LLM-DR: A Novel LLM-Aided Diffusion Model for Rule Generation on Temporal Knowledge Graphs

Kai Chen[1*], Xin Song[1*], Ye Wang[1†], Liqun Gao[1], Aiping Li[1†], Xiaojuan Zhao[2], Bin Zhou[1], Yalong Xie[1]

[1]College of Computer Science and Technology, National University of Defense Technology, Changsha, China.
[2]Information School, Hunan University of Humanities, Science and Technology, Loudi, China.
{chenkai_, songxin, ye.wang, gaoliqun15, liaiping, zhaoxiaojuan18, binzhou}@nudt.edu.cn,
xieyalong@alumni.nudt.edu.cn

Abstract

Among various temporal knowledge graph (TKG) extrapolation methods, rule-based approaches stand out for their explicit rules and transparent reasoning paths. However, the vast search space for rule extraction poses a challenge in identifying high-quality logic rules. To navigate this challenge, we explore the use of generation models to generate new rules, thereby enriching our rule base and enhancing our reasoning capabilities. In this paper, we introduce LLM-DR, an innovative rule-based method for TKG extrapolation, which harnesses diffusion models to generate rules that are consistent with the distribution of the source data, while also amalgamating the rich semantic insights of Large Language Models (LLMs). Specifically, our LLM-DR generates semantically relevant and high-quality rules, employing conditional diffusion models in a classifier-free guidance fashion and refining them with LLM-based constraints. To assess rule efficacy, we meticulously design a coarse-to-fine evaluation strategy that initiates with coarse-grained filtering to eliminate less plausible rules and proceeds with fine-grained scoring to quantify the reliability of the retained. Extensive experiments demonstrate the promising capacity of our LLM-DR.

Introduction

Knowledge graphs (KGs) are essential for structuring and reasoning through complex information (Song et al. 2021; Zhao et al. 2021), serving as a key artificial intelligence (AI) tool. Nowadays, the advent of temporal knowledge graphs (TKGs) has introduced a dynamic element, allowing for the representation of knowledge that changes over time (Trivedi et al. 2017; Goel et al. 2020). By integrating time-varying relations, TKGs transcend the limitations of static KGs, offering a more fluid and temporally nuanced perspective on knowledge representation. This development has sparked interest in TKG extrapolation, a field that focuses on predicting future events based on historical knowledge. By harnessing historical data and trends, TKG extrapolation enables the anticipation of knowledge evolution, thus serving as a powerful instrument for foresight in various domains.

Currently, a multitude of TKG extrapolation methods have been proposed (Jin et al. 2020; Sun et al. 2021; Li et al. 2021), with rule-based methods (Liu et al. 2022; Bai et al. 2023) distinguishing themselves due to their high level of interpretability. By extracting rules from observable historical data, rule-based methods can leverage these patterns to guide reasoning about future, unobservable events (Li et al. 2023). These methods excel in offering transparent reasoning paths (Liu et al. 2022), which is particularly valuable for decision-makers seeking to understand the logic behind predictions. However, the search space for rule extraction is exponentially vast, which complicates the identification of high-quality logic rules (Zhang et al. 2021; Liu et al. 2023).

To this end, recent research (Luo et al. 2023; Wang et al. 2024) has endeavored to capitalize on the strengths of Large Language Models (LLMs) to extend the limited set of extracted rules, seeking to enrich the rule base and broaden reasoning capabilities. Nevertheless, the generation process of LLMs requires meticulously crafted prompts (Zhou et al. 2023), and even slight alterations can lead to significant variability in the generated outcomes, introducing a high degree of uncertainty. Furthermore, the well-documented issue of hallucinations (Ji et al. 2023) in LLMs casts doubt on the stability of the rule quality when using these models for generation. To ensure controllable and reliable rule generation, we turn to a prominent class of probabilistic generative models known as diffusion models. Diffusion models offer a structured approach to data generation by learning the forward process of gradually adding noise to data until it resembles a Gaussian distribution, and then learning to reverse this process to generate new samples (Ho, Jain, and Abbeel 2020). A notable advantage of this method is that the generated rules may be more interpretable, as their creation is based on an understanding of the data distribution rather than being a direct output from a black-box model like an LLM (Pacchiardi et al. 2024). This approach aids in mitigating the risks associated with rules generated by LLMs, providing a more predictable and interpretable framework for rule generation.

On the other hand, considering that the rules obtained may be rife with spurious rules (Yu and Jin 2000; Hahsler and Hornik 2007), it is imperative to rigorously evaluate these rules before utilizing them for reasoning. A prime example from the real-world TKG, ICEWS (Lautenschlager,

[*] Equal contributions.

[†] Corresponding authors.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Shellman, and Ward 2015), is the frequently occurring rule expressed as *Criticize_or_denounce*$(X_0, X_2, T_2)\leftarrow$ *Investigate*$(X_0, X_1, T_0) \wedge$ *Appeal_for_economic_aid*(X_1, X_2, T_1). This rule example, extracted by rule-based method (Liu et al. 2022), is clearly illogical and does not offer a convincing inferential trajectory for TKG extrapolation. Transitioning from the problem, we turn our attention to the solution: Despite the known issue of hallucinations associated with LLMs, they nonetheless possess a substantial capacity to serve as competent evaluators. Recently, some researchers (Zheng et al. 2023) have proved that strong LLM judges like GPT-4 can match both controlled and crowdsourced human preferences well, achieving over 80% agreement, the same level of agreement between humans. Motivated by these insights, we consider the integration of LLMs into our design, for the purpose of rule evaluation.

In this paper, we introduce a novel LLM-aided Diffusion model for Rule generation (**LLM-DR**), which harnesses diffusion models to broaden the rule set for TKG extrapolation, while amalgamating the rich semantic insights of LLMs. Specifically, we employ conditional diffusion models in a classifier-free guidance fashion (Ho and Salimans 2022) to generate new rules that align with the data distribution of the original TKG-extracted rules. Additionally, we leverage LLMs to impose constraints on the generation process, ensuring the semantic relevance and high quality of the generated rules. To assess rule efficacy, we develop a coarse-to-fine evaluation strategy. At the coarse-grained level, we harness LLMs to sift through and discard spurious rules, while at the fine-grained level, confidence scores to quantify rule reliability. This dual-tiered evaluation strategy ensures the selection of the most robust and credible rules for TKG extrapolation, enhancing the overall performance and explainability of the reasoning process. Through rigorous experimentation, we demonstrate that LLM-DR outperforms existing state-of-the-art (SOTA) methods in TKG extrapolation, with ablation studies and case studies substantiating the efficacy of each component of our design.

