BEYOND HOMOGENEOUS RULES: ONTOLOGY-AWARE LOGICAL REASONING FOR TEMPORAL KNOWLEDGE GRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

Extrapolative temporal knowledge graph reasoning (TKGR) requires models to predict facts that have not yet appeared, using history before the query time. Most existing methods organize temporal regularities by relation: rule-based methods index rules by relation and share them among all entities, and generative methods retrieve evidence by relation. Entities barely participate in the parameterization of the regularities. However, our experiments show that the relation-centric temporal rule base cannot be reused at equal accuracy across different semantic types, leading to coverage failure and ranking failure. Meanwhile, what generative reasoning truly relies on is the temporal structure executable on the graph rather than name semantics, and large language models (LLMs) can barely answer correctly when the answer never enters the evidence. In this paper, we propose OnLoR, an **On**tology-aware **Lo**gical **R**easoning method, which applies the established ontology concept from knowledge representation to temporal rules: through semantic-based ontology induction and the ontology-aware rule learning module OnRule, the same temporal rules obtains different confidence values on different ontology classes, while generic rules not bound to any ontology are retained to maintain coverage. We further change the role of the LLM, which no longer generates answers from names, but performs constrained pairwise correction on adjacent candidates over the logical evidence packs provided by OnRule. Comprehensive experiments on four TKGR datasets show that OnLoR improves over both the rule-based and generative methods, achieving state-of-the-art in most settings.

1 INTRODUCTION

Temporal knowledge graphs (TKGs) organize timestamped facts into event sequences that unfold over time, capturing dynamic processes such as international relations, social interactions, etc Bordes et al. (2013). Extrapolative TKGR requires models to predict future facts, using only the history observed before the query time. Unlike completing missing edges within known snapshots, extrapolation should distill transferable regularities from history and answer at future timestamps.

In recent years, rule-guided TKGR is becoming mainstream Liu et al. (2022). Among these methods, temporal logical rules compress history into relation chains executable on the graph and provide verifiable structural evidence, making them particularly important for TKGR Li et al. (2023). Generative methods instead write the retrieved history into evidence packs and hand them to LLMs for prediction Xu et al. (2025). Despite their different implementations, most existing methods still organize regularities by relation: rules are indexed by relation and shared by all query subjects of that relation Li et al. (2025), and generative retrieval also extracts evidence by the query relation Liao et al. (2024). Entities barely participate in the parameterization of the regularities themselves, and they are only instantiated along paths or written into evidence packs.

This practice does not match the division of labor among entities in TKGs. Semantic types such as countries, governments, politicians, and social roles engage in relations like "visit" and "criticize" in markedly different ways, yet the rules do not model these roles differentially. We therefore study: can a relation-centric temporal rule base attain equal accuracy across query subjects of different semantic types? Experiments show that it cannot. Some entity groups systematically fall outside the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

mined temporal chains, exhibiting coverage failure; others are covered but the chains show markedly weaker discrimination, exhibiting ranking failure. Dataset-level average metrics smooth away such inter-group differences. This suggests that entity role differences should be written into the rules, rather than letting all subjects continue to share one homogeneous chain.

Generative TKGR usually relies on off-the-shelf rule mining to construct its inputs, but rarely asks what in evidence packs truly enables correct reasoning. Meanwhile, some embedding-based methods have achieved better accuracy than some LLM-based settings, prompting us to rethink the role of LLM in TKGR. We therefore study: in LLM-based TKGR, how much do the on-graph temporal structure and name semantics each contribute, and can the LLM serve as the primary reasoner that directly produces answers? We find that, when the answer has a similar chance of entering the evidence pack, removing temporal rules causes a large accuracy drop, whereas removing names causes only a small one. Semantics is not useless, but for TKGR task the executable temporal structure matters more. Executing rules without any language generation already matches the accuracy of the generative setting, and when the answer is absent, the LLM can barely answer correctly. Consequently, the LLM should not generate answers directly from names without structural evidence.

To address these challenges, we propose an Ontology-aware Logical Reasoning method, OnLoR. For the failure of rules to be reused equally across semantic types, we apply the ontology concept from knowledge representation to temporal rules: we first perform semantic-based ontology induction and then design an ontology-aware rule learning module, OnRule, which endows same temporal rules with different confidence values on ontologies, while retaining generic rules not bound to any ontology so that sparse ontology classes do not lose coverage entirely. For the finding that the on-graph structure is what truly supports correct reasoning, we change the role of the LLM in reasoning: it no longer generates answers from names; inspired by Pairwise Ranking Prompting in information retrieval, the LLM instead compares the logical evidence packs of adjacent candidates, performing constrained correction within the candidate list. Our contributions are as follows:

- We analyze the transferability of relation-centric temporal rules across entity semantic types, together with roles of structural evidence and name semantics in generative TKGR, and find: the shared rule base cannot be reused at equal accuracy across semantic types, with coverage failure or weak ranking; and for TKGR, the temporal structure matters more than name semantics, and the LLM cannot reason primarily without structural evidence.
- We propose OnLoR, which designs OnRule to apply the ontology concept to temporal rules so that the same chain holds to different degrees according to entity roles, addressing coverage failure and ranking failure simultaneously; and performs pairwise correction of the LLM over logical evidence packs, turning the LLM from the primary reasoner in generative reasoning into a comparator of structural evidence.
- We compare with 22 methods in four categories on four datasets. OnLoR improves over both rule-based and generative methods, achieving state-of-the-art in most settings; further analysis shows the effectiveness of OnRule and LLM pairwise correction.

2 Related Work

For TKGR, embedding methods map entities and relations into vector spaces to capture structural dependencies and historical patterns, such as CENET Xu et al. (2023b). Reinforcement learning formulates multi-hop reasoning as path search: TITer Sun et al. (2021) guides walks with time-shaped rewards. However, these methods leave the reasoning rules invisible and difficult to transfer. **Rule-based methods** provide verifiable reasoning rules. TLogic Liu et al. (2022) extends them into temporal logical rules, mining and applying time-ordered paths indexed by relation. INFER Li et al. (2025) enhances rule application with continuous-time validity and matrix projection. These methods still share rules by relation: entities are only instantiated as variables along the paths, and the rules are not differentiated by semantic types. **LLM-based methods** write quadruples as text. ChapTER Peng et al. (2024) adapts a frozen LLM through historical contrast and prefix tuning. GenTKG Liao et al. (2024) retrieves evidence packs with temporal logical rules and lets an LLM generate answers; AnRe Tang et al. (2025) constructs contexts through analogy replay. However, LLMs are better at pairwise comparing than at producing a global ranking at once. The comprehensive related work can be found in Appendix A.

3 EMPIRICAL ANALYSIS

3.1 PRELIMINARIES

TKG and TKG Reasoning. A TKG is a collection of timestamped facts $\mathcal{G} = \{(s, r, o, t)\} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E} \times \mathcal{T}$, where $\mathcal{E}$, $\mathcal{R}$, and $\mathcal{T}$ denote the sets of entities, relations, and timestamps, respectively, and a quadruple (s, r, o, t) states that subject s and object o are connected by relation r at time t. Equivalently, a TKG can be viewed as a time-ordered sequence of graph snapshots $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_t$. We study extrapolative TKG reasoning: given a query $q = (s_q, r_q, ?, t_q)$, the model can only access the history observed before the query time, $\mathcal{G}_{<t_q} = \{(s, r, o, t) \in \mathcal{G} \mid t < t_q\}$, and is required to predict the missing object entity o_q^*; subject queries are defined symmetrically.

Temporal logical rules. Temporal logical rules of TLogic Liu et al. (2022) of length l takes the form: $\rho : \bigwedge_{i=1}^{l}(E_i, R_i, E_{i+1}, T_i) \rightarrow (E_1, R_h, E_{l+1}, T_{l+1})$, where E_i and T_i are entity and time variables, and the timestamps are required to be non-decreasing and earlier than the rule head: $T_1 \leq T_2 \leq \cdots < T_{l+1}$. The rule body describes a multi-hop path that departs from E_1 and evolves forward in time. $\mathbb{R}(r)$ is the set of rules learned for relation r, i.e., those with head relation $R_h = r$.

Evaluation protocol. For each test query q, let rank_q denote the rank of the gold answer o_q^*. We report: Mean Reciprocal Rank $\mathrm{MRR} = \frac{1}{N}\sum_q \frac{1}{\mathrm{rank}_q}$, and $\mathrm{Hits@}K = \frac{1}{N}\sum_q \mathbb{1}\big[\mathrm{rank}_q \leq K\big]$.

3.2 HETEROGENEITY OF THE TEMPORAL RULE BASE FOR ENTITY GROUP

The temporal rule base $\mathbb{R}(r)$ is a set of temporal rules centered on relation r. This section examines whether the temporal regularities learned for a relation transfer across entities.

Research Question. Logical rules compress history into executable chains over relations, while entities enter only as variables instantiated along the paths. However, entities are not homogeneous: different semantic types engage in relations markedly differently. A question therefore arises: *Does the relation-centric temporal rule base $\mathbb{R}(r)$ transfer equally well across different semantic types?*

Experimental Setup. We evaluate on ICEWS14 and ICEWS18 Boschee et al. (2015). Rule bases are produced by TLogic and are frozen during reasoning. We encode each entity's surface name with BERT Devlin et al. (2019) and cluster with KMeans into 30 groups. For each entity group, we compute MRR and Hits@1 independently. We additionally report the coverage failure rate: the fraction of queries in the group whose gold answer is never placed into the candidate set by any rule.

Observations. ***1) Coverage failure tracks ranking quality, but cannot mask inter-group differences.*** Averages of datasets conceal inter-group differences: the overall Hits@1 and MRR are 33.87 and 43.33 on ICEWS14, and 21.66 and 30.10 on ICEWS18. Per-group Hits@1 on ICEWS14 spans 15.26 of group-29 to 60.67 of group-13, a spread of 45.41. On ICEWS18, the range of Hits@1 is 0.54-36.81, with a spread of 36.27. As is shown in Figure 1, Hits@1 and MRR tend to rise as the failure rate falls; yet groups with nearly identical failure rates can still differ considerably. On ICEWS14, groups-12 and -0 both fail on about 23% of queries but attain Hits@1 of 39.16 versus 18.21. Likewise, ICEWS18's overall coverage failure rate 15% is nearly identical to ICEWS14's 19%, while its overall Hits@1 is markedly lower. ***2) The group gap is not small-sample noise.*** On ICEWS14, the largest group-6 reaches Hits@1 of 30.02, already well above the small groups with 17.81 of group-8 and 15.26 of group-29. On ICEWS18, the largest group-23 attains 23.53, while the smallest group-7 reaches 33.53 and group-15 only 5.71. These differences cannot be attributed to small-sample noise. More results are shown in Appendix B.

