse text matching alone. The declared GraphRAG analogue reaches 0.7903 nDCG@10 and remains the strongest observed configuration. It

is retained as a baseline under its own name. HyDE and the graph configuration are declared analogue implementations; the experiment makes no exact-reproduction claim for the published systems.

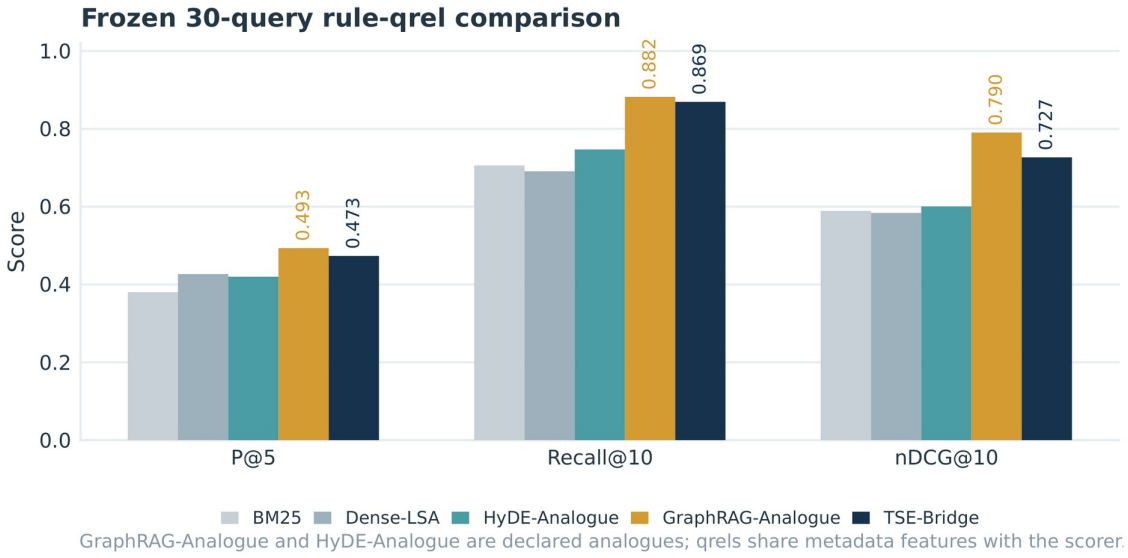


Figure 5: Retrieval comparison under the fixed rule-qrel target. TSE-Bridge improves over sparse and dense text references but does not exceed the declared GraphRAG analogue under the present target.

4.4. RQ3: Component-level code ablation

For component m, the code is rerun after removing that score term and reports $\Delta_m = \mathrm{nDCG}_{\mathrm{full}} - \mathrm{nDCG}_{-m}$. Exact-device alignment is the dominant positive component: its removal lowers nDCG@10 by 0.1009 (95% CI [0.0425, 0.1647], $p < 0.001$). Metric alignment contributes 0.0309. The size and direction of these effects are consistent with the method's emphasis on preserving asset scope.

The current Bridge-prior term behaves differently. Removing it raises nDCG@10 from 0.7266 to 0.8065 ($p = 0.003$), while Bridge Hit Coverage@10 falls from 0.6000 to 0.5667. The prior is therefore exposing more graph-linked evidence while disturbing some exact-scope ordering. This result concerns ranking only: the Bridge records created by Equations (5) and (6) remain unchanged when B is removed. Figure 6 makes the positive and negative contributions visible.

4.5. Observed errors

The main Proposed-DualView run returns 300 pairs, of which 178 are grade 0. The detailed taxonomy uses a separate exploratory configuration, T20_K10_M10_R0, which returns 294 pairs with 169 grade-0 results. Generic-token overmatch (152) and device/family mapping conflict (118) dominate the overlapping diagnostic labels (Figure 7). Signal mapping, evidence routing, and temporal mismatch occur less often. These counts diagnose the implemented retrieval path under a feature-derived target; they do not estimate operational fault prevalence or independently verified semantic error rates.

