Towards Multi-Relational Multi-Hop Reasoning over Dense Temporal Knowledge Graphs

Jian Liu[12*], **Zihe Liu**[2*], **Xueqiang LYU**[13], **Peng Jin**[4], **Jinan Xu**[2]
[1] Beijing Key Laboratory of Internet Culture and Digital Dissemination Research
[2] Beijing Jiaotong University, Beijing, China
[3] Beijing Information Science And Technology University
[4] Key Laboratory of Internet Natural Language Processing of Sichuan Provincial Education Department
Leshan Normal University
{jianliu,23120386,jaxu}@bjtu.edu.cn; lxq@bistu.edu.cn; jandp@pku.edu.cn

Abstract

Temporal knowledge graph reasoning has emerged as a crucial task for answering time-dependent questions within a knowledge graph (KG). Despite tremendous progress, the present research is impeded by the sparsity of a temporal KG and an over-reliance on simple single-relational reasoning patterns. To overcome these challenges, we introduce Mul^{2}Questions, a new temporal KG reasoning benchmark featuring over 200k entities and 960k questions designed to facilitate complex, multi-relational and multi-hop reasoning. Additionally, we propose a new model adept at conducting pattern-aware and time-sensitive reasoning across temporal KGs. The model's efficacy is confirmed through rigorous evaluations, showcasing its effectiveness in sparse data conditions and adeptness at handling questions with long reasoning chains. We have made our benchmark and model publicly accessible at https://github.com/Zihe2003/Mul2Questions.

1 Introduction

Temporal knowledge graph reasoning involves the task of answering questions related to time-dependent facts within a knowledge graph (KG) (García-Durán et al., 2018; Jia et al., 2021; Saxena et al., 2021; Lan et al., 2022; Chen et al., 2023). For example, given two facts with time duration: (Barack Obama, position, President of US, 2009-2017) and (George W. Bush, position, President of US, 2001-2009), for a time-related question "Who was the president of the United States *before* Barack Obama?", a model for this task should infer that the answer is "George W. Bush". Recent benchmarks such as (Saxena et al., 2021) and MultiTQ (Chen et al., 2023) have greatly improved this task and expanded its applicability.

Despite notable progress, the current study on temporal knowledge graph reasoning faces challenges of the sparsity of KGs and oversimplified reasoning patterns. The CronQuestions benchmark (Saxena et al., 2021), for example, uses templates with only five temporal relations to generate questions — this produces a sparse reasoning graph after filtering the original KGs and can significantly ease the reasoning. Furthermore, it is worth noting that a substantial proportion (84.5%) of the questions in previous benchmarks (Saxena et al., 2021; Chen et al., 2023) involve only single-relational reasoning patterns, as shown in the top of Figure 1. In such cases, the reasoning is simplified because only one relation is involved, and the answer is sure to have the same relation shared with the two entities that appear in the questions. There remains a lack of evaluations on complex multi-relational and multi-hop reasoning patterns, as depicted in the bottom of Figure 1.

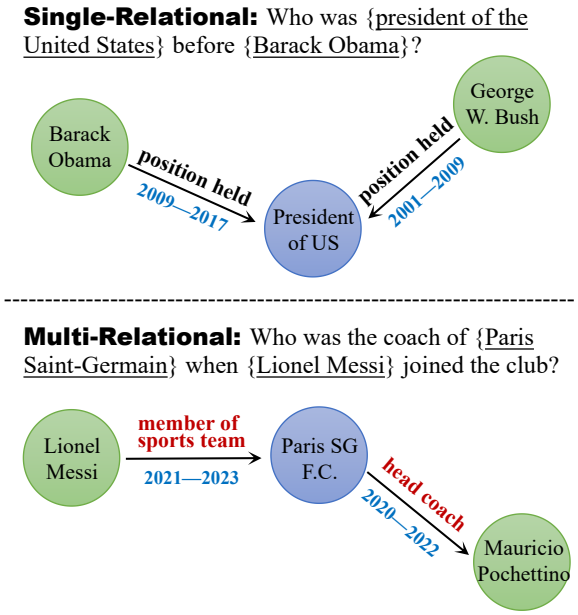

Figure 1: A comparison of single-relational (top) and multi-relational multi-hop (bottom) reasoning patterns.

This work focuses on multi-relational multi-hop reasoning in a densely connected temporal knowledge graph. As our first contribution, we developed

[*]Equal Contribution

Findings of the Association for Computational Linguistics: ACL 2024, pages 14367–14378

a new benchmark named Mul^{2}Questions, which contains over 200k entities spanning a temporal KG that is a strongly connected graph created using a breadth-first graph-expanding strategy. Moreover, the benchmark also includes an extensive collection of temporal questions (over 960k) that require complex multi-relational multi-hop reasoning patterns, as illustrated in Figure 1 (bottom), by incorporating 135 distinct temporal relations. Table 1 shows a statistical comparison between our benchmark and others, which highlights the much higher semantic complexity of our benchmark's questions (as defined in Eq. (1)) — this indicates a heightened challenge in both finding candidate entities and deducing correct answers.

In addition to the data contribution, we present a new model for temporal knowledge graph reasoning that incorporates pattern-aware and time-sensitive joint reasoning methods. Through rigorous evaluations, we show that our methods outperforms existing methods by a substantial margin (+9 points) in overall accuracy. Furthermore, it performs well in data-scarce situations and addressing questions with long reasoning chains.

