EviScope: A Benchmark for Robust Retrieval in Temporal KGQA

Anonymous ACL submission

Abstract

Current benchmarks for Temporal Knowledge Graph Question Answering (TKGQA) retrieval can lead to inconsistent evaluations due to *incomplete supervision* and *miscalibrated difficulty*, potentially misjudging actual model capabilities. To address this, we construct **EviScope**, a large-scale diagnostic benchmark featuring over 130,000 generated question-answer pairs that explicitly map the alternative and distracting evidence space. EviScope provides a calibrated landscape to more faithfully evaluate retrieval models under genuine ambiguity. Furthermore, we propose BS@K, a matching metric designed to evaluate multi-evidence reconstruction while penalizing redundancy. Benchmark evaluations demonstrate that our framework provides a more objective and realistic assessment, effectively illuminating precise directions for future retriever advancements. Our dataset and code are available at `https://anonymous.4open.science/r/EviScope-72B4/`.

1 Introduction

Retrieval is the cornerstone of Temporal Knowledge Graph Question Answering (TKGQA), as answering a question fundamentally depends on identifying temporally consistent evidence from evolving knowledge graphs. With the rise of retrieval-augmented large language models (LLMs), retrieval quality has become a primary bottleneck, directly affecting both answer correctness and reasoning efficiency (Lewis et al., 2020; Li et al., 2024; Ye et al., 2024).

This challenge is particularly pronounced in temporal knowledge graphs, where the interaction between entities, relations, and timestamps significantly enlarges the search space (Li et al., 2024; Wu et al., 2024). As a result, models must distinguish truly relevant facts from a large number of semantically similar but potentially misleading candidates, making reliable retrieval evaluation inherently difficult.

Despite this importance, existing TKGQA benchmarks provide limited retrieval evaluation signals due to *incomplete supervision*. Most benchmarks construct question-answer pairs from a fixed set of supporting facts (e.g., CRONQUESTIONS (Saxena et al., 2021), MultiTQ (Chen et al., 2023), and TimelineKGQA (Sun et al., 2025)), assuming a single ground-truth evidence set. However, real-world knowledge graphs are inherently incomplete and contain multiple semantically equivalent or temporally overlapping facts supporting the same answer (Nickel et al., 2015). Consequently, valid but unannotated evidence is treated as incorrect. This incomplete supervision introduces systematic bias, penalizing models for retrieving legitimate evidence, underestimating retrieval quality, and encouraging shortcut solutions that overfit to annotation artifacts (Ye et al., 2024; Chan et al., 2024).

A second limitation is *miscalibrated difficulty*. Existing benchmarks predominantly define question difficulty by the number of reasoning hops, assuming multi-hop questions are inherently harder (Zhang et al., 2022; Yu et al., 2023). However, recent studies reveal that difficulty is heavily driven by retrieval ambiguity rather than reasoning complexity alone (Liu et al., 2023; Li et al., 2024). High semantic overlap among candidate facts easily misleads retrievers into selecting deceptive evidence, cascading into reasoning failures (Ye et al., 2024; Wu et al., 2024). This miscalibration is exacerbated in temporal KGs due to the dense coexistence of temporally related events. Consequently, current benchmarks fail to capture how variations in confusing alternatives create stark difficulty disparities between structurally identical questions.

To address these limitations, we propose **EviScope**, a diagnostic benchmark that explicitly models the Evidence Space for TKGQA. EviScope rev-

Dataset	Scale	Multi-hop	Evidence Completeness	Ambiguity Awareness	Difficulty Modeling	Temporal Granularity
TempQuestions (Shi et al., 2021)	Small	Limited	×	×	×	Coarse
TimeQuestions (Mohammad et al., 2021)	Small	Limited	×	×	×	Coarse
CRONQUESTIONS (Saxena et al., 2021)	Large	✓	×	×	Structure-based	Medium
MultiTQ (Chen et al., 2023)	Large	✓	Partial	×	Structure-based	Fine
TimelineKGQA (Sun et al., 2025)	Large	✓	Partial	×	Structure-based	Fine
EviScope	Large	✓	✓	✓	**Hybrid (Structure + Ambiguity)**	Fine

Table 1: Comparison of representative TKGQA benchmarks. Existing datasets primarily rely on single ground-truth evidence and structure-based difficulty, while our benchmark explicitly models evidence completeness and retrieval ambiguity.

olutionizes retrieval evaluation through three interconnected core components:

(1) **Evidence Space Reconstruction.** Rather than assuming an idealized single ground-truth evidence set, we systematically augment each question with *substitutable evidence* (valid alternative facts) and *distracting evidence* (temporally or semantically plausible candidates). This layout effectively overcomes incomplete supervision by eliminating false-negative penalties, while simultaneously identifying models that exploit shortcut correlations rather than genuine reasoning.

(2) **Hybrid Difficulty Modeling (HDM).** We introduce HDM, a calibrated formulation that jointly quantifies multi-hop reasoning demands and factual retrieval ambiguity. Instead of blindly escalating structural complexity, HDM offers granular diagnostics regarding a model's precise capacity for distractor filtering.

(3) **Structure-Aware Evaluation Metric (BS@K).** We propose BS@K, a bipartite matching-based metric that evaluates whether the retrieved text can reconstruct the complete downstream reasoning chain while rigorously penalizing uninformative redundancy, yielding a highly robust signal of true retrieval quality.

Extensive experiments demonstrate that EviScope reveals critical performance gaps overlooked by traditional protocols. Our findings highlight that by untangling incomplete supervision and calibrated ambiguity, EviScope provides a more reliable, fine-grained diagnostic framework for retrieval-augmented TKGQA systems.

