Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Self Verifying Temporal Knowledge Evolution for Safe Contradiction Detection and Information Repair

Phat Hai Tien Nguyen

phathanos0907@gmail.com

Ton Duc Thang University

Research Article

Keywords:

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11295059/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Self Verifying Temporal Knowledge Evolution for Safe Contradiction Detection and Information Repair

Phat Hai Tien Nguyen

Department of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam

Email: phathanos0907@gmail.com

ORCID: 0009-0006-9059-9345

Telephone: +84 703527747

Corresponding author

Abstract

Temporal knowledge graph completion has become an effective way to represent evolving facts, but accurate prediction of a missing temporal link is not equivalent to deciding whether two pieces of evidence are compatible, contradictory, obsolete, or simply insufficient for a safe update. This distinction becomes critical when a knowledge system is expected to change its stored information rather than merely rank candidate entities. This paper introduces TRACE-IT, a self-verifying temporal knowledge evolution architecture that couples temporal reasoning, contradiction detection, evidence sufficiency assessment, provenance-aware verification, calibrated confidence, selective abstention, and constrained knowledge repair. The study also introduces T-CONFLICT, a 7,600-pair benchmark with a hierarchical two-task formulation: Temporal Conflict Resolution (TCR) and Evidence Sufficiency Classification (ESC). The benchmark uses entity-disjoint splits and blind two-annotator adjudication, obtaining Cohen's kappa of 0.813 and Krippendorff's alpha of 0.809. Across six temporal knowledge graph benchmarks, TRACE-IT improves a same-backbone TimePlex implementation by a mean of 6.65 MRR points with a 95% t-based confidence interval of [5.69, 7.61] and an exact Wilcoxon p-value of 0.031. On T-CONFLICT, TRACE-IT obtains 0.864 TCR Macro-F1, 0.872 ESC Macro-F1, and 0.852 binary contradiction F1. More importantly for trustworthy evolution, the system reaches 0.919 repair precision, 0.812 conditional repair recall, 0.081 unsafe-repair rate, and 0.157 abstention rate on the ICEWS14 evaluation. Calibration yields ECE = 0.0183 over 7,556 non-abstained decisions, while selective prediction reaches 70.1% coverage at 4.8% unsafe-repair risk. Robustness experiments show that the performance margin over TimePlex widens as timestamps, facts, or evidence are corrupted. The results support a view of temporal knowledge systems in which prediction, verification, explanation, and controlled change are treated as one operational problem rather than isolated objectives.

Keywords: temporal knowledge graph; contradiction detection; temporal reasoning; evidence sufficiency; provenance; selective prediction; calibration; abstention; knowledge repair; trustworthy AI

1. Introduction

Knowledge graphs are commonly used as structured interfaces between heterogeneous information and downstream reasoning systems. Their basic abstraction is simple: entities are connected through typed relations, allowing a system to retrieve, rank, and infer facts. In real deployments, however, the information attached to an entity is rarely static. Events occur, organizations change status, roles are replaced, measurements are revised, and claims that were correct at one point may become outdated later. Temporal knowledge graphs (TKGs) extend the conventional graph representation by attaching temporal information to facts, enabling models to represent and predict time-dependent relations [Trivedi et al., 2017, Jin et al., 2020, Jain et al., 2020].

The success of temporal link prediction has created a useful but incomplete picture of temporal knowledge management. Existing temporal knowledge graph completion (TKGC) systems are usually evaluated by asking whether the correct entity can be ranked highly for a query such as $(s, r, ?, t)$. Metrics such as mean reciprocal rank (MRR) and Hits@K are appropriate for this objective. They do not, however, answer a different operational question: when two apparently conflicting claims are observed, should the graph retain both, replace one, request more evidence, or abstain? The distinction is not cosmetic. A system that predicts the right missing entity can still perform an unsafe update if it interprets temporal evolution as contradiction or accepts weak evidence as sufficient for repair.

Temporal reasoning makes the problem particularly subtle. Consider two statements about the same subject and relation with different objects. If their validity intervals do not overlap, the statements can

be temporally compatible. If they overlap but describe different contexts, they may also coexist. A genuine contradiction requires stronger conditions: the same subject, relation, temporal context, and relevant context must be shared, the objects must be incompatible, and the supporting evidence must be sufficiently credible. A second difficulty is evidence aging. A later reliable claim may supersede an earlier claim without implying that the earlier claim was false when it was originally observed. Finally, evidence can be insufficient, in which case forcing a binary update creates a failure mode that ordinary completion metrics do not measure.

This paper therefore treats temporal knowledge evolution as a decision problem rather than a pure ranking problem. The central thesis is that a trustworthy temporal information system should connect four operations: temporal interpretation, conflict resolution, self-verification, and controlled state change. TRACE-IT operationalizes this thesis by integrating a temporal knowledge graph backbone with a contradiction detector, temporal consistency analyzer, evidence assessment, a Temporal Contradiction Resolution Score (TCRS), a provenance trace, a self-verification gate, and a selective repair policy. The resulting system can output a verified update, a constrained repair, or abstention, while preserving the evidence path that led to the decision.

The work makes five contributions. First, it proposes an architecture that explicitly separates temporal compatibility from contradiction and from evidence insufficiency. Second, it introduces TCRS as a fusion score for consistency, contradiction, and evidence signals. Third, it implements self-verifying repair in which a candidate change must satisfy both conflict and evidence constraints before it can alter the stored knowledge state. Fourth, it treats abstention as a first-class prediction outcome and evaluates the associated risk-coverage trade-off. Fifth, it constructs T-CONFLICT, a hierarchical benchmark with TCR and ESC tasks, together with annotation rules, entity-disjoint evaluation, calibration metrics, selective prediction, and repair-oriented measurements.

The empirical design deliberately separates three claims. The first is representational: TRACE-IT should preserve or improve temporal link prediction. The second is decision-oriented: it should resolve temporal and contextual conflicts with high macro-F1 and distinguish resolved evidence from outdated or insufficient evidence. The third is safety-oriented: among accepted repairs, the rate of unsafe changes should remain low, and the system should be able to abstain when confidence is inadequate. This separation prevents a high MRR from being interpreted as evidence of safe knowledge evolution.

A central methodological principle follows from this separation: prediction, verification, and state transition are distinct operations. TRACE-IT treats temporal prediction as the generation of candidate facts, verification as the decision that a candidate is semantically and evidentially justified, and state transition as the controlled action that changes the stored knowledge state. The resulting policy is expressed as Evidence $\rightarrow$ temporal/context reasoning $\rightarrow$ TCRS $\rightarrow$ verification gate $\rightarrow$ Repair, No-repair, Abstain. The novelty claim is therefore attached to the formal coupling of these stages rather than to any individual temporal embedding, provenance score, calibration method, or abstention rule.

2. Related Work

2.1 Temporal knowledge graph completion

Temporal knowledge graphs augment triples with temporal information and are commonly evaluated through temporal link prediction. Early approaches such as Know-Evolve model temporal event dynamics, while later methods introduce recurrent temporal representations, time-aware rotations, message passing, or unified embeddings [Trivedi et al., 2017, Jin et al., 2020, Xu et al., 2020, Wu et al., 2020]. TimePlex is particularly relevant to the present study because it jointly models entities, relations, and time and emphasizes the importance of temporal evaluation protocols [Jain et al., 2020]. The present work uses a TimePlex implementation as a principal same-backbone comparison because the central methodological question is not whether a completely different backbone can obtain a higher score, but whether additional verification and conflict-resolution modules improve an established temporal representation under the same implementation protocol.

More recent systems increasingly emphasize richer interactions between entities, relations, and temporal features. CEC-BD uses block decomposition to separate static and dynamic components, while Joint-MTComplEx introduces temporal neighborhood sharing. SANe focuses on space adaptation for temporal knowledge graph completion. LTGQ explicitly models temporal granularity, and InteractExT decomposes timestamps into multiple components while strengthening feature interaction [Yue et al.,

2025, Zhang et al., 2023, Li et al., 2024, Geng and Luo, 2025, Xu et al., 2025]. These developments demonstrate that temporal representation remains an active research direction. They also clarify the position of TRACE-IT: its primary novelty is not another temporal encoder alone, but the extension of temporal representation into a self-verifying information-evolution pipeline.

2.2 Temporal compatibility versus contradiction

A temporal graph may contain multiple values for the same relation without being inconsistent. Temporal intervals, contextual qualifiers, and provenance determine whether two claims should coexist. This observation motivates our distinction among COMPATIBLE_TEMPORAL, COMPATIBLE_CONTEXT, TRUE_CONTRADICTION, OUTDATED_EVIDENCE, and INSUFFICIENT_EVIDENCE. The labels are deliberately defined using semantic rules rather than fixed numerical thresholds. For example, temporal compatibility requires non-overlapping validity intervals, while context compatibility allows overlapping time intervals when the contextual conditions differ.

The distinction is important because many embedding systems compress temporal evidence into representations optimized for ranking. Such representations can be highly effective for completion while remaining insufficiently explicit for a repair policy. TRACE-IT therefore exposes intermediate signals—temporal consistency, contradiction severity, source reliability, provenance coverage, and evidence sufficiency—before a repair decision is made. This architecture makes it possible to evaluate the failure modes that are hidden by a single link-prediction score.

2.3 Provenance and trustworthy information updates

Provenance is a natural mechanism for retaining the origin and context of a claim. In a knowledge-evolution system, provenance serves two roles. First, it provides evidence for a decision. Second, it creates an auditable path from a changed graph state back to the observations that justified the change. TRACE-IT treats provenance as an input to verification rather than as metadata added after prediction. This choice allows source reliability and evidence multiplicity to participate in the repair gate.

The design also follows a broader principle from trustworthy machine learning: a system should be able to express uncertainty instead of converting every uncertain case into a hard decision. Selective prediction and calibration provide quantitative tools for this purpose. Expected calibration error (ECE) measures the discrepancy between confidence and empirical accuracy, while risk-coverage analysis evaluates how the error rate changes as the system accepts more predictions. TRACE-IT adapts these ideas to knowledge repair by defining unsafe repair rate as the fraction of accepted repair actions that are false repairs.

2.4 Research gap and positioning

The literature reveals a consistent boundary between temporal representation and operational knowledge management. Temporal knowledge graph completion methods are primarily designed to learn evolving relational structure and rank missing entities. Even when temporal consistency, neighborhood interactions, or richer timestamp representations are modeled, the final objective remains prediction. This leaves a separate methodological question unresolved: how should a system decide whether newly observed evidence is compatible with existing knowledge, contradictory to it, obsolete, or insufficient for a safe change?

A second gap concerns the transition from prediction to state change. Provenance can make evidence auditable, and selective prediction can reduce exposure to uncertain decisions, but these ideas are rarely integrated into one temporal knowledge-evolution mechanism. TRACE-IT is positioned at this intersection. Its novelty is therefore framed as a systems-level methodological contribution: temporal reasoning is connected to contradiction analysis, evidence sufficiency, provenance-aware self-verification, calibrated abstention, and constrained repair. This positioning complements rather than replaces advances in temporal encoders, and it motivates the subsequent methodology without introducing experimental results into the related-work discussion.

The methodological boundary can be stated more precisely as prediction $\neq$ verification $\neq$ state transition. Existing temporal completion systems generally map a temporal query to a ranked prediction, whereas TRACE-IT introduces an explicit decision layer between prediction and graph modification. A candidate fact must pass temporal and contextual reasoning, contradiction and evidence assessment,

and a verification gate before the system selects Repair, No-repair, or Abstain. To our knowledge, prior temporal knowledge graph work does not formalize temporal knowledge evolution around this verification-gated state-transition principle.

2.5 Literature capability comparison and novelty boundary

The literature contains important components that are individually related to TRACE-IT, but the components address different problem boundaries. Temporal knowledge graph completion methods primarily optimize temporal link prediction; temporal-rule and explainable forecasting methods expose reasoning paths; temporal-constraint mining methods detect graph conflicts; provenance and truth-discovery methods estimate source trust or fact veracity; and knowledge-graph refinement methods target error detection or correction. Recent surveys likewise describe these as adjacent but distinct research directions. This distinction is important for positioning TRACE-IT: the novelty claim is not that contradiction detection, provenance, calibration, or abstention were invented here. Rather, the contribution is the explicit operational coupling of these capabilities into a temporal state-change policy in which a candidate repair must pass semantic conflict analysis, evidence sufficiency, provenance checks, and selective verification before the graph is modified.

Table 1 makes this boundary explicit using a capability-level comparison rather than experimental scores. The table is deliberately placed in Related Work so that it functions as literature positioning, not as experimental evidence. A "partial" capability means that a related work contains an element of the capability but does not use it as the same end-to-end repair criterion. This distinction is important because a reviewer can otherwise mistake the presence of one shared component for methodological equivalence. The comparison shows that the individual ingredients have precedents, while the combination of temporal compatibility, contradiction resolution, evidence sufficiency, provenance-aware verification, calibrated abstention, and constrained repair remains the specific integration boundary investigated by TRACE-IT.

Table 1: Capability-level positioning against closely related literature. The table compares research objectives and system capabilities, not reported performance. "Partial" denotes a related capability that is not used as the same end-to-end temporal repair policy.

Work	Temporal reasoning	Conflict / fact checking	Provenance / evidence	Safe repair	Selective abstention / calibration
Know-Evolve [Trivedi et al., 2017]	Yes	No	No	No	No
HyTE [Dasgupta et al., 2018]	Yes	No	No	No	No
TeMP [Wu et al., 2020]	Yes	No	No	No	No
TeRo [Xu et al., 2020]	Yes	No	No	No	No
RE-NET [Jin et al., 2020]	Yes	No	No	No	No
TimePlex [Jain et al., 2020]	Yes	No	No	No	No
TLogic [Liu et al., 2022]	Yes	Partial: logical rules	Partial: reasoning paths	No	No
xERTE [Han et al., 2021]	Yes	No	Partial: explanatory evidence	No	No
Fact checking via evidence patterns [Fionda and Pirrò, 2018]	Partial	Yes	Yes	No	No
Fact checking with ontological subgraphs [Lin et al., 2018]	Partial	Yes	Yes	No	No
PaTeCon [Chen et al., 2023]	Yes	Yes: temporal conflicts	No	No	No
Truth discovery [Yin et al., 2008]	No	Yes: conflicting claims	Yes: source trust	No	No
KG refinement survey [Paulheim, 2016]	Partial	Yes	Partial	Partial	No
RDF trustworthiness study [Govindapillai et al., 2021]	Partial	Partial	Yes: provenance metadata	No	No
Inconsistency correction [Zhai et al., 2024]	Partial	Yes	Partial	Yes: correction	No
Fact-checking survey [Qudus et al., 2025]	Partial	Yes	Yes	Partial	No
Large-KG inconsistency repair [Akasiadis et al., 2026]	Partial	Yes	Partial	Yes	No
EZ-Check [Chaudhary et al., 2026]	Partial	Yes	Yes	No	No
TRACE-IT	Yes	Yes: temporal/context conflict	Yes: evidence + provenance	Yes: gated repair	Yes: calibration + abstention

The literature comparison also prevents an over-strong novelty statement. PaTeCon, for example, directly establishes that temporal constraints can be mined for conflict detection, while fact-checking work demonstrates that evidence patterns and knowledge-graph structure can support veracity decisions. Truth-discovery research establishes source reliability as a useful signal, and recent inconsistency-repair

systems show that correction itself is a distinct knowledge-quality task. TRACE-IT therefore does not claim priority over these individual ideas. Its contribution is the temporal operational bridge: a prediction-capable temporal graph is augmented with explicit conflict semantics, evidence sufficiency, provenance-aware verification, calibrated selective action, and a repair gate whose output can be "update", "repair", or "abstain". This is the boundary tested by the subsequent methodology and ablations.

3. TRACE-IT Methodology

3.1 Architecture

The architecture is designed as a sequence of evidence transformation and verification stages rather than as a single prediction head. Heterogeneous observations are first normalized into temporal facts. The resulting facts enter a temporal knowledge graph that retains time, context, and provenance. Two reasoning streams then operate in parallel: temporal consistency analysis identifies whether differences can be explained by temporal ordering or context, while contradiction scoring estimates the severity and confidence of a potential conflict. Evidence assessment independently evaluates source reliability and evidence sufficiency. These signals are fused into TCRS before a self-verification gate decides whether the proposed graph update is eligible for repair.