In summary, our contributions are three-fold:

- We present Mul^{2}Questions, a new benchmark for temporal knowledge graph reasoning, focusing particularly on multi-relational multi-hop reasoning patterns. We also study its potential extension to multilingual scenarios.
- We introduce a new method with pattern-aware and time-sensitive mechanism for reasoning over temporal KGs; it demonstrates outstanding ability in addressing diversified reasoning patterns.
- We release both the benchmark and model to the public to facilitate further exploration.

2 Related Work

2.1 Temporal KG Reasoning Resources

Temporal KGs are multi-relational graphs with each edge (i.e., relation) marked with time duration information (Dasgupta et al., 2018; García-Durán et al., 2018; Jain et al., 2020). Among the datasets facilitating reasoning over temporal KGs, TempQuestions (Jia et al., 2018b), derived from FreeBase (Bollacker et al., 2008), offers 1,271 questions, and SYGMA (Neelam et al., 2021) utilizes Wikidata to enhance reasoning capabilities for TempQuestions. However, these datasets provide only around a thousand questions, falling short for developing advanced neural models. On the larger scale, TimeQuestions (Jia et al., 2021) gathers 16k time-centric questions from eight KGs, and CronQuestions (Saxena et al., 2021) delivers an extensive dataset with 328k facts and 410k questions from Wikidata. Additionally, MultiTQ (Chen et al., 2023) introduces a dynamic semantic KG from ICEWS05-15 (García-Durán et al., 2018), broadening the scope with rich semantic information. Nonetheless, most benchmarks focus on single-relational reasoning within sparse KGs. By contrast, our approach emphasizes complex multi-relational and multi-hop reasoning over densely populated temporal KGs.

Dataset	#UR	#Tem.	#Ques.	$SC_{\mathcal{Q}}$
TempQ (2018b)	-	-	1.2k	1.12
TimeQ (2021)	-	-	16k	1.00
CronQ (2021)	5	30	410k	1.23
MultiQ (2023)	22	246	500k	1.00
Mul^{2}Questions	**135**	**586**	**960k**	**2.35**

Table 1: A comparison of the number of unique relations (#UR), question templates (#Tem.), the size of the question set (#Ques.), and the semantic complexity of the questions ($SC_{\mathcal{Q}}$) regarding different benchmarks..

2.2 Temporal KG Reasoning Approaches

The exploration of effective reasoning methods over temporal KGs is still in its early stages, with two predominant approaches emerging: semantic parsing-based and embedding-based methods. Semantic parsing-based methods, such as those employed by TEQUILA (Jia et al., 2018a), EXAQT (Jia et al., 2021), and TwiRGCN (Sharma et al., 2023), begin by segmenting the questions into temporal and non-temporal components and then apply temporal constraints to refine the pool of potential answers to find the correct one. Despite their effectiveness, these methods are constrained by the need for manually crafted rules for problem decomposition, limiting their effectiveness in tackling complex queries. In contrast, recent advancements in embedding-based approaches leverage neural networks to capture temporal dynamics and adopt semantic similarity for answer prediction. Specifically, CronKGQA (Saxena et al., 2021) introduces a dynamic, learnable framework for temporal reasoning, avoiding hand-crafted rules. TSQA (Shang

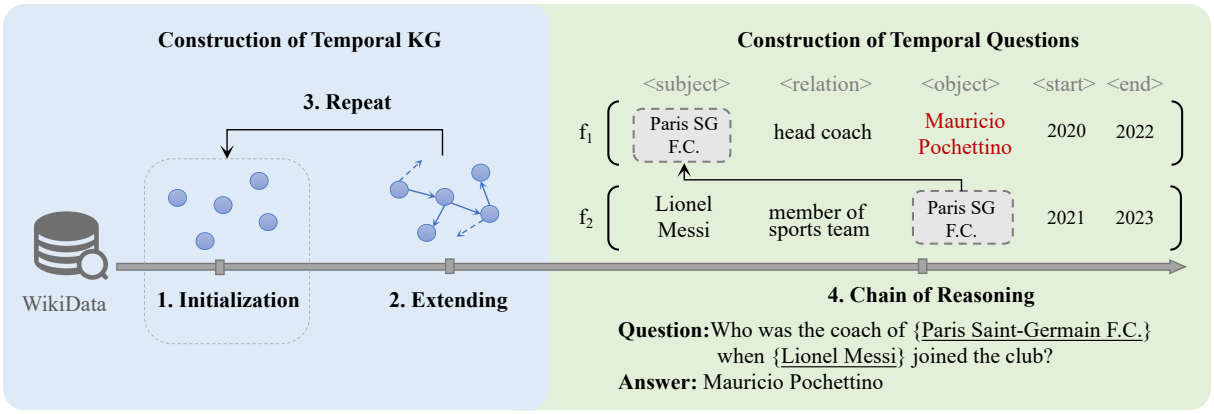

Figure 2: The Mul^{2}Questions benchmark construction process is divided into two parts: temporal KG construction (left) and temporal question construction (right).

et al., 2022) enhances the interaction between time and entities using contrastive learning techniques and time position coding. TempoQR (Mavromatis et al., 2022) integrates time range information and employs entity embeddings for richer question representations. MultiQA (Chen et al., 2023) proposes a new model for particularly multi-granularity inference and establishes a new standard for temporal reasoning tasks. However, limited by the oversimplified reasoning patterns in previous benchmarks, these methods may have difficulties for addressing complex patterns. This paper addresses the challenges by designing a pattern-aware and time-sensitive mechanism to enhance learning.