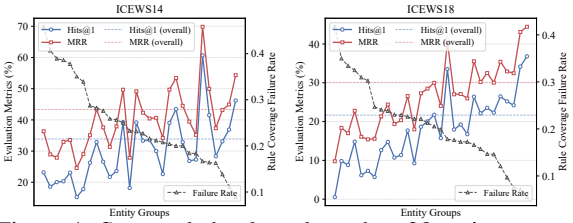

Figure 1: Same rule bank evaluated on 30 entity groups on ICEWS14 and ICEWS18. Groups are ordered by decreasing rule coverage failure rate.

Insight 1. A single temporal rule base cannot be reused at equal accuracy across entity groups. The failure modes of temporal rules are conditioned on entity type: some groups systematically fall outside the mined chains, while others are covered but exhibit weaker chain discrimination.

3.3 Advantage of On-Graph Temporal Structure over Name Semantics for LLM-based TKGR

Generative TKGR puts temporal structures and names of facts into evidence packs and let LLM reason. This section examines how much temporal structure and name semantics contribute to TKGR.

Research Question. Most LLM-based TKGR models retrieve facts through temporal-structural rules and feed them to the LLM, which entangles two signals: temporal chains executable on the graph and the semantics carried by names. This raises a key question: *In LLM-based TKGR, what determines accuracy, the temporal structure or the semantics of names together with the LLM's parametric memory? And can the LLM serve as the primary reasoner to produce answers?*

Experimental Setup. We use ICEWS14 and ICEWS18. The LLM is fixed to Qwen3-8B, and at most 40 historical facts are written into the evidence pack. We additionally report three analysis metrics. Let $\mathrm{Pack}(q)$ denote the evidence pack of query q and o_q^* the gold answer. $\text{Ans-in} = \frac{1}{N}\sum_q \mathbb{1}\big[o_q^* \in \mathrm{Pack}(q)\big]$ is the fraction of queries whose gold answer falls inside the evidence pack. We split the test set into an in-pack subset $Q_{\text{in}} = \{q : o_q^* \in \mathrm{Pack}(q)\}$ and an out-of-pack subset $Q_{\text{out}} = \{q : o_q^* \notin \mathrm{Pack}(q)\}$. InH@1 is Hits@1 computed on Q_{in}, and OutH@1 is computed on Q_{out}. GenTKG Liao et al. (2024) follows the original setting. w/o name keeps the same rule pack but replaces all names with IDs. w/o rule removes temporal-structural rules and keeps facts sharing the query's subject and relation with names retained. TLogic follows the same setting as above.

Table 1: Comparison of GenTKG, w/o name, and w/o rule with Qwen3-8B, and TLogic on ICEWS14 and ICEWS18.

Dataset	Model	Analysis			Evaluation			
		Ans-in	InH@1	OutH@1	MRR	Hits@1	Hits@3	Hits@10
ICEWS14	GenTKG	59.1%	53.94	3.91	40.94	33.47	45.99	53.72
	w/o name	59.1%	50.07	2.51	39.17	30.61	44.89	52.28
	w/o rule	58.5%	30.76	0.33	27.15	18.13	32.38	48.39
	TLogic	80.9%	-	-	43.04	33.56	48.27	61.23
ICEWS18	GenTKG	61.5%	33.01	2.09	30.71	21.09	33.78	43.85
	w/o name	61.5%	30.30	1.01	28.66	19.02	31.86	39.19
	w/o rule	60.6%	27.78	0.36	26.90	16.98	30.25	38.63
	TLogic	85.2%	-	-	29.82	20.54	33.95	48.53

Observations. ***1) Considering similar Ans-in, the accuracy gap is not caused by leaving answers out.*** As is shown in Table 1, on ICEWS14, GenTKG and w/o name both attain an Ans-in of 59.1 and w/o rule attains 58.5; on ICEWS18, the first two reach 61.5 and w/o rule 60.6. The gold answer gets almost the same chance to enter the evidence pack, so the accuracy gap cannot be explained by the shallower settings leaving answers out. ***2) Removing names causes only a small accuracy drop, whereas removing rules causes a large one.*** When the gold answer is already in the pack, InH@1 on ICEWS14 falls from 53.94 of GenTKG to 50.07 of w/o name and further to 30.76 of w/o rule, a loss of 3.87 versus 23.18. On ICEWS18, the metrics follow the same trend. Removing temporal structure therefore hurts accuracy far more than cutting off name semantics, which is a discouraging result for LLMs whose strength is semantic understanding. ***3) Rule execution without an LLM already matches the generative setting for reasoning.*** TLogic performs no generation and only executes temporal rules on TKGs. On ICEWS14, TLogic achieves a comprehensive lead. On ICEWS18, it also attains a Hits@3 of 33.95 and a Hits@10 of 48.53, above the 33.78 and 43.85 of GenTKG; and GenTKG attains an MRR of 30.71 and a Hits@1 of 21.09, above that of TLogic. Rule execution and generative reading each have their strengths, and the generator gains no clear advantage. Therefore, the use of LLMs as the primary reasoner remains to be debated.

Insight 2. What truly matters for TKGR is the on-graph temporal-structural rules rather than the name semantics, and the LLM needs more assistance to serve as the primary reasoner. The evidence packs should give more priority to temporal structures than semantic understanding with LLM.

4 Methodology

4.1 OnLoR Workflow

The two insights motivate the following designs. Insight 1 shows that the shared relation-indexed temporal rules cannot be reused at equal accuracy across different entity semantic types. We therefore lift this role partition into an explicit entity ontology to let rules differentiate across ontologies. Insight 2 shows that for LLMs, the on-graph executable temporal structure matters more than name

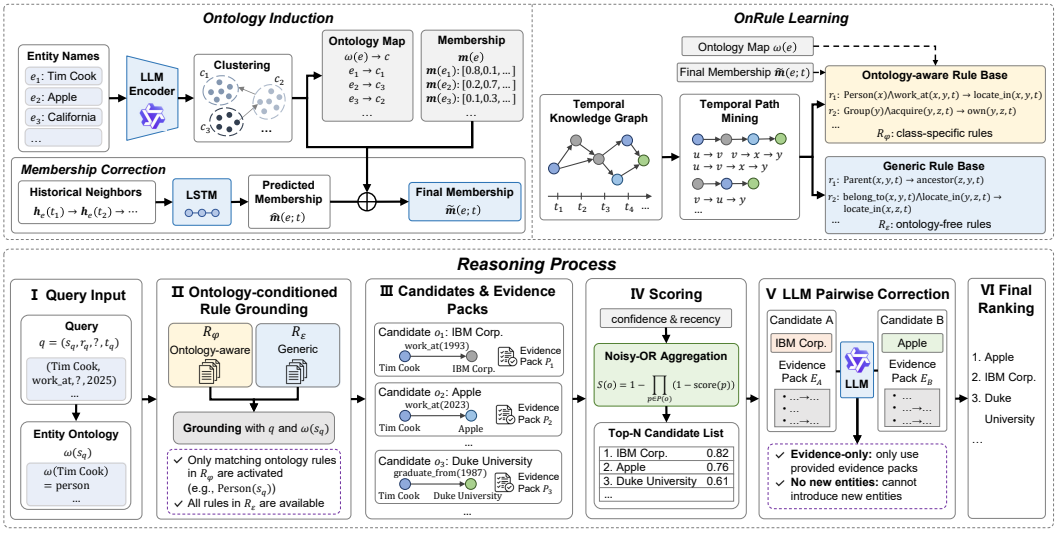

Figure 2: Overview of OnLoR. During rule learning (top), LLM-encoded entity names are clustered into ontologies. One-hop historical neighbors are mean-pooled and passed through an LSTM to correct memberships. OnRule learns an ontology-aware rule base and a generic rule base. During reasoning process (bottom), the query subject's ontology activates matching ontology-aware rules. Paths are scored by confidence and recency, with support aggregated by noisy-OR. Finally, LLM compares adjacent top-N candidates using evidence packs and corrects the candidate list.

semantics of entities and relations. The LLM should correct candidate answers based on the logical evidence packs provided by temporal rules, rather than serve as the primary reasoner to generates answers. Accordingly, we propose OnLoR, which consists of three modules: Semantic-based ontology induction, ontology-aware rule generation, and evidence pack-guided LLM correction in Figure 2.

Stage 1: Semantic-based ontology induction. An LLM partitions the surface names of entities into ontologies with clustering method and dynamic correction, yielding the mapping. The mapping is constructed offline on the training graph; when reasoning, only ontology of query subject is required.

Stage 2: Ontology-aware rule generation. OnRule learns two sets of temporal logical rules on the historical graph: an ontology-aware rule base, and a generic rule base organized by relation only, without ontology binding. The former targets the ranking failure by endowing the same relation chain with class-specific confidence on different ontologies; the latter targets the coverage failure, preventing any ontology from falling entirely outside the reach of rules.

Stage 3: Evidence pack-guided LLM correction. After being executed on the graph, OnRule returns scored object candidates, each attached with a logical evidence pack. The LLM is constrained to never invent entities absent from the packs; instead, it corrects and reranks the candidates solely based on the temporal rule evidence, producing the final ranking.

4.2 Semantic-Based Ontology Induction

The semantic types in Insight 1 serve only as a diagnostic role partition. To condition rules on roles, we set this partition as an explicit entity ontology. Let $\Omega = \{1, \ldots, K\}$ denote the set of ontologies. This section constructs the mapping $\omega : \mathcal{E} \to \Omega$, and retains the membership degree of each entity over all K classes to determine $\omega(e)$ and compute confidence in OnRule.

Semantic clustering. The same frozen LLM encoder maps the surface name of every entity and every relation to $\mathbf{z}_e, \mathbf{z}_r \in \mathbb{R}^d$. Clustering $\{\mathbf{z}_e\}_{e\in\mathcal{E}}$ on the training graph fixes K centers $\{\boldsymbol{\mu}_c\}_{c\in\Omega}$. Static membership is given by its semantic similarity to the class centers:

$$m_{e,c} = \frac{\exp(\mathrm{sim}(\mathbf{z}_e, \boldsymbol{\mu}_c)/\tau)}{\sum_{c'\in\Omega} \exp(\mathrm{sim}(\mathbf{z}_e, \boldsymbol{\mu}_{c'})/\tau)}, \tag{1}$$

where sim denotes cosine similarity and temperature $\tau > 0$. We write $\mathbf{m}_e = (m_{e,1}, \ldots, m_{e,K})$.

