Critic Rule Induction: Improving Temporal Knowledge Graph Forecasting with Generator-Critic Language Models

Yingsong Ning, Fu Zhang*, Jingwei Cheng, Jiashun Peng, Xiaoke Wang
School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China
{2401887, 2472039, 2401911}@stu.neu.edu.cn
{zhangfu, chengjingwei}@neu.edu.cn

Abstract

Temporal knowledge graph (TKG) forecasting aims to infer future facts from historical observations in time-evolving graphs. Traditional rule-based methods often rely on statistical co-occurrences and extensive path enumeration, suffering from rule sparsity and search-space explosion, while recent LLM-based rule reasoning can produce linguistically plausible rules that are weakly constrained by graph evidence and thus may reflect spurious correlations or violate temporal constraints. To address these challenges, we propose **C**ritic-Guided **R**ule **I**nduction (**CRI**), which treats temporal rules as *rule hypotheses* to be examined and adopts a decoupled **Generation–Discrimination** pipeline to induce rules that are both high-coverage and high-precision. CRI first mines seed rules and path evidence from the historical graph and uses an LLM-based generator to abstract and generalize them into broader *raw rule hypotheses*. It then introduces a **Fact-Grounded Rule Evaluator** to perform fact-grounded discrimination of rule hypotheses from complementary perspectives together with necessary temporal and statistical constraints. Finally, CRI performs symbolic reasoning over the *refined rule set* to produce forecasts with traceable reasoning evidence. Experiments on three benchmarks show that CRI outperforms strong baselines, achieving state-of-the-art performance on TKG forecasting.

1 Introduction

Temporal knowledge graphs (TKGs) represent time-evolving facts in the form of $(subject, relation, object, timestamp)$, providing a structured view of how relational events change over time. Forecasting future facts from historical observations in TKGs is critical for many time-sensitive applications such as finance, healthcare, and public opinion monitoring, as well as decision support (Cai et al., 2023).

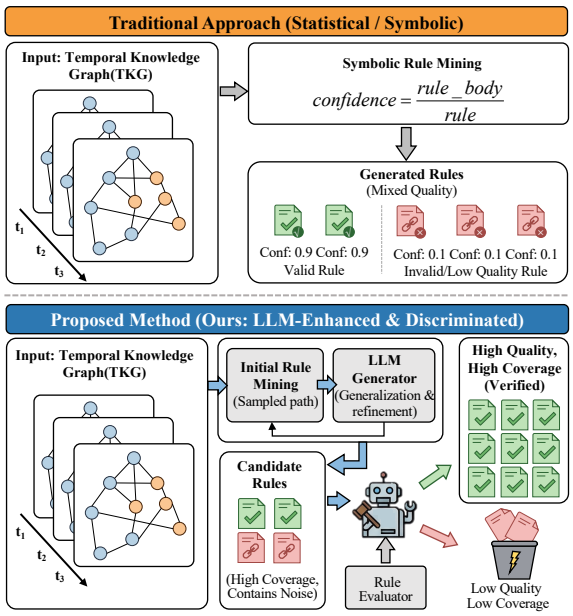

Figure 1: Comparison between traditional rule mining and the proposed LLM-enhanced & discriminated framework for TKG forecasting.

Existing TKG forecasting methods broadly fall into two lines: embedding-based and rule-based methods. *Embedding-based methods* typically treat a TKG as a temporally ordered sequence of graphs and encode historical information using graph neural networks (GNNs) (Li et al., 2021) or recurrent neural networks (RNNs) (Jin et al., 2020). While effective on the TKG forecasting task, their reasoning process is largely implicit in continuous representations, making it difficult to provide verifiable reasoning evidence. In contrast, *rule-based methods* infer future events by mining explicit temporal logical rules from historical facts, offering inherently interpretable reasoning paths (Liu et al., 2022; Li et al., 2023). However, as illustrated in Figure 1, traditional rule mining often relies on statistical co-occurrences and extensive path enumeration, and thus suffers from rule sparsity and search-space explosion. On the one hand, the mined rules of-

* Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 29436–29448

ten cover only a limited set of queries, resulting in insufficient coverage across sparse relations and entities. On the other hand, rule quality can be highly variable: even rules with high confidence may arise from spurious correlations or violate temporal causality, ultimately yielding a rule set with low coverage and mixed quality.

Recent advances in large language models (LLMs) introduce new opportunities for temporal rule reasoning. Compared to the rule mining based on enumeration and co-occurrence, LLMs are better suited to abstracting relation patterns with general semantic regularities from limited historical evidence, and thus have the potential to substantially expand rule coverage. Existing methods such as LLM-DA (Wang et al., 2024) leverage LLMs to mine temporal rules from historical data for TKG forecasting, and further update these rules with the latest events. However, it assesses rule confidence using only a single criterion: a conditional-probability estimate derived from historical facts. Yet LLM-generated rules may be weakly constrained by graph evidence, so they can be linguistically plausible but reflect spurious correlations and violate temporal constraints. As a result, this single confidence measure may not reliably reflect the quality of generated rules, undermining their reliability and ultimately limiting TKG forecasting performance. Therefore, it remains challenging to reliably assess rule confidence and effectively leverage such rules for TKG forecasting.

To this end, we propose **C**ritic-Guided **R**ule **I**nduction (**CRI**), which treats temporal rules as *rule hypotheses* to be examined and adopts a decoupled **Generation–Discrimination** pipeline to induce rules that are both high-coverage and high-precision. CRI first mines seed rules and path evidence from the historical graph, and then employs an LLM-based generator to abstract and generalize them into broader *raw rule hypotheses*, substantially improving recall across queries. CRI further introduces a **Fact-Grounded Rule Evaluator** that performs fact-grounded discrimination of the raw rule hypotheses from multiple complementary perspectives, including temporal causality consistency, semantic self-consistency, and logical coherence. Together with necessary temporal and statistical constraints, the evaluator filters and calibrates the raw rule hypotheses, removing low-quality rules. Finally, CRI performs symbolic reasoning over the *refined rule set*, producing forecasts accompanied by traceable reasoning evidence. Our contributions are summarized as follows:

- We propose CRI, a new Critic-Guided Rule Induction framework that models temporal rules as hypotheses and improves rule coverage and precision via a decoupled generation–discrimination design.
- We design a Fact-Grounded Rule Evaluator that discriminates and filters rules with multi-dimensional evidence (temporal constraints, semantics, and logical coherence), reliably assessing rule confidence and facilitating effective TKG forecasting.
- Experiments on three benchmarks show that CRI outperforms strong embedding/rule- and LLM-based baselines, achieving state-of-the-art (SOTA) performance on TKG forecasting.

2 Related Work

2.1 Traditional TKG Forecasting Methods

TKG reasoning is typically studied under two task settings (Cai et al., 2023): interpolation and extrapolation. TKG interpolation aims to predict missing entities at past timestamps, whereas TKG extrapolation (TKG forecasting) aims to predict missing entities at future timestamps. In this work, we focus on TKG forecasting. The embedding-based methods model TKGs as a temporally ordered sequence of graphs and capture temporal dynamics with neural architectures (Jin et al., 2020; Han et al., 2020; Zhu et al., 2021; Li et al., 2022).