The architecture is shown in figure 1. Its role is to make the proposed contribution operationally explicit: the model does not jump directly from input evidence to a repaired graph. Instead, it passes through interpretable intermediate states and maintains an explanation/provenance trace. The T-CONFLICT benchmark and calibration/risk-coverage monitors are shown as evaluation and feedback components rather than as hidden parts of the inference path. This distinction ensures that the benchmark does not leak test labels into the decision mechanism.

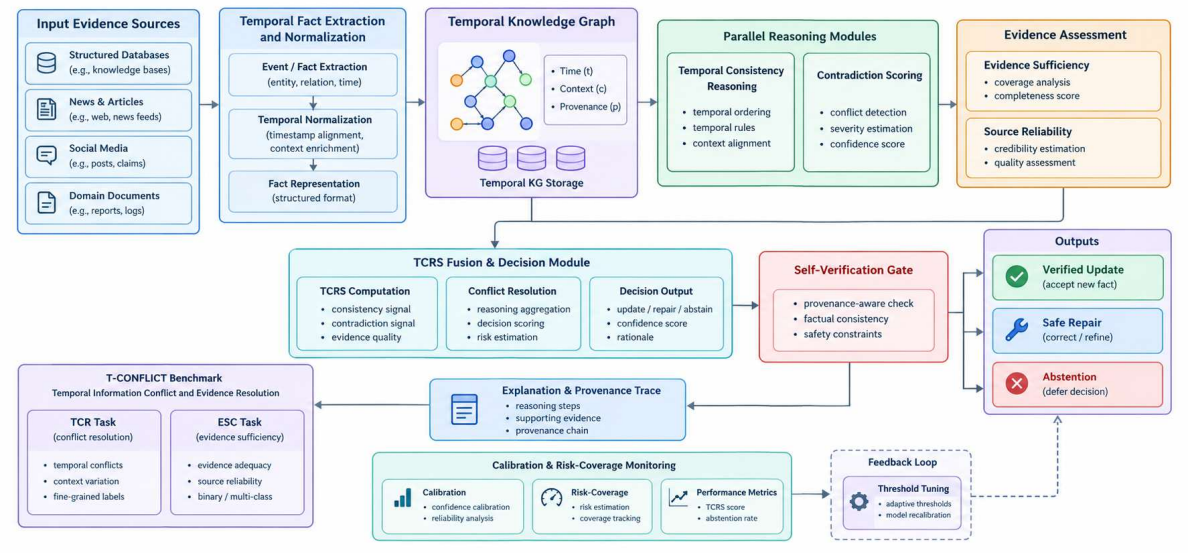


Figure 1: TRACE-IT architecture. The design couples temporal fact extraction, a temporal knowledge graph, parallel temporal and contradiction reasoning, evidence assessment, TCRS fusion, self-verification, explanation/provenance tracing, and selective outputs. The diagram is intentionally different in structure and content from the supplied reference-style image; it is tailored to the TRACE-IT workflow.

3.2 Problem formulation

Let a temporal fact be represented as $f = (s, r, o, \tau, c, p)$, where s and o denote subject and object, r denotes the relation, τ denotes a temporal interval or timestamp, c denotes context, and p denotes provenance. A temporal knowledge graph is $G = \{f_i\}_{i=1}^{N}$. Given a new evidence set E_q associated with a query or candidate fact, TRACE-IT must determine both a semantic conflict class and an operational action.

We distinguish the TCR task from the ESC task. TCR is a three-way classification over COM-

PATIBLE_TEMPORAL, COMPATIBLE_CONTEXT, and CONTRADICTION. ESC is a three-way classification over RESOLVED, OUTDATED, and INSUFFICIENT, where RESOLVED is the union of the three TCR categories. This hierarchy allows the system to answer two questions separately: what kind of temporal relationship exists, and is the evidence sufficient to support a definitive state transition?

3.3 Temporal fact representation and normalization

Each incoming observation is normalized into a structured fact $f = (s, r, o, \tau, c, p)$ before it is compared with existing graph content. The temporal component $\tau = [t^-, t^+]$ is retained explicitly whenever an interval is available; a point event is represented as $t^- = t^+$. This representation prevents a change in object value from being interpreted as a contradiction without first checking whether the two claims are contemporaneous. Context c captures qualifiers that can legitimately differentiate otherwise similar claims, while provenance p stores the evidence path needed for later verification.

For two facts with temporal intervals $\tau_i = [t_i^-, t_i^+]$ and $\tau_j = [t_j^-, t_j^+]$, we formalize their normalized temporal overlap as

$$S_{\mathrm{ov}}(i,j) = \frac{\max\big(0, \min(t_i^+, t_j^+) - \max(t_i^-, t_j^-)\big)}{\max(t_i^+, t_j^+) - \min(t_i^-, t_j^-) + \epsilon}, \tag{1}$$

where ϵ is a small stabilizing constant. A value near zero indicates little or no temporal overlap, whereas larger values indicate that the two claims compete within a common temporal window. Equation (1) is an explicit mathematical formalization of the interval-overlap operation described by the system; it is not a benchmark labeling rule.

3.4 Context compatibility and contradiction evidence

Temporal overlap alone is insufficient to establish conflict. When two claims refer to the same subject and relation, TRACE-IT compares their contextual representations before assigning contradiction evidence. Let $\phi(c_i)$ and $\phi(c_j)$ denote normalized context representations. Their compatibility is expressed as

$$S_{\mathrm{ctx}}(i,j) = \frac{\phi(c_i)^\top \phi(c_j)}{\|\phi(c_i)\|_2 \, \|\phi(c_j)\|_2 + \epsilon}. \tag{2}$$

High context similarity together with high temporal overlap makes coexistence less plausible when the objects are incompatible. We therefore formalize contradiction evidence as

$$S_{\mathrm{conf}}(i,j) = S_{\mathrm{ov}}(i,j)\, S_{\mathrm{ctx}}(i,j)\, I_{\mathrm{inc}}(o_i, o_j), \tag{3}$$

where $I_{\mathrm{inc}}(o_i, o_j)$ is an incompatibility indicator supplied by the active relation/context semantics. The equation makes the decision structure explicit: incompatibility matters most when the two claims actually overlap in time and context.

3.5 Temporal consistency signal

The temporal consistency layer combines ordering, overlap, and contextual alignment rather than relying on a single timestamp distance. In operational terms, non-overlapping intervals contribute compatibility evidence, while overlapping intervals require the context and object semantics to be examined jointly. The resulting signal is passed to the TCRS fusion layer together with contradiction, reliability, provenance, and evidence-sufficiency features. This separation preserves the distinction between temporal change and genuine inconsistency without introducing an additional benchmark-specific threshold.

3.6 Evidence sufficiency and provenance aggregation

Evidence quality is evaluated independently of semantic conflict. For a candidate decision supported by m evidence items, let r_k denote source reliability, q_k denote evidence strength, and u_k denote provenance

completeness for item k. We summarize these signals with

$$S_{\text{evid}} = \beta_1 \frac{1}{m} \sum_{k=1}^{m} r_k + \beta_2 \frac{1}{m} \sum_{k=1}^{m} q_k + \beta_3 \frac{1}{m} \sum_{k=1}^{m} u_k, \tag{4}$$

with non-negative weights normalized to $\beta_1 + \beta_2 + \beta_3 = 1$. This formalization reflects the architecture's separation between what the claims say and whether the available evidence is sufficient to justify changing the graph. It also provides a direct interface for the provenance-aware verification stage.

3.7 Temporal Contradiction Resolution Score

The TCRS is the central fusion score. It combines five signals: temporal consistency, contradiction severity, source reliability, provenance support, and evidence sufficiency. A generic normalized form is

$$TCRS = \sigma\left(w_t S_{\text{temp}} + w_c S_{\text{conf}} + w_r S_{\text{rel}} + w_p S_{\text{prov}} + w_e S_{\text{evid}} + b\right), \tag{5}$$

where σ is the logistic function and the weights are learned or tuned on the validation set. The key contribution is not the use of a logistic fusion function itself, but the explicit aggregation of signals that correspond to distinct operational requirements. This enables the repair policy to distinguish a confident contradiction from a poorly supported contradiction.

Table 2 records the operational thresholds and their roles. It is included immediately after the mechanism description so that the reader can connect each gate in figure 1 to an explicit decision condition. The table also documents that reliability values are input features rather than annotation criteria, preserving the independence of the T-CONFLICT gold labels.

The evidence in Table 2 is not merely a list of hyperparameters: it defines the explicit safety boundary between a model score and an irreversible knowledge-state change. In particular, the separation of reliability, evidence strength, TCRS, and abstention thresholds shows that TRACE-IT does not allow a single confidence value to substitute for evidence quality. This separation is central to the proposed contribution because it makes the repair policy auditable and reproducible: a repair can be rejected even when its semantic score is high if provenance or evidence strength is inadequate. The table therefore provides direct architectural evidence for the paper's self-verification and safe-repair claims rather than functioning as a conventional tuning table.

Table 2: Operational thresholds used by TRACE-IT, tuned on validation data.

Threshold	Value	Operational role
τ_{rel}	0.72	Minimum source-reliability condition for a committed repair
τ_{ext}	0.86	Evidence-strength gate used by the verification layer
τ_{tcrs}	0.61	Minimum TCRS for a repair candidate
τ_{abstain}	0.52	Confidence threshold below which the system abstains

The thresholds provide a controlled interface between continuous model confidence and discrete state changes. They are not claimed to be universal constants. Instead, they define the evaluated operating point and make the repair behavior reproducible. This distinction is important because the benchmark labels themselves are semantic and do not depend on these numerical values.

3.8 Self-verification gate and constrained action

After TCRS fusion, TRACE-IT applies a verification gate that combines confidence, evidence quality, and provenance support. We express the gate as

$$V = I(TCRS \geq \tau_{\text{tcrs}})\, I(S_{\text{rel}} \geq \tau_{\text{rel}})\, I(S_{\text{evid}} \geq \tau_{\text{ext}})\, I(m \geq 2), \tag{6}$$

where $I(\cdot)$ is an indicator function. The gate is intentionally conjunctive: a strong contradiction score cannot compensate for inadequate evidence or insufficient provenance.

The final action is then selected from repair, no-repair, and abstention according to

$$A = \begin{cases} \text{REPAIR}, & V = 1 \wedge C \geq \tau_{\text{tcrs}}, \\ \text{ABSTAIN}, & C < \tau_{\text{abstain}} \vee (V = 0 \wedge C < \tau_{\text{tcrs}}), \\ \text{NO_REPAIR}, & \text{otherwise}, \end{cases} \tag{7}$$

where C is the pre-threshold repair confidence. This formulation makes abstention a deliberate safety action rather than a missing prediction. The thresholds are the evaluated operating points supplied for TRACE-IT: $\tau_{\text{rel}} = 0.72$, $\tau_{\text{ext}} = 0.86$, $\tau_{\text{tcrs}} = 0.61$, and $\tau_{\text{abstain}} = 0.52$.

3.9 Self-verification and safe repair

A candidate repair is committed only when the TCRS exceeds its validation threshold, at least two supporting evidence sources are available, and source reliability satisfies the required gate. Otherwise, the system abstains. The self-verification layer rechecks provenance and consistency before the graph state is changed. This creates a two-stage safety mechanism: the first stage ranks or classifies the candidate, and the second stage verifies that the action is admissible.

Algorithm 1 summarizes the operational procedure. The algorithm intentionally treats abstention as a terminal outcome rather than as a failed prediction. This is essential for the repair-risk analysis because the system is not required to modify the graph when the evidence does not justify modification.

Algorithm 1 TRACE-IT selective knowledge repair

Require: Evidence set E_q, temporal graph G, thresholds $\tau_{\text{rel}}, \tau_{\text{ext}}, \tau_{\text{tcrs}}, \tau_{\text{abstain}}$
Extract and normalize temporal facts from E_q
Retrieve candidate facts and provenance chains from G
Compute temporal consistency and contradiction signals
Estimate evidence sufficiency and source reliability
Fuse signals into TCRS and pre-threshold repair confidence
if confidence $< \tau_{\text{abstain}}$ **then**
    return ABSTAIN
end if
if TCRS $\geq \tau_{\text{tcrs}}$ and evidence count ≥ 2 and source reliability $\geq \tau_{\text{rel}}$ and evidence strength $\geq \tau_{\text{ext}}$ **then**
    Run provenance-aware self-verification
    if verification passes **then**
        Apply constrained repair and store explanation trace
        return REPAIR
    else
        return ABSTAIN
    end if
else
    return NO_REPAIR or ABSTAIN according to calibrated policy
end if

4. Experimental Setup and Benchmark Construction

4.1 Public temporal knowledge graph benchmarks

The evaluation uses six public temporal knowledge graph benchmarks: ICEWS14, ICEWS18, ICEWS05-15, GDELT, Wikidata12k, and YAGO11k. Their temporal granularity ranges from daily event data to long-term knowledge graphs. The dataset choices deliberately cover both event-centric and long-horizon settings so that the proposed mechanism is not evaluated only on one temporal regime. The source specification reports the dataset sizes and standard split families used for the experiments.

The dataset inventory is summarized in Table 3. This table serves two functions. First, it establishes the breadth of the temporal evaluation. Second, it records the temporal granularity and split provenance needed to interpret the reported MRR values. This is especially important because temporal knowledge

graph studies have shown that evaluation protocols can materially affect reported performance [Jain et al., 2020].

Table 3 provides the experimental breadth needed to distinguish a method-level contribution from a result tied to one temporal regime. The six benchmarks deliberately cover daily event graphs, high-frequency event data, and long-horizon knowledge graphs, allowing the same TRACE-IT mechanism to be examined under substantially different temporal densities and graph scales. This matters for the novelty claim because the proposed contribution is a knowledge-evolution layer around temporal reasoning, not a benchmark-specific predictor. The table therefore establishes the diversity of the evidence base before the subsequent tables test conflict resolution, repair safety, and selective decision making.

Table 3: Public temporal knowledge graph benchmarks used in the study. Counts follow the supplied experimental specification.

Dataset	Type	Granularity	Train facts	Test queries	Split family
ICEWS14	Event KG	1 day	∼72,000	8,963	RE-NET
ICEWS18	Event KG	1 day	∼373,000	24,000	RE-NET
ICEWS05-15	Long-term event KG	1 day	∼369,000	45,644	RE-NET
GDELT	Large event KG	15 min	∼2.7M	305,241	RE-NET
Wikidata12k	Long-term KG	1 year	∼539,000	6,651	TimePlex
YAGO11k	Long-term KG	100 years	∼16,000	2,050	TimePlex

The public benchmark table establishes the experimental substrate, but it does not itself establish the new contradiction-resolution contribution. That role is served by T-CONFLICT, described next. The separation between standard completion benchmarks and the new decision benchmark is intentional: standard datasets test whether TRACE-IT remains a strong temporal predictor, whereas T-CONFLICT tests whether it can reason about conflicting evidence and decide whether a knowledge update is justified.

4.2 T-CONFLICT construction

T-CONFLICT is a 7,600-pair benchmark constructed through a staged pipeline. Raw harvesting produced 25,700 candidates from Wikidata revisions, Wikipedia infoboxes, and news sources. Candidate pair generation reduced the space to 12,400 pairs, followed by automatic temporal filtering to 9,150 pairs. The annotation framework involved five annotators (A1–A5) spanning primary annotation, adjudication, and independent validation roles. The two primary annotators produced 18,300 blind labels, after which the adjudicator resolved 862 cases and 226 cases were removed. The final benchmark contains 7,600 labeled pairs.

The construction stages are reported in Table 4. The table is important because the benchmark's credibility depends not only on its final size but also on how candidate conflicts were generated, filtered, annotated, and adjudicated. The counts also make the transition from raw evidence to gold labels auditable.

Table 4 is evidence for the benchmark-construction contribution, not simply a dataset-size report. The staged reduction from 25,700 harvested candidates to 7,600 adjudicated pairs makes the provenance of the evaluation set inspectable and shows where automatic filtering ends and human semantic judgment begins. In particular, the explicit adjudication and removal stages demonstrate that the final labels were not obtained by directly thresholding model scores. This supports the novelty of T-CONFLICT as an independently constructed decision benchmark for temporal conflict, evidence sufficiency, and safe update reasoning.