Context at one timestamp. Let $\mathcal{H}_t(e) = \{(r,u) : (e,r,u,t) \in G \text{ or } (u,r,e,t) \in G\}$ be the set of other endpoints that touch e at time t in $G_{<t_q}$ during reasoning. The context vector is the mean of the frozen name embeddings, which uses only the facts at t and not an average of all past facts:

$$\mathbf{h}_e^{(t)} = \frac{1}{|\mathcal{H}_t(e)|} \sum_{(r,u)\in\mathcal{H}_t(e)} (\mathbf{z}_r + \mathbf{z}_u). \tag{2}$$

Predicting a correction. Let $t_1 < \cdots < t_n < t$ be the latest L_h occurrence times of e before t. The LSTM maps their context vectors to a membership on the fixed centers:

$$\hat{m}_{e,c}^{(t)} \propto \exp\big(\mathrm{sim}(W\mathbf{g}_e^{(t)}, \boldsymbol{\mu}_c)/\tau\big), \quad \text{where } \mathbf{g}_e^{(t)} = \mathrm{LSTM}\big(\mathbf{h}_e^{(t_1)}, \ldots, \mathbf{h}_e^{(t_n)}\big). \tag{3}$$

If e has no earlier occurrence, $\hat{\mathbf{m}}_e^{(t)} = \mathbf{m}_e$. For an occurrence t_i with $i \geq 2$, the target is the normalized mean of $\mathbf{m}_u$ over $\mathcal{H}_{t_i}(e)$; the input stops at t_{i-1}, and the loss is the KL divergence from this target to $\hat{\mathbf{m}}_e^{(t_i)}$. So $\omega(e,t) = \arg\max_c \tilde{m}_{e,c}^{(t)}$ and the membership used downstream is:

$$\tilde{\mathbf{m}}_e^{(t)} = \frac{\mathbf{m}_e + \alpha\hat{\mathbf{m}}_e^{(t)}}{\|\mathbf{m}_e + \alpha\hat{\mathbf{m}}_e^{(t)}\|_1}. \tag{4}$$

4.3 OnRule: Ontology-Aware Temporal Rule Learning

OnRule is the rule learning module of OnLoR. Instead of maintaining a single relational rule base shared by all query subjects, it learns on the training graph two complementary sets of temporal rules: an ontology-aware rule base $\mathbb{R}_\Omega$, which lets the same temporal rule hold to different degrees on different ontologies, and a generic rule base $\mathbb{R}_\varepsilon$, which retains relation-level rules shared across ontologies. Both kinds of rules follow time-ordered path templates.

Ontology-aware rules. A temporal rule of length L is a relation sequence $\pi = (r_1, \ldots, r_L)$. π is realized from e_1 to e_{L+1} if there exist nodes $e_1, \ldots, e_{L+1}$ and timestamps $t_1 \leq \cdots \leq t_L$ such that every hop (e_i, r_i, e_{i+1}, t_i) lies in the graph. An ontology-aware rule is defined: $\rho_\Omega = \langle r, \pi, c\rangle$, where r is the head relation and $c \in \Omega$. For a subject s_q, a candidate object o_{pred}, and a head time t_q, the rule regards $(s_q, r, o_{\mathrm{pred}}, t_q)$ as a candidate it supports only if $\omega(s_q) = c$, π is realized from s_q to o_{pred}, and the last-hop time satisfies $t_L \leq t_q$. Unlike the shared relational chain, the same (r,π) is differentiated into distinct ρ_Ω by ontology. The ontology-aware rule base is defined:

$$\mathbb{R}_\Omega(r) = \{\rho_\Omega = \langle r, \pi, c\rangle\}. \tag{5}$$

During construction, each training path is assigned to the ρ_Ω of its start entity's ontology $c = \omega(e_1)$. Paths with identical head relation, path template, and subject ontology are merged into a single rule, avoiding duplicates introduced by the ontology expansion. Appendix E shows ontology examples.

Generic rule base. With only ontology-aware rules, ontologies whose paths are sparse would likely fall outside the executable chains. Inspired by TLogic Liu et al. (2022), OnRule additionally keeps for each path template a generic rule base shared across ontologies:

$$\mathbb{R}_\varepsilon(r) = \{\rho_\varepsilon = \langle r, \pi\rangle\}. \tag{6}$$

A generic rule ρ_ε can be instantiated for any $\omega(s_q)$ and captures temporal regularities shared across ontologies. At reasoning time, OnRule applies $\mathbb{R}_\varepsilon$ alongside $\mathbb{R}_\Omega$ as a parallel rule base.

Rule confidence. Confidence is computed on timestamp-stripped entity paths, so that the same path is not repeatedly amplified by multiple timestamps. For an ontology-aware rule $\rho_\Omega = \langle r, \pi, c\rangle$, each path is weighted by the membership degree with time t_1 of the first hop on the path:

$$S_{\mathrm{body}}(\rho_\Omega) = \sum_{\xi\in\mathcal{I}(\pi)} \tilde{m}_{e_1(\xi),c}^{(t_1(\xi))}, \tag{7}$$

$$S_{\mathrm{head}}(\rho_\Omega) = \sum_{\xi\in\mathcal{I}(\pi)} \tilde{m}_{e_1(\xi),c}^{(t_1(\xi))} \cdot \mathbb{1}\big[\exists\, t_h > t_L(\xi) :\ (e_1(\xi), r, e_{L+1}(\xi), t_h) \in \mathcal{G}\big], \tag{8}$$

where $\mathcal{I}(\pi)$ denotes the set of deduplicated entity path instances of π on the training graph, and $e_1(\xi)$, $e_{L+1}(\xi)$, and $t_L(\xi)$ denote the start entity, end entity, and last-hop time of a path instance

$\xi \in \mathcal{I}(\pi)$. The ontology-aware confidence is: $\mathrm{conf}(\rho_\Omega) = \frac{S_{\mathrm{head}}(\rho_\Omega)}{S_{\mathrm{body}}(\rho_\Omega)+\lambda_\Omega}$. The confidence of a generic rule is: $\mathrm{conf}(\rho_\varepsilon) = \frac{|\{\xi \in \mathcal{I}(\pi):\text{head holds}\}|}{|\mathcal{I}(\pi)|+\lambda_\varepsilon}$, where "head holds" means there exists $t_h > t_L(\xi)$ with $(e_1(\xi), r, e_{L+1}(\xi), t_h) \in \mathcal{G}$, and $\lambda_\Omega, \lambda_\varepsilon > 0$ are smoothing terms.

In this way, $\mathbb{R}_\Omega$ uses $\tilde{\mathbf{m}}$ to differentiate the strength of the same path on different classes, while $\mathbb{R}_\varepsilon$ preserves coverage through its ontology-free support ratio.

4.4 Evidence Pack-Guided LLM Pairwise Correction

Insight 2 shows that LLM-based TKGR relies more on the on-graph executable temporal structure than on name semantics. OnLoR therefore lets OnRule provide logical evidence on the graph, and the LLM is restricted to this evidence for correction. An example is given in Appendix F.

Candidates and evidence packs. For a candidate object o_{pred} of a query $q = (s_q, r_q, ?, t_q)$, the logical evidence pack is a realized path supporting this object: there exists a triggered rule $\rho \in \mathbb{R}_\Omega(r_q) \cup \mathbb{R}_\varepsilon(r_q)$ whose template $\pi = (r_1, \ldots, r_L)$ is realized from s_q to o_{pred} with timestamps satisfying $t_1 \le \cdots \le t_L \le t_q$. The logical evidence pack is defined as $\mathcal{P}(o_{\mathrm{pred}}) = \big((e_i, r_i, e_{i+1}, t_i)\big)_{i=1}^{L}$, where $e_1 = s_q$ and $e_{L+1} = o_{\mathrm{pred}}$, followed by earlier facts $(s_q, r, o_{\mathrm{pred}}, t)$ with $t < t_q$.

The score of a concrete realization ξ of a rule is given by rule confidence and recency of its last hop:

$$\sigma(\rho, \xi) = \gamma \cdot \mathrm{conf}(\rho) + (1-\gamma) \cdot \exp\big(-(t_q - t_L(\xi))/\beta\big), \tag{9}$$

where $\beta, \gamma > 0$. $\mathrm{conf}(\rho)$ on ρ_Ω incorporates the subject membership weighting. Let $\mathcal{H}_\Omega(o_{\mathrm{pred}})$ and $\mathcal{H}_\varepsilon(o_{\mathrm{pred}})$ denote the sets of realizations in the two rule bases whose endpoint is this object, respectively. The structural score of OnRule aggregates them by noisy-OR aggregation:

$$S(o_{\mathrm{pred}}) = 1 - \left(\prod_{\xi \in \mathcal{H}_\Omega(o_{\mathrm{pred}})} \big(1 - \sigma(\rho_\xi, \xi)\big)\right)\left(\prod_{\xi \in \mathcal{H}_\varepsilon(o_{\mathrm{pred}})} \big(1 - \sigma(\rho_\xi, \xi)\big)\right), \tag{10}$$

where ρ_ξ denotes the rule that triggers the realization ξ. For each o_{pred} entering the list, several realizations with higher σ are written into $\mathcal{P}(o_{\mathrm{pred}})$. OnRule returns a candidate list of length N sorted in descending order of S as: $\mathcal{C}(q) = \big\{\big(o_{\mathrm{pred}}^{(k)},\, S(o_{\mathrm{pred}}^{(k)}),\, \mathcal{P}(o_{\mathrm{pred}}^{(k)})\big)\big\}_{k=1}^{N}$.

LLM pairwise correction. Pairwise Ranking Prompting in information retrieval shows that LLMs cannot stably comprehend listwise or pointwise ranking formats that present all documents at once, whereas reducing the burden to pairwise comparisons is more effective Qin et al. (2024). OnLoR adopts the same strategy. For $o_{\mathrm{A}}, o_{\mathrm{B}} \in \mathcal{C}(q)$, the LLM compares only the two evidence packs $\mathcal{P}(o_{\mathrm{A}})$ and $\mathcal{P}(o_{\mathrm{B}})$, and outputs a pairwise preference:

$$\delta(o_{\mathrm{A}}, o_{\mathrm{B}}) \in \{o_{\mathrm{A}}, o_{\mathrm{B}}\}, \tag{11}$$

where $\delta(\cdot,\cdot)$ denotes the LLM's judgment of more likely ranks of two candidates. Comparisons are carried out on adjacent pairs along OnRule's order, with the winner promoted, aggregating into a sequence $\hat{o}^{(1)}, \ldots, \hat{o}^{(N)}$ of length N. The correction satisfies $\{\hat{o}^{(1)}, \ldots, \hat{o}^{(N)}\} = \{o_{\mathrm{pred}}^{(1)}, \ldots, o_{\mathrm{pred}}^{(N)}\}$: it only permutes existing candidates and never introduces entities outside the list.

4.5 Rule Grounding and Reasoning

In this section, we describe how rules are grounded and how the reasoning proceeds.