In parallel, rule-oriented methods aim for interpretable reasoning by mining temporal rules via random walks, e.g., TLogic (Liu et al., 2022) and TR-Rules (Li et al., 2023). Despite their strong performance, they often rely on confidence-based statistical selection and limited temporal constraints for rule validation, which can yield rule sets with insufficient coverage and mixed quality.

2.2 LLM-based TKG Forecasting Methods

With the recent progress of LLMs in semantic understanding and pattern induction, a growing body of work explores their use for TKG forecasting. One line adopts in-context learning (ICL) as the core inference paradigm, as exemplified by ICL (Lee et al., 2023) and PPT (Xu et al., 2023). Another line enhances LLMs by fine-tuning on historical TKG data (Liao et al., 2024) or exploring higher-order histories (Xia et al., 2024).

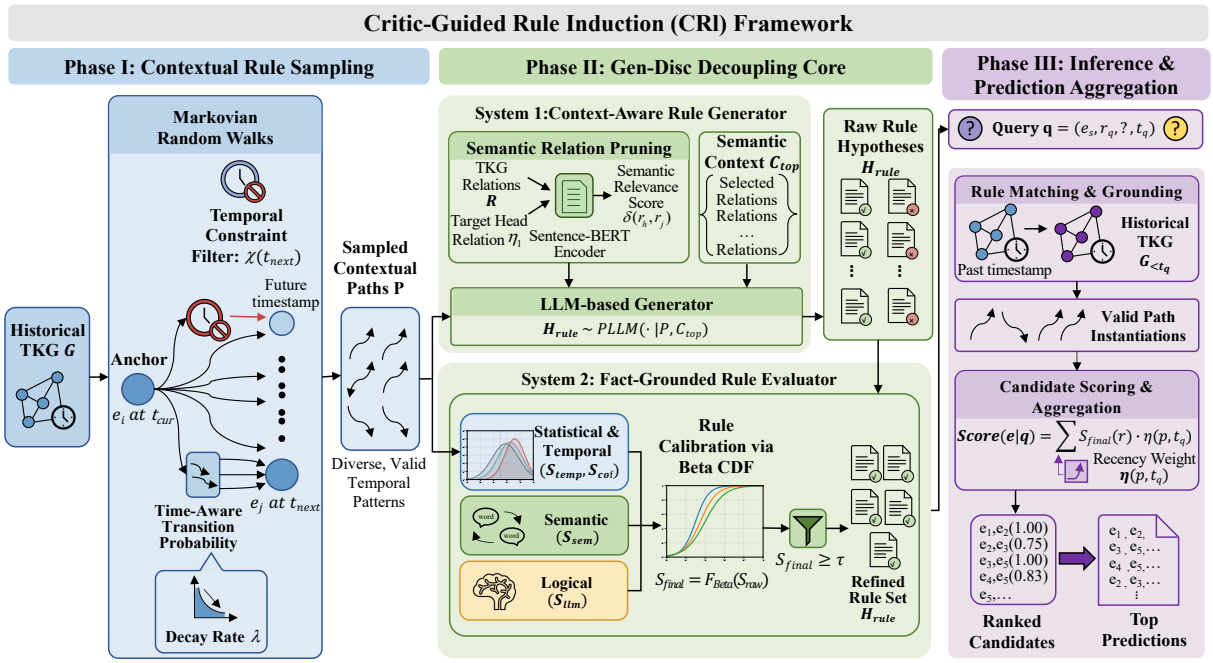

Figure 2: Overview of our CRI framework. It consists of three key phases: (1) Contextual Rule Sampling extracts high-quality historical path evidence via constrained Markovian random walks; (2) The Gen-Disc Decoupling Core synthesizes temporal rules through a generator-discriminator paradigm, where System 1 induces hypotheses and System 2 performs multi-dimensional fact-grounded evaluation; (3) Inference & Prediction Aggregation executes symbolic reasoning by aggregating calibrated rule scores and temporal recency to forecast future entities.

In the static KG setting, ChatRule (Luo et al., 2025) demonstrates that LLMs can mine logical rules by grounding them with KG facts, though it does not address temporal constraints. More recently, some studies explore **combining LLMs with temporal patterns or rules** to further enhance TKG forecasting. G2S follows a general-to-specific generation paradigm, aiming to instantiate abstract temporal patterns into concrete time-evolving fact predictions (Bai et al., 2025). AnRe leverages analogical reasoning by retrieving similar historical event chains to provide query-specific contexts (Tang et al., 2025). LLM-DA (Wang et al., 2024) highlights the advantages of combining rules with LLMs for TKG forecasting. It leverages LLMs to analyze historical data to extract temporal rules and evaluates rule confidence using a conditional-probability estimate computed from historical facts. Despite these advances, LLM-generated rules can be weakly constrained by graph evidence; consequently, they may be linguistically plausible yet reflect spurious correlations or violate temporal constraints. In such cases, this confidence measure may not reliably reflect the true quality of generated rules, undermining their reliability. To address this gap, we propose CRI, which treats temporal rules as evaluable hypotheses and employs a fact-grounded generation–discrimination procedure to filter candidate rules, thereby improving rule reliability and facilitating TKG forecasting.

3 Preliminaries

3.1 Temporal Knowledge Graph Forecasting

A TKG is formalized as $\mathcal{G} = \{(e_s, r, e_o, t) \mid e_s, e_o \in \mathcal{E}, r \in \mathcal{R}, t \in \mathcal{T}\}$, where $\mathcal{E}$, $\mathcal{R}$, and $\mathcal{T}$ denote the sets of entities, relations, and timestamps, respectively. Given a set of observed historical facts $\mathcal{G}_{<t_q}$ (all facts where $t < t_q$), TKG forecasting aims to predict missing facts at a future timestamp t_q. Specifically, for a query $q = (e_s, r_q, ?, t_q)$, the model computes a probability score for each entity in $\mathcal{E}$ and outputs the most plausible object entity. To support bidirectional reasoning, for every fact, we also include its inverse relation (e_o, r^{-1}, e_s, t).

3.2 Temporal Logical Rules

A temporal Horn rule R (Liu et al., 2022) is defined as $Body \rightarrow Head$, which can be formalized as:

$$r_1(e_s, e_1, t_1) \wedge \cdots \wedge r_L(e_{L-1}, e_o, t_L) \rightarrow r_h(e_s, e_o, t_h) \quad (1)$$

where $e_s, e_1, \ldots, e_{L-1}, e_o$ are entity variables. The rule length is the number of body atoms L. A valid temporal rule requires that all body timestamps strictly precede the head timestamp:

$$t_1 \leq t_2 \leq \cdots \leq t_L < t_h \tag{2}$$

4 Methodology

We propose **C**ritic-Guided **R**ule **I**nduction (**CRI**), a novel framework that models temporal rules as hypotheses and improves rule coverage and precision via a decoupled Generation–Discrimination design and a Fact-Grounded Rule Evaluator. Figure 2 illustrates the overall framework.

4.1 Phase I: Contextual Rule Sampling

CRI extracts path instances via a controlled Markovian random walk on $\mathcal{G}$ (Liu et al., 2022). Given an anchor entity e_i at timestamp t_{cur}, the transition to a neighbor e_j via relation r at t_{next} relies on temporal validity and interval weight.