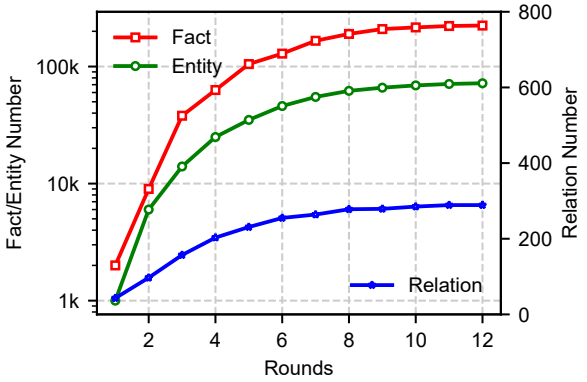

Figure 3: The change in the number of entities, relations, and facts regarding different expanding rounds.

3 The Mul^{2}Questions Benchmark

Our Mul^{2}Questions benchmark targets multi-relational multi-hop reasoning particularly, and we divide the construction process by temporal KG construction (§ 3.1) and temporal question construction (§ 3.2) as visualized in Figure 2.

3.1 Construction of the Temporal KG

We use Wikidata (Vrandečić and Krötzsch, 2014) to build the temporal KG, following previous works (Saxena et al., 2021). Instead of simply filtering out Wikidata to obtain the KG, we employ a breadth-first approach to expand an empty KG gradually. This method guarantees a densely connected KG that includes crucial entities/facts and avoids discontinuities in the graph.

To start, we compile an empty KG and build a seed entity set containing individuals listed as the "TIME 100 Persons" by the *Times* magazine[1]. The underlying assumption for choosing this set is that prominent persons may get involved in the world's important events and, therefore, have a more extensive network of connections in Wikidata. Next, we iterate each entity in the seed set and add entities that have a temporal connection with it in the seed set. Additionally, the relevant facts are added to the temporal KG. We repeat the above process and check each newly added entity until convergence. Finally, we add the important world events, such as (World War II, occurs, 1939-1945), into the temporal KG following Saxena et al. (2021).

Finally, a dense temporal KG consisting of 229k facts, 292 distinct types of temporal relations, and over 200k entities was produced, with Figure 3 visualizing the detailed process. To qualitatively evaluate the constructed KG, we measure its semantic complexity $SC_{\mathcal{KG}}$, defined as the average number of relation types associated with each entity (Chen et al., 2023). The results reveal that our temporal KG has a high $SC_{\mathcal{KG}}$ of 1.97, 49% greater than that of CronQuestions, which is generated by merely filtering the Wikidata set.

[1]https://en.wikipedia.org/wiki/Time_100

Pattern	Template	Contextualized Question
Subject-Join	CTX(s_1, r_1, ?) when/at the time/during CTX(f_2) ?	Which club did {s_1} play for when {s_2} serverd as the coach?
Object-Join	CTX(?, r_1, o_1) when/at the time/during CTX(f_2) ?	Who was the coach of {o_1} when {s_2} play for the club?
Time-Join	CTX(s_1, r_1, ?) when/at the time/during CTX(f_2) ? CTX(?, r_1, o_1) when/at the time/during CTX(f_2) ?	Which club did {s_1} play during World War II? Who was the coach of {o_1} during World War II?

Table 2: Templates with different overlapping patterns for question generation.

3.2 Construction of Temporal Questions

We build a large question set with multi-relational multi-hop reasoning patterns for Mul^{2}Questions by first recognizing time-compatible fact pairs and then converting them into question-answer pairs using templates.

Time-Compatible Fact Pairs. Let f_1 = <s_1, r_1, o_1, t_1> be a fact instance in the temporal KG. We define a time-compatible fact with f_1 as another fact instance f_2 with an overlapped duration time. By definition, f_2 can be either a standard fact represented by f_2 = <s_2, r_2, o_2, t_2> or a world event f_2 = <e_2, t_2> as introduced by (Saxena et al., 2021). We then consider the overlapping patterns of f_1 and f_2 to generate question-answer pairs.

Templates for Question Generation. We define three overlapping patterns for question generation: 1) Subject-Join: f_2 is a standard fact, and the subject s_1 of f_1 is involved in f_2. In this case, we set the answer to o_1 and generate a contextualized question based on the semantics of f_1 and f_2. 2) Object-Join: similar to subject-join type, but the object o_1 of f_1 is involved in f_2. In this case, we set the answer to s_1 and generate a contextualized question. 3) Time-Join: in this case, f_2 is a world event, and we can set the answer to either s_1 or o_1 and generate a contextual question. Figure 2 gives an example of Subject-Join type, and we set the answer to the object "Mauricio Pochettino" and a generated question is *"Who was the coach of Paris SG F.C. when Lionel Messi joined the club?"*. Table 2 shows more examples.

The above describes 2-hop reasoning patterns, but by adding more compatible facts, we can get a longer reasoning chain. For example, when finding another fact f_3 = <Lionel Messi, award received, Ballon d'Or, 2019, 2019>, we can construct a three-hop question, *"Who was the coach of Paris SG F.C. when <the player winning Ballon d'Or at 2019> joined the club?"* However, considering this often gets long and verbose questions, we only use the ten most common 3-hop patterns. In total, we find 586 unique patterns (including ten 3-hop patterns) involving 135 different temporal relations, much larger than previous methods. Then, we employ seven experts to contextualize each pattern and extend them by a generative language model of GPT-4 (OpenAI, 2023), using few-shot prompting techniques (Brown et al., 2020). Finally, we get 6,274 contextualization templates. By propagating each template using compatible fact instances, we finally ended up with 960k questions. We refer to Appendix A for more examples of contextualized templates that we use to generate questions.