2 Related Work

Temporal KGQA Benchmarks. Early benchmarks like TempQuestions (Shi et al., 2021) and TimeQuestions (Mohammad et al., 2021) introduced temporally grounded questions but were limited in scale. To address this, CRONQUESTIONS (Saxena et al., 2021) scaled up TKGQA by generating over 300K questions stratified by structural complexity. MultiTQ (Chen et al., 2023) introduced multi-granularity temporal signals with multiple relevant facts per question, while TimelineKGQA (Sun et al., 2025) proposed a universal timeline-context QA generator. Despite these scaling efforts, all existing benchmarks anchor question-answer pairs to a fixed, explicitly annotated set of supporting facts. This rigid paradigm implicitly assumes a single valid evidence set, failing to account for the inherent incompleteness of real-world KGs and introducing a severe evaluation bias for retrieval-augmented systems.

Multi-hop Reasoning and Retrieval Evaluation. A parallel limitation exists in general multi-hop QA and KGQA. Benchmarks such as MuSiQue (Zhang et al., 2022) and KGT5 (Yu et al., 2023) predominantly characterize question difficulty via topological reasoning structures (e.g., hop counts). While effective for structural composition, this approach isolates reasoning from retrieval, overlooking retrieval-side ambiguity. Recent studies reveal that retrieval ambiguity often exerts a stronger impact on LLM performance than structural complexity alone (Liu et al., 2023; Li et al., 2024). For example, simple single-hop queries become non-trivial when adjacent temporal facts highly overlap with ground-truth evidence. This challenge is amplified in temporal settings where dense, temporally adjacent events create highly deceptive distractors. Existing frameworks lack metrics to capture this interplay, necessitating a benchmark that explicitly models evidence-space ambiguity.

As summarized in Table 1, existing TKGQA benchmarks lack mechanisms for modeling evidence completeness and ambiguity awareness. In contrast, EviScope explicitly characterizes the full evidence space, introduces ambiguity-aware evaluation, and provides fine-grained structural and contextual difficulty annotations, establishing a high-fidelity diagnostic framework.

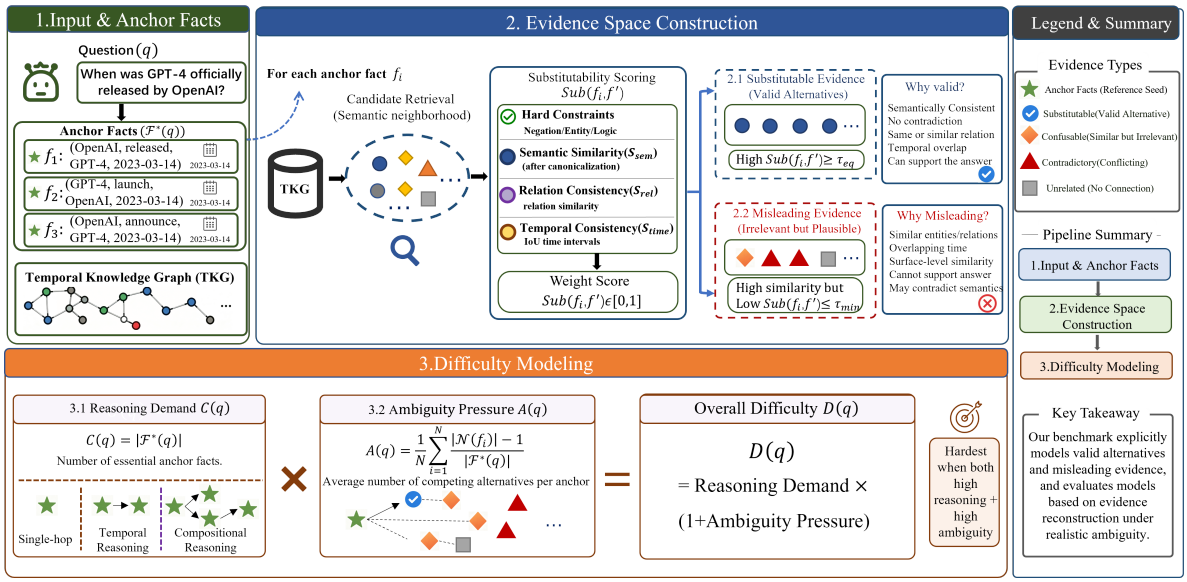

Figure 1: Overview of the EviScope benchmark framework, showing the evidence space construction pipeline: anchor facts are expanded with substitutable and distracting evidence to enable comprehensive retrieval evaluation.

3 Method

3.1 Problem Definition

In this section, we formalize the task setting and highlight the limitations of existing TKGQA benchmarks that motivate our approach. We study retrieval in Temporal Knowledge Graph Question Answering (TKGQA). A temporal knowledge graph is defined as:

$$\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T}, \mathcal{F}),$$

where $\mathcal{E}$, $\mathcal{R}$, and $\mathcal{T}$ denote entities, relations, and timestamps, respectively. Each fact is represented as a quadruple $f = (s, r, o, t)$, where $s, o \in \mathcal{E}$, $r \in \mathcal{R}$, and $t \in \mathcal{T}$.

Given a question q, existing benchmarks typically associate it with a single ground-truth evidence set:

$$\mathcal{F}^*(q) = \{f_1, f_2, \ldots, f_N\}.$$

This formulation implicitly assumes that (i) the ground-truth evidence is complete and unique, and (ii) retrieval quality can be evaluated by measuring overlap with this set. However, both assumptions are problematic in temporal knowledge graphs.

First, evidence incompleteness arises because multiple alternative facts may equally support the answer. For instance, semantically equivalent relations or temporally overlapping events can provide interchangeable support. Second, retrieval ambiguity arises because many irrelevant facts can appear highly similar to relevant ones, especially under temporal and relational variations.

As a result, evaluating retrieval solely based on a single ground-truth set leads to shortcut solutions and unreliable conclusions.