5. Discussion

What the execution establishes. The code completes a traceable path from a numerical interval to a semantic maintenance source for most events. Separate event, semantic-unit, Bridge, and ranking records make each returned item inspectable across the full path. The gain over BM25 shows that explicit temporal–semantic structure contributes retrieval value, and the exact-device ablation identifies asset scope as the main positive component. The result supports a working connection method. It carries no claim that a retrieved record caused the observed event.

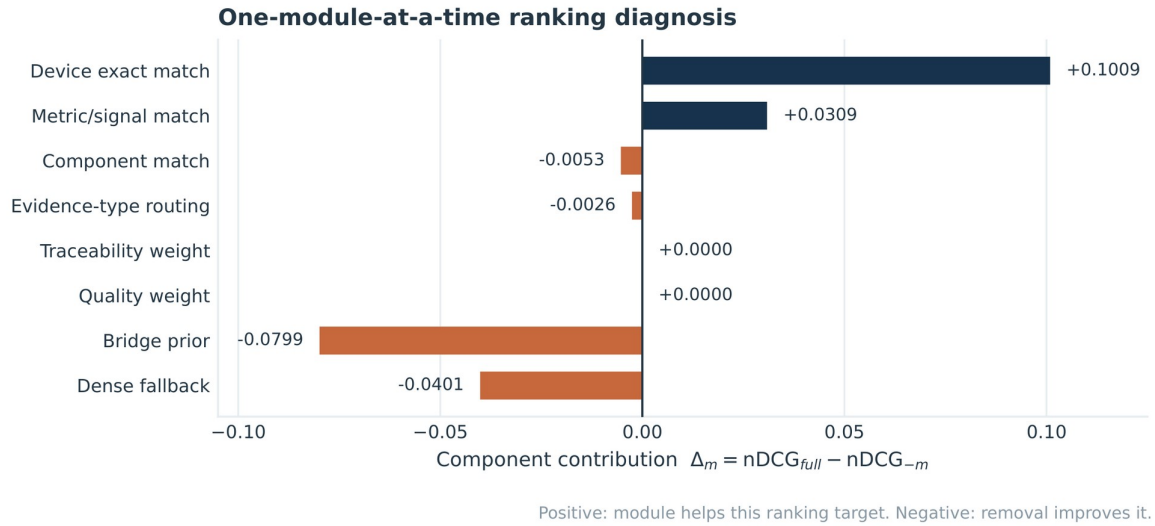


Figure 6: One-component-at-a-time ablation. Positive values mean the module helps the fixed target; negative values mean that its removal improves ranking.

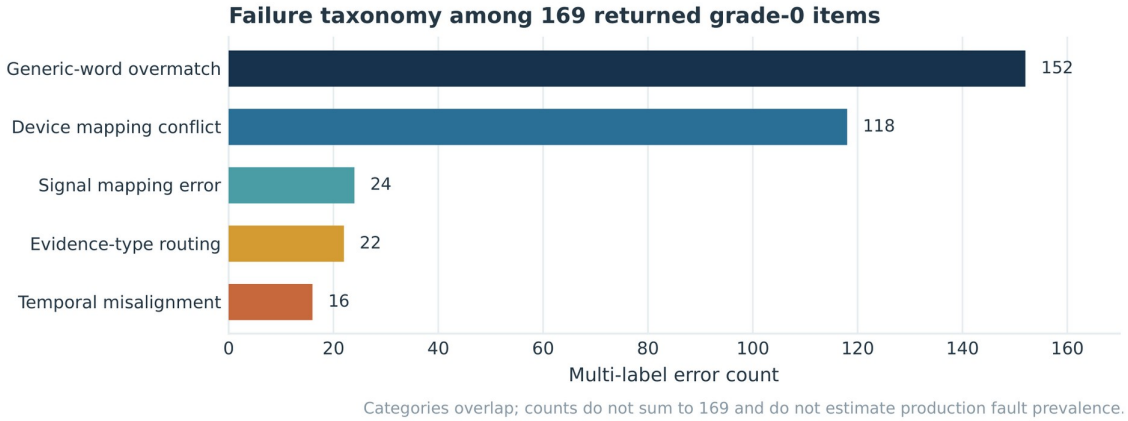


Figure 7: Failure taxonomy for returned grade-0 items. Labels overlap; counts diagnose the implemented retrieval path rather than operational fault frequency.