Semantic Complexity of Questions. Similar to the definition of semantic complexity of a KG, we define semantic complexity of questions($SC_{\mathcal{Q}}$) as the average of the number of relation types involved in each question:

$$SC_{\mathcal{Q}} = \frac{1}{|\mathcal{Q}|} \sum_{q_i \in \mathcal{Q}} N_{q_i}^{r_{type}} \quad (1)$$

where $\mathcal{Q}$ is the question set and $N_{q_i}^{r_{type}}$ is the number of relation types involved in q_i. Consequently, we get a $SC_{\mathcal{Q}}$ = 2.35, which is much higher than that of CronQuestions ($SC_{\mathcal{Q}}$ = 1.21) and MultiTQ ($SC_{\mathcal{Q}}$ = 1.00), where the most of questions involving only single-relational reasoning. Obviously, the larger $SC_{\mathcal{Q}}$, the more relations the question involves and the harder reasoning.

4 Our Temporal Reasoning Model

We propose a new model featuring discerning different reasoning patterns over a temporal KG. As shown in Figure 4, our model consists of three essential parts: 1) reasoning subgraph extraction, 2) reasoning pattern characterization, and 3) pattern-aware joint reasoning.

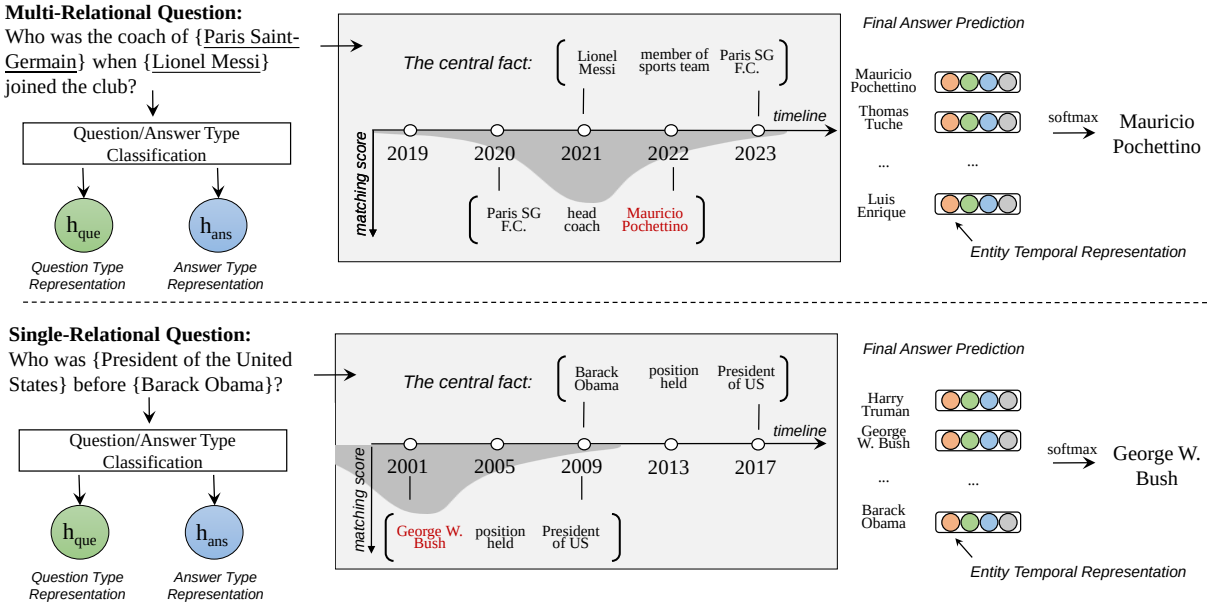

Figure 4: The overview of our approach, featured by discerning different reasoning patterns over a temporal KG.

4.1 Reasoning Subgraph Extraction

Given a question q, we first extract a subgraph to facilitate reasoning. We begin by identifying a key entity/relation set from q using an extra trained entity/relation recognizer[2]. Then, we create a subgraph that includes all facts connected to the entities, and we designate a fact instance that is closest to the entity set as the "central fact". Specifically, the "close" here is indeed a measurement of the overlapping of a fact and the entity/relation set. We then arrange the facts in chronological order, which is used to help decide a region where the answer is more likely to exist.

4.2 Reasoning Pattern Characterization

We motivate our reasoning pattern characterization module by noting that different patterns exist for reasoning. For example, for the multi-relational multi-hop reasoning question, it is more likely that the target fact/entity has a duration overlapping with the "central fact" involved in the question. By contrast, for the single-relational reasoning question, we should focus on the fact adjacent to the "central fact" with the same relation for finding the answer. While for the first/last question (i.e., "Who is the first president of the US?"), we should focus on the fact locating the earliest in the chronological timeline. With the above insights, we build a reasoning pattern characterization module to discern different reasoning patterns to enable reasoning.

We categorize the reasoning patterns into three aspects: 1) The answer type label, which is a label from [entity, time]. 2) The question type label, which is a label from [before/after, first/last, multi-hop] following definitions of (Chen et al., 2023). 3) The timeline distance, which is a scalar value showing the distance from the answer entity to the "central fact". Note that all these patterns are known during training, and we build three prediction models to learn their representations.

