Strikingness-Aware Evaluation for Temporal Knowledge Graph Reasoning

Rikui Huang[1,2,3], **Shengzhe Zhang**[2] and **Wei Wei**[1,2,3,*]

[1]School of Computer Science & Technology, Huazhong University of Science and Technology
[2]Institute of Artificial Intelligence, Huazhong University of Science and Technology
[3]School of Artificial Intelligence & Automation, Huazhong University of Science and Technology
{huangrk, zsz, weiw}@hust.edu.com

Abstract

Temporal Knowledge Graph Reasoning (TKGR) aims at inferring missing (especially future) events from historical data. Current evaluation in TKGR uniformly weights all events, ignoring that most are trivial repetitions, which overestimate the true reasoning ability. Therefore, the rare outstanding events, whose prediction demands deeper reasoning, should be distinguished and emphasized. To this end, we propose a strikingness-aware evaluation framework, which introduces a rule-based strikingness measuring framework (RSMF) to quantify event strikingness by comparing its expected occurrence with peer events derived from temporal rules. Strikingness is then integrated as a weighting factor into metrics like weighted MRR and Hits@k. Experiments on four TKG benchmarks reveal: 1) All representative models perform worse as event strikingness increases, 2) Path-based methods excel on low-strikingness events and representation-based ones on high-strikingness events, 3) We design an ensemble method whose gains stem from fitting trivial events rather than reasoning improvement. Our framework provides a more rigorous evaluation, refocusing the field on predicting outstanding events.

1 Introduction

Recent advances in Temporal Knowledge Graph Reasoning (TKGR) have led to substantial progress, which can be broadly categorized into two classes according to whether they forecast future events: interpolation and extrapolation reasoning [Jin *et al.*, 2020]. The former refers to inferring missing historical facts, while the latter involves predicting future events, also named temporal knowledge graph forecasting [Sun *et al.*, 2021]. This work primarily focuses on extrapolation reasoning, which is essential for many high-risk applications like financial risk control [Aven, 2013].

Despite promising empirical results [Liang *et al.*, 2024], many reported gains may largely arise from data biases, leading to misjudgment of advancements in the field [Kervadec *et al.*, 2021]. A historical parallel exists in the static KGR field, potential data leakage in well-known benchmarks (WN18, FB15k) led to an overestimation of the reasoning capabilities of models [Toutanova and Chen, 2015; Dettmers *et al.*, 2018]. Over 94% and 81% of queries, such as (A, *hypernym*, ?), in WN18 and FB15k can easily be mapped to a training triple (B, *hyponym*, A) if it is known that hyponym is the inverse of hypernym. Recently, a similar phenomenon is emerging in the TKGR field: over 80% of events have occurred in prior history in the ICEWS data [Zhu *et al.*, 2021]. It may enable heuristic-based predictions, inflating state-of-the-art (SOTA) performance, under the existing TKGR evaluation framework, on common events while obscuring poor accuracy on fewer than 10% of truly challenging, striking cases. For example, given a query such as $(A, MakeVisit, ?, T_q)$, they may output an answer B by selecting either the most frequently occurring historical event $(A, MakeVisit, B, T_i)$ [Lee *et al.*, 2023; Xu *et al.*, 2023]. It raises doubts about the predictive quality of current TKGR methods and whether the current evaluation framework could reasonably reflect these models' forecasting capabilities [Gastinger *et al.*, 2024b].

The above flaw in static KGR has been addressed by removing inverse relation triples from their training sets, i.e., creating WN18RR [Dettmers *et al.*, 2018] and FB15k-237 [Toutanova and Chen, 2015]. However, this straightforward removal strategy is fundamentally inapplicable to TKGR, since all historical events, even repetitive ones, constitute essential evidence for forecasting the future. This dilemma raises a critical question: How to construct a more meaningful evaluation framework for TKGR without deleting data?

A principled feasible alternative is to re-weight test-instance gains, rather than treating all instances uniformly. Specifically, trivial events such as (A, *MakeVisit*, B) occurring across different timestamps frequently should be assigned lower weights. In contrast, rarer outstanding events like (A, *Sign Agreement*, B), which require deeper temporal reasoning, should be emphasized. Generally, accurately inferring outstanding events offers far greater practical value than merely predicting numerous trivial ones. However, measuring the weights and automatically identifying outstanding ones from a large volume of trivial events is nontrivial. While there have been some efforts to measure strikingness of facts

*Corresponding author.

in static KGs, there is a notable lack of studies in TKGs field. This gap stems from two challenges: First, beyond the statistical features, a comprehensive TKGR evaluation framework necessitates incorporating both semantic and temporal relevance. Second, since the ground-truth impact of a future event is unknowable in advance, any measure of its strikingness can only be derived from observable historical patterns.

To address this, we propose a **R**ule-based **S**trikingness **M**easuring **F**ramework (**RSMF**) to measure the strikingness of future events based on historical evidence. RSMF first leverages first-order temporal rules to retrieve peer events for the target event. Subsequently, it computes the expected occurrence of candidate events with the semantic confidence of rules, the temporal characteristics of events, and the frequency of event repetition. The strikingness of the future event is derived by contrasting its expected occurrence with that of its peer events. Finally, we construct a strikingness-aware evaluation framework by using strikingness as a weighting factor. Experimentally, we evaluate eight representative baselines under the striking-aware evaluation framework across four widely adopted TKG datasets, including three path-based, three representation-based, and two large language model (LLM)-based approaches. Our contributions and key findings can be summarized as follows:

- We propose RSMF to quantify event strikingness in TKGs, and build a new corresponding striking-aware TKGR evaluation framework that re-weights test instances with their strikingness.
- We show that the reasoning performance of evaluated models decreas as event strikingness increases, i.e., events with higher strikingness are more difficult to predict.
- We find distinct performance patterns across baselines: path-based methods show stronger performance in low-strikingness events, whereas representation-based approaches excel at high-strikingness events.
- We design an ensemble method combining path- and representation-based models, aiming to leverage their complementary strengths. Consequently, it separately obtains significant and marginal gains in the original and our proposed strikingness-aware framework. Analysis reveals that while the method's gains come from dominant low-strikingness events, whereas, performance on rare high-strikingness events decreases.