Rule grounding. For a query $q = (s_q, r_q, ?, t_q)$, the head relation is fixed to r_q. From the ontology-aware rule base, only the rules of the same ontology class as the query subject are activated:

$$\mathbb{R}_\Omega(r_q; s_q) = \big\{\rho_\Omega = \langle r_q, \pi, c\rangle \in \mathbb{R}_\Omega(r_q) : c = \omega(s_q, t_q)\big\}, \tag{12}$$

while generic rule base $\mathbb{R}_\varepsilon(r_q)$ is fully activated for any $\omega(s_q, t_q)$. If there exist nodes $e_1, \ldots, e_{L+1}$ and timestamps $t_1 \le \cdots \le t_L < t_q$ such that $e_1 = s_q$ and every hop $(e_i, r_i, e_{i+1}, t_i) \in \mathcal{G}_{<t_q}$, a realization ξ is obtained, and its endpoint $e_{L+1}(\xi)$ is recorded as the $o_{\mathrm{pred}}(\xi)$ supported by ξ.

Reasoning process. With the structural scores S and evidence packs $\mathcal{P}$ obtained after grounding, the top N objects in descending order of S form the candidate list $\mathcal{C}(q)$. Constrained pairwise correction is then performed to yield the final sequence $\hat{o}^{(1)}, \ldots, \hat{o}^{(N)}$ as the reasoning result of OnLoR.

Table 2: Main results on four TKGR datasets. Four categories are selected: static embedding methods (♠), temporal neural methods (♣), LLM-based methods (♢), and rule-based methods (♡). H@1/3/10 is short for Hits@1/3/10. Best results are in bold and second-best are underlined.

	Method	ICEWS14				ICEWS18				ICEWS05-15				WIKI			
		MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
♠	TransE	18.81	2.49	24.70	35.33	13.60	2.78	19.09	37.23	21.60	4.14	33.91	48.52	46.68	36.19	49.71	51.71
	DistMult	27.67	18.16	31.15	46.96	10.17	4.52	10.33	21.25	28.73	19.33	32.19	47.54	46.12	37.24	49.81	51.38
	ComplEx	30.84	21.51	34.48	49.58	21.01	11.87	23.47	39.87	31.69	21.44	35.74	52.04	47.84	38.15	50.08	51.39
	RotatE	31.69	22.42	35.15	49.97	23.10	14.33	27.61	38.72	33.93	24.30	37.92	54.25	50.88	49.17	51.61	52.70
♣	TTransE	13.43	3.11	17.32	34.55	8.36	1.94	8.71	21.93	15.71	5.00	19.72	38.02	31.74	29.49	36.25	43.45
	TA-DistMult	26.47	17.09	30.22	45.41	16.75	8.61	18.41	33.59	24.31	14.58	27.92	44.21	48.09	45.84	49.51	51.70
	RE-Net	38.28	28.68	41.34	54.52	28.81	19.05	32.44	47.51	42.97	31.26	46.85	63.47	51.97	48.01	52.07	53.91
	TANGO	36.12	27.72	41.88	55.70	24.51	15.83	27.97	41.34	34.15	25.03	38.90	51.41	54.05	51.52	53.84	62.95
	CyGNet	32.73	23.69	36.31	50.67	24.93	15.90	28.28	42.61	34.97	25.67	39.09	52.94	48.73	46.80	49.95	51.98
	EvoKG	30.28	13.60	35.09	58.42	21.23	12.45	27.76	42.00	30.18	23.37	35.70	49.76	49.68	43.79	59.46	61.65
	CENET	42.77	**37.03**	47.85	59.54	27.14	18.58	29.99	44.15	37.16	27.78	41.16	55.49	68.39	68.33	68.36	69.05
	HIP network	42.34	36.81	46.65	55.70	28.66	19.08	31.82	38.13	37.30	27.98	41.77	54.91	54.71	53.82	54.73	56.46
	LogiRL	31.67	22.47	36.25	-	24.37	11.18	29.82	-	-	-	-	-	-	-	-	-
♢	PPT	35.83	28.80	43.25	58.08	25.39	16.28	28.56	43.31	36.51	25.15	39.79	52.63	53.84	51.79	54.36	60.33
	ChapTER	34.31	25.78	37.45	56.45	24.47	15.82	27.66	41.29	35.84	23.96	37.39	50.60	-	-	-	-
	GenTKG	40.94	33.47	45.99	53.72	30.71	21.09	33.78	43.85	41.53	33.56	40.96	51.97	-	-	-	-
	HFL	34.48	27.46	41.01	57.53	27.10	17.83	30.44	45.50	-	-	-	-	-	-	-	-
♡	AnyBURL	29.67	21.26	33.33	46.73	22.77	15.10	25.44	38.91	32.05	23.72	35.45	50.46	-	-	-	-
	TLogic	43.04	33.56	48.27	61.23	29.82	20.54	33.95	48.53	46.97	36.21	53.13	67.43	78.93	73.05	84.97	86.91
	TR-Rules	43.32	33.96	48.55	61.17	30.41	21.10	34.58	48.92	47.64	37.06	53.80	67.57	-	-	-	-
	ILR-IR	43.54	34.38	-	60.30	30.86	21.73	-	49.11	-	-	-	-	64.37	64.11	-	68.39
	INFER	44.08	34.82	48.84	61.74	32.11	22.34	36.26	51.19	48.34	38.07	54.16	67.82	86.07	84.72	87.10	88.53
	OnLoR	**44.81**	34.96	**49.02**	**62.00**	**32.95**	**22.63**	**36.88**	**51.66**	**49.62**	**39.46**	**55.07**	**68.79**	**86.21**	**84.96**	**87.35**	**88.81**

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Datasets and Benchmarks. We evaluate on four TKGR datasets: ICEWS14, ICEWS18, ICEWS05-15 Boschee et al. (2015), and WIKI Leblay & Chekol (2018). We select 22 baselines in four categories. **Static embedding methods**: TransE Bordes et al. (2013), DistMult Yang et al. (2015), ComplEx Trouillon et al. (2016), and RotatE Sun et al. (2019). **Temporal neural methods**: TTransE Leblay & Chekol (2018), TA-DistMult García-Durán et al. (2018), RE-Net Jin et al. (2020), TANGO Han et al. (2021b), CyGNet Zhu et al. (2021), EvoKG Park et al. (2022), CENET Xu et al. (2023b), HIP network He et al. (2024), and LogiRL Chen et al. (2025). **LLM-based methods**: PPT Xu et al. (2023a), ChapTER Peng et al. (2024), GenTKG Liao et al. (2024), and HFL Xu et al. (2025). **Rule-based methods**: AnyBURL Meilicke et al. (2019), TLogic Liu et al. (2022), TR-Rules Li et al. (2023), ILR-IR Mei et al. (2024), and INFER Li et al. (2025). Detailed descriptions are reported in Appendix C.1.

Implementation Details. We select Qwen3-8B Yang et al. (2025) as LLM. OnRule learns temporal paths of length $\{1, 2, 3\}$; ontologies are clustered with K-Means, and the number is set from $[8, 80]$. The generic rule confidence is estimated from 500 sampled instantiations of rule heads and bodies. Each query contains no more than 40 facts and evidence pack to LLM is limited to 12 facts. Evaluation protocol follows xERTE Han et al. (2021a). More details are provided in Appendix C.2.

5.2 PERFORMANCE COMPARISON

Table 2 reports the main results of OnLoR and 22 baselines on four datasets. Overall, OnLoR achieves the best performance on the vast majority of metrics on ICEWS14, ICEWS18, ICEWS05-15, and WIKI. Compared with the rule-based method INFER, OnLoR improves MRR by 0.73, 0.84, and 1.28 on ICEWS datasets; the gains over TLogic are more pronounced: +1.77 on ICEWS14, +3.13 on ICEWS18, and +7.28 on WIKI, which corroborates Insight 1. Among temporal neural methods, CENET attains a Hits@1 of 37.03 on ICEWS14, above the 34.96 of OnLoR. This suggests that the ontology brings smaller gains when relations are extremely sparse, and neural methods can reinforce high-frequency events. Compared with LLM-based methods, GenTKG falls below OnLoR

on ICEWS datasets: its MRR on ICEWS14 is 40.94, below the 44.81 of OnLoR. Other methods lag further behind. This aligns with Insight 2 that letting OnRule reason primarily, with LLM pairwise correction, outperforms settings with LLM generation.

The advantages of OnLoR mainly come from two parts. Compared with the homogeneous rule base, ontology awareness endows rules with different confidence values; compared with generative TKGR, the LLM no longer performs global generation but compares adjacent candidates.

5.3 Ablation Study

Table 3 dissects two core modules of OnLoR on ICEWS14. Removing OnRule drops MRR from 44.81 to 14.42 and Hits@1 from 34.96 to 7.51, showing that the temporal rules reason primarily rather than auxiliary retrieval. Removing the ontology-aware rules yields an MRR/Hits@1 of 42.69/33.24, and removing the generic rules yields 42.06/32.53; both degrade markedly. This corresponds to Insight 1: Ontology-aware rules with different confidence values alleviate ranking failure, and generic rules prevents sparse classes from falling outside rules. On the LLM side, removing LLM correction degrades the results, showing that LLM improves the top ranking. Removing the logical evidence packs drops MRR to 43.16 and Hits@1 to 33.32, and replacing names with IDs drops to 44.62 and 34.83, consistent with Insight 2 that structure matters more than names. Replacing pairwise correction with listwise one further drops Hits@1 to 28.90 and MRR to 40.73. The LLM is therefore more suitable to judge the better candidate than reason primarily.

Table 3: Ablation results of OnLoR on ICEWS14. "w/o": without; "w/": with.

Method	MRR	Hits@1	Hits@3	Hits@10
OnLoR	44.81	34.96	49.02	62.00
w/o OnRule	14.42	7.51	15.80	28.14
w/o ontology-aware rule	42.69	33.24	47.94	60.81
w/o LSTM correction ($\alpha = 0$)	44.74	34.80	48.91	61.78
w/o generic rule	42.06	32.53	47.20	60.27
w/o LLM correction	44.56	34.65	48.78	62.00
w/o evidence pack	43.16	33.32	47.93	62.00
w/o names *w/* IDs	44.62	34.83	48.91	62.00
w/o pairwise *w/* listwise	40.73	28.90	47.78	62.00

The two rule bases of OnRule determine the upper bound of coverage and ranking, and the LLM brings gains when it performs pairwise correction with evidence packs. This is consistent with the division of labor in OnLoR: rules reason on the graph, and the LLM compares structural evidence.

5.4 Analysis

Figure 3 (left) sweeps the number of ontology classes K on ICEWS14. MRR falls within 43.85-44.81 and Hits@1 within 33.93-34.96, and the curves peak at K=70, consistent with the setting in main results. A too-small K merges different roles into the same ontology class, weakening the differentiation; a too-large K makes some classes overly sparse and renders the estimation of ontology-aware rules unstable.