Specifically, given q, we use a BERT encoder to covert it as a continuous representation $\boldsymbol{h} \in \mathcal{R}^d$:

$$\boldsymbol{h} = \text{BERT}(q) \tag{2}$$

where d denotes the dimension of representation. Then, we build two classification models to predict the answer and question type labels. For example, to predict the answer type label, we first transform $\boldsymbol{h}$ into $\boldsymbol{h}_a \in \mathcal{R}^d$:

$$\boldsymbol{h}_a = \boldsymbol{W}_a \boldsymbol{h} \tag{3}$$

where $\boldsymbol{W}_a \in \mathcal{R}^{d \times d}$ is a parameter, and then use a binary classifier to map it into an answer type label. The question type label classification is similar but compute $\boldsymbol{h}_q \in \mathcal{R}^d$ with another parameter $\boldsymbol{W}_q \in \mathcal{R}^{d \times d}$ and a multi-class classifier. For the timeline distance, we compute a scalar number based on the representations of $\boldsymbol{h}$ and the central fact:

$$s_q = \boldsymbol{W}_t(\boldsymbol{h} \oplus \text{BERT}(f)) \tag{4}$$

[2]We use a BERT based recognizer, achieving 92%/85% in F1 for entity/relation recognition in the questions.

where $\boldsymbol{W}_t \in \mathcal{R}^{1\times 2d}$, f stands for the "central fact", and $\oplus$ denotes a concatenation operator. The classification and regression models are trained based on cross-entropy and MSE loss, respectively.

4.3 Pattern-Aware Joint Reasoning

We perform a pattern-aware joint reasoning to locate the answer. Specifically, for an entity e in the subgraph, its representation is built by:

$$\boldsymbol{h}_e = s_{e\to f} * (\boldsymbol{h}_{\text{temp}} \oplus \boldsymbol{h}_a \oplus \boldsymbol{h}_q) \tag{5}$$

where $\boldsymbol{h}_{\text{temp}} \in \mathcal{R}^{d_t}$ is the temporal representation of the entity learned from the temporal KG, following (Saxena et al., 2021), and $s_{e\to f}$ is the timeline matching score of the entity, computed by:

$$s_{e\to f} = \mathcal{N}(|\text{Dis}(e, f) - s_q|; 0, 1) \tag{6}$$

where $\mathcal{N}(; 0, 1)$ denotes a standard Gaussian distribution, and $|\text{Dis}(e, f) - s_q|$ is a measurement for whether the true distance of e and f ($\text{Dis}(e, f)$) matches the predicted timeline distance (s_q).

Finally, we perform a softmax calculation to locate the answer. For example, the probability of e being the answer is computed by:

$$Pr(e|q) = \frac{\exp(\boldsymbol{W}_p \boldsymbol{h}_e)}{\sum_{e'\in\mathcal{G}} \boldsymbol{W}_p \boldsymbol{h}_{e'}} \tag{7}$$

where $\boldsymbol{W}_p \in \mathcal{R}^{1\times d_t+2d}$ denotes a prediction parameter and e' ranges over each entity in the subgraph. For training, we train all parameters by minimizing a cross-entropy loss function:

$$\mathcal{L} = -\sum_{(q,a)\in D} \log Pr(a|q) \tag{8}$$

where (q, a) ranges over each training instance in the training set D. We use Adam (Kingma and Ba, 2014) algorithm for parameter optimization.

5 Experimental Evaluations

5.1 Datasets and Setups

We evaluate our approach using our Mul^{2}Questions benchmark and CronQuestions to allow for different reasoning patterns. For our benchmark, we conduct a split of 8/1/1 for the train/dev/test split, and for CronQuestions, we use the standard settings. As for evaluations, we report results regarding Hit@1 and Hit@10. As for implements, we use $\text{BERT}_{\text{base}}$ as the basic answer/question type classifier and a linear regression model to predict the timeline matching score. For the final optimization, we set the batch size to 20, chosen from 5, 10, 20, 40, and the learning rate as 1e-5, chosen from a set {1e-4, 2e-5, 1e-5} on the dev set by the overall Hits@1 metrics. Our models are implemented by PyTorch and trained using NVIDIA Tesla V100 GPUs. We have released the code at https://github.com/Zihe2003/Mul2Questions to enable further exploration.

5.2 Baseline Methods

We use previous state-of-the-art temporal knowledge graph reasoning models as baselines, including 1) **BERT-Based Models**, which use representations constructed by BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) as entity representations for prediction. 2) **EmbedKGQA** (Saxena et al., 2020), which uses representations designed for static KGs to perform reasoning. To address temporal questions, it ignores timestamps during training and employs random time embeddings. 3) **CronKGQA** (Saxena et al., 2021), which extends EmbedKGQA by incorporating temporal KG embeddings, serving as a standard baseline for evaluation temporal KG reasoning. 4) **TempoQR** (Mavromatis et al., 2022), which proposes an embedding representation for question answers using time-assisted constraints, inspiring further research into time-sensitive question answering . 5) **MultiQA** (Chen et al., 2023), which introduces an innovative approach by incorporating a multi-granularity technique that leverages time-assisted constraints. We refer to our approach as PATKGQA to signify that it is Pattern-Aware.