From connection to selection. Several records may be legitimately connected to the same event. Bridge coverage asks whether a source-bearing path exists; nDCG asks how well a ranking agrees with the present labels. The negative Bridge-prior ablation suggests a practical revision to the downstream policy: preserve valid direct evidence first, then order analogues and descriptions within their own channels. Independent training labels could be used to calibrate Equation (7) without altering the Bridge generator.

Representation and scope. The first-class record in Equation (6) can be implemented as a property-graph edge, a reified assertion, or an evidence node annotated with PROV-O and OWL-Time. The particular Web representation is less important than the fields it keeps queryable. Evidence type, direct or analogue scope, time relation, and source provenance must survive representation and retrieval instead of collapsing into a binary edge.

Evaluation limits and next tests. The 30-query labels share device, metric, component, and evidence-type features with the scorer. This design is useful for reproducibility and ablation, but it does not supply independent relevance judgments. A later protocol should label entity-link validity, direct/analogue scope, temporal applicability, query usefulness, and support for an LLM claim as separate decisions. Parameters can then be frozen for a later-period test. A subsequent answer-level exper-

iment should give identical evidence bundles to live LLMs and assess citation support, contradiction handling, qualified abstention, and over-inference.

For workshop discussion, three issues now appear concrete: (i) which Web representation can preserve Bridge scope and provenance without making retrieval impractical; (ii) whether graph exposure belongs in the ranking as a constraint, a secondary objective, or a diagnostic; and (iii) how human labels should separate valid connection, temporal applicability, and answer-level support.

6. Conclusion

TSE-Bridge provides an executable connection between industrial time-series events and semantic maintenance evidence. The implementation builds 377 events and 1,629 source-bearing relations, connects 333 events, and improves nDCG@10 over BM25 from 0.5891 to 0.7266 under the fixed rule target. Exact-device ablation supports the core alignment mechanism. The negative Bridge-prior result also identifies a specific ranking change that can be tested without rebuilding the cross-view relations. The present study therefore establishes and evaluates the connection stage, while ranking calibration, independent human relevance, and live-LLM faithfulness remain separate next steps.

Data and Artifact Availability

The accompanying revision materials contain an experiment-to-claim map, selected frozen result tables, an executable consistency audit, and an event-ID crosswalk. Raw operational records and the complete semantic catalog are not included. Original upstream programs retain dependencies on unavailable source files and historical workspace paths. The companion audit is intended for the supplied result artifacts and does not imply unrestricted public release of industrial data.

Declaration on Generative AI

Generative AI assistance was used for language editing, artifact-level consistency checks, and figure-layout assistance. The author is responsible for the final manuscript, interpretations, and claims.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 9459–9474.

[2] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, J. Larson, From local to global: A graph RAG approach to query-focused summarization, arXiv preprint arXiv:2404.16130 (2024). doi:10.48550/arXiv.2404.16130.

[3] B. J. Gutiérrez, Y. Shu, Y. Gu, M. Yasunaga, Y. Su, HippoRAG: Neurobiologically inspired long-term memory for large language models, in: Advances in Neural Information Processing Systems, volume 37, 2024, pp. 59532–59569. URL: https://proceedings.neurips.cc/paper_files/paper/2024/hash/6ddc001d07ca4f319af96a3024f6dbd1-Abstract-Conference.html.

[4] X. Zhu, Y. Xie, Y. Liu, Y. Li, W. Hu, Knowledge graph-guided retrieval augmented generation, in: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, 2025, pp. 8912–8924. doi:10.18653/v1/2025.naacl-long.449.