To address these issues, we propose to explicitly model the *evidence space* of each question. Instead of relying on a single evidence set, we construct an augmented evidence set:

$$\widetilde{\mathcal{F}}(q) = \mathcal{F}^*(q) \cup \mathcal{F}^{\text{sub}}(q) \cup \mathcal{F}^{\text{dis}}(q),$$

where $\mathcal{F}^{\text{sub}}(q)$ and $\mathcal{F}^{\text{dis}}(q)$ denote substitutable and distracting evidence, respectively. This formulation enables a more complete and realistic characterization of the retrieval problem. Figure 1 illustrates the overall construction pipeline.

3.2 Evidence Space Construction

In this section, we describe how to construct the evidence space for each question. The key idea is to expand anchor facts into a structured space that includes both valid alternatives and misleading candidates, thereby capturing the full range of retrieval behavior. Figure 2 provides a concrete illustration of the substitutability evaluation.

Anchor Facts We begin by identifying a minimal set of anchor facts:

$$\mathcal{F}^*(q) = \{f_1, f_2, \ldots, f_N\},$$

which are necessary to answer the question q.

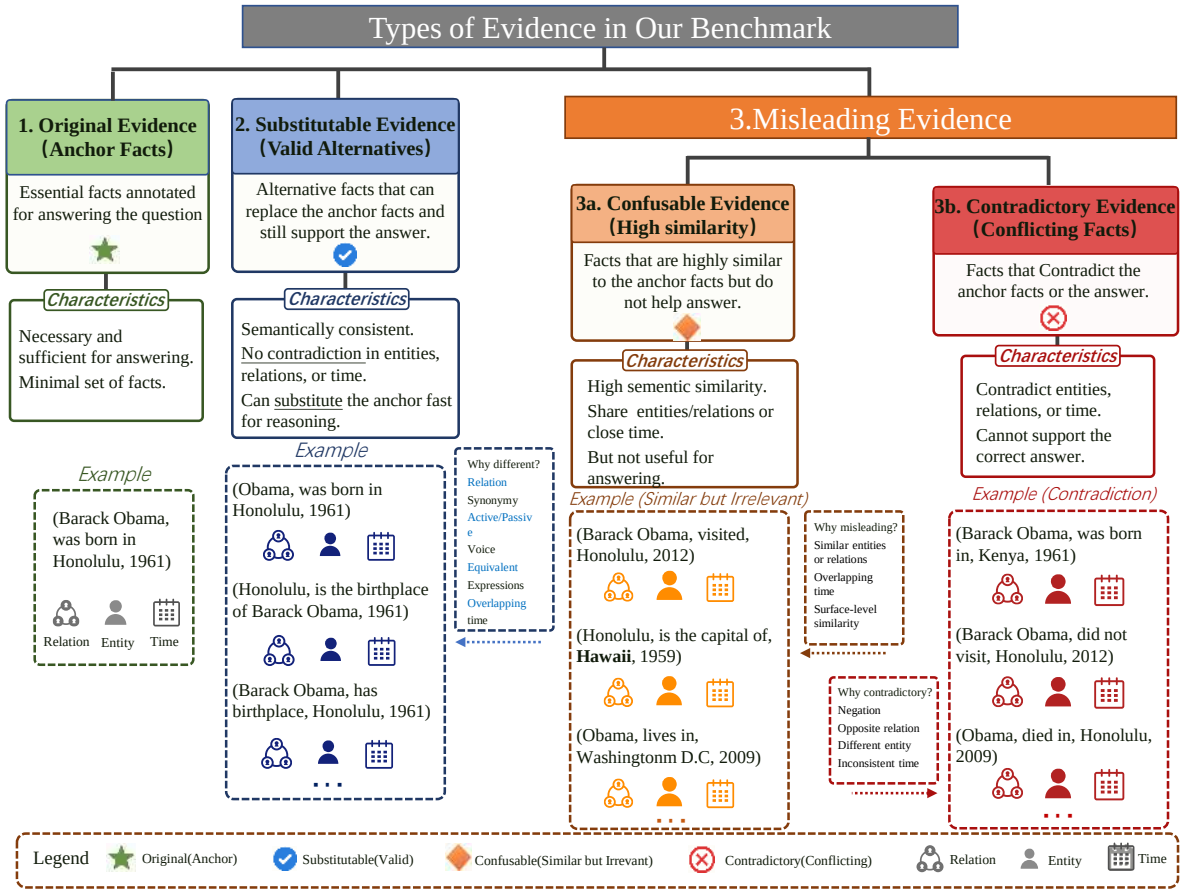

Figure 2: Illustration of the substitutability function, demonstrating how candidate evidence is evaluated for semantic, relational, and temporal consistency with anchor facts.

These anchor facts can be obtained from existing annotations or generated using standard TKGQA pipelines. Importantly, we treat them as reference points rather than exhaustive supervision. Each anchor fact defines a local semantic region in the knowledge graph, which we further expand.

This perspective is crucial: instead of assuming that the annotated facts are the only valid evidence, we use them as seeds to explore a broader evidence space.

Substitutable Evidence We construct substitutable evidence that captures alternative valid ways of supporting the answer.

A central observation is that retrieval should not be penalized for returning evidence that is different in surface form but equivalent in meaning. Therefore, we introduce a *substitutability* function to measure whether a candidate fact can replace an anchor fact in answering the question.

For an anchor fact f_i and a candidate fact f', we define:

$$\mathrm{Sub}(f_i, f') \in [0, 1],$$

which measures the degree to which f' can substitute f_i.

Unlike conventional similarity metrics, substitutability is task-aware: it evaluates whether the candidate fact preserves answerability, rather than simply measuring semantic proximity.

We compute substitutability as:

$$\mathrm{Sub}(f_i, f') = \mathbb{I}_{\text{valid}}(f_i, f') \cdot (\alpha S_{\text{sem}} + \beta S_{\text{rel}} + \gamma S_{\text{time}}),$$

where α, β, γ are weighting coefficients.