2 Related Work

Temporal Knowledge Graph Reasoning Evaluation In recent years, researchers have proposed various extrapolation TKGR methods, including graph neural networks-based [Li *et al.*, 2021; Li *et al.*, 2022; Chen *et al.*, 2024], rule-based [Liu *et al.*, 2022; Huang *et al.*, 2024], reinforcement learning based [Sun *et al.*, 2021; Zheng *et al.*, 2023; Dong *et al.*, 2023], and the increasingly popular large language models-based methods [Lee *et al.*, 2023; Liao *et al.*, 2024; Xia *et al.*, 2024]. Alongside these advancements, the common rank-based evaluation, for link prediction like KGR, methods are undergoing continuous refinement. Initially, to address the issue of multiple answers for a single query, correct answers other than the target answer are filtered out during ranking to avoid underestimating model performance [Bordes *et al.*, 2013]. Subsequently, to accommodate TKGR, time-aware filtering [Han *et al.*, 2021] and time interval prediction evaluation were introduced [Jain *et al.*, 2020]. Considerable efforts have also been made in re-evaluating the performances of various models to establish fair comparisons [Sun *et al.*, 2020; Ruffinelli *et al.*, 2020; Gastinger *et al.*, 2023]. In addition, a valuable baseline named *Recurrency* highlight flaws in datasets and offered significant insights [Gastinger *et al.*, 2024b]. To explore the capability boundaries of TKGR models, some studies attempt to build new benchmark datasets tailored to different scenarios, such as context-aware [Ma *et al.*, 2023], multi-modal [Li *et al.*, 2024], and large-scale settings [Gastinger *et al.*, 2024a]. However, they may also suffer from the above data biases, as the repetitive pattern is an inherent characteristic of TKGs.

Outstanding Facts Mining in Knowledge Graph Outstanding facts (OFs) mining focuses on quantifying the strikingness of facts. Early research focused on extracting OFs from unstructured data (e.g., text) [Angiulli *et al.*, 2009; Hassan *et al.*, 2014; Wu *et al.*, 2012]. Maverick [Zhang *et al.*, 2018] firstly measured event strikingness in static KGs with the specific attribute values of entities. FMINER [Yang *et al.*, 2021] introduced context entity constraints and designed a pattern relevance model to optimize the process of event searching. The robustness of measured outstanding events is further explored using perturbation analysis [Xiao *et al.*, 2024]. To the best of our knowledge, our framework is the first principled extension of the established Outstanding Fact Mining paradigm from static KGs to TKGs. Furthermore, we transform a mining technique into a comprehensive evaluation framework, creating weighted metrics that reorient the TKGR field towards valuing outstanding reasoning.

3 Strikingness-Aware Evaluation

3.1 Preliminaries

Temporal Knowledge Graph Reasoning A TKG can be represented as a sequence of timestamp KGs, denoted as $\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_t\}$. Each KG at a specific timestamp t is defined as $\mathcal{G}_t = (\mathcal{E}, \mathcal{R}, \mathcal{F}_t)$, where $\mathcal{E}$ is the set of entities, $\mathcal{R}$ represents the set of relations, and $\mathcal{F}_t = \{(s, r, o, t)\}$ refers to the set of events observed at timestamp t. Given a query $(s, r, ?, t)$, a reasonable TKGR model is to infer the object o based on the facts observed before t, where s and o are subject and object entities, r is a relation, and t is a timestamp. For instance, the query (*Markieff Morris*, *join*, ?, 2025-02) requires the model to predict *the Lakers* based on events before 2025-02 to validate its forecasting capability in practice.

Strikingness of Events Strikingness quantifies how outstanding a target event $f = (s, r, o, t)$ is, which is a continuous value in the range [0, 1]. The closer the value is to 1, the more outstanding the event, and vice versa. Events with low strikingness can be referred to as *trivial events*, while events with high strikingness can be referred to as *outstanding events*. Since it is not meaningful to compare two entirely

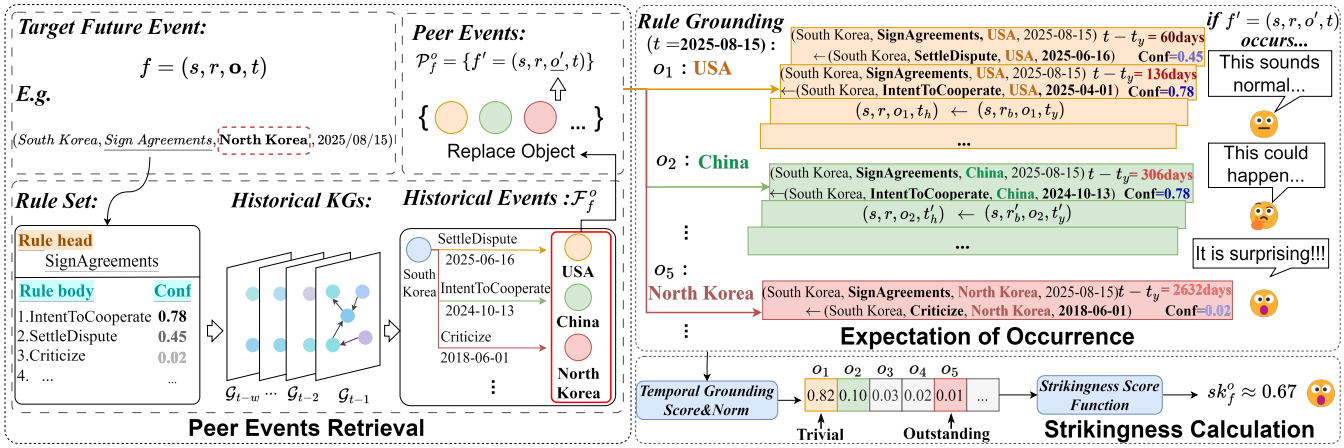

Figure 1: An example of strikingness measuring for target future event $(South\ Korea, Sign\ Agreements, North\ Korea, 2025/08/15)$ (replacing object). In **Peer Events Retrieval**, RMFS retrieves the historical events and constructs the peer events with the rule set. The **Expectation of Occurrence** and **Strikingness Calculation** are calculated by the rule grounding and strikingness scoring function.

unrelated events, such as *Markieff Morris will join the Lakers in 2025* and *The Federal Reserve will cut interest rates in 2026*, the strikingness thus is defined by comparing it with peer events $\mathcal{P}$.

Peer Events A peer event is a related event of the target event generated by replacing entities or relations. For a target future event $f = (s, r, o, t)$, its peer events are defined as $\mathcal{P}_f^s = \{(s', r, o, t) | s' \in \mathcal{E}\}$, $\mathcal{P}_f^o = \{(s, r, o', t) | o' \in \mathcal{E}\}$, and $\mathcal{P}_f^r = \{(s, r', o, t) | r' \in \mathcal{R}\}$.