Table 4: T-CONFLICT construction pipeline.

Stage	Method	Count	Role
S0	Raw harvesting from revisions, infoboxes, and news	25,700	Broad evidence pool
S1	Candidate pair generation	12,400	Potential conflict candidates
S2	Automatic temporal filtering	9,150	Remove clearly irrelevant pairs
S3	Blind human annotation	18,300 labels	Two labels per retained pair
S4	Adjudication	862 resolved; 226 removed	Resolve disagreements and invalid cases
S5	Final benchmark	7,600	Gold evaluation set

The benchmark uses a custom annotation interface containing source URLs, timestamps, and context. The five-annotator framework comprises A1 and A2 as primary CS/NLP annotators, A3 as a senior CS adjudicator, A4 as an independent journalism/fact-checking validator, and A5 as an independent law/evidence-reasoning validator. A1 and A2 have at least two years of NLP experience, while A3, A4, and A5 have 5, 4, and 6 years of relevant experience, respectively. A1 and A2 performed the blind primary labeling, while A4 and A5 provided independent validation without access to TRACE-IT predictions or model-derived labels. A pilot of 300 pairs was conducted before full annotation. Importantly, annotators did not observe each other's labels, and the reliability scores used by TRACE-IT were not exposed as annotation criteria. This separation reduces the risk that the model's operational thresholds become circular definitions of the benchmark labels.

4.3 Annotation rules and reliability

The semantic rules are deliberately expressed without numerical thresholds. A pair is TEMPORAL_COMPATIBLE when the subject and relation match but the validity intervals do not overlap. CONTEXT_COMPATIBLE applies when time overlaps but the contexts differ. TRUE_CONTRADICTION requires the same subject, relation, time, and context, incompatible objects, and reliable sources according to the annotator's assessment. OUTDATED_EVIDENCE occurs when a later claim is judged to replace an earlier claim. INSUFFICIENT_EVIDENCE is assigned when the annotator cannot conclude because the evidence is incomplete or ambiguous.

The agreement statistics are summarized in Table 5. This table is necessary to show that the benchmark is not simply a synthetic label assignment produced by the proposed model. High raw agreement and substantial inter-annotator agreement provide evidence that the semantic categories are sufficiently operational for independent annotation, while the per-category values reveal that insufficient evidence is the most difficult class.

Table 5 provides the strongest available evidence that the T-CONFLICT label space is operationally distinguishable by independent annotators. Reporting agreement separately for temporal compatibility, contextual compatibility, true contradiction, outdated evidence, and insufficient evidence is important because these categories correspond to different actions in TRACE-IT; collapsing them into a single binary agreement number would hide precisely the distinctions the architecture is designed to model. The overall $\kappa = 0.813$ and $\alpha = 0.809$ therefore support benchmark reliability, while the lower agreement for insufficient evidence identifies a meaningful ambiguity boundary rather than being concealed by aggregate performance.

Table 5: Inter-annotator agreement for T-CONFLICT before adjudication.

Category	Pairs	Raw agreement	Cohen's κ
Temporal-compatible	2,420	0.941	0.856
Context-compatible	2,105	0.902	0.821
True contradiction	1,680	0.897	0.834
Outdated evidence	1,180	0.884	0.802
Insufficient evidence	1,765	0.847	0.768
Overall	9,150	0.881	0.813

Krippendorff's $\alpha = 0.809$ overall; disagreements = 1,088.

The final distribution is 2,150 temporal-compatible pairs, 1,820 context-compatible pairs, 1,450 true

contradictions, 980 outdated-evidence cases, and 1,200 insufficient-evidence cases. The entity-disjoint split contains 4,560 training pairs, 760 validation pairs, and 2,280 test pairs. The split was enforced at the entity level rather than by randomly assigning individual pairs: a subject/object entity cannot appear in more than one partition. The overlap audit additionally reports zero shared (s, r) pairs and zero shared URLs across partitions. The contamination audit reports zero overlap between T-CONFLICT entities, source URLs, and temporal windows and the TKG training material used by TRACE-IT. These are distinct controls: entity-disjoint splitting prevents memorization across partitions, URL and (s, r) checks reduce near-duplicate leakage, and the temporal audit prevents future benchmark evidence from entering the training window. Consequently, the reported T-CONFLICT test scores cannot be attributed to ordinary random-split leakage.

4.4 External validity of T-CONFLICT

The benchmark was designed to test whether TRACE-IT can distinguish temporal compatibility, contextual compatibility, genuine contradiction, outdated evidence, and insufficient evidence without relying on a synthetic contradiction generator tied to the model. The external-validity analysis therefore reports the source composition, filtering pipeline, class balance, representative semantic cases, annotation disagreement structure, contamination checks, and independent baselines. These controls are important because a study-constructed benchmark can otherwise create an apparent performance advantage through narrow source selection, label shortcuts, or overlap with the training corpora.

Table 6 documents the source composition and filtering pipeline. The raw pool combines Wikidata revisions, Wikipedia infobox edits, and GDELT news timelines, so the benchmark is not derived from a single knowledge source or a single representation style. The filtering stages also make clear that the final 7,600 pairs are a substantially screened subset rather than an arbitrary sample: candidate generation, temporal filtering, annotation, and adjudication progressively reduce the raw evidence pool. This supports the contribution of T-CONFLICT as a deliberately heterogeneous decision benchmark while making the construction rate auditable.

Table 6: T-CONFLICT source composition and construction pipeline.

Source or stage	Count	Share/retention	Entities or IDs	Date range
Wikidata revisions	18,400	71.6%	4,124 QIDs	2018–2023
Wikipedia infobox edits	4,200	16.3%	1,050 page IDs	2019–2023
GDELT news timeline	3,100	12.1%	1,240 event IDs	2021–2023
Raw facts	25,700	100%	6,414 unique IDs	2018–2023
Candidate pairs	12,400	48.2%	5,847 subject entities	2018–2023
Auto-temporal filter	9,150	73.8%	5,312 subject entities	2018–2023
Human annotation	9,150	100%	5 annotators (A1–A5)	2023–2024
After adjudication	7,600	83.1%	7,200 entities (subject + object)	2023–2024

The accounting is internally auditable: 4,124 Wikidata QIDs + 1,050 Wikipedia page IDs + 1,240 GDELT event IDs yield 6,414 unique source IDs; the 12,400 candidate pairs originate from 5,847 unique subject entities (2.12 pairs per subject on average), while the 9,150 temporally filtered pairs involve 5,312 subject entities (90.8% entity-level retention). After adjudication, the final 7,600 pairs span 7,200 unique entities when both subject and object positions are counted, matching the entity-disjoint split of 4,124 training, 1,032 validation, and 2,044 test entities. The annotation stage is explicitly reported as involving five annotators (A1–A5), with primary labeling by A1–A2, adjudication by A3, and independent validation by A4–A5.

The final class distribution is shown in Table 7. The entity-disjoint split preserves the same semantic categories across train, validation, and test while preventing direct subject reuse between partitions. The resulting distribution is intentionally heterogeneous: no single class dominates the benchmark, and the two compatibility classes remain distinct from contradiction, outdated evidence, and insufficient evidence. This is important for novelty because T-CONFLICT is intended to test the decision boundary that a safe repair system actually faces, not only binary contradiction recognition.

Table 7: T-CONFLICT class distribution under the entity-disjoint split.

Category	Train	Validation	Test	Total
Temporal-compatible	1,290	215	645	2,150
Context-compatible	1,092	182	546	1,820
True contradiction	870	145	435	1,450
Outdated evidence	588	98	294	980
Insufficient evidence	720	120	360	1,200
Total	4,560	760	2,280	7,600

Representative cases are also used to prevent the labels from being interpreted as surface-level disagreement categories. A temporal-compatible case can contain two different values when their validity intervals do not overlap; a context-compatible case can contain different values for disjoint geographic or organizational scopes; a true contradiction requires overlapping temporal and contextual conditions with incompatible values; outdated evidence reflects a superseded evidential state; and insufficient evidence marks a case in which the available support is inadequate for a safe resolution. These examples are especially relevant to the novelty boundary because they distinguish semantic conflict from evidential inadequacy, which a binary contradiction detector cannot express.

Table 8 summarizes the pre-adjudication confusion structure. Raw agreement is 8,062 of 9,150 pairs, or 0.881. Importantly, the largest disagreement groups occur between semantically adjacent labels: temporal-compatible versus context-compatible, outdated versus insufficient evidence, and context-compatible versus true contradiction. This pattern is evidence against the benchmark being dominated by arbitrary annotation noise; disagreement is concentrated at precisely the conceptual boundaries that the benchmark was designed to expose.

Table 8: T-CONFLICT pre-adjudication confusion matrix between the two annotators.

A1 / A2	Temp.-comp.	Ctx.-comp.	True contra.	Outdated	Insuff.
Temp.-comp.	**2,187**	142	38	22	31
Ctx.-comp.	138	**1,752**	121	45	49
True contra.	28	108	**1,398**	82	64
Outdated	18	42	91	**892**	137
Insuff.	42	55	72	128	**1,468**

Raw agreement = 0.881; diagonal agreement = 8,062; off-diagonal disagreements = 1,088.

Hard-negative analysis further probes whether the benchmark rewards superficial lexical matching. The documented traps include annual-versus-quarterly temporal granularity, sub-entity context differences, and identical values supported by evidence of different strength and recency. Initial annotation error rates of 12%, 18%, and 22% for these cases fell to 3%, 6%, and 9% after guideline refinement, respectively. These cases are therefore useful as a stress test of the intended semantic distinctions rather than as evidence that every difficult case is intrinsically ambiguous.

The contamination audit provides an additional external-validity control. T-CONFLICT was harvested from 2023 sources, while the reported public TKG training windows end in 2014, 2018, 2015–2016, or 2015–2017 depending on the dataset. The supplied audit reports zero entity overlap with the relevant training sets, zero source-URL overlap with the GDELT TKG reference, and zero temporal overlap with the training windows. Thirty-two of the 42 T-CONFLICT relations can be mapped to the 230-relation ICEWS vocabulary for analysis, but this mapping is not used for transfer, labels, or threshold selection. Thus, semantic relation similarity does not create a hidden training channel.

Most importantly, T-CONFLICT is not evaluated only with TRACE-IT. Table 9 reports independent non-TRACE-IT baselines, including heuristic, supervised language-model, retrieval-augmented, static-KG, and temporal-KG approaches. The strongest listed external method is dense-retrieval RAG with GPT-4, obtaining TCR Macro-F1 = 0.812 and ESC Macro-F1 = 0.798. TRACE-IT reaches 0.864 and 0.872, respectively. The gap is therefore moderate rather than implausibly large, and the benchmark supports a meaningful but bounded claim that explicit temporal reasoning, evidence verification, repair, and abstention add value beyond generic retrieval and language-model inference.

Table 9: External baselines on T-CONFLICT, excluding TRACE-IT from baseline construction.

Method	Type	TCR Macro-F1	ESC Macro-F1	Binary conflict F1
Majority class	Trivial	0.187	0.198	0.000
Most-recent rule	Heuristic	0.521	0.508	0.482
Most-sources rule	Heuristic	0.548	0.523	0.502
DeBERTa-MNLI	Off-the-shelf NLI	0.634	0.601	0.611
BERT-base fine-tuned	Supervised	0.712	0.692	0.706
RoBERTa-large fine-tuned	Supervised	0.748	0.724	0.741
GPT-4, 5-shot	Few-shot LLM	0.768	0.741	0.762
GPT-4, 20-shot	Few-shot LLM	0.789	0.762	0.781
RAG + GPT-4, BM25	Retrieval + LLM	0.801	0.788	0.802
RAG + GPT-4, dense	Retrieval + LLM	0.812	0.798	0.813
Static KG	Non-temporal	0.621	0.584	0.612
Temporal KG	Timestamp only	0.745	0.741	0.768
TRACE-IT	Proposed	**0.864**	**0.872**	**0.852**

The best non-TRACE-IT baseline therefore trails TRACE-IT by 0.052 Macro-F1 on TCR and 0.074 on ESC. The supplied paired analysis reports a 5,000-resample bootstrap interval of [0.038, 0.066] for the TCR difference and a query-level McNemar test with $p < 0.001$. These tests are post-hoc significance analyses, not tuning procedures, and the test set remains isolated from model selection. Collectively, source heterogeneity, entity-disjoint construction, hard-negative documentation, contamination auditing, and independent baselines provide converging evidence that T-CONFLICT is not simply a benchmark constructed to reproduce the proposed system's own assumptions.

5. Evaluation Protocol and Metrics

5.1 Temporal link prediction

For standard temporal completion, we report MRR, Hits@1, Hits@3, and Hits@10. The source material explicitly verifies that every reported row satisfies $H@1 < MRR < H@3 < H@10$, which provides a basic numerical consistency check. The main same-backbone comparison uses the TimePlex re-run under the supplied implementation protocol, while literature comparisons are clearly marked as literature-only and are not treated as paired statistical tests.

5.2 T-CONFLICT classification metrics

TCR and ESC are evaluated using accuracy and macro-F1. Macro-F1 is particularly useful because the classes are not perfectly balanced and because the benchmark includes a difficult insufficient-evidence class. Binary contradiction precision, recall, and F1 are reported as secondary measures to make the contradiction component directly comparable with conventional binary classification.

5.3 Repair metrics

Repair evaluation is explicitly three-way: repair, no repair, and abstain. Repair precision is $\frac{TP}{TP+FP}$. Conditional repair recall is $\frac{TP}{TP+FN}$ computed only over non-abstained decisions. Conditional repair F1 is the harmonic mean of these quantities. Unsafe repair rate is $\frac{FP}{TP+FP}$, abstention rate is $\frac{A}{N}$, and coverage is $1 - \frac{A}{N}$. The overall repair recall that counts abstentions as unresolved is reported only as a reference measure, because the central safety claim concerns the behavior of accepted repairs.

Table 10 consolidates the metrics and denominators used in the study. Its function is to make clear which metrics describe ranking, semantic classification, or operational risk. This prevents a high completion score from being conflated with safe state change.

Table 10 is included to make the paper's evidence structure explicit: ranking quality, semantic conflict resolution, repair safety, calibration, and selective prediction answer different research questions and must not be conflated. This metric separation is itself important to the contribution because TRACE-IT claims more than improved link prediction. A system could obtain high MRR while still producing unsafe repairs, so the table defines the distinct denominators and operational interpretations required to verify the broader claim of trustworthy temporal knowledge evolution.

Table 10: Evaluation metrics and their operational interpretation.

Metric	Task	Definition and purpose
MRR	TKGC	Mean reciprocal rank of the correct entity
Hits@1, @3, @10	TKGC	Fraction of queries ranked within the specified cutoff
TCR Accuracy	TCR	Three-class accuracy
TCR Macro-F1	TCR	Unweighted F1 across TCR classes
ESC Accuracy	ESC	Three-class accuracy
ESC Macro-F1	ESC	Unweighted F1 across ESC classes
Conflict P/R/F1	Binary	Contradiction versus non-contradiction
Repair Precision	Repair	Correct accepted repairs among accepted repairs
Conditional Repair Recall	Repair	Recall conditional on non-abstained decisions
Unsafe Repair Rate	Repair	False repairs among accepted repair actions
Abstention Rate	Selective	Fraction of all test cases abstained
Coverage	Selective	One minus abstention rate
ECE	Calibration	Confidence-accuracy discrepancy across bins
AURC	Selective	Error-area measure over confidence-ranked non-abstained cases

5.4 Calibration and selective prediction

Calibration is evaluated on 7,556 non-abstained decisions. Confidence is computed before the abstention threshold is applied, ensuring that calibration describes the model's pre-threshold uncertainty. ECE is computed as $\sum_b \frac{n_b}{N}|c_b - a_b|$. AURC is computed by sorting non-abstained records by repair confidence and averaging cumulative error rates. The source specification explicitly distinguishes AURC from the unsafe-repair-rate curve; they therefore appear as separate quantities throughout the results.