We first apply strict filtering to eliminate invalid candidates. Specifically, we define:

$$\mathbb{I}_{\text{valid}}(f_i, f') = \begin{cases} 0 & \text{if constraints are violated} \\ 1 & \text{otherwise} \end{cases}$$

We enforce negation consistency, entity consistency, and logical consistency: facts containing explicit negation that contradict the anchor fact are removed, core entities must match or be resolvable via aliasing, and relations must not introduce semantic contradictions. These constraints ensure that substitutability is not inflated by superficial similarity.

We compute semantic similarity after converting facts into a canonical form to normalize linguistic variations such as active/passive voice. This step reduces bias introduced by surface-level differences. Formally:

$$S_{\text{sem}}(f_i, f') = \cos\big(\phi(\text{canon}(f_i)), \phi(\text{canon}(f'))\big),$$

where $\phi(\cdot)$ is a text encoder that maps a fact into a dense semantic representation. This allows us to capture contextual similarity beyond surface-level token overlap, making the measure robust to paraphrases and syntactic variations.

To further refine substitutability, we incorporate relation and temporal signals. Relation consistency is defined as $S_{\text{rel}}(f_i, f') = \text{sim}_r(r_i, r')$, which captures semantic proximity between relations. Temporal consistency is defined as $S_{\text{time}}(f_i, f') = |t_i \cap t'| \,/\, |t_i \cup t'|$, measuring temporal overlap between facts. Together, these components ensure that substitutable evidence is both semantically and temporally aligned with the anchor facts.

The substitutable evidence set $\mathcal{F}^{\text{sub}}(q)$ collects all facts sufficiently substitutable for any anchor fact, with each candidate weighted by its substitutability score to enable graded evaluation in later stages:

$$\mathcal{F}^{\text{sub}}(q) = \bigcup_{i=1}^{N} \{f' \in \mathcal{F} \mid \text{Sub}(f_i, f') \geq \tau_{\text{eq}}\},$$

where $\mathcal{F}$ is the full fact space of the KG and τ_{eq} is the equivalence threshold for substitutability.

Distracting Evidence While substitutable evidence captures valid alternatives, distracting evidence targets facts that appear relevant but do not contribute to answering the question. Such cases are common in temporal knowledge graphs due to overlapping entities, similar relations, and nearby timestamps.

We define distracting evidence as facts that pass similarity checks and remain structurally plausible, yet fail to support the answer:

$$\mathcal{F}^{\text{dis}}(q) = \left\{ f' \;\middle|\; \begin{array}{l} \text{Sim}(f_i, f') \geq \tau_{\text{sim}} \\ \wedge\, \text{Sub}(f_i, f') < \tau_{\text{min}} \end{array} \right\}.$$

By incorporating distracting evidence, we introduce realistic retrieval interference that tests whether a model can distinguish truly useful evidence from misleading candidates.

3.3 Difficulty Modeling

In this section, we define a difficulty metric that reflects both reasoning requirements and retrieval ambiguity.

Existing benchmarks often approximate difficulty using the number of reasoning steps. However, this ignores an important factor: even simple questions can become difficult when many misleading candidates exist. Conversely, complex reasoning can be relatively easy when ambiguity is low.

To capture this, we model difficulty as a combination of reasoning demand and ambiguity pressure:

$$D(q) = C(q) \cdot (1 + A(q)).$$

Reasoning Demand We quantify the reasoning requirement of a question based on the number of essential facts needed to derive the answer. Intuitively, answering a temporal question requires identifying and composing a sequence of events, and the number of such events directly determines the minimal reasoning complexity.

Formally, we define:

$$C(q) = |\mathcal{F}^*(q)|$$

where $\mathcal{F}^*(q)$ denotes the set of anchor facts associated with question q.

This term reflects the minimum number of facts required to answer the question and provides a lower bound on reasoning complexity. In particular, $C(q)$ is then combined with ambiguity pressure $A(q)$ to form the overall difficulty score, which is discretized into three intervals corresponding to Simple, Medium, and Complex difficulty levels.

Compared to prior work that uses hop count as a proxy for difficulty, our formulation directly operates on the underlying evidence structure. This allows us to decouple reasoning complexity from graph traversal artifacts and provides a more faithful estimate of the intrinsic reasoning demand of a question.

Ambiguity Pressure While reasoning demand captures the structural complexity of a question, it does not account for the difficulty introduced by competing evidence. In practice, retrieval is often challenged not by the absence of relevant facts, but by the presence of many plausible yet misleading alternatives. To capture this phenomenon, we introduce ambiguity pressure as a complementary factor.

We define ambiguity pressure as:

$$A(q) = \frac{1}{N} \sum_{i=1}^{N} \frac{|\mathcal{N}(f_i)| - 1}{|\mathcal{F}^*(q)|},$$

where $\mathcal{N}(f_i)$ denotes the semantic neighborhood of anchor fact f_i, i.e., the set of candidate facts that are semantically similar to f_i under a predefined similarity threshold.

This term captures how many alternative candidates compete with each anchor fact. A larger neighborhood implies higher ambiguity and greater retrieval difficulty, as models must distinguish truly supportive evidence from a dense set of similar but potentially irrelevant facts.

Importantly, ambiguity pressure reflects a dimension of difficulty that is largely overlooked in existing benchmarks. Two questions with identical reasoning demand $C(q)$ can exhibit drastically different levels of difficulty depending on the size and structure of their evidence neighborhoods. By explicitly modeling $A(q)$, we are able to quantify this variability and provide a more comprehensive characterization of retrieval challenges.

Furthermore, ambiguity operates as a multiplicative factor in our overall difficulty formulation, amplifying the effect of reasoning demand when interference is high. This aligns with empirical observations that even simple questions can become difficult when the retrieval space is crowded with confusing candidates.

This formulation provides a more expressive characterization of difficulty: questions with low $C(q)$ but high $A(q)$ are dominated by retrieval ambiguity, those with high $C(q)$ but low $A(q)$ are dominated by reasoning complexity, and questions with both high $C(q)$ and $A(q)$ are the most challenging. Importantly, ambiguity acts as a multiplicative factor, amplifying difficulty beyond structural complexity.