We summarize our main contributions as follows:

1. We present LLM-DR, an innovative rule-based method for TKG extrapolation that harnesses diffusion models to broaden the TKG-extracted rule set, while amalgamating the rich semantic insights of LLMs.
2. We attain controllable rule generation using diffusion models in a classifier-free guidance manner, enhanced by LLM-based constraints to refine the rule quality.
3. We design a coarse-to-fine evaluation strategy to assess rule efficacy and enhance reasoning accuracy.
4. Extensive experiments demonstrate that LLM-DR surpasses SOTA TKG extrapolation methods, with ablation studies and case studies substantiating the efficacy of each component of our design.

Related Work

TKG Extrapolation Methods

Recently, a significant number of methods have been proposed (Han et al. 2021b; Zhu et al. 2021; Liang et al. 2023) for TKG extrapolation. These methods, such as RE-NET (Jin et al. 2020), RE-GCN (Li et al. 2021), HiSMatch (Li et al. 2022), TiRGN (Li, Sun, and Zhao 2022), and LogCL (Chen et al. 2024), adeptly harness neural networks and embeddings to capture the nuanced temporal dynamics within TKGs. Concurrently, there has been a development of methods aimed at bolstering the reasoning explainability. TITer (Sun et al. 2021) delves into potential reasoning paths by employing reinforcement learning, while xERTE (Han et al. 2021a) sheds light on the reasoning process and outcomes by extracting subgraphs and tracing the underlying processes. Rule-based methods are currently in the limelight: TLogic (Liu et al. 2022) and TR-Rules (Li et al. 2023) can automate the learning of rules from data, and deliver explicit reasoning paths. Additionally, there is a burgeoning interest in leveraging the strengths of the currently popular LLMs for reasoning on TKGs (Gao et al. 2024; Luo et al. 2024).

Diffusion Models

Diffusion models (Ho, Jain, and Abbeel 2020) have risen to prominence in the realm of generative modeling, significantly influencing the landscape of sample generation in diverse domains. These models operate by incrementally adding noise to data and then learning to reverse this process for denoising and creating new, coherent samples. Building on this, conditional diffusion models (Tashiro et al. 2021) offer refined control over generation, facilitating tailored applications in visual generation (Saharia et al. 2022), natural language processing (Gong et al. 2023), and multimodal generation (Nichol et al. 2022). Within this framework, classifier guidance (Dhariwal and Nichol 2021) introduces a structured directive via a classifier, while classifier-free guidance (Ho and Salimans 2022) champions an open-ended, flexible approach to generative modeling.

Preliminaries

Task Definition

A TKG comprises millions of temporal facts, represented as $\mathcal{G} \subset \mathcal{E} \times \mathcal{R} \times \mathcal{E} \times \mathcal{T}$. In this context, $\mathcal{E}$ denotes the set of entities, $\mathcal{R}$ the set of relations, and $\mathcal{T}$ the set of timestamps. Each temporal fact within $\mathcal{G}$ is structured as a quadruple (e_s, r, e_o, τ), indicating that a relation $r \in \mathcal{R}$ exists between a subject entity $e_s \in \mathcal{E}$ and an object entity $e_o \in \mathcal{E}$ at a specific time τ. Furthermore, we incorporate inverse relations to enrich temporal facts, yielding (e_o, r^{-1}, e_s, τ) for every quadruple. Consider $\mathcal{O} = \{(e_s, r, e_o, \tau) | \tau \in [\tau_b, \tau_e]\}$ as the collection of observed facts within our accessible time interval $[\tau_b, \tau_e]$. For a given query $\boldsymbol{q} = (e_q, r_q, ?, \tau_q)$, our TKG extrapolation reasoning task is to deduce the missing object entity based on the provided elements. A key requirement for this task is $\tau_q > \tau_e$, which signifies the need for extrapolation to predict future events based on historical data.

Temporal Logical Rule

Definition 1. *We define each temporal logical rule as a first-order Horn clause, consisting of a body of conjunctive rela-*

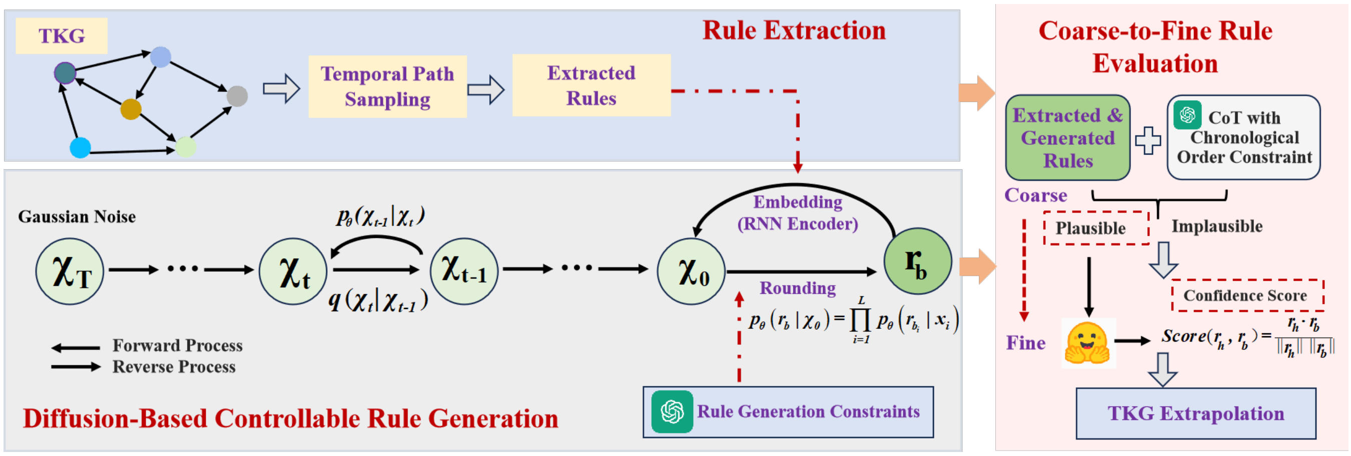

TKG
Temporal Path Sampling
Extracted Rules
Rule Extraction
Gaussian Noise
χ_T
$p_\theta(\chi_{t-1}|\chi_t)$
χ_t
χ_{t-1}
$q(\chi_t|\chi_{t-1})$
χ_0
Embedding (RNN Encoder)
r_b
Rounding
$p_\theta(r_b|\chi_0)=\prod_{i=1}^{L} p_\theta(r_{b_i}|x_i)$
Forward Process
Reverse Process
Diffusion-Based Controllable Rule Generation
Rule Generation Constraints
Coarse-to-Fine Rule Evaluation
Extracted & Generated Rules
CoT with Chronological Order Constraint
Coarse
Plausible
Implausible
Confidence Score
Fine
$Score(r_h, r_b)=\frac{r_h \cdot r_b}{\|r_h\| \|r_b\|}$
TKG Extrapolation