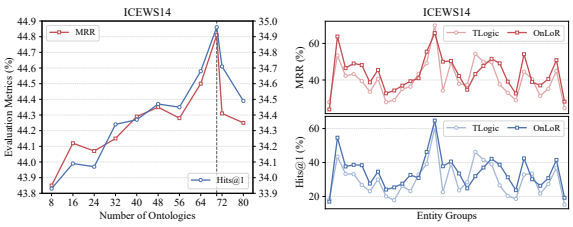

Figure 3: Hyperparameter and group-wise analysis.

Figure 3 (right) compares OnLoR with TLogic on the same 30 groups from Research Question 1. The upper plot shows per-group MRR and the lower one Hits@1. OnLoR surpasses TLogic on most groups, and the originally weaker groups, e.g., groups 14, 21, and 22, improve more markedly. The inter-group variance of MRR decreases from 95.51 to 92.56, and that of Hits@1 from 103.33 to 100.28, reducing uneven performance across entity groups. This indicates that ontology-aware rules and generic rules alleviate the inter-group imbalance. Appendix D, E, and F gives more analyses.

6 Conclusion

We present two empirical findings in TKGR: relation-centric temporal rules cannot be equally reused across semantic types, failing through insufficient coverage or weak ranking; generative reasoning truly relies on temporal structure and LLM should not reason primarily. We propose OnLoR to apply ontology to temporal rules: OnRule reasons with ontology-aware rules and generic rules ; LLM performs pairwise correction on candidates.

AI USE STATEMENT

In this work, we used generative AI tools for [assist with translation, support qualitative and thematic data analysis, interpret results]. We have not used generative AI tools for [help develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat dataset], and [generate synthetic data sets, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs] are not applicable to this work. Additionally, we used generative AI tools for [create or modify scientific figures or images, summarize or analyse existing literature, sourcing/searching for information, edit a research paper to improve readability]. We have reviewed all AI-assisted work. [We checked LLM-generated contents in our paper for potential plagiarism through a manual literature survey. LLM-generated code was verified and tested for correctness by 2 authors.]. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide our codes and datasets used at https://anonymous.4open.science/r/OnLoR-8897 to enhance reproducibility. And the implementation details are provided in Section 5.1 and Appendix C.2.

REFERENCES

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 2787–2795, 2013.

Elizabeth Boschee, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz, and Michael Ward. Icews coded event data. *Harvard Dataverse*, 12(2), 2015.

Qian Chen, Jinyu Zhang, and Ling Chen. Beyond entity correlations: Disentangling event causal puzzles in temporal knowledge graphs. In *The Fourteenth International Conference on Learning Representations*, 2026.

Tingxuan Chen, Liu Yang, Zidong Wang, and Jun Long. A rule-and query-guided reinforcement learning for extrapolation reasoning in temporal knowledge graphs. *Neural Networks*, 185: 107186, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Alberto García-Durán, Sebastijan Dumančić, and Mathias Niepert. Learning sequence encoders for temporal knowledge graph completion. *arXiv preprint arXiv:1809.03202*, 2018.

Zhen Han, Peng Chen, Yunpu Ma, and Volker Tresp. Explainable subgraph reasoning for forecasting on temporal knowledge graphs. In *Iclr*, 2021a.

Zhen Han, Zifeng Ding, Yunpu Ma, Yujia Gu, and Volker Tresp. Learning neural ordinary equations for forecasting future links on temporal knowledge graphs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 8352–8364. Association for Computational Linguistics, 2021b.

Yongquan He, Peng Zhang, Luchen Liu, Qi Liang, Wenyuan Zhang, and Chuang Zhang. HIP network: Historical information passing network for extrapolation reasoning on temporal knowledge graph. *CoRR*, abs/2402.12074, 2024.

Woojeong Jin, Meng Qu, Xisen Jin, and Xiang Ren. Recurrent event network: Autoregressive structure inferenceover temporal knowledge graphs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 6669–6683. Association for Computational Linguistics, 2020.

Julien Leblay and Melisachew Wudage Chekol. Deriving validity time in knowledge graph. In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pp. 1771–1776. ACM, 2018.

Ningyuan Li, E Haihong, Shi Li, Mingzhi Sun, Tianyu Yao, Meina Song, Yong Wang, and Haoran Luo. 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*, pp. 7885–7894, 2023.

Ningyuan Li, Tianyu Yao, Tianyi Hu, Yuhan Li, Haoran Luo, Meina Song, Yifan Zhu, et al. Infer: a neural-symbolic model for extrapolation reasoning on temporal knowledge graph. In *International Conference on Learning Representations*, volume 2025, pp. 91993–92008, 2025.

Zixuan Li, Xiaolong Jin, Saiping Guan, Wei Li, Jiafeng Guo, Yuanzhuo Wang, and Xueqi Cheng. Search from history and reason for future: Two-stage reasoning on 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 (volume 1: long papers)*, pp. 4732–4743, 2021a.

Zixuan Li, Xiaolong Jin, Wei Li, Saiping Guan, Jiafeng Guo, Huawei Shen, Yuanzhuo Wang, and Xueqi Cheng. 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*, pp. 408–417, 2021b.

Ruotong Liao, Xu Jia, Yangzhe Li, Yunpu Ma, and Volker Tresp. Gentkg: Generative forecasting on temporal knowledge graph with large language models. In *Findings of the association for computational linguistics: NAACL 2024*, pp. 4303–4317, 2024.

Yushan Liu, Yunpu Ma, Marcel Hildebrandt, Mitchell Joblin, and Volker Tresp. Tlogic: Temporal logical rules for explainable link forecasting on temporal knowledge graphs. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pp. 4120–4127. AAAI Press, 2022.

Xin Mei, Libin Yang, Xiaoyan Cai, and Zuowei Jiang. An adaptive logical rule embedding model for inductive reasoning over temporal knowledge graphs. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 7304–7316, 2022.

Xin Mei, Libin Yang, Zuowei Jiang, Xiaoyan Cai, Dehong Gao, Junwei Han, and Shirui Pan. An inductive reasoning model based on interpretable logical rules over temporal knowledge graph. *Neural Networks*, 174:106219, 2024.

Christian Meilicke, Melisachew Wudage Chekol, Daniel Ruffinelli, and Heiner Stuckenschmidt. Anytime bottom-up rule learning for knowledge graph completion. In *IJCAI*, volume 2019, pp. 3137–3143, 2019.

Namyong Park, Fuchen Liu, Purvanshi Mehta, Dana Cristofor, Christos Faloutsos, and Yuxiao Dong. Evokg: Jointly modeling event time and network structure for reasoning over temporal knowledge graphs. In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, pp. 794–803. ACM, 2022.

Miao Peng, Ben Liu, Wenjie Xu, Zihao Jiang, Jiahui Zhu, and Min Peng. Deja vu: Contrastive historical modeling with prefix-tuning for temporal knowledge graph reasoning. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pp. 1178–1191. Association for Computational Linguistics, 2024.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, et al. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1504–1518, 2024.

Yuehang Si, Zefan Zeng, Qing Cheng, and Jincai Huang. Frlt: Adaptive reinforcement learning via fuzzy search for temporal knowledge graph reasoning. *Expert Systems with Applications*, pp. 133398, 2026.

Haohai Sun, Jialun Zhong, Yunpu Ma, Zhen Han, and Kun He. Timetraveler: Reinforcement learning for temporal knowledge graph forecasting. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 8306–8319, 2021.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. Rotate: Knowledge graph embedding by relational rotation in complex space. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Guo Tang, Zheng Chu, Wenxiang Zheng, Junjia Xiang, Yizhuo Li, Weihao Zhang, Ming Liu, and Bing Qin. 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)*, pp. 4632–4650, 2025.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 2071–2080. JMLR.org, 2016.

Siheng Xiong, Yuan Yang, Faramarz Fekri, and James Clayton Kerce. Tilp: Differentiable learning of temporal logical rules on knowledge graphs. *arXiv preprint arXiv:2402.12309*, 2024.

Wenjie Xu, Ben Liu, Miao Peng, Xu Jia, and Min Peng. Pre-trained language model with prompts for temporal knowledge graph completion. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 7790–7803. Association for Computational Linguistics, 2023a.

Wenjie Xu, Ben Liu, Miao Peng, Zihao Jiang, Xu Jia, Kai Liu, Lei Liu, and Min Peng. Historical facts learning from long-short terms with language model for temporal knowledge graph reasoning. *Inf. Process. Manag.*, 62(3):104047, 2025.

Yi Xu, Junjie Ou, Hui Xu, and Luoyi Fu. Temporal knowledge graph reasoning with historical contrastive learning. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 4765–4773. AAAI Press, 2023b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Shangfei Zheng, Hongzhi Yin, Tong Chen, Quoc Viet Hung Nguyen, Wei Chen, and Lei Zhao. Dream: Adaptive reinforcement learning based on attention mechanism for temporal knowledge graph reasoning. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pp. 1578–1588, 2023.

Cunchao Zhu, Muhao Chen, Changjun Fan, Guangquan Cheng, and Yan Zhang. Learning from history: Modeling temporal knowledge graphs with sequential copy-generation networks. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on*

Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pp. 4732–4740. AAAI Press, 2021.

A Comprehensive Related Work

A.1 Classic TKGR Methods

Extrapolative TKGR predicts future facts from the history before the query time, in contrast to interpolation, which completes missing edges on known snapshots. Early work is dominated by embeddings: RE-NET Jin et al. (2020) infers event dependencies with a recurrent structure, CyGNet Zhu et al. (2021) introduces a copy-generation mechanism to emphasize recurring history, RE-GCN Li et al. (2021b) encodes snapshot sequences with graph convolutions, CENET Xu et al. (2023b) uses contrastive learning to distinguish whether the current query relies more on historical or non-historical facts, and xERTE Han et al. (2021a) performs explainable reasoning on query subgraphs.

Reinforcement learning formulates multi-hop reasoning as path search: CluSTeR Li et al. (2021a) first retrieves clues and then reasons temporally, TITer Sun et al. (2021) shapes rewards with time, DREAM Zheng et al. (2023) combines attention with adaptive rewards, and FRLT Si et al. (2026) further realizes adaptive reinforcement learning with fuzzy search. HEDRA Chen et al. (2026) decouples non-causality and pseudo-causality at the event level and separates dynamic from static causality.

These approaches improve accuracy, but the regularities settle into vectors, rewards, or event representations: the rationale is invisible to users, and the learned patterns are difficult to transfer as explicit rules across entities and datasets.