5.3 Quantitative Results

Table 3 shows the performance of different models on our Mul^{2}Questions benchmark. Accordingly to the results, our method outperforms baseline models by a significant margin (+9.5% in Hit@1 and +9.7% in Hit@10), demonstrating its usefulness. We also observe that temporal embedding methods outperform static representation methods such as BERT and RoBERTa, emphasizing the importance of introducing temporal embeddings. When comparing different types of questions, we can see that the Time-Join question is easier than the other two. This is expected since the Time-Join pattern often contains a world event, which we can easily determine the timing, and use it as a proxy to locate another fact. In contrast, Subject-Join and Object-Join need us to consider sharing patterns as well as

Model	Hit@1				Hit@10			
	Overall	Reasoning Pattern			Overall	Reasoning Pattern		
		S-Join	O-Join	T-Join		S-Join	O-Join	T-Join
BERT (2019)	0.167	0.158	0.082	0.207	0.505	0.513	0.313	0.514
RoBERTa (2019)	0.160	0.149	0.066	0.209	0.503	0.512	0.329	0.507
EmbedKGQA (2020)	0.322	0.313	0.082	0.390	0.629	0.624	0.309	0.698
CronKGQA (2021)	0.378	0.344	0.148	0.516	0.688	0.682	0.494	0.739
TempoQR (2022)	0.522	0.430	0.473	0.795	0.765	0.706	0.724	0.943
MultiQA (2023)	0.416	0.373	0.107	0.592	0.721	0.691	0.527	0.840
PATKGQA (Ours)	**0.614**	**0.577**	**0.585**	**0.830**	**0.862**	**0.818**	**0.823**	**0.961**

Table 3: Results on the Mul^{2}Questions benchmark, where S-Join, O-Join, and T-Join represent the Subject-Join, Object-Join, and Time-Join patterns used to obtain a question.

Model	Over.	Question Type	
		Simple	Compl.
BERT (2019)	0.243	0.249	0.239
RoBERTa (2019)	0.225	0.237	0.217
EmbedKGQA (2020)	0.288	0.290	0.286
CronKGQA (2021)	0.647	0.987	0.392
TempoQR (2022)	0.918	**0.990**	0.864
MultiQA (2023)	0.764	0.987	0.712
PATKGQA (Ours)	**0.931**	**0.990**	**0.887**

Table 4: Results of Hit@1 on the CronQuestions benchmark, where Simple and Compl. refer to questions that are based on either a single fact or several facts (albeit they are still single-relational questions).

distinct temporal relations, which could make the reasoning more challenge.

Table 4 shows the results of the CRONQuestions benchmark, which has a different reasoning structure than ours. As a result, our methods performs well, indicating that it is effectiveness in addressing different reasoning patterns. Particularly, the improvement stems primarily from addressing complex questions. To demonstrate its effectiveness, consider a question having a (first/last) pattern in the CRONQuestions benchmark: "When did Messi play their first game?". We show that our approach produces a negative timeline matching score, signaling that the answer should appear early in the timeline, which aids in locating the correct entity.

6 Discussion

Ablation Study. We perform an ablation study to validate the effect of each component in Table 5. In particular, we first employ the TempoQR (Mavromatis et al., 2022) architecture as the base model, and then augment it with each component respectively, including subgraph extraction, question and answer type presentations, and the augmentation of timeline matching score, to visualize the impacts. The results show that each component aids learning, with the timeline matching score being the most effective one and leading to the largest improvement. This observation highlights the importance for modeling temporal relations of facts dynamically in the reasoning process, which has often been missed by earlier research.

Method	CronQ.	Mul^{2}Q.
Base Model	0.918	0.522
+ Subgraph Extraction	0.925	0.564
+ QA Type Representations	0.919	0.528
+ Timeline Matching Score	**0.927**	**0.605**

Table 5: Ablation studies of different modules on two datasets, where the Base Model shares a similar implementation as the state-of-the-art baseline TempoQR (Mavromatis et al., 2022).

Type	Question Examples
2-Hop	Who was the coach of Paris SG F.C. when Lionel Messi joined the club?
3-Hop	Who was the coach of Paris SG F.C. when <the player winning Ballon d'Or at 2019> joined the club?
4-Hop	Who was the coach of <the club owed by Qatar Sports Investments> when <the player winning Ballon d'Or at 2019> joined the club?

Table 6: Example questions for multi-hop reasoning.

Method	2-Hop	3-Hop	4-Hop
BERT (2019)	0.172	0.083	0.027
CronKGQA (2021)	0.383	0.129	0.074
TempoQR (2022)	0.541	0.172	0.081
MultiQA (2023)	0.420	0.131	0.074
PATKGQA	**0.623**	**0.375**	**0.312**

Table 7: Results of addressing longer reasoning chains.

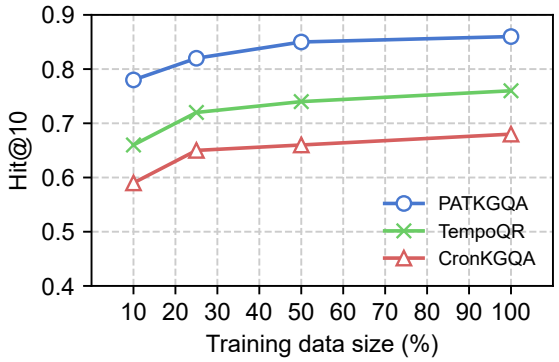

Figure 5: Results in data-scarce scenarios on our Mul^{2}Questions benchmark.