Figure 1: The overview of the LLM-DR framework.

tions and a single head:

$$r_h(e_1, e_{L+1}, \tau_h) \leftarrow \bigwedge_{i=1}^{L} r_{b_i}(e_i, e_{i+1}, \tau_i), \qquad (1)$$

with a ***Chronological Order Constraint*** *met:*

$$\tau_h > \tau_1 \geq \tau_2 \geq \ldots \geq \tau_L. \qquad (2)$$

The left side in Equation 1 is referred to as the **rule head**, which for convenience and brevity, can be denoted as r_h; while the right side constitutes the **rule body**, denoted as $r_b = \bigwedge_{i=1}^{L} r_{b_i}$. For a temporal logical rule, the subject e_1 and the object e_{L+1} serve as the origin and the terminus of the chain-like rule body, respectively, thereby creating a cyclic definition of the logical rule. The **chronological order constraint** (Equation 2) necessitates that all timestamps within the rule body must: 1) be ordered chronologically, and 2) precede the timestamp of the rule head.

Diffusion Models for Generation

The diffusion model (Ho, Jain, and Abbeel 2020) constitutes a probabilistic generative architecture designed to approximate the distribution of target data for the synthesis of samples. This model is fundamentally composed of two complementary processes: a forward process and a reverse process. In the forward process, an initial sample χ_0 drawn from a distribution $q(\chi)$ undergoes a sequential transformation. This transformation involves the incremental introduction of Gaussian noise across a series of time steps, indexed by t ranging from 1 to T. This results in a Markov chain of latent variables, $\chi_1, ..., \chi_T$, where each step t follows a transition probability $q(\chi_t|\chi_{t-1})$. Ultimately, the initial data X_0 is distorted into a standard Gaussian distribution, $\chi_T \sim \mathcal{N}(0, 1)$, at the final time step T. Conversely, the reverse process aims to reconstruct the original data from its noisy corruption. This is achieved by training the diffusion model to predict the preceding state in the sequence, governed by the learnable distribution $p_\theta(\chi_{t-1}|\chi_t)$. Essentially, the reverse process offers precise control over the generative mechanism, enabling the model to discern and replicate the nuanced dependencies and underlying patterns inherent within the data distributions.

Methodology

In this section, we provide a detailed description of our LLM-DR method, as illustrated in Figure 1.

Rule Extraction

Initially, extracting a finite set of rules from TKGs is crucial as it lays the groundwork for further operations. Our methodology for rule extraction is inspired by the approach presented in (Liu et al. 2022), with a particular emphasis on two pivotal stages: the temporal path sampling and the alignment of these paths with temporal logic rules.

Temporal Path Sampling We initially employ a sampling strategy for temporal paths, achieved through initiating a random walk (Kendall 1967; Cheng, Ahmed, and Sun 2023) from a chosen temporal edge (fact), serving as the anchor for the rule's head. Starting with a rule head $r_h(e_1, e_{L+1}, \tau_h)$, we then embark on a journey from node (entity) e_1, sampling a sequence of temporal edges, denoted by $\mathcal{P}_L$, with a predefined length L.

Alignment with Temporal Logic Criteria Post the sampling of temporal paths, we engage in a process of alignment, matching these paths against the criteria of the predefined temporal logic rule (as outlined in Definition 1). This involves assessing the suitability of the sampled paths to form the body of a temporal logic rule. Specifically, we verify if the terminal node of the sampled path $\mathcal{P}_L$ aligns with the anticipated entity e_{L+1}. Beyond mere endpoint congruence, it is imperative that the temporal sequence adheres to the chronological order constraint (Equation 2), ensuring a logical and coherent progression.

Diffusion-Based Controllable Rule Generation

Building upon the finite collection of rules derived from TKGs, we then employ a diffusion model with classifier-free guidance for controllable rule generation, enriching this foundational rule base. Concretely, our rule generation primarily encompasses a forward process and a reverse process.

Forward Process In the forward process, we begin by extracting the embedding of each rule body and subsequently convert the embedding into a Gaussian distribution by strategically injecting noise. The initial state $\boldsymbol{\chi}_0$ is set as the embedding of a L-length rule body r_b, achieved by employing an RNN-based encoder which excels at modeling sequential data; while the rule head r_h is regarded as the category/label of the rule. Then, the model constructs the Markov chain $\boldsymbol{\chi}_{1:T}$ by incrementally introducing Gaussian noise across T steps. The transition from $\boldsymbol{\chi}_{t-1}$ to $\boldsymbol{\chi}_t$ is parameterized as:

$$q(\boldsymbol{\chi}_t|\boldsymbol{\chi}_{t-1}) = \mathcal{N}(\boldsymbol{\chi}_t; \sqrt{1-\beta_t}\boldsymbol{\chi}_{t-1}, \beta_t \boldsymbol{I}), \tag{3}$$

where $t \in 1, \cdots, T$ indexes the diffusion step, $\mathcal{N}$ denotes the Gaussian distribution, and $\beta_t \in (0,1)$ governs the noise scale at step t. As $T \to \infty$, $\boldsymbol{\chi}_T$ approaches a standard Gaussian distribution. Leveraging the reparameterization trick and the additivity of independent Gaussian noises, we derive $\boldsymbol{\chi}_t$ directly from the initial state $\boldsymbol{\chi}_0$. Formally:

$$q(\boldsymbol{\chi}_t|\boldsymbol{\chi}_0) = \mathcal{N}(\boldsymbol{\chi}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{\chi}_0, (1-\bar{\alpha}_t)\boldsymbol{I}), \tag{4}$$

where $\bar{\alpha}_t = \prod_{t'=1}^{t}(1-\beta_{t'})$. And $\boldsymbol{\chi}_t$ is reparameterized as:

$$\boldsymbol{\chi}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{\chi}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I}). \tag{5}$$

To regulate noise addition in $\boldsymbol{\chi}_{1:T}$, a linear noise scheduler is incorporated, utilizing three hyperparameters: s, α_{low}, and α_{up}. The scheduler is defined by:

$$1-\bar{\alpha}_t = s \cdot \left[\alpha_{low} + \frac{t-1}{T-1}(\alpha_{up} - \alpha_{low})\right], \ t \in \{1, \cdots, T\}. \tag{6}$$

Here, $s \in [0,1]$ modulates the noise scale, while $\alpha_{low} < \alpha_{up} \in (0,1)$ establish the noise bounds.