[5] Y. Hu, Z. Lei, Z. Zhang, B. Pan, C. Ling, L. Zhao, GRAG: Graph retrieval-augmented generation, in: Findings of the Association for Computational Linguistics: NAACL 2025, 2025, pp. 4145–4157. doi:10.18653/v1/2025.findings-naacl.232.

[6] C. Mavromatis, G. Karypis, GNN-RAG: Graph neural retrieval for efficient large language model reasoning on knowledge graphs, in: Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 16682–16699. doi:10.18653/v1/2025.findings-acl.856.

[7] X. He, Y. Tian, Y. Sun, N. V. Chawla, T. Laurent, Y. LeCun, X. Bresson, B. Hooi, G-Retriever: Retrieval-augmented generation for textual graph understanding and question answering, in: Advances in Neural Information Processing Systems, volume 37, 2024, pp. 132876–132907. doi:10.52202/079017-4224.

[8] Z. Shen, C. Diao, P. Vougiouklis, P. Merita, S. Piramanayagam, E. Chen, D. Graux, A. Melo, R. Lai, Z. Jiang, Z. Li, Y. Qi, Y. Ren, D. Tu, J. Z. Pan, GeAR: Graph-enhanced agent for retrieval-augmented generation, in: Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 12049–12072. doi:10.18653/v1/2025.findings-acl.624.

[9] W. Tao, X. Li, Y. Lan, W. Qian, TagRAG: Tag-guided hierarchical knowledge graph retrieval-augmented generation, in: Findings of the Association for Computational Linguistics: ACL 2026, 2026, pp. 6434–6456. doi:10.18653/v1/2026.findings-acl.321.

[10] T. Zhou, P. Niu, X. Wang, L. Sun, R. Jin, One fits all: Power general time series analysis by pre-trained LM, Advances in Neural Information Processing Systems 36 (2023).

[11] Z. Yang, X. Du, R. Mao, J. Ni, D. Hershcovich, E. Cambria, EventRAG: Enhancing LLM generation with event knowledge graphs, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, 2025, pp. 16967–16979. doi:10.18653/v1/2025.acl-long.830.

[12] Z. Zhang, Z. Li, Y. Li, B. Ding, B. K. H. Low, Respecting temporal-causal consistency: Entity-event knowledge graph for retrieval-augmented generation, in: Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics, 2026, pp. 2017–2054. doi:10.18653/v1/2026.eacl-long.90.

[13] J. F. Allen, Maintaining knowledge about temporal intervals, Communications of the ACM 26 (1983) 832–843. doi:10.1145/182.358434.

[14] S. Cox, C. Little, Time Ontology in OWL, W3C/OGC Recommendation, World Wide Web Consortium, 2022. URL: https://www.w3.org/TR/owl-time/.

[15] S. Robertson, H. Zaragoza, The probabilistic relevance framework: BM25 and beyond, Foundations and Trends in Information Retrieval 3 (2009) 333–389. doi:10.1561/1500000019.

[16] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W. Yih, Dense passage retrieval for open-domain question answering, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2020, pp. 6769–6781. doi:10.18653/v1/2020.emnlp-main.550.

[17] L. Gao, X. Ma, J. Lin, J. Callan, Precise zero-shot dense retrieval without relevance labels, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2023, pp. 1762–1777. doi:10.18653/v1/2023.acl-long.99.

[18] J. Barmettler, A. Bernstein, L. Rossetto, ConceptFormer: Towards graph-native grounding of large language models via latent concept injection, in: Companion Proceedings of the ACM Web Conference 2026, Association for Computing Machinery, 2026, pp. 587–596. doi:10.1145/3774905.3794653.

[19] N. Wang, X. Han, J. Singh, J. Ma, V. Chaudhary, CausalRAG: Integrating causal graphs into retrieval-augmented generation, in: Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 22680–22693. doi:10.18653/v1/2025.findings-acl.1165.

[20] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A.-C. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. F. Sequeda, S. Staab, A. Zimmermann, Knowledge graphs, ACM Computing Surveys 54 (2021) 1–37. doi:10.1145/3447772.

[21] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: A roadmap, IEEE Transactions on Knowledge and Data Engineering 36 (2024) 3580–3599. doi:10.1109/TKDE.2024.3352100.