3.4 Evaluation Protocol

We adopt three complementary metrics. Hits@K is the standard retrieval metric measuring whether the correct answer appears in the top-K retrieved facts, capturing answer-level recall.

BS@K (Bipartite Matching-based Score) evaluates whether retrieved evidence faithfully reconstructs the reasoning path. Given anchor facts $\mathcal{F}^*(q)$ and retrieved facts $\mathcal{R}_K(q)$, we construct a matching matrix based on embedding similarity:

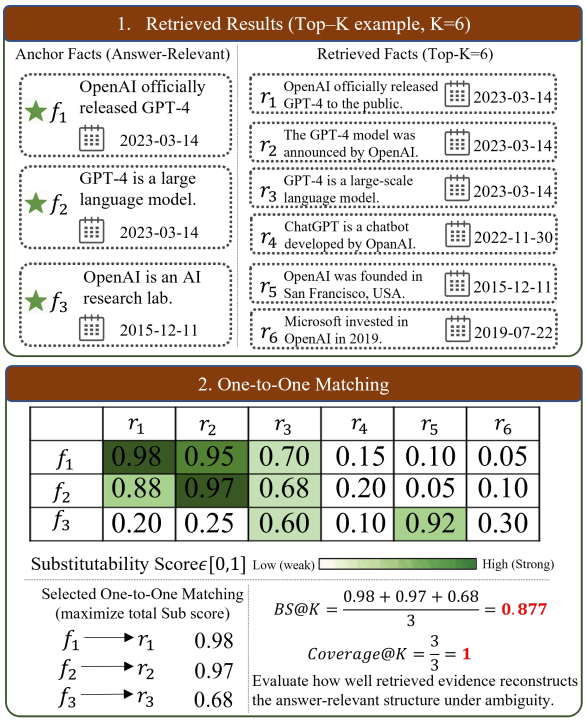

Figure 3: Illustration of BS@K computation: an anchor fact and its top-3 retrievals are matched bilaterally via substitutability scores, demonstrating how BS@K rewards valid alternative evidence that traditional overlap-based metrics would penalize.

$$M_{ij} = \begin{cases} \cos(e(f_i), e(r_j)), & \text{if } \cos(e(f_i), e(r_j)) > \tau \\ 0, & \text{otherwise} \end{cases}$$

where $e(\cdot)$ denotes the fact embedding and τ is a similarity threshold. BS@K then computes the optimal bipartite matching score:

$$BS@K(q) = \frac{1}{|\mathcal{F}^*(q)|} \max_{\pi} \sum_{(i,j)\in\pi} M_{ij}$$

which rewards models that retrieve diverse, complementary evidence covering all anchor facts. We further note Coverage@K as the fraction of anchor facts with at least one matched retrieved fact, providing a diagnostic view of which reasoning steps are missed.

Figure 3 illustrates a concrete computation example. Together, Hits@K and BS@K form complementary signals: Hits@K answers whether the system finds the answer, BS@K evaluates how well it reconstructs the full evidence, and the gap between them exposes shortcut exploitation.

Table 2: EviScope statistics across datasets and difficulty levels.

Statistic	ICEWS				CronQuestions			
	S	**M**	**C**	**All**	**S**	**M**	**C**	**All**
Difficulty Statistics								
Questions	29364	25729	30898	85991	12000	13754	15966	41720
Avg. HDM	4.36	8.73	12.72	8.67	3.83	7.54	9.46	7.20
HDM range	1.69–4.43	4.43–8.87	8.87–13.30	1.69–13.30	1.00–4.43	4.43–8.87	8.87–13.30	1.00–13.30
Evidence-Space Statistics								
Avg. anchor facts	1.00	2.00	3.00	2.02	1.00	2.00	3.00	2.10
Avg. sub. evd.	0.24	0.55	0.81	0.54	0.76	1.82	2.54	1.79
Avg. sub. per anchor	0.24	0.27	0.27	0.27	0.76	0.91	0.85	0.85
Q w/ sub. (%)	14.7	30.5	40.5	28.7	25.9	49.2	63.0	47.8

3.5 Dataset Construction

We construct two benchmark datasets from complementary temporal knowledge graphs for demonstration purposes. The first is derived from the ICEWS Coded Event Data[1], a widely used temporal event dataset that captures time-anchored political interactions with daily-level granularity, yielding 85,991 questions. The second is derived from the CronQuestions knowledge graph (Saxena et al., 2021), a Wikidata-based TKGQA dataset with year-level temporal annotations across diverse domains, yielding 41,720 questions. These two sources provide complementary temporal characteristics: ICEWS represents time-range event data with dense temporal interactions, whereas CronQuestions represents time-point knowledge with broader topical coverage.

Following Section 3, we construct the evidence space for each question by extracting anchor facts from existing annotations and generating substitutable evidence through semantic and temporal consistency filtering. As summarized in Table 2, the resulting datasets exhibit a clear difficulty progression: the average number of anchor facts increases from 1 to 3, while HDM scores rise monotonically from Simple to Complex. Substitutable evidence also becomes increasingly prevalent at higher difficulty levels, with CronQuestions showing a richer alternative evidence space than ICEWS (1.79 vs. 0.54 average substitutable facts per question; 47.8% vs. 28.7% of questions with substitutes). Together, these complementary properties make EviScope a suitable testbed for evaluating retrieval under both dense temporal interactions and broader semantic variability.

[1]ICEWS (Integrated Crisis Early Warning System) is a large-scale event dataset covering global political events with fine-grained timestamps.