3.2 Strikingness Measuring

To measure the strikingness of an event, three key challenges must be addressed: 1) Constructing a set of peer events that can be used for comparison with the target future event, 2) Computing the expectation of occurrence of the target event and its peer events, and 3) Calculating the strikingness score of the target event using the expectation of occurrence. The overall procedure for RSMF is outlined in Figure 1.

Peer Events Retrieval Peer events can be obtained by replacing the entities or relations of the target event. However, direct substitution may generate many meaningless peer events, such as (*Markieff Morris*, *join*, *Microsoft*, 2025-02). Therefore, we utilize temporal rules to constrain the generation of peer events from historical KGs.

For a target future event $f = (s, r, o, t)$, we first obtain the rule set TR corresponding to the relation r through rule mining [Liu *et al.*, 2022]. While higher-order rules could capture more complex patterns, they also introduce exponential computational complexity and risk of overfitting. Thus, we only use the length 1 rules as a practical measure. A temporal rule is defined as follows:

$$(E_1, r_h, E_2, T_2) \leftarrow (E_1, r_b, E_2, T_1) \tag{1}$$

where $T_1 < T_2$, r_h and r_b denote rule head and body relation, E_i and T_i indicate entity and timestamp variables.

For ease of understanding, we take the retrieval of the object entity as an example. We first mask the object entity in the target event to convert it into a query $f_q = (s, r, ?, t)$, and then use the historical KG sequences and temporal rules to search for historical events that support this query. For a rule $tr \in TR$, we ground the rule body in the historical KG sequences $\{\mathcal{G}_i\}_{i=t-w}^{t-1}$ to obtain the grounded historical events:

$$\mathcal{F}_{f,tr}^o = \{(s, r_b, o', t') | t - w \leq t' < t\} \tag{2}$$

where r_b represents the relation of the rule body for tr, while w controls the window of the historical KG sequences. The grounded events of all rule bodies for tr:

$$\mathcal{F}_f^o = \bigcup_{tr \in TR} \mathcal{F}_{f,tr}^o \tag{3}$$

Then, we take the object entities in $\mathcal{F}_f^o$ as the candidate set $\mathcal{C}_f^o = \{o' | (s, r_b, o', t') \in \mathcal{F}_f^o\}$ for object entity replacement. Further, peer object events $\mathcal{P}_f^o = \{(s, r, o', t) | o' \in \mathcal{C}_f^o\}$ of the target event f can be generated by substituting the object entities o. The peer relation events $\mathcal{P}_f^r$ and the peer subject events $\mathcal{P}_f^s$ could be obtained similarly.

Expectation of Occurrence Strikingness is related to human expectations regarding the occurrence of an event. To this end, we employ a rule-based approach to compute the expected scores of the target event and its peer events. For a peer event f', we ground the instances by applying each rule $tr \in TR$ to obtain the rule grounding events pair:

$$rg_{f'}^{tr} = (s, r, o', t_h) \leftarrow (s, r_b, o', t_y) \tag{4}$$

Each $rg_{f'}^{tr}$ consists of a rule body and a rule head means the body event could support the occurrence of head event.

Furthermore, we also consider the impact of event frequency on strikingness measurement. Intuitively, the more frequently the rule grounding observed, the higher the expectation of event f'. Therefore, we iteratively collect rule grounding to obtain the set of rule grounding:

$$RG_{f'}^{tr} = \{(s, r, o', t_{h_i}) \leftarrow (s, r_b, o', t_{y_j})\}_{j=1}^{n},$$
$$with\ t_{y_1} < t_{h_1} \leq t_{y_2} < t_{h_2} \leq ... \leq t_{y_n} < t_{h_n} \tag{5}$$

where $t_{y_1} \geq t - w$, and n represents the number of rule groundings. The temporal constraint is utilized to avoid overlap and redundancy. Additionally, we set $t_{h_n} = t$, indicating that in the temporally closest grounding, only the rule body is a historical event, which provides support for reasoning about the potential future event f'.

After obtaining the rule grounding set, we compute the expectation score of target event and its peer events. Following the rule-based reasoning methods [Ott *et al.*, 2023; Huang *et al.*, 2024], we design the expectation score from two aspects. First, the effectiveness of rule-based reasoning will decay over time, and we use an exponential distribution to model this decay. Second, since different rules contribute variably to expectation, we employ the confidence of each rule to reflect its contribution. The expectation score is calculated as follows:

$$sc_{f'} = \sum_{tr \in TR} \sum_{rg^{tr}_{f'} \in RG^{tr}_{f'}} conf(tr) * e^{-\lambda(t-t_y)} \tag{6}$$

where $conf(tr)$ represents the confidence of rule tr, $\lambda > 0$ is the temporal decay coefficient, and t_y denotes the timestamp of the rule body in the rule grounding $rg^{tr}_{f'}$.

According to the different elements being replaced, we obtain the corresponding score sets of the target event f and transform them to vectors, denoted as $\mathbf{sc}^s_f \in \mathbb{R}^{|\mathcal{C}^s_f|}$, $\mathbf{sc}^r \in \mathbb{R}^{|\mathcal{C}^r_f|}$, and $\mathbf{sc}^o \in \mathbb{R}^{|\mathcal{C}^o_f|}$. These three sets of scores estimate the expectation of events from different perspectives.

Strikingness Calculation The expectation score reflects the prior perception of the likelihood of an event's occurrence. Strikingness measures the degree to which the event exceeds the prior expectation and is thus inversely correlated with the expectation score. That is, the higher the expectation score assigned to the event f', the less striking the occurrence of the event. To constrain strikingness within the range [0, 1], the obtained score sets need to be normalized as follows:

$$\mathbf{sc}^{be}_{norm} = \mathbf{sc}^{be}_f / ||\mathbf{sc}^{be}_f||_2 \tag{7}$$

where body element $be \in \{s, r, o\}$ is a replaced element, and $||\cdot||_2$ is L2 normalization.

Then, we compare the normalized scores of peer events f' to highlight the prominence of the target event f. The strikingness scoring function accounts for both the magnitude of peer event scores and the differences between them. Based on the strikingness measure proposed in [Angiulli *et al.*, 2009], we adopt the following function to calculate the strikingness of the body elements:

$$sk^{be}_f = \sum sc^{be}_{f'} * (sc^{be}_{f'} - sc^{be}_f) * \mathbb{I}(sc^{be}_{f'} > sc^{be}_f) \tag{8}$$

where $sc^{be}_{f'} \in \mathbf{sc}^{be}_{norm}$, and $\mathbb{I}(\cdot)$ is the indicator function that returns the value 1 if the condition is true and 0 otherwise.