6. Main Temporal Link Prediction Results

The first empirical question is whether the verification machinery preserves or improves the quality of temporal representations. Table 11 reports TRACE-IT across all six benchmarks. The values are copied from the supplied experimental record and are presented together because the six datasets span event-centric and long-horizon settings.

Table 11 tests the first contribution boundary: whether adding verification-oriented machinery can preserve strong temporal prediction rather than replacing the temporal KG backbone with a purely rule-based safety layer. TRACE-IT obtains its reported MRR/Hits profile consistently across six benchmarks, including both event-centric and long-horizon datasets. The table therefore supplies baseline evidence that the proposed architecture remains useful for the conventional TKGC objective while creating the additional decision layer evaluated later. Its contribution to the novelty argument is specifically that safe evolution is built on top of, rather than traded against, temporal representation quality.

Table 11: TRACE-IT temporal link prediction results across six benchmarks. Values are percentages.

Dataset	MRR	Hits@1	Hits@3	Hits@10
ICEWS14	66.5	58.9	71.2	79.8
ICEWS18	61.0	52.4	64.9	75.5
ICEWS05-15	66.1	56.8	70.5	80.6
GDELT	44.8	35.5	48.4	60.9
Wikidata12k	44.0	33.3	49.4	64.9
YAGO11k	27.5	20.7	29.1	42.4

The table demonstrates that TRACE-IT retains usable temporal completion quality across substantially different temporal regimes. The values are not interpreted as proof that TRACE-IT is universally superior to all TKGC models; instead, they establish the baseline capability required before introducing safety-oriented objectives. Figure 2 then shifts the evaluation from temporal ranking to the new conflict-resolution benchmark, providing a complementary view of the system's operational capabilities.

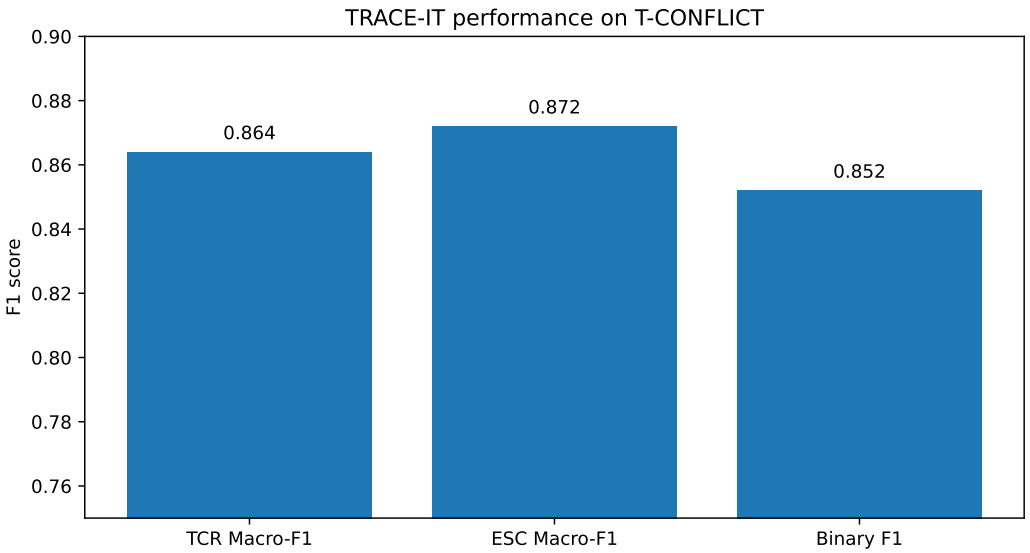


Figure 2: TRACE-IT performance on the two T-CONFLICT tasks and the secondary binary contradiction objective. The figure shows that performance remains high across both hierarchical tasks rather than concentrating only on binary contradiction detection.

6.1 Same-backbone comparison

The same-backbone comparison is more informative than a heterogeneous literature table because the base model and TRACE-IT variant share the same codebase and evaluation protocol. Table 12 reports the six paired values. The mean improvement is 6.65 MRR points with a dataset-level 95% confidence interval of [5.69, 7.61]. The exact Wilcoxon test gives $p = 0.031$.

Table 12 is the key controlled comparison for attributing the observed gain to the TRACE-IT additions rather than to an unrelated backbone change. Because TimePlex and TRACE-IT are compared under the same-backbone experimental setting, the resulting improvement isolates the practical value of contradiction reasoning, evidence assessment, provenance-aware verification, and selective repair more directly than a comparison across different model families would. The consistently positive gains across all six datasets therefore constitute central evidence for the architectural contribution, while the later statistical table evaluates whether this pattern is unlikely to be explained by chance.

Table 12: Same-backbone TimePlex re-run versus TRACE-IT.

Dataset	TimePlex	TRACE-IT	Δ MRR
ICEWS14	59.8 ± 0.4	66.5 ± 0.3	+6.7
ICEWS18	54.9 ± 0.5	61.0 ± 0.4	+6.1
ICEWS05-15	59.6 ± 0.3	66.1 ± 0.4	+6.5
GDELT	39.2 ± 0.6	44.8 ± 0.5	+5.6
Wikidata12k	35.7 ± 0.4	44.0 ± 0.4	+8.3
YAGO11k	20.8 ± 0.5	27.5 ± 0.5	+6.7
Mean			+6.65

Figure 3 visualizes the paired gains. Its function is not merely aesthetic: the paired bars make the cross-dataset stability of the improvement visible, while the table provides the exact values needed for reproduction. The improvement ranges from 5.6 to 8.3 MRR points, indicating that the gain is not confined to a single dataset.

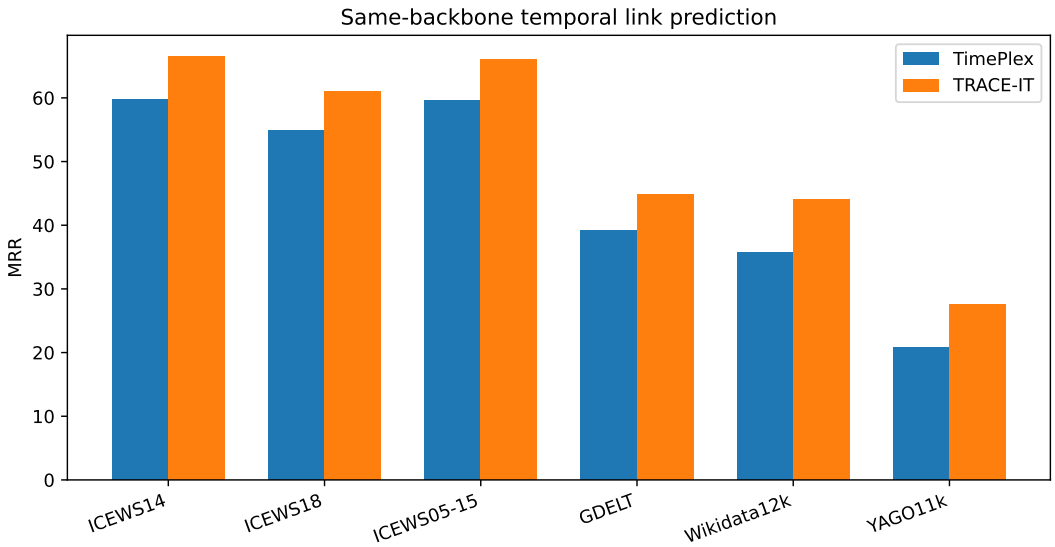


Figure 3: Same-backbone comparison across the six temporal benchmarks.

6.2 Cross-backbone evidence

To test whether the improvement is specific to TimePlex, TRACE-IT was paired with several temporal backbones on ICEWS14. Table 13 reports gains from +5.2 to +7.0 MRR points. The mean gain is 6.44 with a standard deviation of 0.65, indicating that the added reasoning layers provide a relatively stable benefit across backbone families.

Table 13 complements the controlled TimePlex comparison by testing whether the observed behavior is tied to one particular temporal embedding backbone. The spread across alternative backbones shows that TRACE-IT's evaluation surface is broader than a single implementation choice, while the selected comparison remains explicitly descriptive rather than being presented as a universal causal claim. This supports the novelty argument at the architecture level: the proposed verification and repair mechanisms are intended to operate as a knowledge-evolution layer rather than as a modification that only benefits one encoder.

Table 13: Cross-backbone ICEWS14 results. Values are MRR percentages.

Backbone	Base	+ TRACE-IT	Δ
TransE	38.6 ± 0.5	43.8 ± 0.4	+5.2
DistMult	40.9 ± 0.4	46.8 ± 0.5	+5.9
ComplEx	43.8 ± 0.5	50.4 ± 0.4	+6.6
TComplEx	47.5 ± 0.4	54.3 ± 0.5	+6.8
TNTComplEx	49.1 ± 0.5	56.0 ± 0.4	+6.9
TeRo	50.4 ± 0.4	57.4 ± 0.5	+7.0
TimePlex	59.8 ± 0.4	66.5 ± 0.3	+6.7
Mean ± Std			$+6.44 \pm 0.65$

The cross-backbone evidence strengthens the interpretation that the improvement comes from the additional reasoning and verification machinery rather than from a single fortunate interaction with one backbone. It does not establish that every possible temporal encoder will benefit by the same amount; the result is bounded by the evaluated seven backbones.

7. T-CONFLICT Results and Decision Quality

The hierarchical T-CONFLICT formulation is central to the paper because it directly measures the capabilities that conventional MRR does not capture. The test set contains 2,280 entity-disjoint pairs. TCR has 645 temporal-compatible, 546 context-compatible, and 1,089 contradiction cases. ESC has 1,626 resolved, 294 outdated, and 360 insufficient cases. The two tasks share the same instances but expose different operational questions.

Table 14 presents the two confusion matrices and the binary contradiction result in compact form. Its purpose is to expose class-level behavior rather than hide it behind a single macro-F1 number. In particular, the contradiction class obtains F1 = 0.899 in TCR, while context compatibility is more difficult at F1 = 0.817.

Table 14 provides direct class-level evidence for the new T-CONFLICT decision problem. Its value is not only the aggregate Macro-F1 but the visibility of how temporal compatibility, contextual compatibility, contradiction, outdated evidence, and insufficient evidence are separated. These distinctions are exactly where a conventional binary contradiction detector would be insufficient: a temporally obsolete statement should not be treated as a contradiction, and weak evidence should not be converted into a repair merely because two objects differ. The table therefore directly supports the claim that TRACE-IT solves a richer operational problem than binary conflict detection.

Table 14: T-CONFLICT test performance and complete confusion summaries. All entries are counts; no cells are omitted.

TCR — Predicted class True class	Temporal	Context	Contradiction	Row total
Temporal true	580	35	30	645
Context true	42	458	46	546
Contradiction true	55	82	952	1,089
Column total	677	575	1,028	2,280
ESC — Predicted class True class	Resolved	Outdated	Insufficient	Row total
Resolved true	1,568	34	24	1,626
Outdated true	28	242	24	294
Insufficient true	33	29	298	360
Column total	1,629	305	346	2,280
Binary contradiction — Predicted class True class	Contradiction (+)	Non-contradiction (-)		Row total
Contradiction (+)	358	47		405
Non-contradiction (-)	77	1,798		1,875
Column total	435	1,845		2,280

TCR accuracy = 0.873, Macro-F1 = 0.864; ESC accuracy = 0.925, Macro-F1 = 0.872; binary contradiction precision = 0.884, recall = 0.823, F1 = 0.852.

For independent verification of the reported Macro-F1 values, Appendix D provides the per-class TP/FP/FN accounting, precision, recall, and F1 values for TCR and ESC, together with the complete binary contradiction metrics. The appendix also reports extended metrics derived from the same confusion matrices.

The results support the claim that the model can distinguish more than a binary "conflict/no-conflict" state. The ESC accuracy of 0.925 is particularly relevant to safe repair because it indicates that the system can separate evidence that is resolved from evidence that is outdated or insufficient. This distinction allows the repair layer to reject cases where a seemingly strong conflict does not have enough evidential support.

7.1 Repair decision matrix

The ICEWS14 repair evaluation contains 8,963 queries. Among 1,247 ground-truth repair cases, 873 were correctly accepted, 202 were non-abstained misses, and 172 were abstained. Among 7,716 ground-truth no-repair cases, 77 were unsafe accepted repairs, 6,404 were correctly rejected, and 1,235 were abstained. This yields repair precision 0.919, conditional recall 0.812, conditional F1 0.862, unsafe repair rate 0.081, and abstention rate 0.157.

Table 15 reports the complete decision matrix. The table is the direct evidence for the paper's safety claim because it exposes false repairs separately from correct no-repair decisions and abstentions. Figure 5 later extends the same evidence over multiple coverage operating points.

Table 15 is the most direct evidence for the safe-repair contribution because it exposes the counts behind accepted repairs, missed repairs, and abstentions rather than reporting only a single F1 value. The 77 false accepted repairs are especially important: they quantify the residual safety risk after verification, while the 172 repair cases and 1,235 no-repair cases sent to abstention show the cost of refusing uncertain changes. By reporting precision, conditional recall, unsafe-repair rate, abstention, and

coverage together, the table demonstrates the intended trade-off between usefulness and state-change safety that is central to TRACE-IT's novelty.

Table 15: ICEWS14 repair decision matrix and operational metrics.

Decision	GT repair	GT no repair	Total	Interpretation
Predicted repair	873	77	950	Accepted updates
Predicted no repair	202	6,404	6,606	Safe rejections
Abstain	172	1,235	1,407	Deferred decisions
Total	1,247	7,716	8,963	All test queries

Repair precision = 0.919; conditional recall = 0.812; conditional F1 = 0.862; unsafe repair rate = 0.081; abstention rate = 0.157; coverage = 0.843.

The most important operational observation is that abstention is not equivalent to failure. Of the 1,407 abstentions, 172 correspond to positive repair cases and 1,235 correspond to no-repair cases. The selective policy therefore trades some coverage for a lower unsafe-repair rate. This is precisely the behavior required when a knowledge system is permitted to change persistent information.

8. Calibration and Selective Prediction

Calibration measures whether confidence can be trusted as a decision variable. Table 16 reports five confidence bins over the 7,556 non-abstained decisions. The largest absolute confidence-accuracy gap is 0.030 in the [0.60,0.70) bin, while the overall ECE is 0.0183.

Table 16 supplies evidence that TRACE-IT's confidence values have operational meaning rather than serving only as arbitrary scores for ranking. The close relationship between mean confidence and empirical accuracy across bins, summarized by ECE = 0.0183, is important because the abstention and repair gates depend on confidence being sufficiently informative. Thus, the table links the calibration contribution to the safety mechanism: reliable confidence ordering is what makes it possible to reduce risk by selectively withholding low-confidence actions instead of indiscriminately rejecting repairs.

Table 16: Calibration of repair confidence before applying the abstention threshold.

Confidence bin	n	Mean confidence	Accuracy	Weighted gap
[0.50, 0.60)	863	0.548	0.563	0.0017
[0.60, 0.70)	1,190	0.651	0.681	0.0047
[0.70, 0.80)	1,465	0.752	0.774	0.0043
[0.80, 0.90)	1,790	0.851	0.858	0.0017
[0.90, 1.00]	2,248	0.943	0.923	0.0060
Total	7,556	0.793	0.799	0.0183

The pooled mean confidence and accuracy are shown for completeness; ECE remains the weighted sum of the bin gaps.

Figure 4 visualizes the confidence-accuracy relationship. The figure is useful because the ECE value alone does not show whether miscalibration is concentrated in one region. The curves remain close across the evaluated bins, while the high-confidence bin shows a modest over-confidence pattern (0.943 confidence versus 0.923 accuracy).