4 Experiments

4.1 Experimental Setup

To provide a comprehensive benchmark, we select representative TKGQA methods spanning three paradigms from recent years: (1) **embedding-based** approaches that reason over temporal KG representations—CronKGQA (Saxena et al., 2021) and CTRN (Jiao et al., 2022); (2) **subgraph-based** methods that extract temporally constrained subgraphs for interpretable reasoning—SubGTR (Chen et al., 2022) (evaluated under both standard and complex configurations) and TempoQR (Mavromatis et al., 2022) (with both soft and hard supervision variants); and (3) **LLM-augmented** approaches that leverage pre-trained language models for decomposition and reasoning—MultiQA (Chen et al., 2023) and RTQA (Gong et al., 2025). Together these baselines span the dominant TKGQA paradigms from 2021–2025, ensuring broad coverage of contemporary techniques.

All methods are evaluated on CronQuestions under both the original single-ground-truth setting (Orig.) and our EviScope-augmented setting (Evi.). We report Hits@K and BS@K at $K = 1$ and $K = 10$.

4.2 Results and Analysis

Overestimation of Retrieval Robustness. Experimental results (Table 3) indicate that traditional TKGQA benchmarks systematically overestimate retrieval robustness. Upon introducing semantically similar distracting evidence via EviScope, several models maintain their Hits performance while suffering severe degradation in evidence reconstruction. For instance, TempoQR (Hard Supervision) exhibits negligible variance in Hits@10 (0.203 → 0.194), yet its corresponding

Table 3: Main experimental results comparing model performance on the original dataset (Orig.) and our augmented EviScope benchmark (Evi.). The Δ column explicitly quantifies the performance shift. EviScope exposes significant discrepancies between retrieval robustness (Hits) and evidence faithfulness (BS).

Method	**Hits@1**			**Hits@10**			**BS@1**			**BS@10**		
	Orig.	Evi.	Δ	Orig.	Evi.	Δ	Orig.	Evi.	Δ	Orig.	Evi.	Δ
CronKGQA	0.200	0.203	+0.003	0.323	0.330	+0.007	0.335	0.329	-0.006	0.450	0.457	+0.007
CTRN	0.090	0.090	–	0.141	0.141	–	0.428	0.476	+0.048	0.527	0.576	+0.049
SubGTR (wikidata_big)	0.267	0.273	+0.006	0.327	0.323	-0.004	0.534	0.544	+0.010	0.559	0.561	+0.002
SubGTR (w_big_complex)	0.330	0.270	-0.060	0.373	0.357	-0.016	0.537	**0.583**	**+0.046**	0.559	**0.615**	**+0.056**
TempoQR (Soft, CronKGQA)	0.139	0.139	–	0.215	0.228	+0.013	0.410	**0.303**	**-0.107**	0.437	**0.319**	**-0.118**
TempoQR (Soft, TempoQR)	0.031	0.047	+0.016	0.074	0.120	+0.046	0.840	0.828	-0.012	0.882	0.866	-0.016
TempoQR (Hard, TempoQR)	0.068	0.080	+0.012	0.203	**0.194**	**-0.009**	0.381	**0.274**	**-0.107**	0.414	**0.293**	**-0.121**
MultiQA	0.613	0.613	–	0.742	**0.871**	**+0.129**	0.613	0.645	+0.032	0.741	**0.871**	**+0.130**
RTQA	0.020	0.017	-0.003	0.130	**0.120**	**-0.010**	0.680	**0.513**	**-0.167**	0.782	**0.706**	**-0.076**

BS@10 drops sharply from 0.413 to 0.292. RTQA mirrors this trend, maintaining stable Hits@10 but experiencing a noticeable decline in BS@10 ($0.781 \rightarrow 0.705$). These discrepancies demonstrate that existing benchmarks inadvertently reward shortcut exploitation. EviScope effectively exposes retrieval failures completely masked by standard indicators.

Disentangling Correctness from Faithfulness. Furthermore, EviScope successfully disentangles answer correctness from evidence faithfulness. High Hits rates do not guarantee the retrieval of high-quality supporting facts. While TempoQR retrieves correct answers heavily contaminated by distractors, SubGTR (wikidata_big_complex) demonstrates the opposite behavior. Despite a decrease in Hits@10 ($0.373 \rightarrow 0.357$), its BS@10 significantly improves ($0.559 \rightarrow 0.615$) on EviScope. This indicates the model successfully recovers semantically valid substitute evidence—legitimate reasoning paths that traditional benchmarks unfairly penalize due to their rigid single ground-truth assumption.

Diagnostic Capability for Retrieval-Aware Reasoning. The diverse ambiguity sensitivity exhibited by different baselines highlights EviScope's diagnostic utility. MultiQA, rather than degrading, achieves substantial gains across both Hits@10 ($0.742 \rightarrow 0.871$) and BS@10 ($0.741 \rightarrow 0.871$). This confirms that EviScope does not merely inflate task difficulty, but accurately calibrates retrieval ambiguity. Models with robust evidence aggregation thrive, whereas those reliant on shallow shortcuts fail. Consequently, EviScope provides actionable, fine-grained feedback for optimizing future TKGQA retrieval architectures. **More detailed analysis is provided in the appendix.**

5 Conclusion

This paper introduced EviScope, a benchmark for retrieval-augmented TKGQA that explicitly models the evidence space by augmenting each question with substitutable and distracting evidence. Our experiments show that existing benchmarks often blur the distinction between answer correctness and evidence faithfulness, thereby obscuring important differences in model behavior under retrieval ambiguity. Rather than uniformly degrading performance, EviScope reveals a more diagnostic picture: some methods retain competitive Hits@K while suffering noticeable BS@K drops, indicating reliance on shallow shortcuts, whereas others benefit from the enriched evidence space and recover more faithful reasoning paths. Moreover, our ambiguity-aware difficulty modeling yields more consistent difficulty rankings than hop-based definitions, and our similarity analysis identifies a clear threshold effect under high-similarity interference. Together, these findings demonstrate the importance of moving beyond single-ground-truth evaluation and establish EviScope as a more faithful and informative benchmark for retrieval-augmented TKGQA.