Temporal Constraint Filter. To ensure causal logic relative to the target timestamp t_h in Eq. (1), we define a filter $\chi(t_{next})$ that enforces strict past occurrence:

$$\chi(t_{next}) = \begin{cases} 1, & \text{if } t_{next} < t_{cur} \leq t_h \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

This guarantees the sampled path $e_s \xrightarrow{t_1} e_1 \ldots \xrightarrow{t_L} e_L$ follows a reverse chronological order.

Time-Aware Transition Probability. For valid neighbors where $\chi(t_{next}) = 1$, we define the transition probability $Pro(e_j|e_i, t_{cur})$ using exponential decay to favor temporally adjacent facts:

$$Pro(e_j, r, t_{next} \mid e_i, t_{cur}) = \frac{\exp(-\lambda|t_{cur} - t_{next}|)}{\sum_{(e',r',t') \in \mathcal{N}(e_i)} \chi(t') \cdot \exp(-\lambda|t_{cur} - t'|)} \tag{4}$$

where $\mathcal{N}(e_i)$ denotes the set of temporal neighbors, and λ is the decay rate. The resulting paths $\mathcal{P} = \{p_1, \ldots, p_N\}$ serve as input for Phase II.

4.2 Phase II: Gen-Disc Decoupling Core

This phase is central to CRI, comprising two interacting systems: a Context-Aware Rule Generator (System 1) and a Fact-Grounded Rule Evaluator (System 2).

4.2.1 System 1: Context-Aware Rule Generator

To enhance the coverage of sparse patterns, we design a generator featuring a *Semantic Relation Pruning* module to constrain the search space.

Semantic Relation Pruning. We prune the search space by identifying relations semantically relevant to the target r_h. We employ a pre-trained Sentence-BERT encoder (Reimers and Gurevych, 2019) to vectorize r_h and candidates $r_j \in \mathcal{R}$:

$$\mathbf{v}_{r_h}, \mathbf{v}_{r_j} = \text{Encoder}(r_h, r_j) \tag{5}$$

We then compute the semantic relevance score $\delta(r_h, r_j)$ via cosine similarity:

$$\delta(r_h, r_j) = \frac{\mathbf{v}_{r_h} \cdot \mathbf{v}_{r_j}}{\|\mathbf{v}_{r_h}\| \|\mathbf{v}_{r_j}\|} \tag{6}$$

We select the top-k relations with the highest semantic relevance scores to form the context set $\mathcal{C}_{top} = \text{Top-k}(\{\delta(r_h, r_j) \mid r_j \in \mathcal{R}\})$, allowing the LLM to focus on relations semantically aligned with r_h.

Rule Hypothesis Generation. The LLM functions as a pattern inductor using path instances $\mathcal{P}$ (from Phase I) and context $\mathcal{C}_{top}$. It is instructed to generate *rule hypotheses* $\mathcal{H}_{rule}$ using only relations from $\mathcal{C}_{top}$:

$$\mathcal{H}_{rule} \sim P_{\text{LLM}}(\cdot \mid \mathcal{P}, \mathcal{C}_{top}) \tag{7}$$

Grounding generation in statistical evidence $\mathcal{P}$ and semantic constraints $\mathcal{C}_{top}$ ensures structural validity and semantic focus (see Appendix B).

4.2.2 System 2: Fact-Grounded Rule Evaluator

The Discriminator $\mathcal{D}$ (System 2) rigorously evaluates raw hypotheses $\mathcal{H}_{rule}$. Unlike existing methods such as LLM-DA (Wang et al., 2024), which rely solely on body-head co-occurrence, we address the critical issue where LLM-generated rules reflect spurious correlations or violate temporal causality. Thus, we evaluate each hypothesis from *three complementary dimensions*, followed by a non-linear calibration.

Temporal/Statistical Grounding. Valid rules must be supported by historical evidence and respect causal ordering. For a candidate rule $r \in \mathcal{H}_{rule}$, we retrieve matching path instances $\mathcal{I}_{body} = \{p \mid p \vDash Body(r)\}$ from $\mathcal{G}_{<t_q}$. First, to ensure

causality, we enforce a strict temporal constraint $\Phi_{\text{time}}(p) = (t_1 \leq \cdots \leq t_L < t_h)$ relative to Eq. (1). The *temporal constraint score* is calculated as the validity ratio:

$$S_{\text{temp}}(r) = \frac{\sum_{p \in \mathcal{I}_{body}} \mathbb{I}(\Phi_{\text{time}}(p))}{|\mathcal{I}_{body}| + \epsilon} \tag{8}$$

where $\mathbb{I}(\cdot)$ is the indicator function and ϵ is a smoothing term. Second, to evaluate rule confidence, we calculate the *historical confidence score* (Agrawal et al., 1993) as the conditional probability:

$$S_{\text{conf}}(r) = \frac{\text{supp}(Body(r) \wedge Head(r))}{\text{supp}(Body(r)) + \epsilon_{\text{conf}}} \tag{9}$$

where $\text{supp}(\cdot)$ counts occurrences satisfying temporal constraints. Crucially, the denominator $\text{supp}(Body(r))$ captures all premise activations—including counter-examples where the head is absent—thereby penalising rules that are statistically frequent but lack precision.

Semantic Consistency. To capture intrinsic semantic plausibility, we utilize a zero-shot evaluator based on embedding similarity using templates $\mathcal{H}^+$ and $\mathcal{H}^-$ (see Appendix B). We encode the rule r and templates into embeddings $\mathbf{v}_r$ and $\mathbf{v}_t$ using Sentence-BERT. The *semantic consistency score* is:

$$S_{\text{sem}}(r) = \sigma\left(\max_{t \in \mathcal{H}^+} \text{sim}(\mathbf{v}_r, \mathbf{v}_t) - \max_{t \in \mathcal{H}^-} \text{sim}(\mathbf{v}_r, \mathbf{v}_t)\right) \tag{10}$$

LLM-based Logical Coherence. Since embeddings may overlook nuanced logical inconsistencies, we incorporate an *LLM-as-a-Judge* (Zheng et al., 2023) module. The LLM evaluates the rule r on a continuous scale $S_{\text{llm}}(r) \in [0, 1]$, based on criteria such as causal plausibility and generalization ability.

Rule Calibration via Beta CDF. We aggregate the results into S_{raw}. To achieve sharper separation between high- and low-quality rules, we apply a non-linear calibration using the regularized Beta CDF (Kull et al., 2017):

$$\begin{aligned} S_{\text{final}}(r) &= F_{\text{Beta}}(S_{\text{raw}}(r); \alpha, \beta) \\ &= \int_0^{S_{\text{raw}}(r)} \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}\, dx \end{aligned} \tag{11}$$

where $\alpha, \beta > 1$ control the transition steepness. Finally, we retain rules with $S_{\text{final}}(r) \geq \tau$ to form the *refined rule set* $\mathcal{H}'_{rule}$.

4.3 Phase III: Inference & Prediction Aggregation

In this phase, we deploy $\mathcal{H}'_{rule}$ to forecast the query $q = (e_s, r_q, ?, t_q)$ via rule matching and score aggregation.