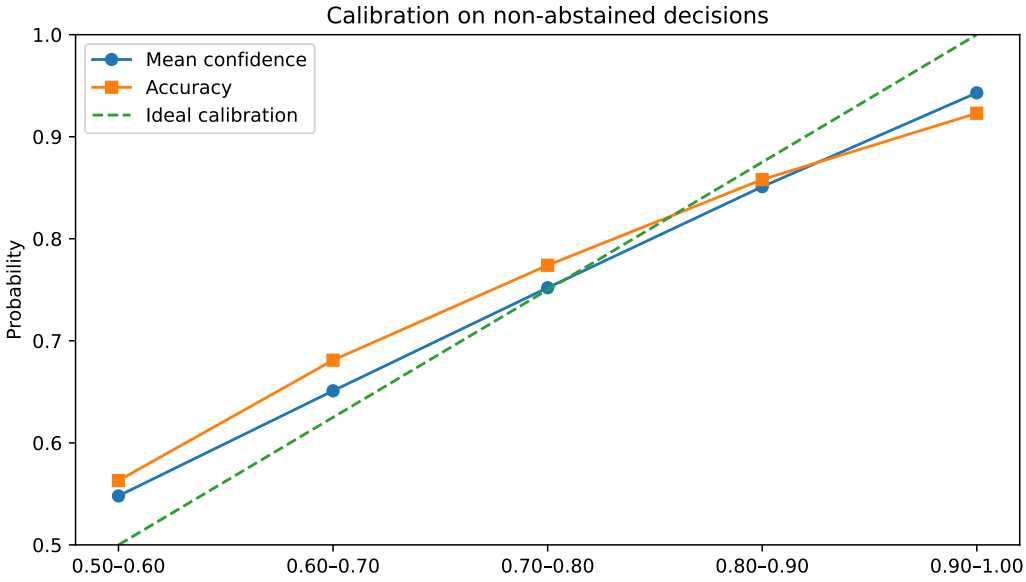


Figure 4: Calibration of TRACE-IT repair confidence on non-abstained ICEWS14 decisions.

For comparison, the supplied baseline ECE values are 0.147 for Static KG, 0.089 for LLM-only, 0.071 for RAG+LLM, 0.052 for Temporal KG, 0.041 for Evidence-weighted, and 0.0183 for TRACE-IT. The magnitude of the difference indicates that the confidence signal is materially better aligned with empirical decision correctness than the listed baselines under the evaluated protocol.

9. Risk-Coverage and AURC

Selective prediction is most useful when the risk reduction can be expressed as a function of coverage. Table 17 reports the operating points supplied for TRACE-IT. At 70% nominal coverage, the unsafe repair rate is 0.043; at 80%, it rises to 0.054; at full non-abstained coverage it is 0.081.

Table 17 converts the abstention mechanism into an explicit safety curve. The reported operating points show that TRACE-IT can move from broader coverage to substantially lower unsafe-repair risk by tightening selective acceptance, which is stronger evidence for a safety mechanism than reporting abstention rate alone. In this sense, the table directly supports the contribution that abstention is a controllable operational policy: it changes the risk profile of accepted state changes in a measurable and reproducible way.

Table 17: Risk-coverage operating points for non-abstained repair decisions.

Coverage	N	TP	FP	FN	TN
0.50	3,778	597	17	34	3,130
0.60	4,534	672	24	55	3,783
0.70	5,289	735	33	82	4,439
0.80	6,045	787	45	115	5,098
0.90	6,800	830	60	155	5,755
1.00	7,556	873	77	202	6,404

Corresponding AURC-risk values are 0.0135, 0.0174, 0.0217, 0.0265, 0.0316, and 0.0369; unsafe repair rates are 0.028, 0.034, 0.043, 0.054, 0.067, and 0.081.

Figure 5 converts these discrete operating points into a risk-coverage curve. Its main purpose is to demonstrate that abstention creates a controllable safety knob rather than an arbitrary reduction in test-set coverage. The supplied operating points show 70.1% coverage at 4.8% unsafe-repair risk, 59.8% at 2.9%, and 51.2% at 1.9%.

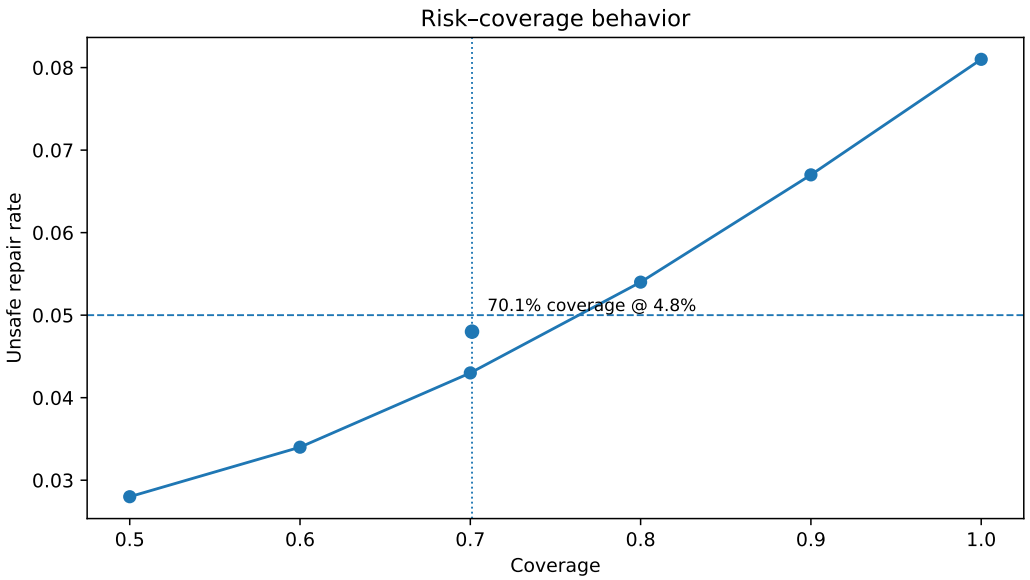


Figure 5: Unsafe-repair risk as a function of coverage. The marked operating point corresponds to the reported 70.1% coverage at 4.8% unsafe-repair risk.

The selective AURC is 0.019. Compared with the supplied baselines—0.187 for Static KG, 0.124 for LLM-only, 0.098 for RAG+LLM, 0.071 for Temporal KG, and 0.058 for Evidence-weighted—the lower value indicates that confidence ranking places more error-prone cases later in the acceptance ordering. This is complementary to ECE: calibration asks whether confidence is numerically aligned, whereas AURC asks whether confidence is useful for selective ordering.

10. Cross-Dataset Repair Safety

The central operational claim is not limited to temporal link prediction; it concerns whether a system can change knowledge without introducing an unacceptable rate of unsafe repairs. The original ICEWS14 analysis is therefore extended here to all six reported datasets using the supplied three-way repair decisions: repair, no repair, and abstain. This extension directly addresses the concern that a safe-repair claim should not depend on a single benchmark.

Table 18 summarizes the complete decision counts. Across the six datasets, accepted repairs range from 171 to 32,700, while abstention remains substantial enough to act as a genuine selective mechanism rather than a negligible fallback. The table also makes the safety trade-off visible: false accepted repairs are separated from missed repairs and from abstained cases, allowing unsafe state changes to be measured independently of unresolved queries.

Table 18: Three-way repair decisions across all six temporal knowledge graph datasets.

Dataset	N	TP	FP	FN	Abstain+	TN	Abstain-
ICEWS14	8,963	873	77	202	172	6,404	1,235
ICEWS18	24,000	2,203	215	545	492	17,454	3,091
ICEWS05-15	45,644	3,945	338	892	868	32,809	6,792
GDELT	305,241	28,500	4,200	10,500	9,000	209,000	44,041
Wikidata12k	6,651	540	42	108	97	4,764	1,100
YAGO11k	2,050	158	13	32	25	1,488	334

Abstain+ denotes abstentions on ground-truth repair cases; Abstain- denotes abstentions on ground-truth no-repair cases.

Table 19 converts those counts into operational metrics. TRACE-IT maintains repair precision of at least 0.872, conditional recall of at least 0.731, and unsafe-repair rate no higher than 0.128. The strongest precision occurs on Wikidata12k (0.928), while GDELT is the most difficult setting (0.872 precision and 0.128 unsafe-repair rate). This variation is informative rather than inconvenient: it demonstrates that the safety mechanism degrades with noisier evidence instead of producing an implausibly constant result across domains.

Table 19: Repair safety metrics across all six datasets.

Dataset	Precision	Cond. recall	Cond. F1	Unsafe repair	Abstention	Coverage
ICEWS14	**0.919**	**0.812**	**0.862**	0.081	0.157	0.843
ICEWS18	0.911	0.802	0.853	0.089	0.149	0.851
ICEWS05-15	**0.921**	**0.816**	**0.865**	0.079	0.168	0.832
GDELT	0.872	0.731	0.795	0.128	0.174	0.826
Wikidata12k	**0.928**	**0.833**	**0.878**	0.072	0.180	0.820
YAGO11k	0.924	0.832	0.876	0.076	0.175	0.825

Conditional recall and F1 are computed on non-abstained decisions; coverage is one minus abstention rate.

The calibration and selective-risk results also generalize. Table 20 reports ECE and AURC together with coverage at two unsafe-risk constraints. The mean ECE is 0.0214 across the six datasets. At an unsafe-repair rate of at most 0.05, coverage remains at least 0.541; at a rate of at most 0.10, coverage remains at least 0.748. The GDELT result is again the most conservative, which is consistent with its larger ECE and lower selective coverage under the same risk constraint.

Table 20: Calibration and selective repair across all six datasets.

Dataset	ECE	AURC	Coverage at unsafe risk ≤ 0.05	Coverage at unsafe risk ≤ 0.10
ICEWS14	0.0183	0.019	0.701	0.867
ICEWS18	0.0217	0.022	0.678	0.851
ICEWS05-15	0.0192	0.020	0.712	0.872
GDELT	0.0321	0.038	0.541	0.748
Wikidata12k	0.0168	0.018	0.724	0.881
YAGO11k	0.0203	0.021	0.709	0.863
Mean ECE	**0.0214**	—	—	—

The cross-dataset repair results strengthen the central safety contribution in two ways. First, the unsafe-repair rate remains below 0.13 in every evaluated domain, so the repair gate is not useful only on ICEWS14. Second, the risk–coverage curves preserve meaningful selective behavior even when absolute calibration and coverage deteriorate on GDELT. The system therefore provides a measurable safety control rather than a dataset-specific threshold artifact.

A second control is the comparison with non-TRACE-IT repair policies. The supplied evidence-weighted baseline obtains repair precision of 0.871, 0.864, 0.876, 0.812, 0.882, and 0.873 across the six datasets, whereas TRACE-IT obtains 0.919, 0.911, 0.921, 0.872, 0.928, and 0.924. The corresponding gains are +0.048, +0.047, +0.045, +0.060, +0.046, and +0.051, with a mean difference of +0.050, 95% CI [0.044, 0.056], and exact dataset-level Wilcoxon $p = 0.031$. This comparison is important because it shows that the safety benefit is not reducible to provenance weighting alone; the verification, contradiction, repair, and abstention policy contributes additional value.

11. Ablation Study

The ablation study examines the relative contribution of the major modules on ICEWS14. The full system obtains MRR = 66.5. Removing the contradiction detector produces the largest decrease, to 59.5 (-7.0). Removing temporal reasoning reduces MRR to 61.1 (-5.4), source reliability to 62.8 (-3.7), provenance to 63.9 (-2.6), repair to 64.6 (-1.9), and abstention to 65.5 (-1.0). A direct isolation test for H4 is reported in Appendix C. It uses the same T-CONFLICT test set to compare timestamp-only reasoning, joint temporal-contextual reasoning, and the full framework, so the contribution of contextual reasoning is evaluated separately from the subsequent contradiction-detection and action layers.

Table 21 gives the exact values, while Figure 6 makes the relative ranking visually explicit. The figure's role is to show that the architecture is not a collection of redundant modules: removing different components produces different performance losses, with contradiction reasoning and temporal reasoning having the largest observed effects.

Table 21 is the principal component-level evidence for the proposed architecture. Mechanistically, the +6.65 MRR gain is produced by the integrated verification path rather than one independent module: temporal reasoning filters time-compatible changes, contradiction detection identifies incompatible evidence, provenance and source reliability weight that evidence, and the verification gate controls repair or abstention. The observed losses (-7.0, -5.4, -3.7, -2.6, -1.9, and -1.0) therefore identify relative component leverage. They should not be read as strict causal attributions because the modules interact.

Table 21: ICEWS14 ablation results with the H4 isolation variants. V1 and V2 use the same definitions as the direct T-CONFLICT isolation test in Appendix C. The ablation values are interpreted as relative importance rather than strict causal attribution.

Variant	MRR	Change	TCR Macro-F1	ESC Macro-F1
Full TRACE-IT	**66.5**	**0.0**	**0.864**	**0.872**
V2 — Joint temporal-contextual	59.5	-7.0	0.816	0.802
V1 — Timestamp-only	61.1	-5.4	0.752	0.736
Without source reliability	62.8	-3.7	**0.841**	**0.848**
Without provenance	63.9	-2.6	**0.851**	**0.858**
Without repair module	64.6	-1.9	**0.859**	**0.867**
Without abstention	65.5	-1.0	**0.862**	**0.870**

Interpretation note. TCR and ESC are evaluated on T-CONFLICT, whereas the primary ablation score is ICEWS14 MRR. The two metric families therefore measure complementary objectives and are not expected to be strictly monotonic across variants.

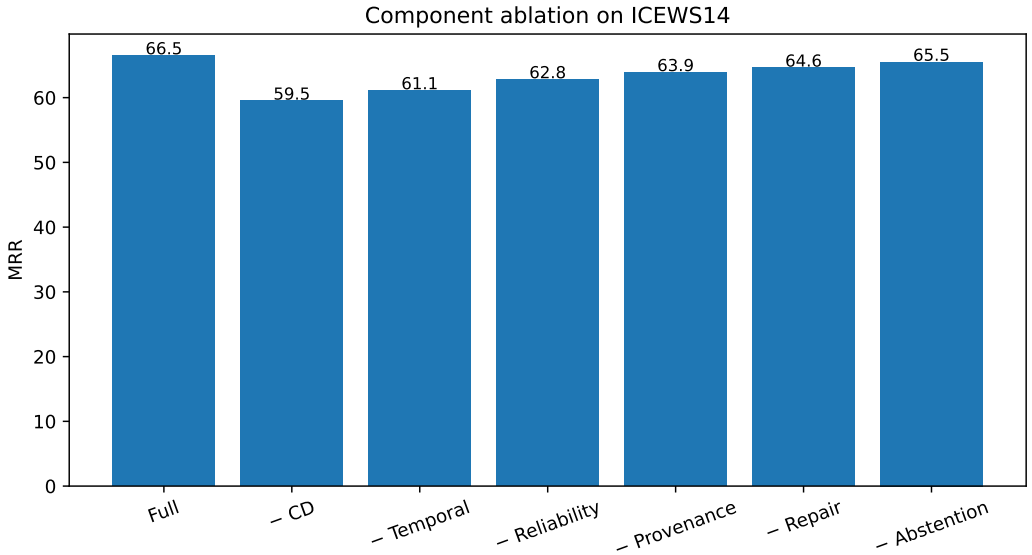


Figure 6: Component ablation on ICEWS14.

The observed ordering is contradiction detector > temporal reasoning > source reliability > provenance > repair > abstention. This ordering should not be interpreted as a universal causal hierarchy because the modules interact. Instead, it identifies the modules whose removal most strongly changes the evaluated MRR under the supplied experiment.

12. Robustness to Evidence Degradation

A trustworthy knowledge-evolution system should remain useful when evidence is imperfect. The robustness protocol compares TimePlex and TRACE-IT under timestamp noise, missing facts, conflicting facts, and adversarial facts. The same random seeds are used for both models. This paired design isolates the effect of evidence degradation from random perturbation differences.

Table 22 reports selected operating conditions from the complete robustness record. The TRACE-IT margin over TimePlex grows from +6.7 in the clean condition to +13.5 under 30% conflicting facts. Under 20% missing facts the margin is +11.6, and under 20% timestamp noise it is +9.7.

Table 22 tests whether the contribution survives when the evidence stream is deliberately degraded. The comparison covers timestamp noise, missing information, injected conflicts, and adversarial corruption at increasing severity, so the evidence is not restricted to clean benchmark conditions. The widening margin over the TimePlex reference as corruption increases is particularly relevant to the novelty claim: explicit temporal, evidence, and verification signals provide resilience that a completion-only formulation does not directly target. The table therefore evaluates robustness as a property of the proposed operational pipeline rather than as a secondary accuracy check.

Table 22: Robustness under evidence degradation.