Reverse Process Following the completion of the forward process, the inverse process is engaged to meticulously reconstruct the rule body representation $\boldsymbol{\chi}_0$ from a pure Gaussian noise $\boldsymbol{\chi}_T$. The denoising transition is characterized by:

$$p_\theta(\boldsymbol{\chi}_{t-1}|\boldsymbol{\chi}_t) = \mathcal{N}(\boldsymbol{\chi}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{\chi}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{\chi}_t, t)). \tag{7}$$

The neural networks $\mathcal{N}$, parameterized by θ, are employed to calculate the mean $\boldsymbol{\mu}_\theta(\boldsymbol{\chi}_t, t)$ and covariance $\boldsymbol{\Sigma}_\theta(\boldsymbol{\chi}_t, t)$ of the resulting Gaussian distribution. To achieve a conditional generation, we adopt the classifier-free guidance (Ho and Salimans 2022) approach, and the initial mean $\boldsymbol{\mu}_\theta(\boldsymbol{\chi}_t, t)$ will be adjusted to:

$$\tilde{\boldsymbol{\mu}}_\theta(\boldsymbol{\chi}_t, t) = \boldsymbol{\mu}_\theta(\boldsymbol{\chi}_t, t) + w\left(\boldsymbol{\mu}_\theta(\boldsymbol{\chi}_t, t, r_h) - \boldsymbol{\mu}_\theta(\boldsymbol{\chi}_t, t)\right), \tag{8}$$

where w is the guidance weight parameter that controls the balance between the conditional and unconditional generations. Thus, we get an adjusted denoising transition:

$$p_\theta(\boldsymbol{\chi}_{t-1}|\boldsymbol{\chi}_t, r_h) = \mathcal{N}(\boldsymbol{\chi}_{t-1}; \tilde{\boldsymbol{\mu}}_\theta(\boldsymbol{\chi}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{\chi}_t, t)). \tag{9}$$

To obtain the specific members of the rule body, we take a *rounding* step $p_\theta\left(r_b|\boldsymbol{\chi}_0\right) = \prod_{i=1}^{L} p_\theta\left(r_{b_i}|x_i\right)$, where $p_\theta\left(r_{b_i}|x_i\right)$ is a softmax distribution, to map the final generated rule body representations to discrete relations r_{b_i}.

LLM-based Rule Generation Constraints By leveraging the advanced language comprehension capabilities of LLMs, we strategically apply constraints to the rule generation process, ensuring that the potential relations within the rule body are relevant and of high quality. The advantage of this approach is that it prevents the diffusion models from producing rule bodies that are unrelated to the rule head, thus maintaining coherence and enhancing the overall quality of the generated rules. For example, given a rule head such as *Consult*, the relations *Engage in negotiation* and *Express intent to meet or negotiate* are semantically more plausible candidates for inclusion in the rule body than a relation like *Forgive*. Specifically, given a rule head r_h, we select the k most semantically relevant relations from the relation set $\mathcal{R}$ based on their semantic similarity. This strategy ensures a higher quality of rule generation by focusing on relevant and meaningful relations. A simple example of the relation selection prompt is shown in the following prompt box.

Prompt for Rule Generation Constraints

Task: Given a relation $\{r_h\}$ from the ICEWS datasets, Please select $\{k\}$ relations from the Candidate relation set, which are semantically close or have progressive relations with the given relation $\{r_h\}$.
Candidate relation set: Relation set
Background knowledge: These relations are derived from the ICEWS datasets and serve to characterize a spectrum of interactions and social events among nations and political figures. If possible, combine your knowledge of social-political events to assist in the above analysis.

Please provide the selection result in list form.
Selection results:

It is well-known that LLMs often exhibit inconsistencies in their responses (Elazar et al. 2021; Wang et al. 2023b). Therefore, we conduct multiple queries and then collect responses. Further, we design a voting prompt to allow LLM to vote on multiple responses, selecting the most confident answer as the final response. We view them as the constraints on the rule body during the rule generation process.

Optimization To optimize our generation model, we aim to maximize the Evidence Lower Bound (ELBO) of the likelihood of the rule body representation $\boldsymbol{\chi}_0$. We follow (Wang et al. 2023a; Jiang et al. 2024) and get the final optimization objective:

$$\mathcal{L}(\boldsymbol{\chi}_0, \theta, r_h) = \mathbb{E}_{t \sim \mathcal{U}(1,T)}\mathcal{L}_t, \tag{10}$$

with a uniform distribution $t \sim \mathcal{U}(1,T)$. And $\mathcal{L}_t$ is defined as follows:

$$\mathcal{L}_t = \mathbb{E}_{q(\boldsymbol{\chi}_t|\boldsymbol{\chi}_0)}\left[\frac{1}{2}\left(\frac{\bar{\alpha}_{t-1}}{1-\bar{\alpha}_{t-1}} - \frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}\right)||\hat{\boldsymbol{\chi}}_\theta(\boldsymbol{\chi}_t, t, r_h) - \boldsymbol{\chi}_0||_2^2\right], \tag{11}$$

where $\hat{\boldsymbol{\chi}}_\theta(\boldsymbol{\chi}_t, t)$ is the predicted $\boldsymbol{\chi}_0$ based on $\boldsymbol{\chi}_t$, t and r_h, and we utilize neural networks, specifically employing a Multi-Layer Perceptron (MLP), to implement calculations.

Coarse-to-Fine Rule Evaluation

Upon obtaining the rules, our focus shifts to evaluate rule plausibility and calculate rule scores from both qualitative and quantitative perspectives. First, from a coarse-grained perspective, we apply an LLM with chain-of-thought (CoT) to evaluate the logical plausibility of rules. It primarily evaluates the logical and semantic validity of rules, sorting them into plausible or implausible categories. Subsequently, advancing to a fine-grained perspective, we utilize a pre-trained BERT (Devlin et al. 2019) model to meticulously compute the semantic proximity between the rule head and the rule body for those rules identified as plausible.

Coarse-Grained Rule Evaluation (Based on LLMs) After the controllable rule generation module, we can obtain a set of rules for a target relation r_h. Existing methods typically use statistical scores to filter out rules. However, these approaches have two drawbacks: (1) Due to the sparsity of TKGs, some generated rules may have low or even zero statistical scores, yet they are logically and semantically reasonable; (2) Due to potential noise in the TKG, some rules with high statistical scores may be logically implausible, potentially leading to the spurious rules. Recently, extensive works(Wei et al. 2022; Huang and Chang 2023) have explored the ability of LLMs to spontaneously decompose the complex multi-step problem into intermediate reasoning steps through CoT prompting. It is elicited by a simple prompt like “Let’s think step by step” or well-designed demonstrations with human-annotated rationales.