Towards Reasoning over Longer Chains. Then, we investigate the ability of different models to handle longer reasoning chains. Specifically, We generated and sampled 10k 2-hop, 3-hop, and 4-hop questions[3] (same examples are shown in Table 6) respectively and utilize a model for answer prediction. The results in Table 7 indicate that our method performs well when handling longer reasoning chains. The reason for this is that in questions with longer reasoning chain, the answer often has an in-direct connection to entities appearing in the question, which is difficult to handle in prior methods that do not recognize the reasoning pattern. In contrast, our approach utilizes answer/question type representations and timing matching score mechanisms to assist in locating the answer.

Results in Data-Scarce Scenarios. We examine the the ability of our approach for learning in scenarios with limited data, and the results are illustrated in Figure 5. Compared to earlier methods, our methodology exhibits better performance in situations where data is limited, where the underlying reason might be that the answer/question type representations and timeline matching score offer more informative guidance for locating the answer.

[3]Note that we only included a small number of 3-hop questions and no 4-hop questions in our benchmark because the majority of them are verbose and impractical

Method	CronQ.	Mul^{2}Q.
LLama 2 13B (ZERO)	0.326	0.228
LLama 2 13B (RAG)	0.937	0.719
GPT-4 (ZERO)*	0.441	0.295
GPT-4 (RAG)*	**0.954**	**0.742**
PATKGQA*	0.937	0.621
PATKGQA	0.931	0.614

Table 8: Comparison to LLMs based models, where * indicates results based on a sampled set of questions.

Language	#Facts	#Questions	Hit@1
English	228k	960k	0.614
French	196k	290k	0.603
Chinese	235k	350k	0.487

Table 9: Results of multilingual extension.

Comparison to Large Language Models. Our method is evaluated against large language models (LLMs) in Table 8, namely the open-source LLM LLama 2 13B (Touvron et al., 2023) and the closed-source LLM GPT-4 (OpenAI, 2023). We use two settings: zero-shot (ZERO), where a LLM based model directly predicts an answer without relying on extra clues, and RAG, which incorporates the subgraph into the prompt to enhance reasoning. Following prior RAG studies on KBQA (Kim et al., 2023; Jiang et al., 2023), our approach first retrieving a subgraph and then organizing them in chronological order. We finally construct a prompt like *"Given the facts: [facts], Please answer [question]"* and query an LLM for the answer. Acocording to the results, LLM-based models do perform well at this task; even in a zero-shot setting, they are accurate about 30% of the time, and their performance is much better when a reasoning sub-graph is added. One reason for the good performance of LLM is that the questions in the benchmark are all fact-based questions that an LLM can master at the pre-training stage, while we note there is still room for addressing more difficult multi-hop reasoning patterns.

Multilingual Extension. In addition, we explore the feasibility of applying our benchmark construction approach to other languages. Table 9 shows a comparison of the corpus size in French and Chinese, demonstrating its practicality. However, in

comparison to English, the performance of our model in French and Chinese are marginally reduced, which may due to a less compact KG.

7 Conclusion

In this research, we investigate multi-relational multi-hop reasoning over a dense temporal knowledge graph. We introduce Mul^{2}Questions, a new benchmark that includes over 200k entities and 960k questions with multi-relational multi-hop patterns. In addition, we present a new model capable of performing pattern-aware and time-sensitive joint reasoning over temporal KGs. The effectiveness has been verified by extensive evaluations. In the future, we would examine its applicability in a broader range of domains/scenarios.

Acknowledgements

This work has been supported by the Fundamental Research Funds for the Central Universities 2023JBMC058, the National Nature Science Foundation of China (No.62106016), the opening project of Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, and the Key Lab of Internet Natural Language Processing of Sichuan Provincial Education Department (No. INLP202305).

The authors would like to thank the anonymous reviewers for their valuable suggestions.

Limitations

One limitation is that, like many benchmarks, our benchmark's questions are completely factual and can be easily addressed using modern large language models (LLMs) trained on texts with world knowledge. To comprehensively test temporal reasoning, we may employ more specific questions that do not contain real-world knowledge. Second, the questions are generated using customized templates. However, handcrafted templates are typically based on the exporter's personal knowledge and experience, which may add subjective biases. Furthermore, the KG employed in the study is based on a single domain, Wikidata, and we hope to expand it to other domains in the future.

References

Kurt D. Bollacker, Colin Evans, Praveen K. Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD Conference*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ziyang Chen, Jinzhi Liao, and Xiang Zhao. 2023. Multi-granularity temporal question answering over knowledge graphs. In *Annual Meeting of the Association for Computational Linguistics*.

Shib Sankar Dasgupta, Swayambhu Nath Ray, and Partha Talukdar. 2018. HyTE: Hyperplane-based temporally aware knowledge graph embedding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2001–2011, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Prachi Jain, Sushant Rathi, Mausam, and Soumen Chakrabarti. 2020. Temporal Knowledge Base Completion: New Algorithms and Evaluation Protocols. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3733–3747, Online. Association for Computational Linguistics.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018a. TEQUILA: temporal question answering over knowledge bases. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, pages 1807–1810. ACM.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018b. Tempquestions: A benchmark for temporal question an-

swering. In *Companion Proceedings of the The Web Conference 2018*, pages 1057–1062.

Zhen Jia, Soumajit Pramanik, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Complex temporal question answering on knowledge graphs. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 792–802.

Jinhao Jiang, Kun Zhou, Wayne Xin Zhao, Yaliang Li, and Ji-Rong Wen. 2023. Reasoninglm: Enabling structural subgraph reasoning in pre-trained language models for question answering over knowledge graph. *arXiv preprint arXiv:2401.00158*.