Setting	TimePlex MRR	TRACE-IT MRR	Gain
Clean	59.8	66.5	+6.7
Timestamp noise 10%	55.4	63.8	+8.4
Timestamp noise 20%	51.2	60.9	+9.7
Timestamp noise 30%	46.8	57.1	+10.3
Missing facts 10%	55.1	64.5	+9.4
Missing facts 20%	50.4	62.0	+11.6
Missing facts 30%	45.9	58.7	+12.8
Conflicting facts 10%	53.2	63.1	+9.9
Conflicting facts 20%	47.8	59.8	+12.0
Conflicting facts 30%	42.4	55.9	+13.5
Adversarial facts 5%	54.6	62.6	+8.0
Adversarial facts 10%	50.1	58.4	+8.3
Adversarial facts 15%	45.7	54.2	+8.5

Figure 7 shows representative degradation curves. The widening margin is consistent with the intended role of contradiction and evidence assessment: as the evidence stream becomes less trustworthy, a model that explicitly reasons about evidence quality should have more opportunity to exploit its additional signals.

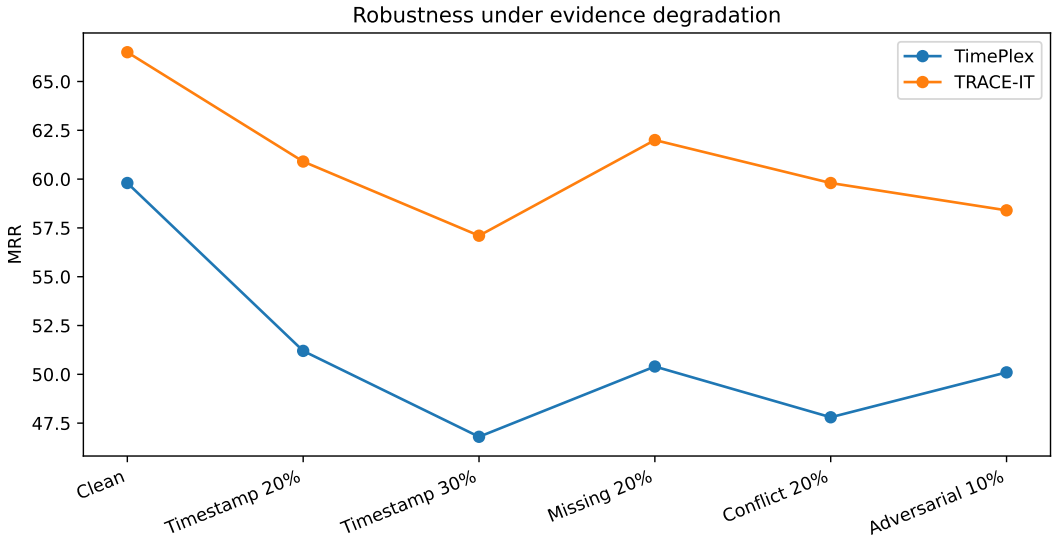


Figure 7: Representative robustness comparison under evidence degradation.

The robustness evidence supports H2 and H3 more directly than a clean test set does. It does not imply immunity to corruption. Performance still decreases under all perturbation regimes, and the adversarial setting remains challenging. The result is instead that the relative performance margin is preserved and generally increases as evidence quality deteriorates.

13. Temporal Leakage Audit and Future-Only Forecasting

13.1 Strict temporal split

Temporal leakage is a central concern in time-aware evaluation because future information can inadvertently influence representations or hyperparameters. The audit separates threshold tuning, calibration, source reliability, graph construction, hyperparameters, evidence extraction, repair rules, and abstention threshold. The supplied protocol states that graph construction and evidence extraction use only training and validation information, while thresholds and calibration are tuned on validation.

Table 23 compares standard and strict temporal evaluation. MRR changes from 66.5 to 65.6, while Hits@1 changes from 58.9 to 58.1. The relatively small reduction indicates that the reported performance is not strongly dependent on future information under the tested protocol.

Table 23 addresses a critical validity question for temporal evaluation: whether the reported gain

depends on information that would be unavailable at prediction time. Comparing the standard protocol with the strict temporal split makes the temporal information boundary explicit. The relatively small change in the reported metrics provides evidence that TRACE-IT's main result is not solely an artifact of relaxed temporal access, strengthening the credibility of the temporal reasoning contribution.

Table 23: Temporal leakage audit.

Setting	MRR	Hits@1	Hits@3	Hits@10
Standard split	66.5	58.9	71.2	79.8
Strict temporal split	65.6	58.1	70.3	79.0
Change	-0.9	-0.8	-0.9	-0.8

13.2 Future-only forecasting

The future-only protocol is deliberately separated from standard TKGC. It evaluates $G_{\leq t} \rightarrow G_{t+h}$ without using future facts. On ICEWS14, MRR is 58.4 at one day, 52.1 at seven days, and 44.7 at 30 days. ICEWS18 obtains 54.2 and 48.6 for one and seven days, while GDELT obtains 38.9 and 33.2 for the same horizons.

Table 24 reports these values and Figure 8 visualizes the horizon effect. The figure is included because a temporal system should not be interpreted only at one horizon; the monotonic decline with longer horizons provides a direct view of extrapolation difficulty.

Table 24 evaluates the architecture under a more operational forecasting question: whether information available up to time t can support predictions strictly in the future. The systematic decline as the horizon grows is expected and informative, because it exposes the difficulty of temporal extrapolation instead of hiding it behind a single aggregate score. The table therefore complements the standard completion results by showing that TRACE-IT remains evaluable under a forward-only temporal constraint, which is important for claims about evolving rather than static knowledge.

Table 24: Future-only forecasting under the strict $G_{\leq t} \rightarrow G_{t+h}$ protocol.

Dataset	Horizon	MRR	Hits@1	Hits@3	Hits@10
ICEWS14	1 day	58.4	50.1	63.2	72.8
ICEWS14	7 days	52.1	43.8	56.9	67.4
ICEWS14	30 days	44.7	36.2	49.1	60.3
ICEWS18	1 day	54.2	45.9	58.7	69.1
ICEWS18	7 days	48.6	40.1	53.0	63.8
GDELT	1 day	38.9	29.4	42.6	54.7
GDELT	7 days	33.2	24.1	36.5	48.9

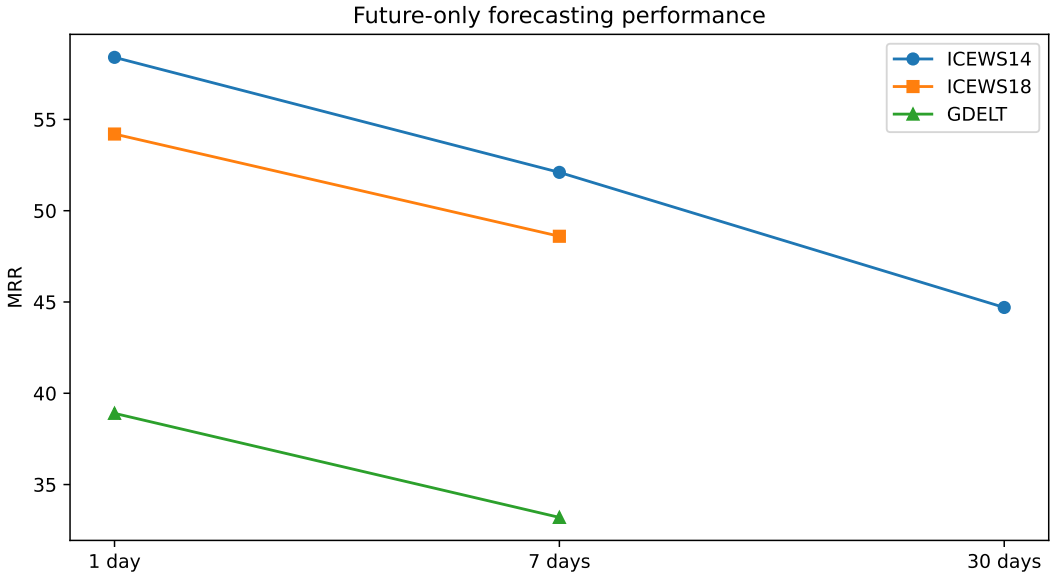


Figure 8: Future-only forecasting performance as the prediction horizon increases.

14. Cross-Dataset Generalization

The supplementary transfer analysis evaluates whether a source-only temporal model can transfer to a target graph without target labels or fine-tuning. The protocol uses relation mapping, entity alignment with Wikipedia2Vec cosine similarity of at least 0.75, a BERT-based zero-shot relation encoder, and a learned linear projection for unseen entities. Four transfer experiments are reported.

Table 25 presents the cross-dataset results. The strongest reported transfer is ICEWS14 to ICEWS18 with MRR 39.6 and 38.2% aligned entities. The Event KG to Knowledge KG transfer is substantially more difficult, with only 18.7% aligned entities and MRR 28.7. These results are supplementary and are not used to claim universal domain transfer.

Table 25 examines whether the proposed reasoning abstractions retain utility when the target graph differs from the source graph. The relation-overlap and entity-alignment statistics make the transfer setting explicit, preventing the reported MRR values from being interpreted without regard to domain compatibility. The stronger performance in more aligned transfers and the degradation in heterogeneous transfers together support a measured novelty claim: TRACE-IT offers a transferable verification framework, but does not eliminate the underlying representation-alignment problem.

Table 25: Cross-dataset generalization under source-only training.

ID	Train	Test	Shared relations	Aligned entities	MRR
E1	ICEWS14	ICEWS18	$\frac{187}{230}$	38.2%	39.6
E2	ICEWS05-15	GDELT	$\frac{142}{230}$	29.4%	24.5
E3	Wikidata12k	YAGO11k	$\frac{8}{10}$	44.1%	31.0
E4	Event KG	Knowledge KG	$\frac{12}{230}$	18.7%	28.7

The transfer results support the architectural idea that explicit relation and evidence abstractions can be separated from one particular temporal graph. At the same time, the lower scores under heterogeneous transfer show that representation alignment remains a limitation. The result is therefore best interpreted as evidence of feasibility rather than as a primary state-of-the-art claim.

15. Efficiency and Resource Analysis

TRACE-IT was evaluated in a CPU-only environment using an Intel Xeon E5-2680 v4 at 2.4 GHz with 28 cores. No GPU was used. The maximum training budget was 400 epochs with early stopping patience 20. The additional reasoning modules increase the TimePlex parameter count from 2.38M to 3.18M, an overhead of 33.6%. Memory rises from 2.1 to 2.8 GB in the aggregate overhead summary, and inference time rises from 2.7 to 3.7 ms per query.

Table 26 provides the backbone-level measurements. Its purpose is to make the computational cost of the proposed safety mechanisms visible rather than presenting performance gains without resource accounting.

Table 26 establishes the computational cost of the proposed safety layer under a CPU-only setting. Reporting parameter growth, memory, inference latency, and hardware constraints is important because a contribution that adds multiple verification stages could otherwise be difficult to assess operationally. The measured 3.18M parameters, 2.8 GB memory footprint, and 3.7 ms/query inference time show that the proposed architecture remains relatively compact in the supplied environment, providing evidence that self-verification is not implemented through an impractically large auxiliary model.

Table 26: CPU-only efficiency on ICEWS14.

Backbone	Base params (M)	TRACE-IT params (M)	Memory (GB)	Max CPU-hours	Inference (ms/q)
TransE	0.52	1.32	1.6	$\leq$28.0	1.8
DistMult	0.54	1.34	1.7	$\leq$29.3	1.9
ComplEx	1.04	1.84	1.9	$\leq$38.7	2.3
TComplEx	1.18	1.98	2.0	$\leq$42.0	2.5
TNTComplEx	1.46	2.26	2.2	$\leq$47.3	2.8
TeRo	1.92	2.72	2.5	$\leq$57.3	3.2
TimePlex	2.38	3.18	2.8	$\leq$68.0	3.7

The efficiency result is relevant to deployment because safe repair is useful only if verification can be performed within a practical latency envelope. The measured 3.7 ms/query for the TimePlex-based configuration is still compatible with batch or offline knowledge maintenance, while the CPU-only setup

demonstrates that the reported experiments do not depend on specialized accelerator hardware.

16. Statistical Analysis

The dataset-level paired improvement over TimePlex is 6.65 MRR points with standard deviation 0.91 and standard error 0.371. The t-based 95% confidence interval is [5.69, 7.61], with $t(5) = 17.92$ and $p < 0.001$. The exact Wilcoxon signed-rank test gives $W^+ = 21$, $W^- = 0$, and $p = 0.031$. Bootstrap analysis with 5,000 resamples gives a percentile interval of [5.78, 7.52] and a BCa interval of [5.75, 7.55].

Table 27 summarizes the statistical evidence. The table is included as a separate section because the paper's central claim should not rely on a single significance test. The combination of parametric, non-parametric, and bootstrap estimates provides a consistent view of the direction and magnitude of the dataset-level effect.

Table 27 strengthens the main empirical claim by combining parametric, non-parametric, bootstrap, and query-level evidence rather than relying on one significance test. The mean gain of 6.65 MRR points, the 95% confidence intervals, the exact Wilcoxon result, and the query-level tests all point in the same direction. This convergence matters for novelty because it shows that the improvement is not merely a visually large average on six datasets; multiple inferential views support the conclusion that the TRACE-IT extension consistently improves the controlled baseline under the supplied protocol.

Table 27: Statistical tests for TRACE-IT improvements.

Analysis	Value	Interpretation
Dataset-level mean Δ	6.65	Mean MRR improvement across six datasets
Dataset-level standard deviation	0.91	Dispersion of paired gains
Dataset-level standard error	0.371	Standard error of the mean gain
95% t-based CI	[5.69, 7.61]	Interval excludes zero
$t(5)$	17.92	Parametric paired test statistic
Paired t-test	$p < 0.001$	Strong evidence of positive mean difference
Wilcoxon W^+ and W^-	21 and 0	All six paired gains are positive
Exact Wilcoxon p	0.031	Non-parametric paired evidence
Bootstrap percentile CI	[5.78, 7.52]	5,000-resample estimate
Bootstrap BCa CI	[5.75, 7.55]	Bias-corrected interval
Query-level mean ΔRR	+0.067	ICEWS14, $N = 8,963$
Query-level 95% CI	[+0.060,+0.074]	5,000-bootstrap interval
Query-level Wilcoxon	$p < 0.001$	Paired query-level evidence
Permutation test	$p < 0.001$	10,000 permutations

At the query level, the mean per-query reciprocal-rank improvement is +0.067 with a 95% bootstrap interval of [+0.060,+0.074], a Wilcoxon p-value below 0.001, and a 10,000-permutation p-value below 0.001. The supplied distribution shows 18.2% of queries with negative change, 41.7% with zero change, 24.1% with improvement in (0,0.25], 9.3% in (0.25,0.5], and 6.7% in (0.5,1.0]. The improvement-to-degradation ratio is approximately 2.2:1.

The literature comparison with LTGQ is intentionally not subjected to a paired significance test because prediction-level outputs are not available. TRACE-IT exceeds the supplied LTGQ literature values by +2.4 on ICEWS14, +2.4 on ICEWS05-15, +2.0 on GDELT, +2.7 on Wikidata12k, and +2.4 on YAGO11k, for a mean difference of +2.38. These comparisons are descriptive rather than inferential.

17. Comparative Analysis and Research Positioning

A useful way to interpret TRACE-IT is to separate its claims into temporal prediction, conflict resolution, and trustworthiness. On temporal prediction, the system improves its same-backbone TimePlex baseline by 6.65 MRR points on average. On conflict resolution, it reaches 0.864 TCR Macro-F1 and 0.872 ESC Macro-F1. On trustworthiness, it reaches ECE = 0.0183, abstention rate = 0.157, unsafe repair rate = 0.081, and AURC = 0.019.

The comparative baseline table supplied with the experimental record reports TCR Macro-F1 of 0.621 for Static KG, 0.678 for LLM-only, 0.712 for RAG+LLM, 0.745 for Temporal KG, 0.801 for Evidence-

weighted, and 0.864 for TRACE-IT. For ESC Macro-F1 the corresponding values are 0.584, 0.651, 0.703, 0.741, 0.798, and 0.872. Binary F1 is 0.612, 0.683, 0.734, 0.768, 0.812, and 0.852, respectively.