4.3.1 Rule Matching

We retrieve rules $r \in \mathcal{H}'_{rule}$ matching r_q and ground them in $\mathcal{G}_{<t_q}$. Valid paths p starting from e_s must satisfy the rule structure and temporal constraints (events occurring before t_q).

4.3.2 Candidate Scoring and Aggregation

Valid paths point to candidate entities e. We aggregate the contributions of rules and paths to calculate the confidence score:

$$Score(e \mid q) = \sum_{r \in \mathcal{R}_q} \sum_{p \in \mathcal{M}(r,e)} S_{\text{final}}(r) \cdot \eta(p, t_q) \tag{12}$$

where $\mathcal{M}(r, e)$ is the set of paths of rule r leading to e, and $S_{\text{final}}(r)$ is the calibrated score in Eq. (11). The time decay $\eta(p, t_q)$ is defined as:

$$\eta(p, t_q) = \exp(-\lambda |t_q - t_{last}(p)|) \tag{13}$$

where $t_{last}(p)$ is the timestamp of latest event in path p, and λ is the decay rate. This weighting reflects the assumption that paths grounded in more recent evidence are more predictive for future queries. We rank candidates by their aggregated scores to select top predictions. The CRI algorithm is provided in Appendix E.

5 Experiments

5.1 Experimental Settings

Datasets. We evaluate our framework on three standard TKG benchmarks: ICEWS14 (García-Durán et al., 2018), ICEWS18 (Jin et al., 2020), and ICEWS05-15 (García-Durán et al., 2018). Detailed statistics are provided in Appendix A.

Baselines. We compare CRI against comprehensive state-of-the-art methods, categorized into three groups. **Embedding-based Methods:** encode temporal dynamics into continuous vector spaces, including RE-NET (Jin et al., 2020), RE-GCN (Li et al., 2021), and TiRGN (Li et al., 2022). **Rule-based Methods:** mine logical rules for TKG forecasting, including TLogic (Liu et al., 2022) and TR-Rules (Li et al., 2023). **LLM-based & Hybrid Methods:** leverage the power of LLMs, including ICL-based GPT-NeoX (Lee et al., 2023) and PPT

Type	Models	Train	ICEWS14				ICEWS05-15				ICEWS18			
			MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
♠	RE-NET	✓	0.383	0.287	0.413	0.545	0.430	0.313	0.469	0.635	0.288	0.191	0.324	0.475
	RE-GCN	✓	0.425	0.320	0.476	0.627	0.478	0.371	0.535	0.682	0.326	0.224	0.368	0.527
	TiRGN	✓	0.441	0.341	0.497	0.650	0.495	0.389	0.559	0.703	0.336	0.232	0.380	0.542
	TLogic	✓	0.390	0.295	0.437	0.573	0.459	0.360	0.518	0.646	0.298	0.205	0.340	0.485
	TR-Rules	✓	0.433	0.340	0.486	0.612	0.476	0.371	0.538	0.676	0.304	0.211	0.346	0.489
♣	PPT	✓	0.384	0.289	0.425	0.570	0.389	0.286	0.434	0.586	0.266	0.169	0.306	0.454
	Llama-2-7b-CoH	✓	–	0.349	0.470	0.591	–	0.386	0.541	0.699	–	0.219	0.361	0.520
	Vicuna-7b-CoH	✓	–	0.328	0.457	0.656	–	0.392	0.546	0.707	–	0.206	0.344	0.531
	GPT-NeoX	✗	–	0.334	0.460	0.565	–	–	–	–	–	0.192	0.313	0.414
	Mixtral-8x7B-CoH	✗	0.439	0.331	0.496	0.649	0.497	0.380	0.564	0.713	–	–	–	–
◊	G2S	✓	–	0.383	0.541	0.686	–	–	–	–	–	0.230	0.353	0.466
	LLM-DA (RE-GCN)	✗	0.461	0.356	0.515	0.662	0.501	0.394	0.568	0.710	–	–	–	–
	LLM-DA (TiRGN)	✗	0.471	0.369	0.526	0.671	0.521	0.416	0.586	0.728	–	–	–	–
	AnRe	✗	0.474	0.369	0.511	0.657	0.509	0.391	0.580	0.696	0.355	0.260	0.392	0.567
Ours	**CRI**	✗	**0.498**	**0.381**	**0.551**	**0.702**	**0.531**	**0.407**	**0.593**	**0.751**	**0.388**	**0.274**	**0.421**	**0.578**

Table 1: Performance comparison of TKG forecasting on ICEWS14, ICEWS05-15, and ICEWS18. The best results are **bolded** and the second-best results are underlined. ♠ denotes traditional embedding/rule-based methods, ♣ represents pure LLM-based approaches, and ◊ indicates Hybrid methods.

(Xu et al., 2023); chain-of-history (CoH)-based (Xia et al., 2024); and further combine LLMs with temporal patterns G2S (Bai et al., 2025) and AnRe (Tang et al., 2025); as well as the rule-enhanced method LLM-DA (Wang et al., 2024).

Implementation Details. For the *Context-Aware Rule Generation* and *LLM-based Logical Coherence* modules, we utilize `gpt-3.5-turbo-0125` via the OpenAI API to ensure cost-effectiveness and reproducibility. We evaluate performance using standard metrics: *Mean Reciprocal Rank* (MRR) and *Hits@N* ($N = 1, 3, 10$) under the *time-aware filtered* setting. For hyperparameters, we set the temporal decay rate $\lambda = 0.1$ in Phase I and III. The discriminator threshold τ is empirically set to 0.5. All experiments are conducted on a single NVIDIA RTX 3090 GPU.

We provide the **parameter sensitivity analysis** in Appendix C and the **cost and efficiency analysis** in Appendix D.

5.2 Main Results

Table 1 presents the TKG forecasting performance on all three datasets. Overall, **CRI** achieves consistent state-of-the-art performance, outperforming both strong embedding baselines and competitive LLM-based approaches.

Superiority over Rule-based Baselines. Compared to the traditional rule-based TLogic and TR-Rules, CRI demonstrates significant improvements, particularly in the Hit@10 metric (**+12.9%** over TLogic on ICEWS14). Traditional methods suffer from rule sparsity as they rely on strict statistical path matching. In contrast, CRI leverages the semantic generalization capability of the LLM Generator to induce high-coverage rule hypotheses, effectively capturing long-tail and semantic patterns that rigid rule enumeration misses.

Superiority over LLM-based Methods. Pure LLM methods like GPT-NeoX often yield suboptimal performance due to ungrounded reasoning. Even compared to the strong hybrid LLM & rule baseline LLM-DA, CRI achieves notable gains (**+2.7%** MRR on ICEWS14). While LLM-DA relies on a single confidence score, CRI's *Fact-Grounded Rule Evaluator* filters noise via multi-dimensional discrimination (temporal, semantic, and logical). This ensures that the generated rules are not only high-coverage but also high-precision, leading to more reliable TKG forecasting.

5.3 Impact of Model Backbones and Scale

To evaluate the robustness and scalability of CRI, we examine the impact of different LLM backbones and scales on forecasting quality. We select five representative models, ranging from efficient open-source models (Qwen2.5-7B, InternLM2.5-7B, Llama-3-8B) to larger-scale or closed-source models (GPT-3.5-Turbo, Llama-3.3-70B).