Jiho Kim, Yeonsu Kwon, Yohan Jo, and Edward Choi. 2023. Kg-gpt: A general framework for reasoning on knowledge graphs using large language models. *arXiv preprint arXiv:2310.11220*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2022. Complex knowledge base question answering: A survey. *IEEE Transactions on Knowledge and Data Engineering*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

Costas Mavromatis, Prasanna Lakkur Subramanyam, Vassilis N. Ioannidis, Adesoji Adeshina, Phillip R. Howard, Tetiana Grinberg, Nagib Hakim, and George Karypis. 2022. Tempoqr: Temporal question reasoning over knowledge graphs. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 5825–5833. AAAI Press.

Sumit Neelam, Udit Sharma, Hima P. Karanam, Shajith Ikbal, Pavan Kapanipathi, I. Abdelaziz, Nandana Mihindukulasooriya, Young-Suk Lee, Santosh K. Srivastava, Cezar Pendus, Saswati Dana, Dinesh Garg, Achille Fokoue, G. P. Shrivatsa Bhargav, Dinesh Khandelwal, Srinivas Ravishankar, Sairam Gurajada, Maria Chang, Rosario A. Uceda-Sosa, Salim Roukos, Alexander G. Gray, Guilherme LimaRyan Riegel, Francois P. S. Luus, and L. V. Subramaniam. 2021. Sygma: System for generalizable modular question answering overknowledge bases. *ArXiv preprint*, abs/2109.13430.

OpenAI. 2023. Gpt-4 technical report. *ArXiv preprint*, abs/2303.08774.

Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. 2021. Question answering over temporal knowledge graphs. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6663–6676, Online. Association for Computational Linguistics.

Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. 2020. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4498–4507, Online. Association for Computational Linguistics.

Chao Shang, Guangtao Wang, Peng Qi, and Jing Huang. 2022. Improving time sensitivity for question answering over temporal knowledge graphs. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8017–8026, Dublin, Ireland. Association for Computational Linguistics.

Aditya Sharma, Apoorv Saxena, Chitrank Gupta, Mehran Kazemi, Partha Talukdar, and Soumen Chakrabarti. 2023. TwiRGCN: Temporally weighted graph convolution for question answering over temporal knowledge graphs. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2049–2060, Dubrovnik, Croatia. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

A Appendix

In our study, we provided detailed instructions to participants listed below:

Annotation Instructions

Understanding Patterns: Begin by familiarizing yourself with the unique patterns. Each pattern represents a reasoning structure within the knowledge graph.

Template Structure:

- Craft clear, concise question templates that can be filled with specific entities or attributes.
- Ensure the template directly relates to the pattern it's designed for.

Creativity and Variety:

- Use creative wording to cover the breadth of possible scenarios within a pattern.
- Include a variety of question types, e.g., "What is the...?", "Who was the...?", "When did...?".

Clarity and Simplicity:

- Ensure questions are straightforward and understandable.
- Avoid complex or confusing phrasing.

Inclusivity:

- Use inclusive and respectful language.
- Avoid bias or potentially offensive language.

Review and Edit:

- Review questions for grammatical accuracy and clarity.
- Edit to ensure adherence to the pattern.

Submission: Submit the created templates according to the provided guidelines, ensuring they are well-organized and labeled according to the corresponding pattern.

Compensation: Annotators will receive fair compensation, determined by task complexity and their demographic location.

Consent and Data Usage: By participating, you consent to the use of your contributions in our projects, with the possibility of inclusion in research publications.

Ethics and Approval: This project has ethics review board approval, adhering to standards for privacy, consent, and risk management.

Support: For questions or clarification, contact the project coordinators.

We recruited annotators through our institutions, employing seven individuals to create question templates for 586 unique patterns. Each annotator received fair compensation (about 100$ on the average), determined by the task's complexity and their outputs. We obtained informed consent from all participants, clearly explaining data usage and their rights. Additionally, our data collection protocol was approved by an ethics review board from our institution, affirming our adherence to ethical standards in privacy, consent, and risk management. For better exploration, we give several of the most common question templates in Table 10.

Relation 1	**Relation 2**	**Template Examples**
position held	significant event	Who held the position of $\{o_1\}$ after $\{s_2\}$ Who was the $\{o_1\}$ during $\{s_2\}$?
educated at	rector	Who was the rector of $\{o_1\}$ when $\{s_1\}$ was a student there? When $\{s_1\}$ graduated from $\{o_1\}$, the president stepped down in which year?
award received	spouse	When did the recipient of the $\{t_1\}$ $\{o_1\}$ Award wed $\{o_2\}$? Name the spouse of $\{s_1\}$ when he secured $\{o_1\}$.
nominated for	award received	Which honors did $\{s_1\}$ receive in the year $\{o_1\}$ nominated it? In what year was $\{s_1\}$ nominated by $\{o_1\}$ and awarded the $\{o_2\}$ award?
employer	chairperson	Who occupied the chairman's seat during the time $\{s_1\}$ was employed by $\{o_1\}$? During the year $\{t_1\}$, who was the president of the agency responsible for hiring $\{s_1\}$?
country	head of government	Who held the position of head of government at the time $\{s_1\}$ obtained $\{o_1\}$ citizenship? When did the head of government step down from office after $\{s_1\}$ obtained $\{o_1\}$ citizenship?

Table 10: Example templates for question generation.