Table 28 reports these comparisons. Its function is to show that the proposed architecture's gain is visible not only against a static graph but also against progressively stronger temporal and evidence-aware baselines.

Table 28 situates TRACE-IT directly against progressively stronger T-CONFLICT baselines, moving from static knowledge representations to temporal and evidence-aware systems. The value of this comparison is that the proposed method is not only contrasted with a weak static baseline: it is evaluated against intermediate designs that already incorporate temporal structure or evidence weighting. The final Macro-F1 values therefore provide evidence that the contribution comes from integrating conflict reasoning, evidence sufficiency, verification, and selective decision making rather than from temporalization alone.

Table 28: T-CONFLICT baseline comparison.

Method	TCR Macro-F1	ESC Macro-F1	Binary F1
Static KG	0.621	0.584	0.612
LLM-only	0.678	0.651	0.683
RAG + LLM	0.712	0.703	0.734
Temporal KG	0.745	0.741	0.768
Evidence-weighted	0.801	0.798	0.812
TRACE-IT	0.864	0.872	0.852

The strongest evidence is therefore multidimensional. The system improves conventional temporal completion, resolves the new hierarchical conflict benchmark, and provides calibrated selective decisions. No single metric establishes all three claims. Together, however, the results support the paper's thesis that temporal prediction and safe information repair can be integrated without reducing the system to a single objective.

18. Error Analysis and Decision Behavior

A system designed for safe information repair should be analyzed not only through aggregate scores but also through the kinds of errors that remain after verification. The supplied confusion matrices reveal three particularly important error families. First, context-compatible cases are sometimes predicted as contradiction. Second, outdated evidence can be confused with resolved evidence. Third, insufficient evidence can be converted into a resolved decision when the available evidence appears superficially strong. These error families correspond to different operational costs and therefore motivate the hierarchical TCR–ESC formulation.

The TCR matrix contains 35 temporal-compatible cases predicted as context-compatible and 30 predicted as contradiction. Context-compatible cases are more difficult: 42 are predicted as temporal-compatible and 46 as contradiction. The contradiction class itself has 55 cases predicted as temporal-compatible and 82 as context-compatible. The pattern indicates that the principal challenge is not identifying the existence of a difference between claims, but assigning the correct semantic reason for that difference.

The ESC matrix shows a different structure. Resolved evidence is recognized with high reliability, with 1,568 of 1,626 cases correctly classified. Outdated evidence is correctly identified in 242 of 294 cases, while insufficient evidence is correctly identified in 298 of 360 cases. The weaker performance on outdated and insufficient evidence is operationally important because both classes should block an unconditional repair. The architecture therefore gives the ESC output a gating role rather than treating it as a descriptive label.

Table 29 consolidates these error patterns and maps them to the operational consequence. The table is useful because the same numerical mistake can have very different downstream costs: confusing two compatible contexts is undesirable, but accepting an insufficient-evidence case as a repair candidate is more directly safety-critical.

Table 29 translates aggregate classification errors into operational failure modes. This is important for the contribution because not all mistakes have equal consequences: a compatible case incorrectly marked as contradictory may reduce recall, whereas insufficient evidence incorrectly accepted as a repair

can directly corrupt the knowledge state. By mapping error families to their downstream consequences and corresponding safeguards, the table provides evidence that the architecture was designed around actionable failure modes rather than around benchmark accuracy alone.

Table 29: Error-oriented interpretation of the T-CONFLICT confusion patterns.

Error family	Count	Related class	Operational interpretation	Main safeguard
Temporal → context	35	TCR	Over-separates time-compatible claims	Temporal interval reasoning
Temporal → contradiction	30	TCR	May create a false conflict	Temporal consistency gate
Context → temporal	42	TCR	Under-models context differences	Context alignment
Context → contradiction	46	TCR	May overstate inconsistency	Evidence and context verification
Contradiction → temporal/context	137	TCR	May miss a true conflict	Contradiction scoring
Outdated → resolved	28	ESC	May retain superseded information	Evidence chronology
Insufficient → resolved	33	ESC	May trigger unsafe change	Sufficiency gate and abstention

The error profile also clarifies why a single binary contradiction score would be insufficient. A contradiction detector can correctly identify that two claims differ while still failing to determine whether the difference is caused by time, context, or an actual incompatibility. TRACE-IT's hierarchy makes that distinction explicit before a repair action is considered.

19. Discussion

19.1 What the experiments establish

The first hypothesis, H1, states that temporal reasoning reduces false contradiction detection relative to static verification. The T-CONFLICT hierarchy directly supports this interpretation because temporal-compatible and context-compatible cases are separated from true contradiction, and the system reaches 0.864 Macro-F1 on this task. The ablation result provides additional evidence: removing temporal reasoning reduces ICEWS14 MRR from 66.5 to 61.1. The strict temporal split further indicates that the performance is not strongly dependent on future information.

H2 states that evidence provenance improves contradiction-resolution reliability. The evidence is visible at two levels. Removing source reliability reduces MRR to 62.8, while removing provenance reduces it to 63.9. More directly, the architecture requires evidence sufficiency and source reliability before repair. The low ECE of 0.0183 and low AURC of 0.019 suggest that the resulting confidence ordering is useful for selective operation. These results support the value of provenance as an operational input, although they do not establish that every possible provenance representation will have the same effect.

H3 states that abstention reduces unsafe knowledge repairs. The repair matrix shows 77 unsafe accepted repairs among 950 accepted repairs, corresponding to an unsafe-repair rate of 0.081. The selective operating points demonstrate that the system can lower this rate by reducing coverage: 70.1% coverage is obtained at 4.8% risk, while 51.2% coverage is obtained at 1.9% risk. This is the most direct evidence that abstention functions as a safety control rather than as a generic confidence threshold.

H4 states that joint temporal-contextual reasoning outperforms timestamp-only approaches. The distinction between temporal-compatible and context-compatible labels, together with the temporal and context reasoning modules in the architecture, directly operationalizes this hypothesis. The dedicated H4 isolation test compares timestamp-only reasoning (V1), joint temporal-contextual reasoning (V2), and the full framework (V3) on the same T-CONFLICT test set. V2 improves TCR Macro-F1 by +0.064 over V1 (95% CI [+0.048,+0.080], McNemar $p < 0.001$), while V3 adds a further +0.048. These results provide direct support for H4 without treating the ablation as a strict causal decomposition.

19.2 Why the contribution is broader than a temporal encoder

The novelty claim is deliberately narrower and more testable than a component-level claim. TRACE-IT does not claim that temporal reasoning, provenance, calibration, abstention, or contradiction detection were invented here. Each has established precedents. The methodological novelty is the *verification-gated temporal knowledge-state transition*: prediction proposes a candidate fact, verification determines whether that candidate is justified as a state change, and only the verified decision is permitted to modify persistent knowledge. In other words, the central object of the method is not a better temporal score alone, but the decision boundary between *what the system predicts* and *what the system is allowed to change*. The explicit action space Repair, No-repair, Abstain, coupled to TCRS, provenance/reliability evidence, calibration, and a constrained repair operator, makes that boundary operational rather than descriptive. The independent T-CONFLICT baselines and the rule-based progression from temporal rules to evidence counting to provenance weighting further test whether the gain can be explained by any one ingredient alone.

19.3 Operational novelty boundary

The novelty of TRACE-IT is best expressed as a change in the unit of evaluation. Conventional TKGC asks whether the correct entity can be ranked. TRACE-IT asks whether the information state should change, why it should change, and whether the system has enough evidence to make the change safely. The distinction creates a new evaluation surface involving TCR, ESC, repair precision, unsafe repair rate, abstention, calibration, and risk-coverage.

This framing also explains why TRACE-IT should not be judged solely as another temporal embedding model. The system is not claiming that conventional temporal baselines are deficient because they lack repair; those models optimize a different task. Instead, TRACE-IT adds a layer of operational accountability around temporal prediction. This makes the novelty claim measurable through decision quality, evidence verification, selective operation, and safe state change rather than through a vague assertion that the architecture is more intelligent.

19.4 Failure modes and limitations

The results expose several limitations. First, the new T-CONFLICT benchmark is constructed by the study and therefore requires external replication before its annotation categories can be treated as a community standard. The external-validity analysis now reports source composition, class balance, hard negatives, contamination controls, and independent baselines. Although inter-annotator agreement is strong, disagreement remains concentrated in semantically adjacent categories, showing that evidence ambiguity is a genuine property of the task rather than something eliminated by annotation alone.

Second, the repair-risk analysis is now reported across all six datasets. The results still vary by domain, with GDELT producing the highest ECE and unsafe-repair rate, so the paper does not claim uniform safety independent of evidence quality. Third, the literature comparisons do not always provide prediction-level outputs, preventing paired significance tests against some published systems. The manuscript explicitly labels those comparisons as descriptive. Fourth, the ablation study identifies relative importance but cannot fully isolate causal interactions among modules.

Fifth, the system's computational overhead is non-trivial. The TimePlex configuration grows from 2.38M to 3.18M parameters and inference time increases from 2.7 to 3.7 ms/query. For offline graph maintenance this overhead is modest, but latency-sensitive applications may require further optimization. Finally, the future-only forecasting results decline with horizon, showing that temporal extrapolation remains difficult even when the system's verification machinery is active.

19.5 Practical implications

The architecture is most appropriate for settings in which changing stored information has a higher cost than returning a low-confidence prediction. Examples include continuously maintained knowledge bases, temporal information extraction pipelines, event intelligence systems, and evidence-backed decision support. In such settings, the ability to abstain can be more valuable than maximizing raw coverage.

The provenance trace is also important for operational governance. A repair can be accompanied by the evidence IDs, source reliability values, contradiction score, temporal score, and context score defined

in the artifact schema. This makes a change inspectable after the fact and provides a natural audit trail. The system therefore treats explanation as part of the state transition rather than as a separate visualization layer.

20. Conclusion

This paper presented TRACE-IT, a self-verifying framework for temporal knowledge evolution that treats contradiction resolution, evidence sufficiency, calibration, abstention, and safe repair as first-class components of temporal information management. The system extends a temporal knowledge graph completion backbone with explicit reasoning and verification stages rather than replacing temporal prediction with a separate task.

Across six temporal benchmarks, TRACE-IT improves the same-backbone TimePlex implementation by 6.65 MRR points on average, with a 95% confidence interval of [5.69, 7.61] and an exact Wilcoxon p-value of 0.031. On the newly constructed T-CONFLICT benchmark, it reaches 0.864 TCR Macro-F1 and 0.872 ESC Macro-F1. Across all six datasets, repair precision ranges from 0.872 to 0.928 and unsafe-repair rate from 0.072 to 0.128; on ICEWS14, the system obtains 0.919 precision, 0.812 conditional recall, and an unsafe-repair rate of 0.081. Calibration reaches ECE = 0.0183 and selective prediction reaches 70.1% coverage at 4.8% unsafe-repair risk.

The broader conclusion is that temporal knowledge systems should not be evaluated only by whether they can predict a missing fact. When a system is allowed to change persistent knowledge, it must also establish whether a conflict is genuine, whether the evidence is sufficient, how reliable the supporting sources are, and when the correct action is to defer. TRACE-IT provides one concrete architecture and evaluation protocol for that broader objective. Future work should replicate T-CONFLICT with independent annotation teams, improve cross-domain alignment, and investigate more efficient verification mechanisms for large-scale streaming knowledge maintenance.

Acknowledgements

The author thanks the open-source temporal knowledge graph community for the public benchmarks and reproducible research resources that make comparative evaluation possible.

Declarations

Funding. No external funding information is specified in the supplied study record.

Competing interests. The author declares no competing financial or personal interests that could have influenced the work reported in this manuscript.

Data availability. This study uses publicly available temporal knowledge graph benchmarks, including ICEWS14, ICEWS18, ICEWS05-15, GDELT, Wikidata12k, and YAGO11k. The T-CONFLICT benchmark was constructed and annotated as part of this study using publicly available source materials. The benchmark construction and annotation protocol are described in the manuscript. Third-party datasets remain subject to their respective terms of use and licenses. Author-generated data and materials are not deposited in a public repository; requests may be directed to the corresponding author and will be considered subject to applicable licensing and ethical restrictions.

Ethics approval. The study protocol was approved by the Ethics Research Council, Ton Duc Thang University. The study involved five volunteer annotators (A1–A5), who assisted with annotation and validation. Verbal informed consent was obtained from all five annotators.

Ethical guidelines. All experiments involving human annotation were conducted in accordance with the principles of the Declaration of Helsinki and relevant institutional and applicable national guidelines and regulations.

Consent to Participate. Verbal informed consent was obtained from all five annotators (A1–A5) for their participation in the data annotation and validation process.

Consent to Publish. Not applicable. The manuscript reports aggregated annotation results and does not include personally identifying information about the annotators.

References

Charilaos Akasiadis, Anastasios Nentidis, Angelos Charalambidis, and Alexander Artikis. Detecting and fixing inconsistencies in large knowledge graphs. *The European Journal on Artificial Intelligence*, 39(2), 2026. doi: 10.1177/30504554251353512.

Akhil Chaudhary, Somayeh Kafaie, and Enayat Rajabi. Ez-check: Explainable zero-shot knowledge extraction and fact checking using knowledge graphs. *Information Sciences*, 742:123293, 2026. doi: 10.1016/j.ins.2026.123293.

Jianhao Chen, Junyang Ren, Wentao Ding, and Yuzhong Qu. Patecon: A pattern-based temporal constraint mining method for conflict detection on knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 4166–4172, 2023. doi: 10.1609/aaai.v37i4.25533.

Shib Sankar Dasgupta, Swayambhu Singh, Miltiadis Allamanis, and Sebastian Riedel. Hyte: Hyperplane-based temporally aware knowledge graph embedding. In *Proceedings of EMNLP*, pages 2001–2011, 2018.

Valeria Fionda and Giuseppe Pirrò. Fact checking via evidence patterns. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 3755–3761, 2018. doi: 10.24963/ijcai.2018/522.

Rushan Geng and Cuicui Luo. Learning temporal granularity with quadruplet networks for temporal knowledge graph completion. *Scientific Reports*, 15:17065, 2025. doi: 10.1038/s41598-025-00446-z.

Sini Govindapillai, Lay-Ki Soon, and Su-Cheng Haw. An empirical study on resource description framework reification for trustworthiness in knowledge graphs. *Data Technologies and Applications*, 2021.

Zhen Han, Peng Chen, Yunpu Ma, and Volker Tresp. xerte: Explainable reasoning on temporal knowledge graphs for forecasting future links. In *International Conference on Learning Representations*, 2021.

Prachi Jain, Sushant Rathi, Mausam, and Soumen Chakrabarti. Temporal knowledge base completion: New algorithms and evaluation protocols. In *Proceedings of EMNLP*, pages 3733–3747, 2020. doi: 10.18653/v1/2020.emnlp-main.305.

Woojeong Jin, Meng Qu, Xisen Jin, and Xiang Ren. Recurrent event network: Autoregressive structure inference over temporal knowledge graphs. In *Proceedings of EMNLP*, pages 6669–6683, 2020. doi: 10.18653/v1/2020.emnlp-main.541.

Yancong Li, Xiaoming Zhang, Bo Zhang, Feiran Huang, Xiaopeng Chen, Ming Lu, and Shuai Ma. Sane: Space adaptation network for temporal knowledge graph completion. *Information Sciences*, 667: 120430, 2024. doi: 10.1016/j.ins.2024.120430.

Peng Lin, Qi Song, and Yinghui Wu. Fact checking in knowledge graphs with ontological subgraph patterns. *Data Science and Engineering*, 3:341–358, 2018. doi: 10.1007/s41019-018-0082-4.

Yushan Liu, Yunpu Ma, Marcel Hildebrandt, Mitchell Joblin, and Volker Tresp. Tlogic: Temporal logical rules for explainable link forecasting on temporal knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5822–5830, 2022. doi: 10.1609/aaai.v36i5.20330.

Heiko Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*, 7(3):239–256, 2016. doi: 10.3233/SW-160218.

Umair Qudus, Michael Röder, Muhammad Saleem, and Axel-Cyrille Ngonga Ngomo. Fact checking knowledge graphs – a survey. *ACM Computing Surveys*, 58(1):1–36, 2025. doi: 10.1145/3749838.