Finally, we weight the strikingness of all body elements to obtain the final strikingness of potential future event f:

$$sk_f = \alpha^s sk^s_f + \alpha^o sk^o_f + \alpha^r sk^r_f \tag{9}$$

where $\alpha^s; \alpha^o; \alpha^r \in [0, 1]$ are weights of body elements and $\alpha^s + \alpha^o + \alpha^r = 1$.

3.3 Striking-aware Evaluation Framework

Given a test event $f = (s, r, o, t)$, its rank of score is computed by corrupting candidate entities. Specifically, the object entity would be replaced by another candidate e_c, and the candidate event $f_c = (s, r, c, t)$ would be scored. By adding the reverse events, i.e., $f' = (o, r^{-1}, s, t)$ and replacing s by c, the subject entity prediction could be achieved. The two metrics are widely used to evaluate model performance:

Mean Reciprocal Rank (MRR) evaluates ranking quality by averaging the inverse rank of the correct result:

$$MRR = \frac{1}{|\mathcal{F'}_{test}|} \sum_{f \in \mathcal{F'}_{test}} \frac{1}{rank_f} \tag{10}$$

Hits@*k* (H@*k*) measures the proportion of queries where the correct result appears in the top-k positions:

$$Hits@k = \frac{1}{|\mathcal{F'}_{test}|} \sum_{f \in \mathcal{F'}_{test}} \mathbb{I}\{rank_f \leq k\} \tag{11}$$

where $\mathcal{F'}_{test} = \mathcal{F}_{test} \cup \{(o, r^{-1}, s, t) | (s, r, o, t) \in \mathcal{F}_{test}\}$, and $\mathbb{I}\{True\} = 1$ and $\mathbb{I}\{False\} = 0$.

However, the approach assigns equal weight to all future events, making it unable to capture the model's ability to predict outstanding events. To address this limitation, we propose a striking-aware evaluation framework to evaluate existing TKGR baselines. Specifically, the computed strikingness scores are used as weighting factors to calculate the Weighted MRR (WMRR) and Weighted Hits@k (WHits@k) metrics, as described below:

$$\text{WMRR} = \frac{\sum_{i=1}^{|N|}(s_i + b) * \frac{1}{rank_i}}{\sum_{i=1}^{|N|}(s_i + b)} \tag{12}$$

$$\text{WHits@k} = \frac{\sum_{i=1}^{|N|}(s_i + b) * \mathbb{I}(rank_i \leq k)}{\sum_{i=1}^{|N|}(s_i + b)} \tag{13}$$

where $|N|$ is the size of the test set and s_i is the strikingness of the event, which ensures that high-strikingness events contribute more to the metric. Setting b is equivalent to assigning a higher cost of mis-prediction to outstanding events in the evaluation. Specifically, with $b = 0.1$, an event with strikingness $sk = 1$ receives approximately ten times the weight of an event with $sk = 0$ in the metric calculation, since $(1 + b)/(0 + b) = 11$. This reflects our value judgment that the utility of correctly predicting a critical outstanding event far outweighs that of correctly predicting a routine trivial one. The parameter b thus quantifies and incorporates this value judgment into the evaluation framework.

In addition, we introduce a simple ensemble method. Instead of constructing complex networks, we follow [Meilicke *et al.*, 2021; Liu *et al.*, 2023; Wang *et al.*, 2024] to combine the output scores of the existing path- and representation-based methods straightforwardly to obtain the final score:

$$\mathbf{y}_{ensem} = \eta \mathbf{y}_{path} + (1 - \eta)\mathbf{y}_{representation} \tag{14}$$

where $\eta \in [0, 1]$ is a hyperparameter. We perform a hyperparameter search for the η on the validation set, and subsequently apply it to the test set. **Our purpose in constructing the ensemble method is not to pursue higher performance. Throughout the paper, we focus on investigating the boundaries of TKGR models' reasoning capabilities.**

4 Experiments

4.1 Implement Settings

Datasets Extensive experiments are conducted on four TKG datasets: ICEWS14, ICEWS18, ICEWS05-15, and GDELT. The datasets are divided in chronological order.

Baselines We conducted comparisons under a unified experimental framework, focusing on reproducible approaches, including path-based methods: Recurrency [Gastinger *et al.*, 2024b], TITer [Sun *et al.*, 2021], and TLogic [Liu *et al.*, 2022], representation-based methods: RE-GCN [Li *et al.*, 2021], TiRGN [Li *et al.*, 2022], and LogCL [Chen *et al.*, 2024], and LLM-based methods: ICL [Lee *et al.*, 2023] and GenTKG [Liao *et al.*, 2024]. We follow the time-aware filtering settings described in [Gastinger *et al.*, 2023][1]. The code and strikingness weights are available[2].

Implementation Details For the ensemble method, we perform a grid search for the hyperparameter η over the range $[0, 1]$ with a step size of 0.1, and determine its optimal value by evaluating performance on the validation set of each dataset. We set λ as 0.1 following TLogic [Liu *et al.*, 2022]. And α^s, α^o, α^r are set to 0.4, 0.4, 0.2. The recommended parameters setting is provided to establish a well-calibrated evaluation framework to facilitate reproducibility and promote community adoption. We employ the rule mining framework from TLogic [Liu *et al.*, 2022] to learn and extract temporal rules, with the distinction that we focus exclusively on rules of length 1.

4.2 TKGR Performance on Group Strikingness

Performance Across Groups Distinct from previous works that only report overall metrics for model performance, we group the data based on strikingness and compute the average performance within each group. Figure 2 presents the grouped MRR results for six baseline models and our proposed ensemble model on the ICEWS14. It could be observed that the volume of events decreases as strikingness increases. That is, low-strikingness events dominate the test set, whereas high-strikingness events are scarce, which aligns with human intuition. On performance, all models exhibit a decline with increasing strikingness, indicating that events with higher strikingness are more difficult to predict.

In overall trends, we find that path-based methods excel at predicting events with low strikingness, while representation-based methods demonstrate superior performance on high-strikingness events. This indicates that different categories of methods have distinct advantages in forecasting future events with different levels of strikingness. Based on this finding, we attempt to investigate whether a simple ensemble method can combine the advantages of both methods. Results on ICEWS0515 and GDELT exhibit consistent conclusions.