A.2 Rule-based TKGR Methods

Rule-based methods provide readable reasoning chains. On static graphs, AnyBURL Meilicke et al. (2019) learns rules bottom-up in an anytime fashion. On the temporal side, TLogic Liu et al. (2022) mines cyclic temporal logical rules through temporal random walks, with rules indexed by head relation and shared by all subjects. TR-Rules Li et al. (2023) points out that temporal redundancy inflates confidence, aggregates rule bodies with a window-based confidence, and mines acyclic rules. ALRE-IR Mei et al. (2022) learns rule embeddings to evaluate causal logic; ILR-IR Mei et al. (2024) models relational paths and entity interaction preferences jointly, whose rule part can transfer to datasets sharing the relation vocabulary without depending on specific entities, but paths alone cannot distinguish different entities along the same historical relation chain. TILP Xiong et al. (2024) learns temporal logical rules in a differentiable manner. INFER Li et al. (2025) replaces binary fact validity with continuous-time validity and substitutes expensive path matching with rule projection on GPUs.

These models improve confidence estimation, rule forms, and application efficiency, but do not differentiate the same chain by semantic types. When applied to query subjects of different roles, sparse groups may systematically fall outside the mined chains, and covered groups may also suffer weaker discrimination due to the shared confidence.

A.3 LLM-based TKGR Methods

Language models convert quadruples into text. PPT Xu et al. (2023a) targets completion at known timestamps and formulates the task as masked prediction with prompts. ChapTER Peng et al. (2024) feeds text with historical context into a pseudo-Siamese encoder and freezes the pre-trained model with prefix tuning. HFL Xu et al. (2025) samples facts from long- and short-term history in multiple views. In generative extrapolation, GenTKG Liao et al. (2024) retrieves evidence packs with temporal logical rules and lets an LLM generate answers; AnRe Tang et al. (2025) retrieves analogous events by semantic clustering and performs training-free analogy replay.

These methods improve the modeling capacity on the text side, but the evidence pack mixes on-graph structure and name semantics, and the model usually performs one global generation over the

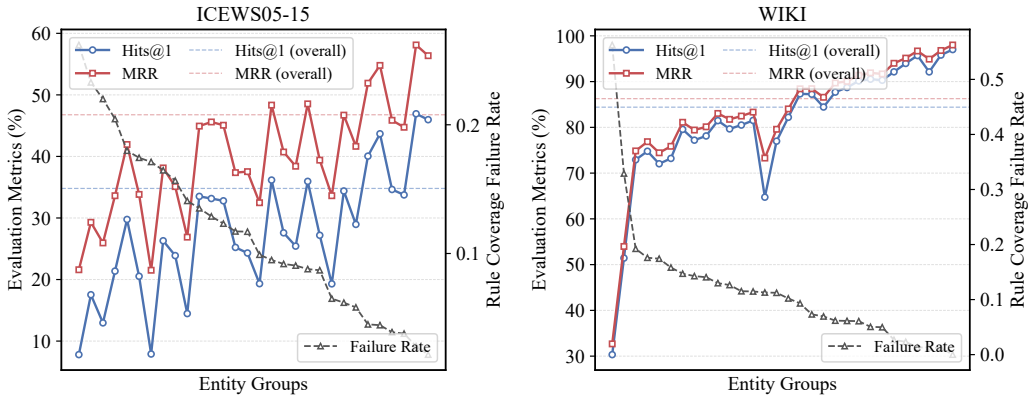

Figure 4: Same rule bank evaluated on fixed entity groups on ICEWS05-15 and WIKI. Groups are ordered by decreasing rule coverage failure rate.

entire list. What truly works for extrapolation is the temporal structure executable on the graph, and LLMs are also better at judging the relative quality of a pair of evidence packs than at serving as the primary reasoner for global ranking.

B Supplementary Results for Research Question 1

This section supplements Research Question 1 with results showing that the same relation-centric rule base performs unevenly across entity groups. The main text presents the line plots for ICEWS14 and ICEWS18: test queries are partitioned by the group of their subjects, and groups are sorted in descending order of the rule coverage failure rate, so that coverage failure can be observed together with the changes in MRR, Hits@1, Hits@3, and Hits@10.

Figure 4 applies the same evaluation protocol to ICEWS05-15 and WIKI: the rule base is learned by relation on the training graph and frozen, and test queries are partitioned into 30 entity groups by their subjects. Due to space limitations, and to avoid duplicating, per-group numerical tables for these two datasets are omitted. The two curves exhibit the same pattern as ICEWS14 and ICEWS18: groups with higher coverage failure rates attain lower Hits@1 and MRR, while groups with similar failure rates can still differ markedly in ranking. The coexistence of coverage failure and ranking failure is reproduced on both event-based data (ICEWS05-15) and knowledge-base data (WIKI), consistent with Insight 1 that the relation-centric temporal rule base cannot be reused at equal accuracy across entity groups.

C Detailed Experimental Setup

C.1 Dataset Statistics and Description of Baselines

This section describes the datasets and baseline setups in detail. The sizes of the four extrapolation benchmarks are shown in Table 4.

We divide the datasets into event-based and knowledge-base ones. ICEWS14, ICEWS18, and ICEWS05-15 all come from the Integrated Crisis Early Warning System (ICEWS) [1], recording political and social events among subjects such as countries, governments, and politicians, with a temporal granularity of 24 hours. WIKI [2] consists of temporal facts from encyclopedic knowledge bases, with a temporal granularity of one year.

- **ICEWS14** Boschee et al. (2015) covers the whole year of 2014 with a moderate scale of entities and facts, and is commonly used for fast extrapolation evaluation.

[1] https://dataverse.harvard.edu/dataverse/icews
[2] https://en.wikipedia.org/wiki

Table 4: Statistics of the four extrapolative TKGR benchmarks.

Dataset	Entities	Relation	Time	Training	Validation	Test	Interval
ICEWS14	7,128	230	365	63,685	13,823	13,222	24 hours
ICEWS18	23,033	256	365	373,018	45,995	49,545	24 hours
ICEWS05-15	10,488	251	4,017	322,958	69,224	69,147	24 hours
WIKI	12,554	24	232	539,286	67,538	63,110	1 year

- **ICEWS18** Boschee et al. (2015) covers 2018, with significantly more entities and training facts and denser queries.
- **ICEWS05-15** Boschee et al. (2015) spans 2005–2015 and contains far more timestamps than the former two, testing extrapolation over long event streams.
- **WIKI** Leblay & Chekol (2018) has 24 relation types but many training facts, dominated by year-to-year evolution.

To verify the effectiveness of OnLoR, we compare it with 22 representative TKGR methods in four categories. The models are described as follows.

Static embedding methods:

- **TransE** Bordes et al. (2013) treats a relation as a translation between head and tail entities.
- **DistMult** Yang et al. (2015) scores triples with a diagonal bilinear factorization.
- **ComplEx** Trouillon et al. (2016) models asymmetric relations in the complex space.
- **RotatE** Sun et al. (2019) represents relations as rotations in the complex plane.

Temporal Neural methods:

- **TTransE** Leblay & Chekol (2018) incorporates time into the translation of TransE.
- **TA-DistMult** García-Durán et al. (2018) extends DistMult with time tokens.
- **RE-Net** Jin et al. (2020) performs autoregressive inference over event sequences with a recurrent network.
- **TANGO** Han et al. (2021b) uses neural ordinary differential equations to capture the continuous evolution of entities and relations over time.
- **CyGNet** Zhu et al. (2021) emphasizes recurring historical facts with a copy-generation mechanism.
- **EvoKG** Park et al. (2022) models the dynamic evolution of knowledge graphs.
- **CENET** Xu et al. (2023b) uses contrastive learning to determine whether the current query relies more on historical or non-historical facts.
- **HIP network** He et al. (2024) propagates information along historical paths to support temporal prediction.
- **LogiRL** Chen et al. (2025) introduces rule signals into reinforcement learning to search extrapolation paths on the graph.

LLM-based methods:

- **PPT** Xu et al. (2023a) formulates completion as masked prediction with entity, relation, and time prompts, targeting completion at all timestamps.
- **ChapTER** Peng et al. (2024) feeds text with historical context into a pseudo-Siamese encoder and freezes the pre-trained model with prefix tuning.
- **GenTKG** Liao et al. (2024) retrieves evidence packs with temporal logical rules and then generates answers with an LLM.
- **HFL** Xu et al. (2025) samples facts from long- and short-term history for the language model to predict entities.

Rule-based methods:

- **AnyBURL** Meilicke et al. (2019) learns rules bottom-up in an anytime fashion on static knowledge graphs.
- **TLogic** Liu et al. (2022) mines cyclic temporal logical rules through temporal random walks and shares them by relation across all subjects.
- **TR-Rules** Li et al. (2023) handles temporal redundancy and mines acyclic rules beyond cyclic ones.
- **ILR-IR** Mei et al. (2024) combines interpretable relational paths with entity interaction preferences to support inductive transfer.
- **INFER** Li et al. (2025) replaces binary fact validity with continuous-time validity and applies rules on the graph with matrix projection.

C.2 Implementation Details

OnLoR is implemented in PyTorch, and all experiments are conducted on a single 32GB NVIDIA GPU. For each test quadruple (s_q, r_q, o_q, t_q), we perform both object prediction $(s_q, r_q, ?, t_q)$ and subject prediction $(?, r_q, o_q, t_q)$, and report the MRR and Hits@k computed on the merged results. Evaluation follows the time-aware filtering protocol of xERTE Han et al. (2021a): other correct objects that have already appeared at the query time t_q are filtered out of the ranking, a protocol more suitable for the extrapolation setting than static filtering.

Ontology induction. We adopt Qwen3-8B Yang et al. (2025) as the LLM to encode the surface names of entities and relations, cluster the entity embeddings to obtain K ontology classes and their centers. The numbers of ontology classes on the four datasets are: $K = 70$ for ICEWS14, $K = 72$ for ICEWS18, $K = 64$ for ICEWS05-15, and $K = 72$ for WIKI.

OnRule. Temporal paths of length $\{1, 2, 3\}$ are learned on the training graph, yielding the ontology-aware rule base and the generic rule base. Rule confidence is estimated by sampling 500 instantiations of the rule body and counting how often the rule head holds.

Evidence packs and LLM correction. We also adopt Qwen3-8B Yang et al. (2025) as the LLM. At most 40 facts are written into the evidence for each query, consistent with the default length setting of TLogic, and each evidence pack fed to the LLM contains at most 12 pieces of evidence. The frozen LLM compares adjacent candidates pairwise along OnRule's default order, only permuting objects within the list and never inventing entities outside the list. The LLM correction method is a single-round bubble sort from the back of the sorted list to the front.

Baselines. For most baselines, we mainly quote the results reported by INFER Li et al. (2025), and re-run and verify some of them. OnLoR, TLogic Liu et al. (2022), and the generative counterparts are evaluated under the same data split and filtering protocol.

Other details. Since we follow the default setting of TLogic, which considers at most 40 pieces of evidence per query, and further limit the number of evidence items in each pack to 12 during LLM pairwise correction, we compare against "INFER (with vc)", the second version of the INFER baseline with the same setting as OnLoR, instead of the best-performing setup that requires 60 rules for a single query.