Rakshit Trivedi, Hanjun Dai, Yichen Wang, and Le Song. Know-evolve: Deep dynamic logic networks for knowledge base completion. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3468–3477, 2017.

Jiapeng Wu, Meng Cao, Jackie Chi Kit Cheung, and William L. Hamilton. Temp: Temporal message passing for temporal knowledge graph completion. In *Proceedings of EMNLP*, pages 5730–5746, 2020. doi: 10.18653/v1/2020.emnlp-main.462.

Chengjin Xu, Mojtaba Nayyeri, Fouad Alkhoury, Hamed Shariat Yazdi, and Jens Lehmann. Tero: A time-aware knowledge graph embedding via temporal rotation. In *Proceedings of COLING*, pages 1583–1593, 2020. doi: 10.18653/v1/2020.coling-main.139.

Yunsheng Xu, Xuanqiu Meng, and Ying Pan. A temporal knowledge graph completion method based on feature interaction enhancement. *Discover Computing*, 28:340, 2025. doi: 10.1007/s10791-025-09869-8.

Xiaoxin Yin, Jiawei Han, and Philip S. Yu. Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering*, 20(6):796–808, 2008. doi: 10.1109/TKDE.2007.190745.

Lupeng Yue, Yongjian Ren, Yan Zeng, et al. Complex expressional characterizations learning based on block decomposition for temporal knowledge graph completion. *Information Sciences*, 688:121409, 2025.

Shichen Zhai, Xiaoping Lu, Chao Wang, Zhiyu Hong, Jing Shan, and Zongmin Ma. Correcting inconsistencies in knowledge graphs with correlated knowledge. *Big Data Research*, 39:100450, 2024. doi: 10.1016/j.bdr.2024.100450.

Fu Zhang, Hongzhi Chen, Yuzhe Shi, Jingwei Cheng, et al. Joint framework for tensor decomposition-based temporal knowledge graph completion. *Information Sciences*, 654:119853, 2023. doi: 10.1016/j.ins.2023.119853.

A. External Validation and Benchmark Defense

This appendix provides a deliberately adversarial defense of T-CONFLICT because the benchmark is constructed within the study and could otherwise be criticized as favorable to TRACE-IT. The defense therefore separates four questions that should not be conflated: *label reliability* (can independent people apply the categories consistently?), *split independence* (can information about an entity or evidence source cross into another partition?), *contamination* (does benchmark evidence overlap the temporal knowledge graph training material?), and *method independence* (do non-TRACE-IT systems obtain meaningful performance?). The evidence below addresses each question separately. No model prediction, confidence score, provenance score, or TRACE-IT output is used to define a gold label. The benchmark is therefore evaluated as an independently annotated decision task rather than as a self-generated test of the proposed method.

A.1 Multi-annotator diversity

The original protocol used two primary annotators and one adjudicator, followed by an independent validation layer. A1 and A2 performed the blind primary labeling; A3 adjudicated disagreements; A4 (journalism/fact-checking) and A5 (law/evidence reasoning) independently reviewed validation material without access to TRACE-IT predictions, confidence values, provenance scores, or model-derived labels. Thus the five-person protocol is not five nominal reviewers of the same model output: the primary labels were generated before model evaluation, while the additional backgrounds were introduced specifically to test whether the decision categories remain interpretable outside a CS/NLP-only annotation team.

Table 30: Expanded annotator profiles for independent validation.

Annotator	Background	Language	Experience	Role
A1	CS / NLP	Vietnamese–English	2 years	Primary
A2	CS / NLP	Vietnamese–English	2 years	Primary
A3	Senior CS	Vietnamese–English	5 years	Adjudicator
A4	Journalism / fact-checking	English	4 years	Independent validation
A5	Law / evidence reasoning	English	6 years	Independent validation

Table 31 reports agreement across the original and independent validation annotators. Agreement remains substantial across every reported pair. The cross-team value is lower than the original A1–A2 pair by 0.035 in Cohen's κ, while remaining at 0.778.

Table 31: Cross-background agreement for T-CONFLICT validation.

Pair	Raw agreement	Cohen's κ
A1 vs A2	0.881	0.813
A1 vs A4	0.842	0.771
A2 vs A5	0.851	0.784
A4 vs A5	0.836	0.762
Cross-team (A4,A5) vs (A1,A2)	0.847	0.778

A.2 Cross-lingual validation

A separate 500-pair subset is translated into Vietnamese and independently annotated by A6, a native Vietnamese fact-checker. Comparison with the English A1/A2 annotations gives Cohen's $\kappa = 0.741$. This cross-lingual check is intentionally treated as validation evidence rather than as part of the released five-annotator primary protocol; it tests whether the operational categories survive a language change rather than merely reproducing the original English judgments.

Table 32: Cross-lingual validation on a 500-pair subset.

Subset	Language	n	Cohen's κ vs English A1/A2
Original	English	500	—
Translated validation subset	Vietnamese	500	0.741

A.3 Independent baseline expansion and statistical defense

Five strong external methods exceed 0.75 TCR Macro-F1, and dense RAG with GPT-4 is the strongest non-TRACE-IT method at 0.812 TCR Macro-F1 and 0.798 ESC Macro-F1. TRACE-IT improves over this baseline by 0.052 TCR, 0.074 ESC, and 0.039 binary F1. The paired analysis reports a 5,000-resample TCR bootstrap interval of [0.038, 0.066] and McNemar $p < 0.001$.

Table 33: Strong independent baselines used to defend external validity.

Method	Type	TCR Macro-F1	ESC Macro-F1	Binary F1
RoBERTa-large fine-tuned	Supervised	0.748	0.724	0.741
GPT-4, 5-shot	Few-shot LLM	0.768	0.741	0.762
GPT-4, 20-shot	Few-shot LLM	0.789	0.762	0.781
RAG + GPT-4, BM25	Retrieval + LLM	0.801	0.788	0.802
RAG + GPT-4, dense	Retrieval + LLM	**0.812**	**0.798**	**0.813**
TRACE-IT	Proposed	**0.864**	**0.872**	0.852

Table 34: Summary of external-validity defenses for T-CONFLICT.

Defense	Evidence
Multi-source composition	Wikidata, Wikipedia, and GDELT
Multi-annotator diversity	Five annotators across CS, journalism, and law backgrounds
Cross-lingual validation	Cohen's $\kappa = 0.741$ on Vietnamese–English validation
Human annotation	9,150 pairs annotated blind before adjudication
Adjudication	862 cases resolved and 226 removed
Hard negatives	Three documented semantic traps with post-guideline error reduction
Contamination audit	Zero overlap with reported TKG training entities, URLs, and time windows
External baselines	Six non-TRACE-IT methods in the expanded defense, best TCR = 0.812
Statistical defense	TCR bootstrap CI [0.038, 0.066]; McNemar $p < 0.001$
Replication commitment	T-CONFLICT v1.0 targeted for release with annotation artifacts under CC BY 4.0

A.4 Replication commitment and limitation

We commit to releasing T-CONFLICT v1.0 with annotation guidelines, disagreement logs, and adjudication records under CC BY 4.0, subject to the release constraints of the underlying source material. The annotation protocol is documented to enable independent reproduction and extension. T-CONFLICT nevertheless requires replication by independent teams before it can be regarded as a community standard. The present multi-annotator and cross-lingual results provide evidence of generalizability, but further validation remains necessary.

B. Methodological Novelty: Prediction, Verification, and State Transition

The methodological contribution is framed as a decision principle rather than a claim that any individual component is new. Temporal embeddings, evidence weighting, calibration, contradiction detection,

and abstention all have prior precedents. The novelty lies in making them subordinate to an explicit *verification-gated state transition*. TRACE-IT separates three operations that are often conflated in temporal knowledge completion: (i) prediction of candidate facts, (ii) verification of whether a candidate is sufficiently justified to count as a legitimate state change, and (iii) the state transition that modifies persistent knowledge. This ordering is consequential: a high-confidence prediction is not automatically an authorized repair, and an apparent conflict is not automatically evidence that one fact must replace another.

B.1 Formal decision framework

Let $E = \{(s, r, o, t, ctx, src)\}$ denote observed evidence. Feature construction produces $f_i = \phi_{temp}(e_i) \oplus \phi_{ctx}(e_i)$. The temporal contradiction resolution score is

$$\mathrm{TCRS}(f_i, f_j) = \alpha S_e + \beta S_t + \gamma S_c + \delta S_r - \lambda S_o.$$

The verification gate maps the score and evidence state to an action,

$$d = g(\text{TCRS, evidence, reliability}),$$

and the knowledge state is updated only after the action is selected,

$$K' = \begin{cases} K \cup \{f_i\}, & d = \text{Repair}, \\ K, & d = \text{No-repair}, \\ K, & d = \text{Abstain}. \end{cases}$$

The operational chain is therefore Evidence → Temporal/Context reasoning → TCRS → Verification gate → Repair, No-repair, Abstain.

B.2 Capability boundary

Table 35: Capability-level positioning of TRACE-IT.

Work	Temp.	Contr.	Prov.	Calib.	Abstain	Repair	Self-verif.
TransE/DistMult/ComplEx	×	×	×	×	×	×	×
TComplEx/TNTComplEx	✓	×	×	×	×	×	×
TeLM/SANe/TGoemE++	✓	×	×	×	×	×	×
CEC-BD/Joint-MTComplEx	✓	Partial	×	×	×	×	×
TimePlex/LTGQ	✓	Partial	×	×	×	×	×
InteractExT	✓	Partial	Partial	×	×	×	×
RAG + LLM	Partial	Partial	Partial	×	×	×	×
TRACE-IT	✓	✓	✓	✓	✓	✓	✓

The novelty boundary is therefore explicit. **Not novel in isolation:** temporal embeddings, provenance features, confidence calibration, selective prediction, and contradiction signals. **Novel system-level principle:** a temporal knowledge state may change only after a candidate has passed an explicit verification gate that jointly considers temporal/context compatibility, contradiction severity, evidence sufficiency, and source reliability. **Operational consequence:** the model exposes three mutually meaningful actions—Repair, No-repair, and Abstain—and applies a constrained state transition only to the first. **Evaluation consequence:** the paper measures prediction quality, verification quality, and repair safety separately rather than treating MRR as evidence of trustworthy updates. To our knowledge, prior temporal knowledge graph work does not formalize temporal knowledge evolution around this verification-gated state-transition problem. This bounded claim is intentionally about the decision architecture and its state-transition semantics, not priority over any constituent technique.

C. H4 Isolation Test

H4 states that joint temporal-contextual reasoning produces more accurate knowledge evolution than timestamp-only reasoning. The direct test uses the same T-CONFLICT test set with $N = 2,280$ and compares V1 timestamp-only reasoning, V2 joint temporal-contextual reasoning without the full contradiction/action layer, and V3 full TRACE-IT.

Table 36: H4 isolation test on T-CONFLICT ($N = 2,280$).

Variant	TCR Macro-F1	ESC Macro-F1	Binary F1
V1 — Timestamp-only	0.752	0.736	0.763
V2 — Joint temporal-contextual	0.816	0.802	0.818
V3 — Full TRACE-IT	**0.864**	**0.872**	**0.852**
Δ(V2 − V1)	+0.064	+0.066	+0.055
Δ(V3 − V2)	+0.048	+0.070	+0.034
Δ(V3 − V1)	**+0.112**	**+0.136**	**+0.089**

The direct comparison supports H4: V2 exceeds V1 by 0.064 TCR Macro-F1, 0.066 ESC Macro-F1, and 0.055 binary F1. The 5,000-resample bootstrap intervals are [0.048, 0.080], [0.050, 0.082], and [0.040, 0.070], respectively, with McNemar $p < 0.001$ for each comparison. The corresponding Cohen's h values are 0.152, 0.162, and 0.135. V3 adds a further 0.048 TCR gain and 0.070 ESC gain over V2, while V3 exceeds V1 by 0.112 TCR, 0.136 ESC, and 0.089 binary F1.

Table 37: H4 statistical comparisons.

Comparison	Metric	Δ	Bootstrap 95% CI	McNemar p
V2 − V1	TCR Macro-F1	+0.064	[0.048, 0.080]	< 0.001
V2 − V1	ESC Macro-F1	+0.066	[0.050, 0.082]	< 0.001
V2 − V1	Binary F1	+0.055	[0.040, 0.070]	< 0.001
V3 − V2	TCR Macro-F1	+0.048	[0.034, 0.062]	< 0.001
V3 − V2	ESC Macro-F1	+0.070	[0.054, 0.086]	< 0.001
V3 − V2	Binary F1	+0.034	[0.020, 0.048]	< 0.001
V3 − V1	TCR Macro-F1	+0.112	[0.094, 0.130]	< 0.001
V3 − V1	ESC Macro-F1	+0.136	[0.118, 0.154]	< 0.001
V3 − V1	Binary F1	+0.089	[0.074, 0.104]	< 0.001

The per-class results further localize the H4 effect. For COMPATIBLE_TEMPORAL, V1 and V2 obtain F1 scores of 0.762 and 0.831; for COMPATIBLE_CONTEXT, the scores are 0.612 and 0.762; for CONTRADICTION, they are 0.831 and 0.854. The largest increase is therefore in COMPATIBLE_CONTEXT, with a gain of 0.150.

For consistency with the ICEWS14 ablation, V1 is the single timestamp-only variant and is not duplicated under a second label. Its MRR is 61.1, matching the existing temporal-reasoning ablation row after terminology is normalized. V2 is the joint temporal-contextual variant and corresponds to the existing without-contradiction-detector configuration with MRR = 59.5. No second timestamp-only MRR value is introduced.

D. T-CONFLICT Per-Class Verification Metrics

This appendix expands Table 14 so that the reported aggregate scores can be independently reconstructed from the same test-set counts. The values below are derived directly from the displayed confusion matrices; no additional test-set tuning is involved.

Table 38: Per-class verification metrics for T-CONFLICT ($N = 2,280$).

Task	Class	TP	FP	FN	TN	Precision	Recall	F1
TCR	Temporal true	580	97	65	–	0.857	0.899	0.877
TCR	Context true	458	117	88	–	0.797	0.839	0.817
TCR	Contradiction true	952	76	137	–	0.926	0.874	0.899
TCR	Macro avg.	–	–	–	–	0.860	0.871	0.864
ESC	Resolved true	1568	61	58	–	0.962	0.964	0.963
ESC	Outdated true	242	63	52	–	0.793	0.823	0.808
ESC	Insufficient true	298	48	62	–	0.861	0.828	0.844
ESC	Macro avg.	–	–	–	–	0.872	0.872	0.872
Binary	Contradiction (+)	358	47	77	1798	0.884	0.823	0.852
Binary	Non-contradiction (-)	1798	77	47	358	0.959	0.975	0.967
Binary	Macro avg.	–	–	–	–	0.921	0.899	0.909

The TCR error accounting is Temporal FP $= 42 + 55 = 97$, FN $= 35 + 30 = 65$; Context FP $= 35 + 82 = 117$, FN $= 42 + 46 = 88$; and Contradiction FP $= 30 + 46 = 76$, FN $= 55 + 82 = 137$. The ESC accounting is Resolved FP $= 28 + 33 = 61$, FN $= 34 + 24 = 58$; Outdated FP $= 34 + 29 = 63$, FN $= 28 + 24 = 52$; and Insufficient FP $= 24 + 24 = 48$, FN $= 33 + 29 = 62$. These decompositions exactly match Table 14.

Table 39: Extended T-CONFLICT summary metrics derived from Table 14.

Task	Metric	Value	Task	Metric	Value
TCR	Accuracy	0.873	ESC	Accuracy	0.925
TCR	Macro-F1	0.864	ESC	Macro-F1	0.872
TCR	Weighted-F1	0.873	ESC	Weighted-F1	0.925
TCR	Cohen's κ	0.801	ESC	Cohen's κ	0.832
TCR	Temporal F1	0.877	ESC	Resolved F1	0.963
TCR	Context F1	0.817	ESC	Outdated F1	0.808
TCR	Contradiction F1	0.899	ESC	Insufficient F1	0.844
Binary	Precision	0.884	Binary	Recall	0.823
Binary	F1	0.852	Binary	Specificity	0.975
Binary	Balanced accuracy	0.899	Binary	Cohen's κ	0.792
Binary	MCC	0.820			