Limitations

Our study is conducted on two datasets, ICEWS and CronQuestions, which cover complementary but still limited temporal granularities and domains. Extending the benchmark to other TKGQA settings, such as future-time forecasting or multi-

granularity temporal reasoning, remains an important direction for future work. In addition, although BS@K provides a more informative view of evidence reconstruction than standard answer-only metrics, it still depends on the quality of the substitutability function and may not fully capture cases where retrieved evidence is only partially supportive or insufficient for complete reasoning. More broadly, while EviScope makes evidence-space ambiguity explicit, developing evaluation metrics that jointly characterize answer correctness, evidence faithfulness, and reasoning sufficiency remains an open challenge.

References

Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. 2024. Rq-rag: Learning to refine queries for retrieval augmented generation. In *ACL Findings*.

Ziyang Chen, Jinzhi Liao, and Xiang Zhao. 2022. Temporal knowledge graph question answering via subgraph reasoning. *Knowledge-Based Systems*, 251:109134.

Ziyang Chen, Jinzhi Liao, and Xiang Zhao. 2023. Multi-granularity temporal question answering over knowledge graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 11378–11392.

Zhaoyan Gong, Juan Li, Zhiqiang Liu, Lei Liang, Huajun Chen, and Wen Zhang. 2025. Rtqa: Recursive thinking for complex temporal knowledge graph question answering with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9865–9882.

Songlin Jiao, Zhenfang Zhu, Wenqing Wu, and Zicheng Zuo. 2022. An improving reasoning network for complex question answering over temporal knowledge graphs. *Applied Intelligence*, 53(10):1–14.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*.

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. Retrieval augmented generation or long-context llms? a comprehensive study and hybrid approach. In *EMNLP (Industry Track)*.

Nelson F. Liu and 1 others. 2023. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*. ArXiv:2307.03172.

Costas Mavromatis, Prasanna Lakkur Subramanyam, Vassilis N. Ioannidis, Adesoji Adeshina, Phillip R. Howard, Tetiana Grinberg, Nagib Hakim, and George Karypis. 2022. Tempoqr: Temporal question reasoning over knowledge graphs. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI)*.

Saeed M. Mohammad and 1 others. 2021. Timequestions: A benchmark for temporal question answering. In *ACL*.

Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2015. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*.

Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. 2021. Question answering over temporal knowledge graphs. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 6663–6676.

Jiaxin Shi, Chenliang Li, and 1 others. 2021. Answering complex temporal questions with implicit constraints. In *ACL*.

Qiang Sun, Sirui Li, Du Huynh, Mark Reynolds, and Wei Liu. 2025. Timelinekgqa: A comprehensive question-answer pair generator for temporal knowledge graphs. In *Companion Proceedings of the ACM Web Conference 2025*, pages 1–4, New York, NY, USA. ACM.

Shangyu Wu, Ying Xiong, and 1 others. 2024. Retrieval-augmented generation for natural language processing: A survey. *arXiv preprint arXiv:2407.13193*.

Fuda Ye, Shuangyin Li, Yongqi Zhang, and Lei Chen. 2024. R^2ag: Incorporating retrieval information into retrieval augmented generation. In *Findings of ACL: EMNLP*.

Wenhao Yu and 1 others. 2023. Kgt5: Scaling up knowledge graph question answering with large language models. In *ACL*.

Harsh Trivedi Zhang, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554. ArXiv:2108.00573.

A Importance of Substitutable Evidence in Mitigating Distractor Interference

The instability of reasoning results in the face of highly similar distractors is the core motivation for constructing and introducing the substitutable evidence mechanism. Traditional benchmarks relying on single ground-truth evidence often fail to capture the nuanced semantic variations in real-world knowledge graphs, leaving models vulnerable to attribution confusion when encountering high-similarity interference.

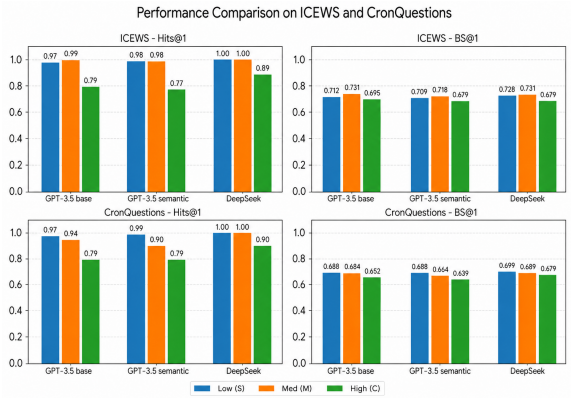

Figure 4: Model performance under different levels of distractor semantic similarity.

Semantic Traps in High-Similarity Scenarios. As shown in Figure 4, both $Hits@1$ and $BS@1$ experience significant degradation under high-similarity conditions. For instance, GPT-3.5_base on ICEWS drops from 0.97 (Low) to 0.79 (High) in $Hits@1$, representing an 18.5% decline. Similarly, $BS@1$ falls from 0.712 to 0.695, indicating reduced evidence faithfulness. These patterns reveal that models struggle to distinguish subtle semantic differences, as superficially similar yet logically contradictory evidence misleads retrieval and disrupts decision-making logic. Such evidence creates a "threshold effect": models maintain robustness against low/medium similarity distractors but suffer catastrophic performance drops when interference crosses a critical similarity threshold.

Substitutable Evidence as a Mitigation Mechanism. Introducing substitutable evidence addresses this challenge by enabling multi-dimensional verification logic. Unlike single ground-truth approaches that rely on isolated evidence points, substitutable evidence allows models to cross-validate across multiple valid reasoning paths. For example, when a high-similarity distractor threatens one evidence path, alternative substitutable facts provide corroborating signals that counteract the bias. This mechanism fundamentally enhances reasoning fairness and objectivity by reducing over-reliance on any single piece of evidence. On CronQuestions, deepseek demonstrates this advantage: while its $BS@1$ declines from 0.699 (Low) to 0.679 (High), the magnitude of degradation is noticeably smaller than models without such multi-path validation capabilities.