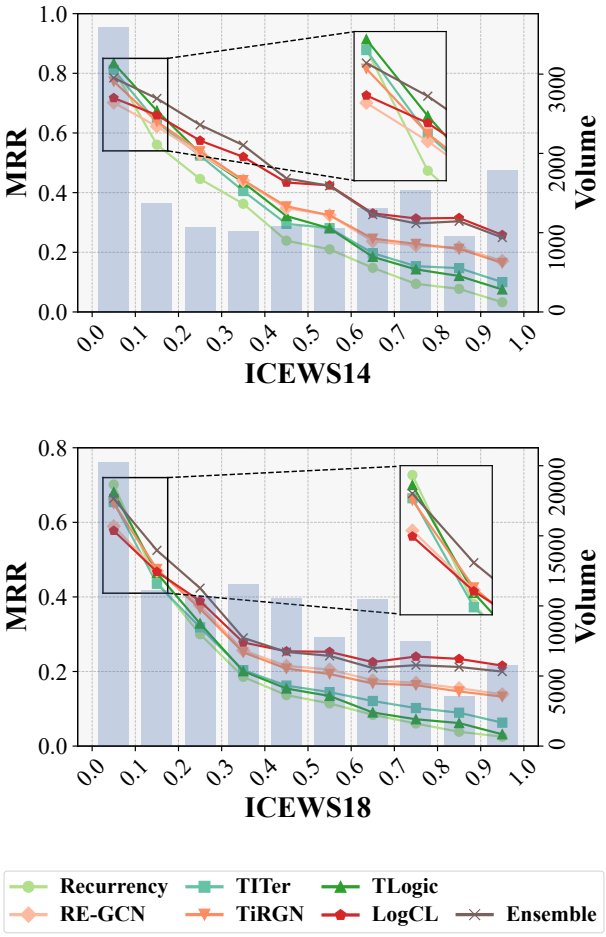

Figure 2: Group performances on ICEWS14 and ICEWS18. In each group, the bars denote the number of test events, while the lines indicate the average performance.

For the ensemble method, we observe two phenomena: **1) Trade-off**: For events with $sk < 0.1$ or $sk > 0.5$, the performance of the ensemble method lies between those of the individual methods. **2) Enhancement**: For events with $0.1 < sk < 0.5$, the ensemble method outperforms both individual methods. Based on these observations, it can be inferred that some models that leverage the ensemble method mainly improve the performance for low-strikingness events, which are predominant. For instance, comparing RE-GCN and TiRGN, which incorporates additional repeated events information, TiRGN significantly outperforms RE-GCN on low-strikingness events, while achieving similar performance on high-strikingness events.

Predictability To better understand the predictability of high-strikingness events, we introduce Neighborhood Overlap (NO_f), a structural metric that quantifies the richness of historical interactions between the subject and object entities.

[1]https://github.com/nec-research/TKG-Forecasting-Evaluation
[2]https://github.com/PersimmonZ1/RSMF

Model	Type	ICEWS14			
		$S(0.6, 0.7)$	$S(0.7, 0.8)$	$S(0.8, 0.9)$	$S(0.9, 1.0)$
Recurrency	High NO_f	19.10	13.46	12.05	5.23
	Low NO_f	12.28	5.77	5.36	1.65
TITer	High NO_f	27.61	21.92	22.89	16.67
	Low NO_f	17.51	11.41	9.13	7.44
TLogic	High NO_f	28.96	21.28	19.08	14.60
	Low NO_f	15.57	10.90	8.73	4.55
RE-GCN	High NO_f	38.36	31.03	30.92	24.57
	Low NO_f	17.81	19.23	17.86	14.98
TiRGN	High NO_f	35.97	31.15	29.12	21.78
	Low NO_f	19.31	20.26	18.25	14.88
LogCL	High NO_f	49.40	42.56	43.98	33.45
	Low NO_f	27.10	28.46	29.56	25.83
Ensemble	High NO_f	49.70	40.64	42.77	32.60
	Low NO_f	26.65	25.90	28.37	24.69

Table 1: The Hits@3 metric of High and Low NO_f events within the high-strikingness groups on ICEWS14.

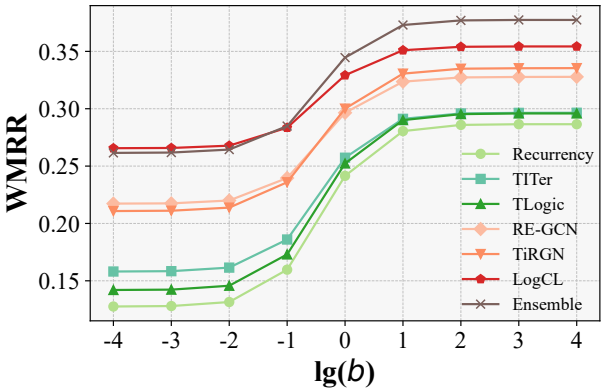

Figure 3: WMRR with different bias b on ICEWS18.

The NO_f is defined as follows:

$$NO_f = \frac{||N_s \cap N_o||}{||N_s \cup N_o||} \tag{15}$$

where $N_s = \{o'|(s, r', o', t')\} \cup \{o'|(o', r', s, t')\}$ and $N_o = \{s'|(s', r', o, t')\} \cup \{s'|(o, r', s', t')\}$ denote the set of neighboring entities of s and o, and $t - w \leq t' < t$ is consistent with the window for calculating strikingness.

As shown in Table 1, within each high-strikingness interval, events with higher NO_f (richer historical evidence) are consistently more predictable than those with lower NO_f. This confirms that even among outstanding events, those supported by sufficient historical evidence remain learnable. The findings validate that our strikingness measure aligns not only with event rarity but also with predictive difficulty rooted in evidence scarcity.

4.3 Strikingness-Aware Evaluation for TKGR

To mitigate the low-strikingness bias in the dataset and more comprehensively evaluate the ability of TKGR models to forecast future events, we propose a striking-aware evaluation framework. We evaluate existing models using weighted MRR and weighted Hits@k. The hyperparameter b determines the extent to which the evaluation results emphasize the model's ability to predict outstanding events. By adjusting b, the framework can place more or less weight on events, thereby controlling the balance performance between overall events and outstanding events. Figure 3 shows the WMRR of the models on ICEWS18 under different b. A small b indicates that WMRR focuses on the model's ability to predict outstanding events. When the b becomes very large, i.e., $b \geq 100$, the results of WMRR are close to the original MRR. Additionally, the comparison of model performance reverses as the value of b changes. When $b > 0.1$, Ensemble outperforms LogCL, and TiRGN outperforms RE-GCN. However, when $b < 0.1$, the results change such that LogCL surpasses Ensemble, and RE-GCN outperforms TiRGN. This demonstrates that LogCL has a superior ability to predict outstanding events compared to the Ensemble method, and a similar conclusion holds for RE-GCN and TiRGN.