Table 2 summarizes the performance across three datasets. First, regarding model architectures, CRI demonstrates strong robustness across different backbones. Efficient 7B-parameter models like Qwen2.5 and Llama-3-8B achieve competitive results comparable to GPT-3.5-Turbo, indicating that our framework effectively activates the reasoning potential of smaller models. Second, regarding

Model Backbone	ICEWS14				ICEWS05-15				ICEWS18			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
InternLM2.5-7B	48.7	36.9	54.1	69.5	52.1	39.4	58.0	74.2	37.5	26.3	41.0	56.8
Qwen2.5-7B	49.6	37.8	55.0	70.3	52.9	40.4	59.1	74.9	39.0	27.6	42.0	57.6
Llama-3-8B	49.3	37.5	54.8	69.9	52.6	40.1	58.8	74.7	38.2	26.9	41.7	57.3
GPT-3.5-Turbo	49.8	38.1	55.1	70.2	53.1	40.7	59.3	75.1	38.8	27.4	42.1	57.8
Llama-3.3-70B	**60.6**	**51.1**	**64.6**	**78.1**	**65.9**	**55.8**	**71.5**	**84.4**	**56.4**	**48.4**	**58.7**	**70.2**

Table 2: Impact of LLM backbones and model scale on forecasting performance. The table compares standard-scale models (∼7B) with a large-scale model (70B), demonstrating both the stability of CRI on smaller backbones and its scalability with larger models.

model scale, increasing the parameter count yields performance gains. Within the Llama-3 family, scaling from 8B to 70B results in a substantial improvement (e.g., +18.2% MRR on ICEWS18). This trend suggests that CRI can effectively leverage the superior reasoning capabilities of large-scale models to capture complex temporal patterns, demonstrating the framework's potential for scalability.

5.4 Ablation Study

To assess the contribution of each component in CRI, we conduct a comprehensive ablation study on the ICEWS14 dataset. We compare the full model against several variants by removing key modules. The results are summarized in Table 3.

System Variant	MRR	H@1	H@3	H@10
CRI (Full Model)	**0.498**	**0.381**	**0.551**	**0.702**
w/o Generator (Sys. 1)	0.390	0.295	0.437	0.573
w/o Discriminator (Sys. 2)	0.462	0.358	0.518	0.665
w/o Rule Calibration	0.489	0.373	0.542	0.696

Table 3: System-level ablation study. "w/o Generator" replaces the LLM with traditional path mining (TLogic); "w/o Discriminator" bypasses the evaluation phase; "w/o Calibration" uses linear score aggregation.

Impact of Rule Generator (System 1). Replacing the LLM-based Generator with traditional constrained random walks (essentially reverting to TLogic (Liu et al., 2022)) results in the most significant performance drop (**-21.7%** in MRR). This confirms that purely symbolic methods suffer from severe sparsity and fail to capture complex patterns. The LLM's ability to abstract generalized rules from limited paths is the primary driver of CRI's high coverage.

Impact of Fact-Grounded Discriminator (System 2). Removing the Discriminator and treating all LLM-generated rules as valid leads to a 7.2% drop in MRR. This underscores the necessity of rigorous verification: while LLMs excel at abstracting patterns, they frequently generate rules that are linguistically plausible yet lack historical grounding or violate temporal causality. System 2 effectively filters out such spurious noise, ensuring high precision.

Effectiveness of Calibration. Replacing the nonlinear Beta CDF calibration with simple linear aggregation causes a slight performance decline. The Beta CDF helps to sharpen the decision boundary, effectively separating high-quality rules from ambiguous ones, which linear summation fails to achieve efficiently.

5.5 Effectiveness of Multi-Dimensional Discrimination

To verify the contribution of each evaluation dimension defined in the Methodology, we perform a fine-grained ablation within System 2. We selectively disable each of the four scoring metrics: Temporal Constraint (S_{temp}), Statistical Confidence (S_{conf}), Semantic Consistency (S_{sem}), and Logical Coherence (S_{llm}). The results in Table 4 demonstrate that all dimensions are complementary.

Ablated Metric	MRR	H@1	H@3	H@10
CRI (Full Metrics)	**0.498**	**0.381**	**0.551**	**0.702**
w/o Temporal (S_{temp})	0.482	0.365	0.535	0.688
w/o Confidence (S_{conf})	0.485	0.368	0.538	0.690
w/o Semantic (S_{sem})	0.490	0.375	0.544	0.695
w/o Logical (S_{llm})	0.493	0.378	0.548	0.699

Table 4: Fine-grained ablation of the four scores within the Fact-Grounded Evaluator. Each metric addresses a specific type of noise in generated rules.

Temporal & Statistical Grounding (S_{temp}, S_{conf}). The removal of the Temporal Constraint (S_{temp}) results in the most significant performance drop,

confirming that strict causal ordering is critical to prevent information leakage. Similarly, excluding Statistical Confidence (S_{conf}) impairs the distinction between robust historical patterns and rare coincidences, leading to notable degradation.

Semantic & Logical Consistency ($S_{\text{sem}}, S_{\text{llm}}$). Ablating Semantic Consistency (S_{sem}) or Logical Coherence (S_{llm}) causes moderate declines. While statistical metrics target factual errors, these filters specifically eliminate logically unsound patterns—such as tautologies or irrelevant associations—that may coincidentally possess data support. Their integration ensures that retained rules are both empirically grounded and interpretably valid.

Overall, removing any single dimension of the critic degrades performance. Notably, removing the temporal constraint causes a substantial drop, highlighting the importance of strictly enforcing causal ordering in TKG forecasting. The statistical and semantic scores act as complementary filters, ensuring rules are both empirically grounded and logically sound.

5.6 Performance on Sparse Data

A key advantage of CRI is its ability to handle data sparsity via LLM-based generalization. We divide the relations in ICEWS14 into three groups based on their frequency: *Sparse* (< 50 instances), *Medium* ($50-500$), and *Frequent* (> 500). We compare the Hits@10 performance of CRI against TLogic (rule-based baseline) and RE-GCN (embedding-based baseline).

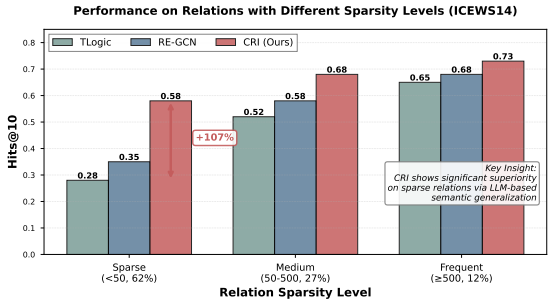

Figure 3: Hits@10 performance across relation groups with different sparsity levels.

As illustrated in Figure 3, traditional symbolic methods like TLogic perform poorly on sparse relations due to the lack of sufficient path evidence for rule mining. Embedding methods also struggle to learn distinct representations for rare relations. In contrast, CRI outperforms baselines significantly on the *Sparse* group. This confirms that our *Generator* successfully transfers semantic knowledge from frequent patterns to rare ones, constructing high-quality rules even with limited observations.