[22] S. Wu, S. Zhao, M. Yasunaga, K. Huang, K. Cao, Q. Huang, V. N. Ioannidis, K. Subbian, J. Zou, J. Leskovec, STaRK: Benchmarking LLM retrieval on textual and relational knowledge bases, in: Advances in Neural Information Processing Systems, volume 37, 2024, pp. 127129–127153. doi:10.52202/079017-4037.

[23] S. Knollmeyer, M. U. Akmal, L. Koval, S. Asif, S. G. Mathias, D. Grossmann, Document knowledge graph to enhance question answering with retrieval augmented generation, in: 2024 IEEE 29th International Conference on Emerging Technologies and Factory Automation, 2024, pp. 1–4. doi:10.1109/ETFA61755.2024.10711054.

[24] M. S. Gill, J. Vyas, A. Markaj, F. Gehlhoff, M. Mercangöz, Leveraging LLM agents and digital twins for fault handling in process plants, in: 2025 IEEE 30th International Conference on Emerging Technologies and Factory Automation, 2025, pp. 1–8. doi:10.1109/ETFA65518.2025.11205597.

[25] A. Haller, K. Janowicz, S. Cox, M. Lefrançois, K. Taylor, D. L. Phuoc, R. García-Castro, J. Lieberman, C. Stadler, Semantic Sensor Network Ontology, W3C Recommendation, World Wide Web Consortium, 2017. URL: https://www.w3.org/TR/vocab-ssn/.

[26] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, J. Zhao, PROV-O: The PROV Ontology, W3C Recommendation, World Wide Web Consortium, 2013. URL: https://www.w3.org/TR/prov-o/.

[27] J. J. Carroll, C. Bizer, P. Hayes, P. Stickler, Named graphs, provenance and trust, in: Proceedings of the 14th International Conference on World Wide Web, 2005, pp. 613–622. doi:10.1145/1060745.1060835.

[28] R. Albertoni, A. Isaac, J. Debattista, L. L. de Santana Nunes, V. Alexiev, M. Dekkers, A. Perego, P. Winstanley, Data on the Web Best Practices: Data Quality Vocabulary, W3C Recommendation, World Wide Web Consortium, 2016. URL: https://www.w3.org/TR/vocab-dqv/.

[29] H. Knublauch, D. Kontokostas, Shapes Constraint Language (SHACL), W3C Recommendation, World Wide Web Consortium, 2017. URL: https://www.w3.org/TR/shacl/.

[30] D. Zhou, Y. Zhu, X. Wang, H. Zhou, Y. He, J. Chen, S. Staab, E. Kharlamov, What breaks knowledge graph based RAG? benchmarking and empirical insights into reasoning under incomplete knowledge, in: Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics, 2026, pp. 2522–2538. doi:10.18653/v1/2026.eacl-long.114.

[31] S. Es, J. James, L. Espinosa-Anke, S. Schockaert, RAGAS: Automated evaluation of retrieval augmented generation, arXiv preprint arXiv:2309.15217 (2023). doi:10.48550/arXiv.2309.15217.

[32] D. Ru, L. Qiu, X. Hu, T. Zhang, P. Shi, S. Chang, C. Jiayang, C. Wang, S. Sun, H. Li, Z. Zhang, B. Wang, J. Jiang, T. He, Z. Wang, P. Liu, Y. Zhang, Z. Zhang, RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation, in: Advances in Neural Information Processing Systems, volume 37, 2024, pp. 21999–22027. doi:10.52202/079017-0692.

[33] C. Niu, Y. Wu, J. Zhu, S. Xu, K. Shum, R. Zhong, J. Song, T. Zhang, RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 10862–10878. doi:10.18653/v1/2024.acl-long.585.

[34] J. Saad-Falcon, O. Khattab, C. Potts, M. Zaharia, ARES: An automated evaluation framework for retrieval-augmented generation systems, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2024, pp. 338–354. doi:10.18653/v1/2024.naacl-long.20.