We empirically select $b = 0.1$ to provide a unified evaluation result and facilitate fair comparison. This setting balances the metric's emphasis between the model's ability to reason events with varying levels of strikingness. Table 2 shows the experimental results of the models under both the original and the striking-aware evaluation frameworks. Obviously, the absolute values of the striking-aware metrics are significantly lower than those of the original evaluation framework, which aligns more closely with the recognized challenges of the future event forecasting task. Nevertheless, it is important to emphasize that our motivation is not to lower the scores purposely. Instead, we aim to reveal the models' comprehensive predictive capabilities through the differences between the original and striking-aware metrics. Furthermore, adjustments to other hyperparameters have a negligible impact on model ranking evaluations, demonstrating the robustness of the strikingness-aware evaluation framework.

As shown, path-based methods exhibit a reduction of over 30% across all datasets, with the heuristic baseline Recurrency demonstrating a particularly striking decrease of 50%. In contrast, methods based on evolutionary representations experience a notably smaller decline, with all reductions remaining below 30%. Among them, LogCL achieves the most robust performance, with its reduction consistently maintained at around 20%. This indicates that the existing path-based methods have not demonstrated the claimed multi-hop reasoning on TKGs. Methods based on LLMs exhibit similar behavior to path-based approaches because the context window size constrains their reasoning, typically limited to first-order neighborhood information as input.

For the ensemble method, it still achieves state-of-the-art performance under the striking-aware metrics. However, the improvement is limited compared to the original metrics. This is because the ensemble approach primarily enhances the model's predictive performance on events with low strikingness, whereas the striking-aware evaluation framework focuses on the model's ability to predict outstanding events. By restricting the improvements from low-strikingness events, the striking-aware evaluation allows researchers to more objectively assess progress in the field.

Dataset	Model	(W)MRR			(W)Hits@1			(W)Hits@3			(W)Hits@10		
		ORG ↑	SK ↑	Δ ↓	ORG ↑	SK ↑	Δ ↓	ORG ↑	SK ↑	Δ ↓	ORG ↑	SK ↑	Δ ↓
ICEWS14	ICL	-	-	-	32.40	15.95	50.77 %	45.94	26.18	43.01 %	56.59	36.59	35.34 %
	GenTKG	-	-	-	37.04	18.28	50.65 %	48.43	28.07	42.04 %	53.62	33.96	36.67 %
	Recurrency	37.12	19.47	47.55 %	29.69	13.62	54.13 %	40.75	21.49	47.26 %	51.26	30.75	40.01 %
	TITer	41.87	25.46	39.19 %	32.97	17.55	46.77 %	46.45	28.39	38.88 %	58.31	40.72	30.17 %
	TLogic	42.52	24.92	41.39 %	33.19	16.68	49.74 %	47.63	28.53	40.10 %	60.27	41.41	31.29 %
	RE-GCN	42.43	29.92	29.48 %	31.90	19.76	38.06 %	47.59	33.86	28.85 %	62.74	49.90	20.47 %
	TiRGN	44.45	30.43	31.54 %	33.77	20.30	39.89 %	49.57	34.09	31.23 %	64.89	50.69	21.88 %
	LogCL	48.84	38.17	**21.85 %**	37.76	26.88	**28.81 %**	54.60	43.23	**20.82 %**	70.43	60.78	**13.70 %**
	Ensemble	**51.35**	**38.67**	24.69 %	**40.23**	**27.18**	32.44 %	**57.60**	**44.07**	23.49 %	**72.56**	**61.49**	15.26 %
ICEWS18	ICL	-	-	-	19.27	9.42	51.12 %	31.35	17.87	43.0 %	43.97	28.58	35.00 %
	GenTKG	-	-	-	21.36	11.04	48.31 %	33.51	20.35	39.27 %	40.03	26.68	33.35 %
	Recurrency	28.66	15.97	44.28 %	20.77	10.22	50.79 %	32.25	18.05	44.03 %	43.54	26.83	38.38 %
	TITer	29.65	18.60	37.27 %	21.58	12.10	43.93 %	33.06	20.46	38.11 %	44.98	31.32	30.37 %
	TLogic	29.59	17.30	41.53 %	20.42	10.21	50.00 %	33.60	19.61	41.64 %	48.06	32.05	33.31 %
	RE-GCN	32.78	23.96	26.91 %	22.54	14.87	34.03 %	36.91	26.83	27.31 %	52.74	42.03	20.31 %
	TiRGN	33.54	23.59	29.67 %	22.92	14.21	38.00 %	38.09	26.67	29.98 %	54.38	42.25	22.31 %
	LogCL	35.43	28.35	**19.98 %**	24.09	**17.79**	**26.15 %**	40.22	32.11	**20.16 %**	58.04	49.66	**14.44 %**
	Ensemble	**37.74**	**28.49**	24.51 %	**26.19**	17.77	32.15 %	**42.84**	**32.31**	24.58 %	**60.70**	**50.15**	17.38 %
ICEWS0515	Recurrency	44.39	26.66	39.94 %	35.68	18.89	47.06 %	49.26	30.05	39.00 %	60.54	41.61	31.27 %
	TITer	48.03	31.39	34.65 %	38.61	22.43	41.91 %	53.03	34.85	34.28 %	65.42	48.81	25.39 %
	TLogic	46.56	30.65	34.17 %	35.48	20.50	42.22 %	53.27	35.65	33.08 %	67.25	50.69	24.62 %
	RE-GCN	47.93	33.73	29.63 %	37.32	23.39	37.33 %	53.78	38.33	28.73 %	68.16	54.23	20.44 %
	TiRGN	49.90	35.01	29.84 %	38.95	24.12	38.07 %	56.13	40.08	28.59 %	70.69	56.52	20.05 %
	LogCL	56.95	45.39	**20.30 %**	45.88	33.62	**26.72 %**	63.73	51.75	**18.80 %**	77.79	68.38	**12.10 %**
	Ensemble	**58.53**	**46.10**	21.24 %	**47.48**	**34.17**	28.03 %	**65.43**	**52.61**	19.59 %	**79.23**	**69.41**	12.39 %
GDELT	Recurrency	24.37	16.53	32.17 %	**16.43**	9.96	39.38 %	26.79	17.98	32.89 %	39.70	29.10	26.70 %
	TITer	20.17	13.16	34.75 %	14.23	8.18	42.52 %	21.98	14.06	36.03 %	30.67	22.07	28.04 %
	TLogic	19.77	12.44	37.08 %	12.23	6.53	46.61 %	21.67	13.43	38.02 %	35.62	25.01	29.79 %
	RE-GCN	19.73	14.39	27.07 %	12.50	7.85	37.20 %	20.96	15.21	27.43 %	33.89	27.10	20.04 %
	TiRGN	21.25	15.41	27.48 %	13.27	8.38	36.85 %	22.81	16.38	28.19 %	37.01	29.20	21.10 %
	LogCL	23.74	19.47	**17.99 %**	14.62	**10.79**	**26.20 %**	25.57	20.88	**18.34 %**	42.33	37.14	**12.26 %**
	Ensemble	**25.26**	**19.69**	22.05 %	15.60	10.71	31.35 %	**27.58**	**21.37**	22.52 %	**45.03**	**38.00**	15.61 %