5.7 Case Study

Case 1: Improved Coverage (vs. TLogic). *Query:* (South Korea, `Sign_agreement`, ?, 2014-05-12). *Target:* Japan.

TLogic fails to answer this query because the specific path pattern has low support in history. However, CRI's Generator successfully induces a generalized rule:

$$r_{gen}: \text{Visit}(A, B, t_1) \wedge \text{Express\_intent\_coop}(B, A, t_2) \rightarrow \text{Sign\_agreement}(A, B, t_3) \quad (t_1 < t_2 < t_3)$$

Although the exact path sequence was rare, the semantic pattern "Visit $\rightarrow$ Intent $\rightarrow$ Agreement" is generalized by the LLM. CRI retrieves this rule and correctly predicts *Japan*.

Case 2: Improved Precision (vs. LLM-DA). *Query:* (Citizen, `Protest`, ?, 2014-08-01). *Target:* Government. LLM-DA generates a high-confidence rule based on spurious correlation: "If *Protest* happens, then *Appeal* happens". While statistically frequent, the temporal order is often reversed in the graph. CRI's Discriminator detects this causality violation ($S_{temp} \approx 0$) and rejects the rule. Instead, CRI uses a strictly causal rule:

$$r_{cri}: \text{Impose\_Sanction}(Gov, Cit, t_1) \rightarrow \text{Protest}(Cit, Gov, t_2) \quad (t_1 < t_2)$$

This rule respects temporal constraints, leading to a correct prediction, whereas LLM-DA's hallucinatory rule leads to an incorrect entity.

5.8 Error Analysis

We identify three dominant failure modes of CRI by examining incorrect predictions on ICEWS14.

Entity Ambiguity. The gap between Hits@10 (0.702) and Hits@1 (0.381) reveals that the correct entity frequently appears in the candidate set but is not ranked first. This occurs when multiple entities share similar relational patterns under the matched rules, making top-1 disambiguation difficult. Rule-based scoring aggregates structural evidence but lacks entity-level features to distinguish semantically similar candidates.

Insufficient Path Evidence. While the LLM Generator mitigates sparsity for rare *relations* (Figure 3), rare *entities* that appear infrequently in the training graph still lack sufficient grounding paths for precise scoring. When $|\mathcal{I}_{body}|$ is small, the discriminator scores become unreliable, and the temporal recency weighting $\eta(p, t_q)$ has limited signal to differentiate candidates.

Semantic Pruning Misses. The top-k semantic pruning may exclude relations with low surface-level similarity to the query relation but valid causal connections. Relaxing k increases recall but introduces noise, suggesting an adaptive pruning strategy as a promising direction for future work.

6 Conclusion

In this paper, we propose **C**ritic-Guided **R**ule **I**nduction (CRI), a novel framework designed to enhance Temporal Knowledge Graph (TKG) forecasting by combining the generative capabilities of LLMs with rigorous, fact-grounded discrimination. Specifically, CRI first employs a generator to abstract broad rule hypotheses from historical paths, effectively addressing the rule sparsity issue. To mitigate spurious correlations, we design a multi-dimensional discriminator that rigorously evaluates rules by assessing temporal causality, semantic coherence, and statistical support. This decoupled paradigm ensures that the induced rules cover a wide range of scenarios (high-coverage) while maintaining accuracy and relevance (high-precision). Experimental results demonstrate that CRI achieves state-of-the-art performance against strong baselines.

Limitations

Despite CRI's superior performance, we acknowledge limitations common to LLM-based frameworks. First, relying on LLMs for induction introduces higher computational costs compared to embedding-based baselines. Although offline induction mitigates online overhead, the initial API resource consumption remains significant. Second, rule quality is bounded by the backbone model. While our discriminator filters noise, challenges persist if the LLM generates spurious content or lacks specific domain knowledge for niche events. Finally, our constrained random walk prioritizes local temporal dynamics. Patterns requiring global structural awareness or extremely long-term dependencies might be less effectively captured. Future work will explore distilling rules into smaller models to reduce API dependency.

Acknowledgments

The authors sincerely thank the anonymous reviewers for their valuable comments and suggestions, which have greatly improved this paper. This work is supported by the National Natural Science Foundation of China (62276057).

References

Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216.

Long Bai, Zixuan Li, Xiaolong Jin, Jiafeng Guo, Xueqi Cheng, and Tat-Seng Chua. 2025. G2S: A general-to-specific learning framework for temporal knowledge graph forecasting with large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20927–20938, Vienna, Austria. Association for Computational Linguistics.

Borui Cai, Yong Xiang, Longxiang Gao, He Zhang, Yunfeng Li, and Jianxin Li. 2023. Temporal knowledge graph completion: a survey. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6545–6553.

Alberto García-Durán, Sebastijan Dumančić, and Mathias Niepert. 2018. Learning sequence encoders for temporal knowledge graph completion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4816–4821, Brussels, Belgium. Association for Computational Linguistics.

Zhen Han, Peng Chen, Yunpu Ma, and Volker Tresp. 2020. Explainable subgraph reasoning for forecasting on temporal knowledge graphs. In *International conference on learning representations*.

Woojeong Jin, Meng Qu, Xisen Jin, and Xiang Ren. 2020. Recurrent event network: Autoregressive structure inferenceover temporal knowledge graphs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6669–6683, Online. Association for Computational Linguistics.

Meelis Kull, Telmo Silva Filho, and Peter Flach. 2017. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial intelligence and statistics*, pages 623–631. PMLR.

Dong-Ho Lee, Kian Ahrabian, Woojeong Jin, Fred Morstatter, and Jay Pujara. 2023. Temporal knowledge graph forecasting without knowledge using in-

context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 544–557, Singapore. Association for Computational Linguistics.

Ningyuan Li, E Haihong, Shi Li, Mingzhi Sun, Tianyu Yao, Meina Song, Yong Wang, and Haoran Luo. 2023. Tr-rules: Rule-based model for link forecasting on temporal knowledge graph considering temporal redundancy. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 7885–7894.

Yujia Li, Shiliang Sun, and Jing Zhao. 2022. Tirgn: Time-guided recurrent graph network with local-global historical patterns for temporal knowledge graph reasoning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 2152–2158. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Zixuan Li, Xiaolong Jin, Wei Li, Saiping Guan, Jiafeng Guo, Huawei Shen, Yuanzhuo Wang, and Xueqi Cheng. 2021. Temporal knowledge graph reasoning based on evolutional representation learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 408–417.

Ruotong Liao, Xu Jia, Yangzhe Li, Yunpu Ma, and Volker Tresp. 2024. GenTKG: Generative forecasting on temporal knowledge graph with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4303–4317, Mexico City, Mexico. Association for Computational Linguistics.

Yushan Liu, Yunpu Ma, Marcel Hildebrandt, Mitchell Joblin, and Volker Tresp. 2022. Tlogic: Temporal logical rules for explainable link forecasting on temporal knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 4120–4127.