Motivated by this, we leverage an LLM to evaluate the plausibility of rules through step-by-step reasoning. Please note that we execute the step-by-step process on the rule body based on the Chronological Order Constraint (see Definition 1). Moreover, we assess the plausibility of rules by logical and semantic reasoning, compensating for the shortcomings of purely statistical methods when sufficient statistical support is lacking. A simple prompt example is shown in the following prompt box.

Prompt for Coarse Rule evaluation with CoT

Task: Given the following rules derived from the ICEWS dataset. Please evaluate their logical plausibility based on commonsense knowledge in realistic scenarios and analyze step-by-step followed by the temporal order whether the rule body logically can lead to the rule head.
Rules: {Generated rules }
Background knowledge: ICEWS (Integrated Crisis Early Warning System) is a large time-series social event dataset focused on international relations and political events, encompassing a range of activities from politics, economics to military events.

You can provide a final assessment of whether the given rule is logically plausible based on the analyzed steps, Return the final evaluation results as plausible or implausible and also explain why.
Predicate:

Fine-Grained Rule Scoring (Based on BERT) After the coarse-grained filtration process, our objective is to meticulously ascertain the fine confidence scores for plausible rules, conducting evaluation from a semantic standpoint. Although LLMs with a CoT prompt can employ specific prompts to generate confidence scores or use answer consistency as a confidence indicator, these can have poor calibration performance or cannot provide fine-grained confidence scores (Xiong et al. 2023; Tian et al. 2023).

Therefore, we consider building a rule-scoring function based on representation learning, rather than relying solely on statistical rule instances or LLM with prompts. Specifically, we leverage a pre-trained small language model such as the BERT model(Devlin et al. 2019), to embed the rule head r_h and its corresponding rule bodies r_b into the vector space, obtaining the embedding vectors $\mathbf{r_h}$ and $\mathbf{r_b}$:

$$\mathbf{r}_h, \mathbf{r_b} = \text{BERT}(r_h, r_b) \tag{12}$$

Then, we calculates the semantic similarity score Score(r_h, r_b) through the cosine similarity function:

$$Score(r_h, r_b) = \frac{\mathbf{r_h} \cdot \mathbf{r_b}}{\|\mathbf{r_h}\| \|\mathbf{r_b}\|} \tag{13}$$

where $\cdot$ denotes the dot product operation, and $\|$ denotes the norm of the vector. For each query, we can leverage the obtained temporal logical rules and their confidence scores to provide suitable answers.

Experiments

Experimental Setup

Benchmark Datasets Four TKG benchmark datasets are leveraged to evaluate our LLM-DR, including ICEWS14 (García-Durán, Dumancic, and Niepert 2018), ICEWS18 (Jin et al. 2020), ICEWS05-15 (García-Durán, Dumancic, and Niepert 2018), and GDELT (Jin et al. 2020).

Evaluation Protocol The evaluation for TKG extrapolation involves the adoption of a link prediction task. This task focuses on inferring incomplete time-wise facts that contain a missing entity, represented as either $(e_s, r, ?, t)$ or $(?, r, e_o, t)$. We use the ground truths for extrapolation, as is the case with many previous methods (Jin et al. 2020; Li et al. 2021). And we use the time-wise filtered setting (Goel et al. 2020) to report the experimental results. The performance is reported on the standard evaluation metrics: the proportion of correct triples ranked in the top 1, 3 and 10 (Hits@1, Hits@3, and Hits@10), and Mean Reciprocal Rank (MRR). All the metrics are the higher the better. For all experiments, we report averaged results across 5 runs, and we omit the variance as it is generally low. Across various modules and ablation variants of our model, we consistently employ the GPT-3.5-turbo version as the LLM.

Baselines We compare with twelve up-to-date TKG extrapolation baseline methods, encompassing a diverse range of approaches. Our comparison includes **embedding-based** methods such as CyGNet (Zhu et al. 2021), TITer (Sun et al. 2021), RE-GCN (Li et al. 2021), xERTE (Han et al. 2021a), TiRGN (Li, Sun, and Zhao 2022), HiSMatch (Li et al. 2022),

	Model	ICEWS14				ICEWS18				ICEWS05-15				GDELT			
		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
♠	CyGNet [2021]	35.1	25.7	39.0	53.6	24.9	15.9	28.3	42.6	36.8	26.6	41.6	56.2	18.5	11.5	19.6	32.0
	TITer [2021]	40.9	32.1	45.5	57.6	30.0	22.1	33.5	44.8	47.7	38.0	52.9	65.8	20.2	14.1	22.2	31.2
	RE-GCN [2021]	40.4	30.7	45.0	59.2	32.6	22.4	36.8	52.7	48.0	37.3	53.9	68.3	19.8	12.5	21.0	34.0
	xERTE [2021a]	40.0	32.1	44.6	56.2	30.0	22.1	33.5	44.8	46.6	37.8	52.3	63.9	18.9	12.3	20.1	30.3
	TiRGN [2022]	44.0	33.8	49.0	63.8	33.7	23.2	38.0	54.2	50.0	39.3	56.1	70.7	21.7	13.6	23.3	37.6
	HiSMatch [2022]	46.4	35.9	51.6	66.8	34.0	23.9	37.9	53.9	52.9	42.0	59.1	73.3	22.0	14.5	23.8	36.6
	RPC [2023]	-	-	-	-	34.9	24.3	38.7	55.9	51.4	39.9	57.0	71.8	22.4	14.4	24.4	38.3
	LogCL [2024]	48.9	37.8	54.7	**70.3**	35.7	24.5	40.3	**57.7**	57.0	46.1	63.7	**77.9**	23.8	14.6	25.6	42.3
♣	TLogic [2022]	42.5	33.2	47.6	60.3	29.6	20.4	33.6	48.1	47.0	36.2	53.1	67.4	19.8	12.2	21.7	35.6
	TR-Rules [2023]	43.3	34.0	48.6	61.2	30.4	21.1	34.6	48.9	47.6	37.1	53.8	67.6	-	-	-	-
◊	Llama-2-7b-CoH [2024]	-	34.9	47.0	59.1	-	22.3	36.3	52.2	-	38.6	54.1	69.9	-	-	-	-
	Vicuna-7b-CoH [2024]	-	32.8	45.7	65.6	-	20.9	34.7	53.6	-	39.2	54.6	70.7	-	-	-	-
	LLM-DR (ours)	**50.5**	**40.6**	**55.8**	67.0	**38.2**	**30.4**	**40.9**	55.6	**58.9**	**50.5**	**64.8**	75.3	**30.7**	**22.5**	**33.4**	**42.6**

Table 1: Performance (in percentage) for link prediction on four benchmarks with time-aware metrics, where "H@" represents "Hits@". ♠ denotes embedding-based baselines, ♣ denotes rule-based baselines, and ◊ denotes LLM-based baselines.