Table 2: Performance comparison of original ('ORG') and striking-aware ('SK') evaluation, the higher value means better performance (↑). Δ represents the relative performance decrease across two evaluation settings, and the smaller value indicates that the model is less affected by repetitive bias (↓). The best results are bolded, and the second-best results are underlined.

4.4 Human Evaluation

To validate the effectiveness of the proposed RSMF in measuring event strikingness, we conducted a human evaluation study involving six volunteers. The six volunteers are all graduate students (Master's or Ph.D. candidates) in artificial intelligence, with research backgrounds spanning temporal knowledge graphs, knowledge graphs, relation extraction, and named entity recognition. Specifically, we randomly sampled 3000 events and a peer event for each target event, providing the contextual information for all events. The volunteers were asked to evaluate which event in each pair exhibited higher strikingness based on the given context.

H1	H2	H3	H4	H5	H6	Average
0.683	0.726	0.698	0.667	0.703	0.696	0.696

Table 3: Cohen's Kappa between humans and RSMF.

Table 3 reports the Cohen's Kappa coefficients between the evaluations of individual annotators and the proposed RSMF. Cohen's Kappa is a widely used measure for inter-rater agreement, with values ranging from -1 to 1, where -1 denotes "less than chance agreement" and 1 represents "almost perfect agreement." As shown in Table 3, the average Cohen's Kappa coefficient between RSMF and human evaluators is 0.696, indicating that RSMF achieves "substantial agreement" with human evaluators on the strikingness of events.

5 Conclusion

We observe that the current TKGR evaluation overweights trivial repetitive events, overshadowing models' ability to predict rare yet meaningful ones. To rectify this, we introduce a strikingness-aware evaluation framework that quantifies event strikingness through rule-based peer-event comparison and incorporates it as a dynamic weight into ranking metrics. Experiments on four benchmarks demonstrate: 1) a consistent performance drop as event strikingness increases, 2) a clear divide between path-based methods (strong on low-strikingness events) and representation-based methods (superior on high-strikingness events), and 3) we design a simple ensemble method and find it mainly improves prediction of repetitive events, with limited gains on rare, striking events. Our framework recenters evaluation on the forecasting of outstanding events, offering a rigorous and meaningful benchmark for TKGR, and calls for future work to prioritize reasoning beyond repetition.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62276110 and in part by the fund of Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL). The authors would also like to thank the anonymous reviewers for their comments on improving the quality of this paper.

References

[Angiulli *et al.*, 2009] Fabrizio Angiulli, Fabio Fassetti, and Luigi Palopoli. Detecting outlying properties of exceptional objects. *ACM Transactions on Database Systems*, 34(1):7:1–7:62, 2009.

[Aven, 2013] Terje Aven. On the meaning of a black swan in a risk context. *Safety science*, 57:44–51, 2013.

[Bordes *et al.*, 2013] Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *In Proceedings of the 27th Conference on Neural Information Processing Systems (NeurIPS)*, 2013.

[Chen *et al.*, 2024] Wei Chen, Huaiyu Wan, Yuting Wu, Shuyuan Zhao, Jiayaqi Cheng, Yuxin Li, and Youfang Lin. Local-global history-aware contrastive learning for temporal knowledge graph reasoning. In *40th IEEE International Conference on Data Engineering (ICDE)*, 2024.

[Dettmers *et al.*, 2018] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. In *Proceedings of the 32nd Conference on Artificial Intelligence, (AAAI)*, 2018.

[Dong *et al.*, 2023] Hao Dong, Zhiyuan Ning, Pengyang Wang, Ziyue Qiao, Pengfei Wang, Yuanchun Zhou, and Yanjie Fu. Adaptive path-memory network for temporal knowledge graph reasoning. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI)*, 2023.

[Gastinger *et al.*, 2023] Julia Gastinger, Timo Sztyler, Lokesh Sharma, Anett Schuelke, and Heiner Stuckenschmidt. Comparing apples and oranges? on the evaluation of methods for temporal knowledge graph forecasting. In *Machine Learning and Knowledge Discovery in Databases: Research Track - European Conference (ECML-PKDD)*, volume 14171 of *Lecture Notes in Computer Science*, pages 533–549. Springer, 2023.

[Gastinger *et al.*, 2024a] Julia Gastinger, Shenyang Huang, Michael Galkin, Erfan Loghmani, Ali Parviz, Farimah Poursafaei, Jacob Danovitch, Emanuele Rossi, Ioannis Koutis, Heiner Stuckenschmidt, Reihaneh Rabbany, and Guillaume Rabusseau. TGB 2.0: A benchmark for learning on temporal knowledge graphs and heterogeneous graphs. In *In Proceedings of 38th Conferenceon Neural Information Processing Systems (NeurIPS)*, 2024.

[Gastinger *et al.*, 2024b] Julia Gastinger, Christian Meilicke, Federico Errica, Timo Sztyler, Anett Schülke, and Heiner Stuckenschmidt. History repeats itself: A baseline for temporal knowledge graph forecasting. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

[Han *et al.*, 2021] Zhen Han, Peng Chen, Yunpu Ma, and Volker Tresp. Explainable subgraph reasoning for forecasting on temporal knowledge graphs. In *9th International Conference on Learning Representations (ICLR)*, 2021.