Linhao Luo, Jiaxin Ju, Bo Xiong, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. 2025. Chatrule: Mining logical rules with large language models for knowledge graph reasoning. In *Pacific-asia conference on knowledge discovery and data mining*, pages 314–325. Springer.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Guo Tang, Zheng Chu, Wenxiang Zheng, Junjia Xiang, Yizhuo Li, Weihao Zhang, Ming Liu, and Bing Qin. 2025. AnRe: Analogical replay for temporal knowledge graph forecasting. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4632–4650, Vienna, Austria. Association for Computational Linguistics.

Jiapu Wang, Kai Sun, Linhao Luo, Wei Wei, Yongli Hu, Alan W Liew, Shirui Pan, and Baocai Yin. 2024. Large language models-guided dynamic adaptation for temporal knowledge graph reasoning. *Advances in Neural Information Processing Systems*, 37:8384–8410.

Yuwei Xia, Ding Wang, Qiang Liu, Liang Wang, Shu Wu, and Xiaoyu Zhang. 2024. Chain-of-history reasoning for temporal knowledge graph forecasting. *Preprint*, arXiv:2402.14382.

Wenjie Xu, Ben Liu, Miao Peng, Xu Jia, and Min Peng. 2023. Pre-trained language model with prompts for temporal knowledge graph completion. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7790–7803, Toronto, Canada. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Cunchao Zhu, Muhao Chen, Changjun Fan, Guangquan Cheng, and Yan Zhang. 2021. Learning from history: Modeling temporal knowledge graphs with sequential copy-generation networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 4732–4740.

A Dataset Details

We evaluate our proposed framework on three standard TKG benchmarks: **ICEWS14** (García-Durán et al., 2018), **ICEWS18** (Jin et al., 2020), and **ICEWS05-15** (García-Durán et al., 2018). These datasets are subsets of the *Integrated Crisis Early Warning System (ICEWS)*, recording timestamped political events between international entities. ICEWS14 and ICEWS18 cover events in 2014 and 2018 respectively, while ICEWS05-15 spans a longer period from 2005 to 2015, providing a robust testbed for long-term temporal reasoning. Table 5 presents the detailed statistics.

B Prompts and Semantic Templates

B.1 Rule Generation Prompt

Table 6 details the specific prompt structure used to instruct the LLM for temporal rule induction. We employ a few-shot prompting strategy to guide the model in understanding the temporal Horn rule format.

Dataset	Schema	# Entity	# Relation	# Train	# Valid	# Test	Granularity
ICEWS14	CAMEO	7,128	230	74,845	8,514	7,371	1 day
ICEWS18	CAMEO	23,033	256	373,018	45,995	49,545	1 day
ICEWS05-15	CAMEO	10,488	251	368,868	46,302	46,159	1 day

Table 5: Statistics of the datasets. "Entity/Relation" denotes the number of entities/relations, and "Train/Valid/Test" denotes the number of facts in each set.

Component	**Content / Template**
System Role	You are an expert in temporal logic reasoning. Your task is to abstract general logical rules from specific historical paths.
Task Instruction	Given a set of temporal paths starting from entity e_s, identify relational patterns that logically lead to the target relation r_q. Ignore specific entity names and focus on relations and temporal order.
Few-shot Examples	**Input Path:** (US, negotiate, Iran, t_1) $\rightarrow$ (Iran, sign_deal, US, t_2) **Extracted Rule:** negotiate(A, B, T_1) $\wedge$ sign_deal(B, A, T_2) $\rightarrow$ improve_relations(A, B, T_3) where $T_1 \leq T_2 < T_3$.
Context Input	**Query:** $(e_s, r_q, ?, t_q)$ **Sampled Paths:** $\mathcal{P} = \{p_1, \ldots, p_N\}$
Output Format	Generate a list of temporal Horn rules: $r_1, r_2, \ldots$

Table 6: Prompt template for rule generation.

B.2 Semantic Consistency Templates

To rigorously evaluate the semantic plausibility of the generated rules without requiring labeled ground truth, we construct two contrasting sets of rule templates: *High-Quality Templates* ($\mathcal{H}^+$) and *Low-Quality Templates* ($\mathcal{H}^-$). Table 7 provides representative examples from both categories. During the evaluation phase, the semantic score $S_{\text{sem}}(r)$ is computed by measuring the contrastive similarity between the embedding of a generated rule candidate $\mathbf{v}_r$ and these templates.

C Parameter Sensitivity Analysis

We investigate the impact of two key hyperparameters on ICEWS14: the filtering threshold τ and the temporal decay rate λ. First, the threshold τ governs the strictness of the Discriminator. As illustrated in Figure 4, performance peaks at $\tau = 0.5$; lower thresholds admit excessive noise (false positives), whereas higher values (> 0.7) aggressively discard valid but less confident rules, thereby reducing recall. Second, regarding the decay rate λ, which balances the importance of recent versus distant history, we observe optimal performance at $\lambda = 0.1$. Deviating from this optimum either over-penalizes historical context (larger λ) or fails to capture the temporal drift of events (smaller λ).

D Cost and Efficiency Analysis

Financial Cost. CRI adopts a "generate-once, inference-many" paradigm. The rule induction (Phase I & II) is a one-time offline process. Using `gpt-3.5-turbo`, the total cost to induce rules for the ICEWS14 dataset (approx. 7k entities, 230 relations) is approximately $28.50 (based on current OpenAI pricing: $0.50/1M input tokens, $1.50/1M output tokens, with $\sim$200 sampled paths per relation across 5 refinement iterations).

For organizations with GPU infrastructure, open-source models can be deployed locally, though the *GPU inference cost* varies significantly. Specifically, utilizing efficient models like Llama-3-8B or Qwen2.5-7B incurs approximately $24–$26 in estimated GPU compute costs, which is comparable to GPT-3.5 API fees. In contrast, the larger Llama-3.3-70B costs approximately $95 due to its massive parameter count and slower inference speed. While this is the most expensive option, it achieves the highest MRR (60.6%) and may be cost-effective for production deployments where rule quality is critical.

Crucially, baseline methods like TiRGN, RE-Net, and TLogic are purely graph-based or symbolic approaches that do *not* use LLMs, thus incurring zero LLM-related costs. Their expenses arise entirely from GPU training time (for embedding methods) or CPU rule mining (for symbolic methods). CRI's LLM cost is a one-time investment—once generated, rules can be reused indefinitely without additional inference calls.

Time Efficiency. Unlike embedding-based methods that require extensive GPU hours for iterative

High-Quality Templates ($\mathcal{H}^+$)	
Pattern Type	**Template Rule Description**
Causal Chain	If entity A negotiates with B at time T_1, and B signs an agreement with A at T_2 ($T_1 < T_2$), then A likely cooperates with B at a future time T_3 ($T_2 < T_3$).
Temporal Precedence	If country A imposes sanctions on B at T_1, and B protests against A at T_2 ($T_1 < T_2$), then diplomatic relations between A and B deteriorate at T_3 ($T_2 < T_3$).
Transitive Inference	If organization A funds project B at T_1, and B collaborates with institution C at T_2 ($T_1 < T_2$), then A effectively establishes a partnership with C at T_3.
Conditional Pattern	If leader A visits country B at T_1, and an economic agreement follows at T_2 ($T_1 \leq T_2$), then trade volume between A's nation and B increases at T_3.
Low-Quality Templates ($\mathcal{H}^-$)	
Pattern Type	**Template Rule Description**
Tautology	If entity A has relation R with B at time T_1, then A implies relation R with B at the same time T_1 (Redundant information).
Circular Logic	If A cooperates with B at T_1, implying B cooperates with A at T_1, then A cooperates with B at T_1 (Reasoning in circles).
Temporal Violation	If event E occurs at time T_2, and E is claimed to cause outcome O at an earlier time T_1 where $T_1 < T_2$ (Effect precedes cause).
Semantic Incoherence	If Country A signs a treaty with Country B at T_1, then Country A is topologically located inside Country B at T_2 (Semantic type mismatch / Illogical implication).
Overgeneralization	If any entity A interacts with any entity B at time T_1, then a specific relation R_{target} holds between A and B at T_2 (Lack of specific conditions).
Self-contradiction	If A provides military aid to B at T_1 and A imposes sanctions on B at T_1, then A maintains a neutral stance toward B at T_2 (Contradictory premises).