Model	ICEWS14				ICEWS18			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
LLM-DR	**50.5**	**40.6**	**55.8**	**67.0**	**38.2**	**30.4**	**40.9**	**55.6**
-Cons	49.4	39.7	55.2	66.5	36.7	29.0	39.8	54.1
-Diff	43.8	34.3	49.5	62.9	31.2	22.3	35.1	49.8
Diff ⇒ LLM	46.6	36.3	52.2	66.6	35.7	25.5	39.3	54.9
-Coarse	48.8	39.1	54.9	66.5	36.6	29.6	39.2	54.7
BERT ⇒ LLM	46.2	35.8	52.6	65.1	34.8	26.1	37.6	52.9

Table 2: Ablation study results on ICEWS14 and ICEWS18.

RPC (Liang et al. 2023), and LogCL (Chen et al. 2024). Additionally, we have evaluated **rule-based** methods like TLogic (Liu et al. 2022) and TR-Rules (Li et al. 2023), as well as **LLM-based** methods including Llama-2-7b-CoH (Luo et al. 2024) and Vicuna-7b-CoH (Luo et al. 2024).

Performance Comparison

The experimental results obtained from four different datasets are presented in Table 1. The datasets employed in our evaluation exhibit significant differences in scale, number of entities, and number of relations. Our findings demonstrate the effectiveness of our proposed LLM-DR in performing efficient TKG extrapolation of varying sizes and complexity levels. One notable observation is that our LLM-DR outperforms all baselines across all four datasets, underscoring its superior reasoning capabilities and affirming its potential for complex TKG reasoning tasks.

Furthermore, an intriguing and candid observation is that not solely LLM-DR, but also other rule-based methods such as TLogic and TR-Rules, demonstrate comparatively weaker performances in terms of the Hits@10 metric. This mirrors a common challenge among rule-based methods: while they excel at identifying the best answer in reasoning tasks, they fall short compared to embedding-based methods when the requirement shifts to generating a broader spectrum of potential candidate answers. And the marked improvement of LLM-DR over TLogic and TR-Rules signifies that our method can mitigate this particular shortcoming.

Ablation Study

Table 2 presents the ablation studies on ICEWS14 and ICEWS18 datasets, to isolate and quantify the contribution of each component to the reasoning performance.

Rule Generation In addition to the original LLM-DR model, we introduce three sub-models for comparative analysis within the rule generation phase, including (1) a version of LLM-DR devoid of the generation constraints, denoted as "-Cons", (2) a variant lacking rules produced by the diffusion models, denoted as "-Diff", and (3) an adaptation that employs LLMs in lieu of diffusion models for rule generation, denoted as "Diff ⇒ LLM". Performance comparisons among sub-models reveal that the generation constraints exert a certain degree of influence on the outcomes, while the rules generated by diffusion models significantly enhance the effectiveness. Furthermore, there is a noticeable decline in performance when the rule generator is switched from diffusion models to LLMs, underscoring the advantage of utilizing diffusion models for rule generation in our approach.

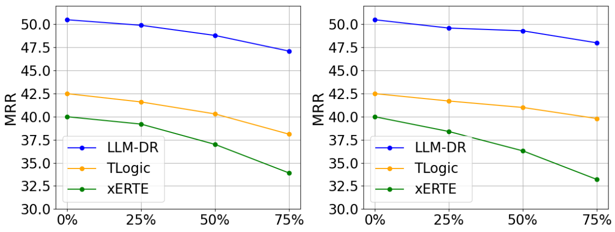
Rule Evaluation Within the rule evaluation phase, we also introduce two sub-models, including (4) a version of LLM-DR devoid of the coarse-grained module, denoted as "-Coarse", and (5) a variant replacing BERT with LLMs within the fine-grained evaluation module, denoted as "BERT ⇒ LLM". The results indicate that the coarse-grained module has a measurable impact on the inference performance. As for the fine-grained evaluation module, substituting BERT with LLMs leads to a significant performance degradation, which suggests that LLMs may have disadvantages when it comes to the nuanced quantification required at a finer level of evaluation detail.

Case Study

Table 3 presents the case studies conducted to elucidate the benefits of our LLM-DR in rule generation and evaluation. Our investigations have yielded the following insights: a) In both Case 1 and Case 2, our LLM-DR adeptly generates novel rules from existing extracted ones, demonstrating our method's capacity to transcend mere reliance on data,

Case	Rule Head	Type	Rule Body	Evaluation	
				Coarse	Fine
1	$Make\_pessimistic\_comment(X_0, X_2, \tau_2)$	**E**	$Praise\_or\_endorse(X_0, X_1, \tau_0) \wedge Veto^{-1}(X_1, X_2, \tau_1)$	Plausible	0.724
		G	$Make\_a\_visit(X_0, X_1, \tau_0) \wedge Reject^{-1}(X_1, X_2, \tau_1)$	Plausible	0.726
2	$Accuse(X_0, X_3, \tau_3)$	**E**	$Make\_an\_appeal\_or\_request(X_0, X_1, \tau_0)$ $\wedge Provide\_aid(X_1, X_2, \tau_1)$ $\wedge Provide\_aid^{-1}(X_2, X_3, \tau_2)$	Implausible	0.713
		G	$Express\_intent\_to\_provide\_economic\_aid(X_0, X_1, \tau_0)$ $\wedge Make\_statement^{-1}(X_1, X_2, \tau_1)$ $\wedge Make\_an\_appeal\_or\_request(X_2, X_3, \tau_2)$	Implausible	0.638

Table 3: Cases of rules from ICEWS05-15, where type **E** denotes "Extracted" and **G** denotes "Generated" rule bodies, indicating their respective source types. And the Chronological Order Constraint is met among $\tau_0, \tau_1, \tau_2, \cdots$.



(a) Data sparsity (b) Data noise

Figure 2: Robustness evaluation on ICEWS14.