[Hassan *et al.*, 2014] Naeemul Hassan, Afroza Sultana, You Wu, Gensheng Zhang, Chengkai Li, Jun Yang, and Cong Yu. Data in, fact out: Automated monitoring of facts by factwatcher. *Proceedings of the VLDB Endowment*, 7(13):1557–1560, 2014.

[Huang *et al.*, 2024] Rikui Huang, Wei Wei, Xiaoye Qu, Shengzhe Zhang, Dangyang Chen, and Yu Cheng. Confidence is not timeless: Modeling temporal validity for rule-based temporal knowledge graph forecasting. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[Jain *et al.*, 2020] Prachi Jain, Sushant Rathi, Mausam, and Soumen Chakrabarti. Temporal knowledge base completion: New algorithms and evaluation protocols. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[Jin *et al.*, 2020] 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.

[Kervadec *et al.*, 2021] Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. Roses are red, violets are blue... but should VQA expect them to? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[Lee *et al.*, 2023] Dong-Ho Lee, Kian Ahrabian, Woojeong Jin, Fred Morstatter, and Jay Pujara. Temporal knowledge graph forecasting without knowledge using in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[Li *et al.*, 2021] 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 *The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2021.

[Li *et al.*, 2022] Yujia Li, Shiliang Sun, and Jing Zhao. Tirgn: Time-guided recurrent graph network with local-global historical patterns for temporal knowledge graph reasoning. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*, 2022.

[Li *et al.*, 2024] Haoxuan Li, Zhengmao Yang, Yunshan Ma, Yi Bin, Yang Yang, and Tat-Seng Chua. Mm-forecast: A multimodal approach to temporal event forecasting with large language models. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM)*, pages 2776–2785, 2024.

[Liang *et al.*, 2024] Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, Xinwang Liu, Fuchun Sun, and Kunlun He. A survey of knowledge graph reasoning on graph types: Static, dynamic, and multi-modal. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 46(12):9456–9478, 2024.

[Liao *et al.*, 2024] 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.

[Liu *et al.*, 2022] Yushan Liu, Yunpu Ma, Marcel Hildebrandt, Mitchell Joblin, and Volker Tresp. Tlogic: Temporal logical rules for explainable link forecasting on temporal knowledge graphs. In *36th Conference on Artificial Intelligence (AAAI)*, 2022.

[Liu *et al.*, 2023] Zhengtao Liu, Lei Tan, Mengfan Li, Yao Wan, Hai Jin, and Xuanhua Shi. Simfy: A simple yet effective approach for temporal knowledge graph reasoning. In *Findings of the Association for Computational Linguistics (EMNLP)*, 2023.

[Ma *et al.*, 2023] Yunshan Ma, Chenchen Ye, Zijian Wu, Xiang Wang, Yixin Cao, and Tat-Seng Chua. Context-aware event forecasting via graph disentanglement. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1643–1652, 2023.

[Meilicke *et al.*, 2021] Christian Meilicke, Patrick Betz, and Heiner Stuckenschmidt. Why a naive way to combine symbolic and latent knowledge base completion works surprisingly well. In *3rd Conference on Automated Knowledge Base Construction (AKBC)*, 2021.

[Ott *et al.*, 2023] Simon Ott, Patrick Betz, Daria Stepanova, Mohamed H. Gad-Elrab, Christian Meilicke, and Heiner Stuckenschmidt. Rule-based knowledge graph completion with canonical models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM)*, 2023.

[Ruffinelli *et al.*, 2020] Daniel Ruffinelli, Samuel Broscheit, and Rainer Gemulla. You CAN teach an old dog new tricks! on training knowledge graph embeddings. In *8th International Conference on Learning Representations (ICLR)*, 2020.

[Sun *et al.*, 2020] Zhiqing Sun, Shikhar Vashishth, Soumya Sanyal, Partha P. Talukdar, and Yiming Yang. A re-evaluation of knowledge graph completion methods. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[Sun *et al.*, 2021] 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 (EMNLP)*, 2021.

[Toutanova and Chen, 2015] Kristina Toutanova and Danqi Chen. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pages 57–66, 2015.

[Wang *et al.*, 2024] Jiapu Wang, Kai Sun, Linhao Luo, Wei Wei, Yongli Hu, Alan Wee-Chung Liew, Shirui Pan, and Baocai Yin. Large language models-guided dynamic adaptation for temporal knowledge graph reasoning. In *In Proceedings of 38th Conferenceon Neural Information Processing Systems (NeurIPS)*, 2024.

[Wu *et al.*, 2012] You Wu, Pankaj K. Agarwal, Chengkai Li, Jun Yang, and Cong Yu. On "one of the few" objects. In *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1487–1495, 2012.

[Xia *et al.*, 2024] Yuwei Xia, Ding Wang, Qiang Liu, Liang Wang, Shu Wu, and Xiaoyu Zhang. Chain-of-history reasoning for temporal knowledge graph forecasting. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 16144–16159. Association for Computational Linguistics, 2024.

[Xiao *et al.*, 2024] Hanhua Xiao, Yuchen Li, Yanhao Wang, Panagiotis Karras, Kyriakos Mouratidis, and Natalia Rozalia Avlona. How to avoid jumping to conclusions: Measuring the robustness of outstanding facts in knowledge graphs. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 3539–3550, 2024.

[Xu *et al.*, 2023] Yi Xu, Junjie Ou, Hui Xu, and Luoyi Fu. Temporal knowledge graph reasoning with historical contrastive learning. In *37th Conference on Artificial Intelligence (AAAI)*, 2023.

[Yang *et al.*, 2021] Yueji Yang, Yuchen Li, Panagiotis Karras, and Anthony K. H. Tung. Context-aware outstanding fact mining from knowledge graphs. In *The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 2006–2016, 2021.

[Zhang *et al.*, 2018] Gensheng Zhang, Damian Jimenez, and Chengkai Li. Maverick: Discovering exceptional facts from knowledge graphs. In *Proceedings of the 2018 International Conference on Management of Data (SIGMOD)*, pages 1317–1332, 2018.

[Zheng *et al.*, 2023] 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 (SIGIR)*, pages 1578–1588, 2023.

[Zhu *et al.*, 2021] Cunchao Zhu, Muhao Chen, Changjun Fan, Guangquan Cheng, and Yan Zhang. Learning from history: Modeling temporal knowledge graphs with sequential copy-generation networks. In *35th Conference on Artificial Intelligence (AAAI)*, 2021.