Table 7: Representative examples of high-quality ($\mathcal{H}^+$) and low-quality ($\mathcal{H}^-$) rule templates.

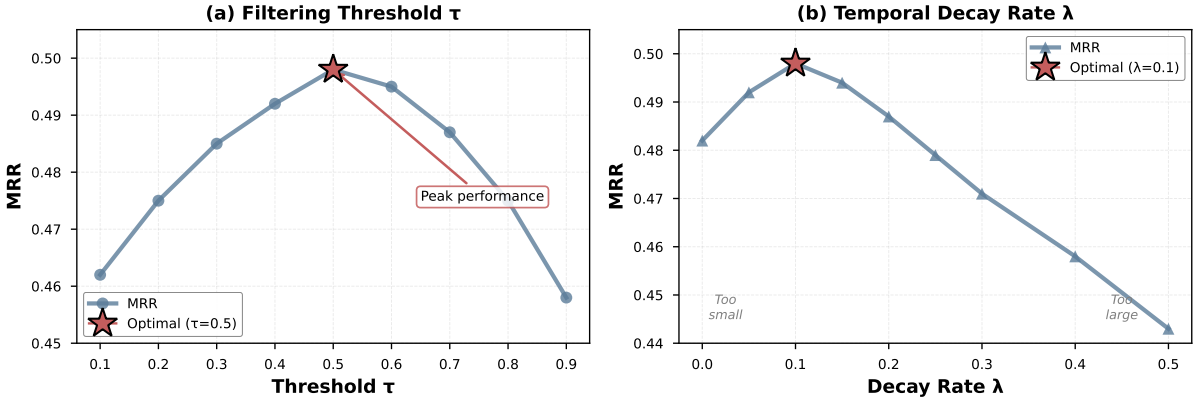

Figure 4: Sensitivity analysis of the Filtering Threshold τ and Decay Rate λ on forecasting performance.

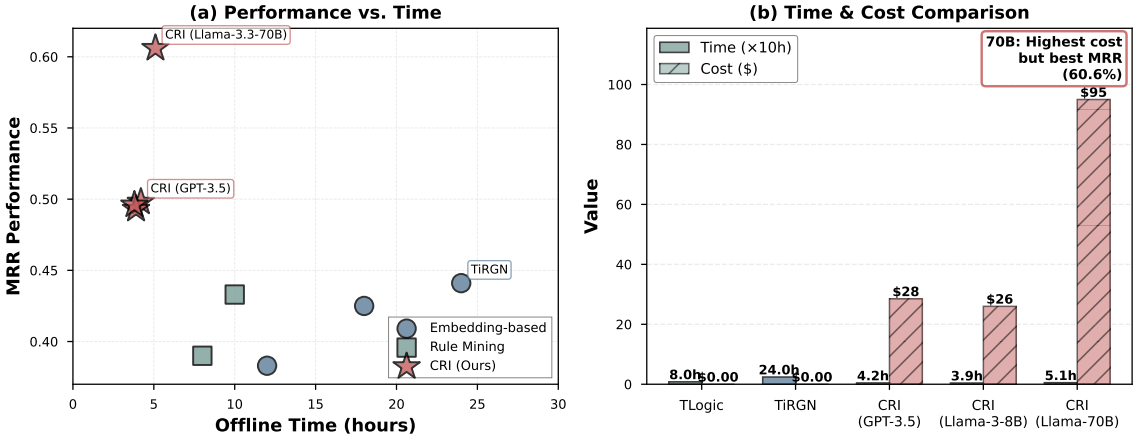

Figure 5: Comparison of MRR performance vs. Offline Time Cost and Financial Cost.

gradient descent training, CRI is training-free. The Phase I sampling and Phase II generation take approximately 4–5 hours on a standard CPU/GPU workstation for ICEWS14. Crucially, the online inference (Phase III) is highly efficient, taking only milliseconds per query using simple symbolic matching, as illustrated in Figure 5. This makes CRI suitable for scenarios requiring transparent and rapid decision-making.

E Algorithm of CRI

Algorithm 1 outlines the computational procedure of CRI, which proceeds in three sequential phases.

Algorithm 1 Critic-Guided Rule Induction (CRI)

Require: Historical TKG $\mathcal{G}_{<t}$, Query $q = (e_s, r_q, ?, t_q)$, Hyperparams λ, k, τ
Ensure: Top-K predicted entities for q

Phase I: Contextual Sampling
$\mathcal{P} \leftarrow \bigcup_{e_i \in \mathcal{E}} \text{ConstrainedWalk}(e_i, \mathcal{G}_{<t}, \chi(t_{next}))$ {Sample paths ensuring $t_{next} < t_{cur}$}
Phase II: Gen-Disc Decoupling
$\mathcal{C}_{top} \leftarrow \text{Top-k}(\{\delta(r_q, r_j) \mid r_j \in \mathcal{R}\})$ {Semantic Pruning}
$\mathcal{R}_{hyp} \leftarrow \text{LLM.Generate}(\mathcal{P}, \mathcal{C}_{top})$ {Hypothesis Induction}
$\mathcal{R}_{refined} \leftarrow \emptyset$
for each rule $r \in \mathcal{R}_{hyp}$ **do**
    Ground r in $\mathcal{G}_{<t}$ to obtain path instances $\mathcal{I}_{body}$
    $S_{raw} \leftarrow \text{Agg}(S_{temp}, S_{conf}, S_{sem}, S_{llm})$ {Multi-dim Evaluation}
    $S_{final} \leftarrow F_{\text{Beta}}(S_{raw})$ {Decisiveness Calibration}
    if $S_{final} \geq \tau$ **then**
        $\mathcal{R}_{refined} \leftarrow \mathcal{R}_{refined} \cup \{(r, S_{final})\}$
    end if
end for
Phase III: Inference & Aggregation
$Score(e) \leftarrow 0$ for all $e \in \mathcal{E}$
for each rule $r \in \mathcal{R}_{refined}$ matching r_q **do**
    for each valid path $p \in \mathcal{I}_{body}$ leading to e **do**
        $\eta(p) \leftarrow \exp(-\lambda |t_q - t_{last}(p)|)$ {Time Decay}
        $Score(e) \leftarrow Score(e) + S_{final}(r) \cdot \eta(p)$
    end for
end for
return Top-K$(Score)$