$\mathcal{G}_{\text{train}}$	$\mathcal{G}_{\text{test}}$	Model	MRR	H@1	H@3	H@10
ICEWS05-15	ICEWS14	TLogic	27.4	20.5	30.6	40.6
		LLM-DR	32.9	25.6	34.3	44.8
ICEWS05-15	ICEWS18	TLogic	20.9	14.9	23.9	33.0
		LLM-DR	25.3	16.7	28.5	38.1
ICEWS18	ICEWS05-15	TLogic	35.3	26.6	40.2	51.8
		LLM-DR	39.4	31.7	43.2	52.9
ICEWS18	ICEWS14	TLogic	26.3	19.7	29.4	38.9
		LLM-DR	31.8	22.7	37.8	47.3
ICEWS14	ICEWS05-15	TLogic	43.7	33.0	49.8	64.2
		LLM-DR	45.6	35.7	51.5	64.9
ICEWS14	ICEWS18	TLogic	20.7	16.0	23.4	29.6
		LLM-DR	22.8	17.3	26.1	33.8

Table 4: Inductive setting where rules learned on $\mathcal{G}_{\text{train}}$ are transferred and applied to $\mathcal{G}_{\text{test}}$.

thereby yielding a more expansive array of potential rules. b) The coarse-grained evaluation phase serves as a pivotal filter, enabling us to discern whether a rule is "plausible" or "implausible". Notably in case 2, where two rules deemed "implausible" in the coarse-grained evaluation still managed to score highly in the fine-grained evaluation. This scenario illuminates the potential shortcomings of relying solely on fine-grained evaluation, which might not suffice to mitigate the influence of spurious rules on the reasoning process. This observation underscores the distinct advantage of our coarse-to-fine rule evaluation design, which offers a more nuanced and layered approach to rule assessment.

Robustness Analysis

To assess the robustness of LLM-DR, we conduct experiments under conditions of data sparsity and noise, manipulating the ICEWS14 dataset by randomly introducing deletions and additions of training facts. In our methodology, we implement random deletions and insertions of training data, ranging from 25% to 75% of the facts, to simulate varying intensities of sparsity and noise. The performance comparison in Figure 2 contrasts LLM-DR with two explainable methods: the rule-based TLogic and the subgraph-based xERTE. LLM-DR exhibits superior performance across the board, regardless of the severity of sparsity or noise, showcasing its robustness. Notably, even under extreme conditions of 75% data alteration, LLM-DR sustains a high MRR value of over 47, underscoring its proficiency in coping with both sparse and noisy datasets effectively. This robustness highlights LLM-DR's reliability in real-world applications where data imperfections are the norm.

Inductive Scenarios

Rule-based methods offer the advantage of generalizability to new datasets that encompass common relations, allowing for the easy application of validated rules to similar datasets (Liu et al. 2022). Utilizing rule transfer (Wang et al. 2022), we establish an inductive scenario by applying rules learned from one dataset's training set to another's test set, as shown in Table 4. Our LLM-DR's consistent outperformance over TLogic in inductive contexts highlights its ability to generate a broader set of rules beyond the dataset constraints, which enriches our model's generalization and adaptability, enabling it to skillfully handle various inductive situations.

Conclusion

In this paper, we introduce LLM-DR, an innovative rule-based method for TKG extrapolation that harnesses diffusion models to broaden the rule set, while integrating the strengths of LLMs. Our method generates semantically relevant and high-quality rules, employing conditional diffusion models guided in a classifier-free manner and refining them with LLM-based constraints. To assess rule efficacy, we meticulously design a coarse-to-fine evaluation strategy that initiates with coarse-grained filtering to eliminate less plausible rules, and proceeds with fine-grained scoring to quantify the reliability of the retained. Extensive experiments demonstrate the promising capacity of our LLM-DR.

Acknowledgements

This work is particularly supported by the National Key Research and Development Program of China (No. 2022YFB3104103), the National Natural Science Foundation of China (No. 62302507), and the Provincial Natural Science Foundation of Hunan (No. 2024JJ7249).

References

Bai, L.; Yu, W.; Chai, D.; Zhao, W.; and Chen, M. 2023. Temporal knowledge graphs reasoning with iterative guidance by temporal logical rules. *Information Sciences*, 621: 22–35.

Chen, W.; Wan, H.; Wu, Y.; Zhao, S.; Cheng, J.; Li, Y.; and Lin, Y. 2024. Local-Global History-Aware Contrastive Learning for Temporal Knowledge Graph Reasoning. In *40th IEEE International Conference on Data Engineering, ICDE 2024, Utrecht, The Netherlands, May 13-16, 2024*, 733–746. IEEE.

Cheng, K.; Ahmed, N. K.; and Sun, Y. 2023. Neural Compositional Rule Learning for Knowledge Graph Reasoning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023.*

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.

Dhariwal, P.; and Nichol, A. Q. 2021. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 8780–8794.

Elazar, Y.; Kassner, N.; Ravfogel, S.; Ravichander, A.; Hovy, E. H.; Schütze, H.; and Goldberg, Y. 2021. Measuring and Improving Consistency in Pretrained Language Models. *Trans. Assoc. Comput. Linguistics*, 9: 1012–1031.

Gao, Y.; Qiao, L.; Kan, Z.; Wen, Z.; He, Y.; and Li, D. 2024. Two-stage Generative Question Answering on Temporal Knowledge Graph Using Large Language Models. *CoRR*, abs/2402.16568.

García-Durán, A.; Dumancic, S.; and Niepert, M. 2018. Learning Sequence Encoders for Temporal Knowledge Graph Completion. In *EMNLP 2018.*

Goel, R.; Kazemi, S. M.; Brubaker, M. A.; and Poupart, P. 2020. Diachronic Embedding for Temporal Knowledge Graph Completion. In *AAAI 2020.*

Gong, S.; Li, M.; Feng, J.; Wu, Z.; and Kong, L. 2023. DiffuSeq: Sequence to Sequence Text Generation with Diffusion Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023.*

Hahsler, M.; and Hornik, K. 2007. New probabilistic interest measures for association rules. *Intelligent Data Analysis*, 11(5): 437–455.

Han, Z.; Chen, P.; Ma, Y.; and Tresp, V. 2021a. Explainable Subgraph Reasoning for Forecasting on Temporal Knowledge Graphs. In *ICLR 2021.*

Han, Z.; Ding, Z.; Ma, Y.; Gu, Y.; and Tresp, V. 2021b. Learning Neural Ordinary Equations for Forecasting Future Links on Temporal Knowledge Graphs. In *EMNLP 2021.*

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.*

Ho, J.; and Salimans, T. 2022. Classifier-Free Diffusion Guidance. *CoRR*, abs/2207.12598.

Huang, J.; and Chang, K. C. 2023. Towards Reasoning in Large Language Models: A Survey. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, 1049–1065. Association for Computational Linguistics.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.; Madotto, A.; and Fung, P. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, 55(12): 248:1–248:38.

Jiang, Y.; Yang, Y.; Xia, L.; and Huang, C. 2024. DiffKG: Knowledge Graph Diffusion Model for Recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining, WSDM 2024, Merida, Mexico, March 4-8, 2024*, 313–321. ACM.

Jin, W.; Qu, M.; Jin, X.; and Ren, X. 2020. Recurrent Event Network: Autoregressive Structure Inference over Temporal Knowledge Graphs. In *EMNLP 2020*, 6669–6683.