We reserve judgment on some of related works that exist only as preprints without released code, as their results cannot be verified under our evaluation protocol.

D Sensitivity analysis of LLMs on ontology encoding and pairwise correction

We show the results of different LLMs on ontology induction and on pairwise correction on ICEWS14 in Table 5. The upper part replaces the name encoder used to build the ontology. The lower part keeps that ontology fixed and changes only the model that performs pairwise correction.

The upper part replaces only the frozen encoder used for ontology induction. Qwen-3-8B and Llama-3.1-8B-Instruct are nearly tied and are the strongest encoders, with MRR 44.81 and 44.76,

Table 5: Results of different LLMs on ontology induction and on pairwise correction on ICEWS14.

Frozen LLM Encoder	**MRR**	**Hits@1**	**Hits@3**	**Hits@10**
BERT	43.92	34.14	48.68	61.62
Gemma-3-1B-it	43.84	34.10	48.61	61.57
Mistral-7B-Instruct-v0.3	44.52	34.48	48.80	61.77
Qwen-3-8B	**44.81**	**34.96**	**49.02**	**62.00**
Llama-3.1-8B-Instruct	44.76	34.93	48.95	61.96
LLM Pairwise Correction	**MRR**	**Hits@1**	**Hits@3**	**Hits@10**
Gemma-3-1B-it	44.41	34.54	48.75	**62.00**
Mistral-7B-Instruct-v0.3	44.63	34.69	48.92	**62.00**
Qwen-3-8B	**44.81**	**34.96**	**49.02**	**62.00**
Llama-3.1-8B-Instruct	44.75	34.87	48.97	**62.00**

Hits@1 34.96 and 34.93. Mistral-7B-Instruct-v0.3 is lower, which are 44.52 and 34.48. BERT is close to each other, with 43.92 and 34.14. Gemma-3-1B-it is the weakest, with 43.84 and 34.10. Hits@10 also moves in this block, from 61.57 to 62.00, so the encoder changes which objects OnRule places in the top-10 rather than only reordering a fixed list.

The lower part replaces only the LLM that performs pairwise correction. Hits@10 stays at 62.00 for every model, since correction only permutes OnRule's top-10. Relative to OnRule without pairwise correction, Qwen-3-8B gains 0.25 MRR and 0.31 Hits@1, and Llama-3.1-8B-Instruct gains 0.19 MRR and 0.22 Hits@1. Mistral-7B-Instruct-v0.3 gains slightly, with 44.63 MRR, 34.69 Hits@1 and 48.92 Hits@3. Gemma-3-1B-it falls below that reference on every metric that correction can change, with 44.41, 34.54 and 48.75: its swaps remove more correct top-1 predictions than they add. Pairwise correction therefore helps only when the LLM is large enough to judge the evidence packs, and the 1B model is not.

E Examples of Learned Ontologies on ICEWS14

To demonstrate the ontologies produced by semantic-based ontology induction, we list 30 classes with clearer identities out of the 70 learned on ICEWS14 in Table 6. Each class is shown with 10 entities. Entity names follow the spelling of the dataset, with underscores replaced by spaces, and highly similar or duplicate entries within the same class are avoided. The Ontology column gives the class index, and n_{ent} is the total number of entities in the class. The Defined Label is the meaning we summarized from the member entities, not a label directly generated by the model.

These classes roughly correspond to stable event domains in ICEWS, rather than a plain "country/person" dichotomy. Class 42 groups sovereign states with foreign-policy actors such as Barack Obama, John Kerry, Pope Francis, and UN Security Council, fitting cross-border interactions like "visit" and "criticize". Class 28 centers on Chinese institutions, and class 65 on Indian parties and politicians; the two share similar relation vocabularies, but the subject roles differ. Classes 54, 55, and 12 revolve around armed and security actors in Yemen, Israel and Palestine, and Afghanistan, respectively. Class 20, by contrast, gathers Kenyan ethnic groups and local media, showing that the ontology also separates social roles within a country instead of only splitting by national borders.

Not all 70 classes are equally clean: some mix the countries of a region with its officials and organizations, e.g., class 47 contains Fumio Kishida and Tony Tan Keng Yam, and some geographically neighboring countries are merged into one domain, e.g., class 69 contains Lesotho. We therefore regard the Defined Label as a post-hoc interpretation: OnRule only uses the class indices and membership degrees, and never relies on these textual labels. Overall, this shows that semantic clustering recovers a role partition consistent with the event structure of ICEWS, providing inspectable grounds for the ontology-aware rules.

Table 6: Examples of learned ontology classes on ICEWS14. Defined labels are human summaries, not model outputs.

Ontology	n_{ent}	Defined Label	Entities
42	97	Sovereign states and foreign-policy actors	China; Japan; Iraq; Canada; Malaysia; Barack Obama; John Kerry; Pope Francis; UN Security Council; Human Rights Watch
47	80	East and South-east Asian states	Vietnam; Myanmar; Laos; Fiji; North Korea; South Korea; New Zealand; Fumio Kishida; Tony Tan Keng Yam; Association of Southeast Asian Nations
3	43	Horn of Africa	Ethiopia; Djibouti; South Sudan; Al-Shabaab; African Union; Uhuru Muigai Kenyatta; William Ruto; Riek Machar; Somali National Army; Doctors Without Borders
34	72	Sudan and South Sudan	Sudan; Dinka; Messiria; Umar al Bashir; David Yau-Yau; James Wani Igga; Sudan People's Liberation Army; Sudan News Agency; Party of National Unity; Indigenous People (South Sudan)
19	79	Thailand and Cambodia	Thailand; Cambodia; Hun Sen; Thaksin Shinawatra; Prayuth Chan-ocha; Abhisit Vejjajiva; Royal Thai Navy; Royal Thai Police; National Council for Peace and Order of Thailand; Chulalongkorn University
61	52	Tanzania and the Swahili coast	Tanzania; Zanzibar; Burundi; John Pombe Magufuli; Jakaya Mrisho Kikwete; Civic United Front; East African Community; Tanzania Investment Bank; Police (Tanzania); Indigenous People (Tanzania)
7	112	Nigeria and the Lake Chad basin	Cameroon; Boko Haram; Muhammadu Buhari; Sani Abacha; Bukola Saraki; Nigerian Navy; Nigerian National Petroleum Corporation; State Security Service; Fula people; Daily Trust
20	49	Kenyan communities	Kikuyu; Kalenjin; Maasai; Kamba; Kisii; Pokot; Mwai Kibaki; Daniel Arap Moi; Kenya Television Network; The Standard
69	62	South Africa	Lesotho; Julius Malema; Kgalema Petrus Motlanthe; Lindiwe Nonceba Sisulu; ANC Youth League; Congress of South African Trade Unions; South African National Defence Force; University of South Africa; Sunday Times; Mogoeng Thomas Reetsang Mogoeng
38	42	Brazil	Luiz Inacio Lula Da Silva; Fernando Henrique Cardoso; Gilberto Kassab; Joaquim Barbosa; Globo Television; Senate (Brazil); First Command of the Capital; Universal Church of the Kingdom of God; Labor Union (Brazil); Public Prosecutor (Brazil)
49	53	Argentina	Nestor Kirchner; Carlos Saul Menem; Fernando de la Rúa; Sergio Massa; Axel Kicillof; Justicialist Party; Argentine Industrial Union; Radio Mitre; Civic Coalition; Julio Cobos
63	78	Mexico and northern Latin America	Mexican Army; Sinaloa Cartel; Drug Gang (Mexico); Rafael Correa; Miguel Ángel Osorio Chong; National Liberation Army; Guerrilla (Revolutionary Armed Forces of Colombia); Yoani Sánchez; Belize; Indigenous People (Guatemala)
41	22	Canada	Justin Trudeau; Kathleen Wynne; Jason Kenney; Philippe Couillard; Bill Blair; Tim Hudak; Liberal Party (Canada); Police (Canada); Labor Union (Canada); Ministry of Public Security (Canada)
36	38	United Kingdom	Theresa May; Boris Johnson; Nicola Sturgeon; Alex Salmond; 10 Downing Street; Scotland Yard; Conservative Party (United Kingdom); Labor Party (United Kingdom); Northern Ireland; Provisional Irish Republican Army
51	70	France and Francophone Africa	Air France; Nathalie Kosciusko-Morizet; Francois Molins; Christian Noyer; Communist Party (France); Catholic Priest (France); Abdoulaye Wade; Guillaume Soro; Aminata Touré; Karim Wade
59	50	Greece	Alexis Tsipras; Antonis Samaras; New Democracy; Coalition of the Radical Left; Panhellenic Socialist Movement; European Central Bank; Hellenic Armed Forces; Evangelos Venizelos; Yannis Stournaras; Bank (Greece)
57	23	Romania	Klaus Johannis; Ion Iliescu; Dan Diaconescu; National Liberal Party; Democratic Union of Hungarians in Romania; National Union for the Progress of Romania; Senate (Romania); Constitutional Court (Romania); Legislature (Romania); Strabag
46	70	Ukraine	Yulia Tymoshenko; Viktor Yushchenko; Vitali Klitschko; Verkhovna Rada; Party of Regions; Security Service of Ukraine; Ukrainian National Guard; Svoboda; Ukrinform; Arsen Avakov
32	58	Korean peninsula	Kim Jong-Un; Roh Moo Hyun; Kim Daejung; Samsung; Korean Workers Party; Grand National Party; Seoul High Court; Chongwadai; Police (South Korea); Military (North Korea)
28	56	Chinese officials and institutions	Bank (China); Legislature (China); City Mayor (China); Information / Communication / Transparency Ministry (China); Central Military Commission; Straits Exchange Foundation; Mainland Affairs Council; Uighur (China); Dissident (China); East Turkistan Islamic Movement
65	104	Indian parties and politicians	Amit Shah; Sharad Pawar; Smriti Irani; Aam Aadmi Party; Shiv Sena; Rajya Sabha; Rashtriya Swayamsevak Sangh; Communist Party of India; Border Security Force; Raghuram Rajan
18	58	Pakistan	Imran Khan; Pervez Musharraf; Benazir Bhutto; Raheel Sharif; Pakistan Tehreek-e-Insaf; Pakistan People's Party; Tehrik-i-Taliban Pakistan; Inter-Services Intelligence; Geo TV; Sindh High Court
5	42	Egypt	Abdel Fattah Al-Sisi; Adli Mansour; Al Jazeera; Egyptian Army; Al-Azhar University; Cairo University; Supreme Constitutional Court; Central Security Forces; Peter Greste; Hamdeen Sabahi
66	48	Turkey	Justice and Development Party; Republican People's Party; Nationalist Movement Party; Devlet Bahçeli; Süleyman Demirel; Mehmet Simsek; Turkish Coast Guard; Berkin Elvan; Militant (Kurdistan Workers' Party); Public Prosecutor (Turkey)
54	39	Yemen	Ali Abdullah Saleh; Ansar ash-Shari'a; Yemeni Congregation for Reform; Militia (Houthis); Combatant (Houthis); Armed Gang (Al Qaeda); Police (Yemen); Scholar (Yemen); Governor (Yemen); Military Police (Yemen)
55	44	Israel and Palestine	Benny Gantz; Mohammed Deif; Yehuda Glick; Arab Israelis; Al Arabiya; Jew (Israel); Resistance Movement (Hamas); Combatant (Izz ad-Din al-Qassam Brigades); Military Wing (Palestinian Territory;
12	73	Afghanistan	Ashraf Ghani Ahmadzai; Abdullah Abdullah; Gulbuddin Hekmatyar; Mohammed Qasim Fahim; Afghan National Army; National Directorate of Security; Hezbe Wahdat; Voice of Jihad; Militant (Taliban); Mujahideen (Afghanistan)
25	99	Philippines	Abu Sayyaf; Moro Islamic Liberation Front; Moro National Liberation Front; Franklin Drilon; Loren Legarda; University of the Philippines; Philippine Military Academy; National Police; Eastern Mindanao Command; Islamic Militant (Philippines)
11	71	Iraqi and Iranian political military actors	Nuri al-Maliki; Haider Al-Abadi; Fuad Masum; Massoud Barzani; Quds Force; Mahmoud Ahmadinejad; Iranian Parliament; Militia (Iraq); Iranian Kurds; Terrorist (Islamic State of Iraq and the Levant)
52	78	Australia	Julia Gillard; Kevin Rudd; Malcolm Bligh Turnbull; Bob Hawke; Liberal National Party; Australia Greens; House of Representatives (Australia); Police (Australia); Hell's Angels; Nicholas "Nick" Xenophon