Rethinking Reliable Reasoning with Substitutable Evidence. The experimental results demonstrate that single-retrieval-based reasoning is insufficient for reliable TKGQA under high ambiguity. Models relying solely on direct matching fail to recognize fine-grained logical deviations when distractors share surface-level similarities. Substitutable evidence serves dual purposes: it enables error correction when initial retrieval is compromised, and more importantly, it provides a core mechanism for models to objectively resolve highly confusable information. By incorporating alternative valid facts, EviScope equips models with the means to reconstruct genuine reasoning processes, effectively addressing the limitations of traditional single-ground-truth evaluation and paving the way for more robust retrieval-augmented TKGQA systems.

B Difficulty Calibration Analysis

We evaluate whether the proposed ambiguity-aware difficulty calibration provides a more reliable measure of reasoning difficulty than traditional hop-based definitions.

We sample 100 queries per difficulty level (Simple, Medium, Complex) from ICEWS Actor and CronQuestions, and evaluate GPT-3.5 and DeepSeek under different classification schemes. Due to saturation at @10, we report Top-1 performance to better reveal reasoning bottlenecks.

Results show that **hop-based difficulty is poorly calibrated**. As shown in Table 4, model performance does not consistently decrease with increasing structural complexity. On CronQuestions, several models even achieve higher $Hits@1$ on "Complex" than on "Medium" queries (e.g., GPT-3.5-serial: 0.75 $\rightarrow$ 0.90), yielding an anomalous negative Δ_{S-C}. This indicates that hop counts fail to capture retrieval ambiguity, making them unreliable as a difficulty signal for TKGQA.

In contrast, **EviScope exhibits strong calibration quality. Under our difficulty definition**, all models show a clear monotonic decline in performance. For example, GPT-3.5-serial on CronQuestions decreases consistently in $Hits@1$, with $\Delta_{S-C} = 0.13$. Similar trends appear in $BS@1$, which avoids the sharp and uninformative collapse observed under hop-based grouping and instead shows a smooth progression (e.g., 0.871 $\rightarrow$ 0.543 $\rightarrow$ 0.397). These results demonstrate that EviScope provides a more faithful and diagnos-

Table 4: Performance comparison across different difficulty classifications (S: Simple, M: Medium, C: Complex). Δ_{S-C} represents the performance drop from Simple to Complex. The Original Difficulty often presents a miscalibrated trend (e.g., negative Δ_{S-C} in CronQuestions).

Dataset	Method	Original Difficulty (Hop-based)								Designed Difficulty (EviScope)							
		Hits@1				*BS*@1				*Hits*@1				*BS*@1			
		S	M	C	Δ_{S-C}	S	M	C	Δ_{S-C}	S	M	C	Δ_{S-C}	S	M	C	Δ_{S-C}
ICEWS Actor	GPT3.5_base	0.94	0.91	0.92	0.02	0.998	0.498	0.998	0.000	0.94	0.93	0.92	0.02	0.998	0.488	0.339	0.659
	GPT3.5_serial	0.94	0.94	0.91	0.03	0.998	0.495	0.332	0.666	0.92	0.95	0.87	0.05	0.978	0.491	0.328	0.650
	deepseek	0.93	0.96	0.92	0.01	0.992	0.499	0.332	0.660	0.93	0.95	0.92	0.01	0.992	0.491	0.339	0.653
CronQuestions	GPT3.5_base	0.89	0.82	0.85	0.04	0.996	0.492	0.331	0.665	0.90	0.86	0.76	**0.14**	0.871	0.547	0.398	0.473
	GPT3.5_serial	0.85	0.75	0.90	-0.05	0.976	0.488	0.332	0.644	0.88	0.84	0.75	**0.13**	0.871	0.543	0.397	0.474
	deepseek	0.87	0.83	0.86	0.01	0.992	0.494	0.331	0.661	0.87	0.90	0.80	**0.07**	0.870	0.548	0.402	0.468

tically meaningful difficulty scale for evaluating retrieval and reasoning in TKGQA.

C Case Study

We illustrate the limitation of single-ground-truth evaluation with one example from CronQuestions: *"Who was a member of Burnley F.C. before the 2015 Tour de Wallonie?"* In the original annotation, the answer is uniquely labeled as *Jim Bellamy*, supported by *(Jim Bellamy, member of sports team, Burnley F.C., 1912–1914)* and the temporal anchor *(2015 Tour de Wallonie, significant event, occurrence, 2015)*.

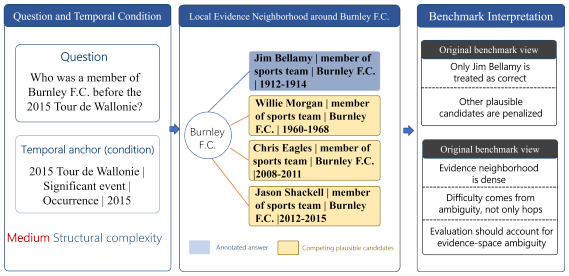

Figure 5: Evidence-space ambiguity illustrated on a CronQuestions example. Although only *Jim Bellamy* is originally annotated, multiple alternate facts co-occur "before 2015." This case proves that dense competing evidence, rather than mere structural hops, drives actual temporal retrieval difficulty.

However, the local evidence neighborhood of *Burnley F.C.* also contains other plausible candidates satisfying the same temporal condition, such as *Willie Morgan* (1960–1968), *Chris Eagles* (2008–2011), and *Jason Shackell* (2012–2015). As a result, the evidence space is not unique: a retriever may return a semantically valid candidate from the same neighborhood, yet it would still be penalized under rigid single-answer evaluation.

Although this question is labeled as *medium* in the original dataset, it is clearly more difficult than many structurally similar cases because the main challenge lies in dense competing evidence rather than hop complexity alone. This example shows why ambiguity-aware benchmarking is necessary for faithful evaluation in TKGQA.