Kendall, D. 1967. PRINCIPLES OF RANDOM WALK. *Journal of the London Mathematical Society*, s1-42(1): 377–378.

Lautenschlager, J.; Shellman, S.; and Ward, M. 2015. ICEWS Events and Aggregations. *Harvard Dataverse*, 3.

Li, N.; E, H.; Li, S.; Sun, M.; Yao, T.; Song, M.; Wang, Y.; and Luo, H. 2023. TR-Rules: Rule-based Model for Link Forecasting on Temporal Knowledge Graph Considering Temporal Redundancy. In *EMNLP 2023 Findings, Singapore, December 6-10, 2023*, 7885–7894. Association for Computational Linguistics.

Li, Y.; Sun, S.; and Zhao, J. 2022. TiRGN: Time-Guided Recurrent Graph Network with Local-Global Historical Patterns for Temporal Knowledge Graph Reasoning. In *IJCAI 2022.*

Li, Z.; Hou, Z.; Guan, S.; Jin, X.; Peng, W.; Bai, L.; Lyu, Y.; Li, W.; Guo, J.; and Cheng, X. 2022. HiSMatch: Historical Structure Matching based Temporal Knowledge Graph Reasoning. In *EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, 7328–7338.

Li, Z.; Jin, X.; Li, W.; Guan, S.; Guo, J.; Shen, H.; Wang, Y.; and Cheng, X. 2021. Temporal Knowledge Graph Reasoning Based on Evolutional Representation Learning. In *SIGIR 2021.*

Liang, K.; Meng, L.; Liu, M.; Liu, Y.; Tu, W.; Wang, S.; Zhou, S.; and Liu, X. 2023. Learn from Relational Correlations and Periodic Events for Temporal Knowledge Graph Reasoning. In *SIGIR 2023*.

Liu, J.; Mao, Q.; Lin, C.; Song, Y.; and Li, J. 2023. LATENTLOGIC: Learning Logic Rules in Latent Space over Knowledge Graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, 4578–4586. Association for Computational Linguistics.

Liu, Y.; Ma, Y.; Hildebrandt, M.; Joblin, M.; and Tresp, V. 2022. TLogic: Temporal Logical Rules for Explainable Link Forecasting on Temporal Knowledge Graphs. In *AAAI 2022*.

Luo, L.; Ju, J.; Xiong, B.; Li, Y.; Haffari, G.; and Pan, S. 2023. ChatRule: Mining Logical Rules with Large Language Models for Knowledge Graph Reasoning. *CoRR*, abs/2309.01538.

Luo, R.; Gu, T.; Li, H.; Li, J.; Lin, Z.; Li, J.; and Yang, Y. 2024. Chain of History: Learning and Forecasting with LLMs for Temporal Knowledge Graph Completion. *CoRR*, abs/2401.06072.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, 16784–16804. PMLR.

Pacchiardi, L.; Chan, A. J.; Mindermann, S.; Moscovitz, I.; Pan, A. Y.; Gal, Y.; Evans, O.; and Brauner, J. M. 2024. How to Catch an AI Liar: Lie Detection in Black-Box LLMs by Asking Unrelated Questions. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Saharia, C.; Chan, W.; Chang, H.; Lee, C. A.; Ho, J.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022. Palette: Image-to-Image Diffusion Models. In *SIGGRAPH '22, Vancouver, BC, Canada, August 7 - 11, 2022*, 15:1–15:10. ACM.

Song, Y.; Li, A.; Tu, H.; Chen, K.; and Li, C. 2021. A Novel Encoder-Decoder Knowledge Graph Completion Model for Robot Brain. *Frontiers Neurorobotics*, 15: 674428.

Sun, H.; Zhong, J.; Ma, Y.; Han, Z.; and He, K. 2021. TimeTraveler: Reinforcement Learning for Temporal Knowledge Graph Forecasting. In *EMNLP 2021*, 8306–8319.

Tashiro, Y.; Song, J.; Song, Y.; and Ermon, S. 2021. CSDI: Conditional Score-based Diffusion Models for Probabilistic Time Series Imputation. In *Advances in Neural Information Processing Systems*, 24804–24816.

Tian, K.; Mitchell, E.; Zhou, A.; Sharma, A.; Rafailov, R.; Yao, H.; Finn, C.; and Manning, C. D. 2023. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, 5433–5442. Association for Computational Linguistics.

Trivedi, R.; Dai, H.; Wang, Y.; and Song, L. 2017. Know-Evolve: Deep Temporal Reasoning for Dynamic Knowledge Graphs. In *ICML 2017*.

Wang, J.; Sun, K.; Luo, L.; Wei, W.; Hu, Y.; Liew, A. W.; Pan, S.; and Yin, B. 2024. Large Language Models-guided Dynamic Adaptation for Temporal Knowledge Graph Reasoning. *CoRR*, abs/2405.14170.

Wang, W.; Xu, Y.; Feng, F.; Lin, X.; He, X.; and Chua, T. 2023a. Diffusion Recommender Model. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, 832–841. ACM.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q. V.; Chi, E. H.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023b. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

Wang, Y.; Cai, H.; Jiang, L.; Shen, B.; Wan, B.; Hu, P.; and Xiong, X. 2022. RTGAN: An Novel GAN-Based Rule Transfer Learning Method for Scalable Industrial Inference Engine. In *ICEBE 2022, Bournemouth, United Kingdom, October 14-16, 2022*, 24–25. IEEE.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*.

Xiong, M.; Hu, Z.; Lu, X.; Li, Y.; Fu, J.; He, J.; and Hooi, B. 2023. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. *CoRR*, abs/2306.13063.

Yu, F.; and Jin, W. 2000. An effective approach to mining exception class association rules. In *Web-Age Information Management: First International Conference, WAIM 2000 Shanghai, China, June 21–23, 2000 Proceedings 1*, 145–150. Springer.

Zhang, J.; Chen, B.; Zhang, L.; Ke, X.; and Ding, H. 2021. Neural, symbolic and neural-symbolic reasoning on knowledge graphs. *AI Open*, 2: 14–35.

Zhao, X.; Jia, Y.; Li, A.; Jiang, R.; Chen, K.; and Wang, Y. 2021. Target relational attention-oriented knowledge graph reasoning. *Neurocomputing*, 461: 577–586.

Zheng, L.; Chiang, W.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*.

Zhou, Y.; Muresanu, A. I.; Han, Z.; Paster, K.; Pitis, S.; Chan, H.; and Ba, J. 2023. Large Language Models are Human-Level Prompt Engineers. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

Zhu, C.; Chen, M.; Fan, C.; Cheng, G.; and Zhang, Y. 2021. Learning from History: Modeling Temporal Knowledge Graphs with Sequential Copy-Generation Networks. In *AAAI 2021*.