F AN EXAMPLE OF LLM PAIRWISE CORRECTION ON ICEWS14

Figure 5 shows a real ICEWS14 object-prediction case from the full test split. The query is whether "Police_(Hong_Kong)" will "Physically_assault" some object at timestamp 334 (2014-12-01); the gold object is "Protester_(Hong_Kong)".

OnRule ranks "Student_(Hong_Kong)" first and "Protester_(Hong_Kong)" second. We attach to each candidate an answer-conditioned evidence pack: historical facts that would support that entity as the missing object, capped at 12 facts per side. All 12 pieces of evidence for "Student_(Hong_Kong)" are one-hop, and none of them involves the query relation "Physically_assault".

For "Protester_(Hong_Kong)", the first two pieces are multi-hop: "Police_(Hong_Kong)" first establish contact with "Other_Authorities_/_Officials_(Hong_Kong)", and "Protester_(Hong_Kong)" then demonstrate against these officials. The remaining 10 pieces are one-hop, and the first 6 of them are exactly "Physically_assault", with the most recent one occurring on day 291, closer to the query time than the most recent one on the student side, which is day 289.

The LLM does not reorder the full shortlist in one shot. It performs an adjacent pairwise comparison between the current rank-1 and rank-2 candidates, using the prompt in the figure. The higher OnRule candidate is presented as A and is kept unless the evidence clearly favors B. The LLM replies B, so the two names are swapped and the gold object moves to rank 1.

This example shows that the OnRule places a plausible neighbor first, while the evidence packs make the query relation itself more specific to the gold object.

Example of LLM Pairwise Correction on ICEWS14

Query and gold answer.
Query: `[Police_(Hong_Kong), Physically_assault, ?, 334]`
Gold answer: `Protester_(Hong_Kong)`

OnRule ranking and evidence packs.
1. `6476.Student_(Hong_Kong)`
   E1. `[Student_(Hong_Kong), Make_statement, 5449.Police_(Hong_Kong), 289]`
   E2. `[Student_(Hong_Kong), Protest_violently,_riot, 5449.Police_(Hong_Kong), 287]`
   E3. `[Police_(Hong_Kong), fight_with_small_arms_and_light_weapons, 6476.Student_(Hong_Kong), 285]`
   E4. `[Student_(Hong_Kong), Accuse, 5449.Police_(Hong_Kong), 277]`
   E5. `[Student_(Hong_Kong), Accuse, 5449.Police_(Hong_Kong), 276]`
   E6. `[Police_(Hong_Kong), Use_conventional_military_force, 6476.Student_(Hong_Kong), 270]`
   E7. `[Student_(Hong_Kong), Protest_violently,_riot, 5449.Police_(Hong_Kong), 270]`
   E8. `[Student_(Hong_Kong), Use_unconventional_violence, 5449.Police_(Hong_Kong), 270]`
   E9. `[Student_(Hong_Kong), Make_statement, 5449.Police_(Hong_Kong), 268]`
   E10. `[Police_(Hong_Kong), Impose_administrative_sanctions, 6476.Student_(Hong_Kong), 266]`
   E11. `[Police_(Hong_Kong), Use_conventional_military_force, 6476.Student_(Hong_Kong), 265]`
   E12. `[Student_(Hong_Kong), Use_unconventional_violence, 5449.Police_(Hong_Kong), 265]`
2. `5665.Protester_(Hong_Kong)`
   E1. `[Other_Authorities_/_Officials_(Hong_Kong), Make_statement, 5449.Police_(Hong_Kong), 287]`
   E2. `[Protester_(Hong_Kong), Demonstrate_or_rally, 5167.Other_Authorities_/_Officials_(Hong_Kong), 288]`
   E3. `[Police_(Hong_Kong), Physically_assault, 5665.Protester_(Hong_Kong), 291]`
   E4. `[Police_(Hong_Kong), Physically_assault, 5665.Protester_(Hong_Kong), 290]`
   E5. `[Police_(Hong_Kong), Physically_assault, 5665.Protester_(Hong_Kong), 289]`
   E6. `[Police_(Hong_Kong), Physically_assault, 5665.Protester_(Hong_Kong), 288]`
   E7. `[Police_(Hong_Kong), Physically_assault, 5665.Protester_(Hong_Kong), 287]`
   E8. `[Police_(Hong_Kong), Physically_assault, 5665.Protester_(Hong_Kong), 286]`
   E9. `[Police_(Hong_Kong), Make_statement, 5665.Protester_(Hong_Kong), 309]`
   E10. `[Police_(Hong_Kong), Use_tactics_of_violent_repression, 5665.Protester_(Hong_Kong), 309]`
   E11. `[Protester_(Hong_Kong), Demonstrate_or_rally, 5449.Police_(Hong_Kong), 309]`
   E12. `[Protester_(Hong_Kong), Protest_violently,_riot, 5449.Police_(Hong_Kong), 309]`
3. ...

Full prompt to the LLM.
System:

You are an expert in Temporal Knowledge Graph Reasoning. You compare two candidate objects for one temporal knowledge graph query. A symbolic rule reasoner already ranked them. You receive the evidence that supports A as the missing object, and the evidence that supports B as the missing object. You must abide by the following rules:
– Use only the two evidence packs. Do not use outside knowledge or invent facts.
– Prefer the object that appears more recently with the query subject, especially with the query relation.
– If the evidence does not clearly favor B, keep A (the reasoner's currently higher-ranked candidate).
– Reply with exactly one letter: A or B. No other text.

User:
Query: `[Police_(Hong_Kong), Physically_assault, ?, 334]`
Subject: `Police_(Hong_Kong)`; Relation: `Physically_assault`; Time: `334`.
A. (OnRule rank 1): `6476.Student_(Hong_Kong)`. Evidence if A is the answer:
E1. `[Student_(Hong_Kong), Make_statement, 5449.Police_(Hong_Kong), 289]`
E2. `[Student_(Hong_Kong), Protest_violently,_riot, 5449.Police_(Hong_Kong), 287]`
E3. `[Police_(Hong_Kong), fight_with_small_arms_and_light_weapons, 6476.Student_(Hong_Kong), 285]`
E4. `[Student_(Hong_Kong), Accuse, 5449.Police_(Hong_Kong), 277]`
E5. `[Student_(Hong_Kong), Accuse, 5449.Police_(Hong_Kong), 276]`
E6. `[Police_(Hong_Kong), Use_conventional_military_force, 6476.Student_(Hong_Kong), 270]`
E7. `[Student_(Hong_Kong), Protest_violently,_riot, 5449.Police_(Hong_Kong), 270]`
E8. `[Student_(Hong_Kong), Use_unconventional_violence, 5449.Police_(Hong_Kong), 270]`
E9. `[Student_(Hong_Kong), Make_statement, 5449.Police_(Hong_Kong), 268]`
E10. `[Police_(Hong_Kong), Impose_administrative_sanctions, 6476.Student_(Hong_Kong), 266]`
E11. `[Police_(Hong_Kong), Use_conventional_military_force, 6476.Student_(Hong_Kong), 265]`
E12. `[Student_(Hong_Kong), Use_unconventional_violence, 5449.Police_(Hong_Kong), 265]`
B. (OnRule rank 2): `5665.Protester_(Hong_Kong)`. Evidence if B is the answer:
E1. `[Other_Authorities_/_Officials_(Hong_Kong), Make_statement, 5449.Police_(Hong_Kong), 287]`
E2. `[Protester_(Hong_Kong), Demonstrate_or_rally, 5167.Other_Authorities_/_Officials_(Hong_Kong), 288]`
E3. `[Police_(Hong_Kong), Physically_assault, 5665.Protester_(Hong_Kong), 291]`
E4. `[Police_(Hong_Kong), Physically_assault, 5665.Protester_(Hong_Kong), 290]`
E5. `[Police_(Hong_Kong), Physically_assault, 5665.Protester_(Hong_Kong), 289]`
E6. `[Police_(Hong_Kong), Physically_assault, 5665.Protester_(Hong_Kong), 288]`
E7. `[Police_(Hong_Kong), Physically_assault, 5665.Protester_(Hong_Kong), 287]`
E8. `[Police_(Hong_Kong), Physically_assault, 5665.Protester_(Hong_Kong), 286]`
E9. `[Police_(Hong_Kong), Make_statement, 5665.Protester_(Hong_Kong), 309]`
E10. `[Police_(Hong_Kong), Use_tactics_of_violent_repression, 5665.Protester_(Hong_Kong), 309]`
E11. `[Protester_(Hong_Kong), Demonstrate_or_rally, 5449.Police_(Hong_Kong), 309]`
E12. `[Protester_(Hong_Kong), Protest_violently,_riot, 5449.Police_(Hong_Kong), 309]`
Which object is better supported by its evidence? A or B:

Ranking after LLM Pairwise Correction.
Model output: `B`
1. `Protester_(Hong_Kong)`; 2. `Student_(Hong_Kong)`; ...

Figure 5: A real object-prediction example on ICEWS14.