TBEN: BENCHMARKING AND TESTING THE RULE BASED TEMPORAL LOGIC REASONING ABILITY OF LARGE LANGUAGE MODELS WITH DATALOGMTL

Anonymous authors

Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly adopted for a variety of tasks, including multi-hop question answering, knowledge probing, and symbolic commonsense reasoning. While LLMs have advanced the state-of-the-art in these areas, their ability to explicitly solve temporal logic reasoning problems—a complex cognitive process involving the understanding, representation, and manipulation of temporal information such as events, their durations, and relation-ships—remains largely unexplored. To enhance understanding of LLM performance in this common task widely explored in the traditional symbolic AI field, we have developed a new set of synthetic benchmark for rule-based temporal logic reasoning named TBEN. TBEN is developed within the context of DatalogMTL, a powerful knowledge representation language designed for reasoning about the properties of systems that evolve over time. Notably, this benchmark defined six levels of rule complexity and provides flexible configurations, allowing for the customization of temporal rules and task complexity to suit various needs. We evaluated the close-sourced GPT-40 and the open-sourced Llama-3¹ using

026 three common prompting settings—zero-shot, few-shot, and zero-shot-CoT—on our synthetic benchmark. Our key findings are as follows: (i) Without generating 028 the reasoning process (chain-of-thought), even advanced LLMs like GPT-40 ex-029 hibited nearly random performance on these rule-based temporal logic reasoning 030 tasks. However, with chain-of-thought prompting, LLMs demonstrated prelim-031 inary temporal logical reasoning abilities; (ii) Both GPT-40 and Llama-3 were 032 unable to solve temporal logical reasoning problems involving recursion, indicating a lack of advanced complex reasoning capabilities in understanding symbolic 034 representations involving time; (iii) There is significant room for improvement in leveraging large language models to address problems widely explored in the traditional logic-based AI domain. Prompts and datasets are available in the appendix, and a datasheet for TBEN is also provided.

038

005 006

007

008 009 010

011

013

014

015

016

017

018

019

021

023

025

040 1 INTRODUCTION

041 Time is a crucial and pervasive concept in our daily life and the real world that requires management 042 in various ways: events happen at specific moments, certain facts are valid over a time period, and 043 temporal relationships exist between facts and events Ozsoyoglu & Snodgrass (1995). Time must 044 be considered when representing information within computer-based systems, querying information about the temporal features of the real world, and reasoning about time-oriented data. In traditional 046 logic-based AI (also known as symbolic AI), researchers have developed a variety of rich knowledge representation rule languages, such as Linear Temporal Logic(LTL) Huth & Ryan (2004), Signal 047 Temporal Logic (STL) Donzé (2013) and DatalogMTL Brandt et al. (2018) to represent and reason 048 these timestamped events. he syntax, semantics, and complexity of these rule languages, along with their corresponding solvers, have been well developed over the past several decades. 050

In recent years, there is evidence that current large language models (LLMs) can perform fundamen tal data analysis tasks Reid et al. (2024); Bai et al. (2023) such as visualization Maddigan & Susnjak

¹We experimented with Meta-Llama-3-8B-Instruct.



Figure 1: Overview of two approaches (symbolic reasoner based and LLMs based) for solving the temporal reasoning. The former can use an existing symbolic reasoner (e.g., MeTeoR Wang et al. (2022)) to derive temporal facts, which are then used to judge whether the given facts are entailed. In contrast, the latter leverages the power of large language models to complete the reasoning process and provide the answer. In particular, we need to provide some background knowledge about the semantics of used notations, which is included in the prompt.

(2023); Cheng et al. (2023), commonsense reasoning Wang & Zhao (2023a); Zhao et al. (2024) and mathematical reasoning Ahn et al. (2024). However, little effort has been made to evaluate the more challenging aspects of rule-based temporal logic reasoning, a long-standing research problem in traditional logical artificial intelligence. While some studies have benchmarked or evaluated the temporal reasoning abilities of LLMs Wang & Zhao (2023b); Xiong et al. (2024), they primarily focus on reasoning over temporal data expressed in natural language, without addressing the temporal logic, which is typically represented as logical rules with well-established syntax and semantics.

090 In this paper, we evaluate the rule-based temporal logic reasoning abilities of large language models 091 (LLMs) from a logical perspective, considering the semantics of temporal logic. Our aim is to bridge 092 the gap between traditional logic-based AI and LLMs. Specifically, we focus on the classic tempo-093 ral reasoning problem known as *fact entailment* Cheng (1996); Brandt et al. (2018); Wałęga et al. (2023b). In our evaluation task, we consider temporal data of the symbolic form $P(a_1,\ldots,a_n)@\rho$, where P denotes a predicate (relation), a_i is an entity, n denotes the arity² and ρ represents a punctual time point or time interval. Given a set of temporal rules and a target temporal fact, the task is 096 to determine whether the fact is entailed by the temporal data and rules. To provide better intuition, 097 we use Example 1 togeter with Figure 2 to describe the problem. 098

Example 1. There is growing evidence that individuals develop COVID-19 immunity if they were infected within the last 6 months (discounting the last ten days when they had no symptom) Feikin et al. (2022). The condition can be captured by a DatalogMTL program Π_{ex} with the following rule:

102 103

076

077

078

079

081

103 104

 $Immune(x) \leftarrow \Leftrightarrow_{(10,183)} Infected(x) \land \exists_{[0,10]} NoSympt(x)$

¹⁰⁵ 106 107

²If the arity is 0, then P is treated as a statement that is either true or false. This differs from temporal knowledge graphs, which consist solely of quadruples (arity=2).

The above rule checks whether an individual infected at some point in the last six months excluding the last 10 days (operator $\diamond_{(10,183]}$) remained continuously without symptoms in the last 10 days (using the 'box past' operator $\boxminus_{[0,10]}$).

Then, we assume a dataset contains some historical data about a person called Ben in the form of facts stamped with validity intervals, where the first day of the year is given by the interval (0, 1], the second day by (1, 2], and so on. Ben got vaccinated at July 19 (represented as 199). Moreover, Ben had no symptoms since July 1 (i.e., 181) until August 30 (i.e., 242). This is represented by a dataset \mathcal{D}_{ex} with the following facts:

Infected(Ben)@199, NoSympt(Ben)@(181,242]

119 If we want to know whether Ben is immune between September 8 and September 9, represented 120 as a temporal fact Ben@(251,252], we can formulate this as a fact entailment problem: Is 121 Ben@(251,252] entailed by \mathcal{D}_{ex} and Π_{ex} ?

Traditionally, a symbolic reasoner is used to check entailment by applying temporal rules to tempo-123 ral data, deriving new facts, and verifying if the given fact is among the derived ones. As shown in 124 Figure 1, this process involves several steps before reaching a conclusion. We are exploring whether 125 large language models (LLMs) can solve such problems by providing correct answers along with 126 the reasoning process and human-readable textual explanations. On one hand, evaluating the tempo-127 ral reasoning capabilities of LLMs could complement evaluations of other reasoning skills, such as 128 mathematical and commonsense reasoning. On the other hand, using LLMs as a tool or a comple-129 mentary tool for temporal reasoning alongside symbolic reasoners could significantly contribute to 130 the development of traditional logic-based AI.

- 131 Our contributions are summarized as follows:
 - We propose a new benchmark, TBEN, to test the rule-based temporal logic reasoning abilities of large language models through synthetic tasks. Our benchmark generator allows for the creation of benchmarks with varying data sizes and rule structures of differing complexity, enabled by flexible configurations.
 - We conduct preliminary experiments using two common large language models on our constructed benchmarks. Our results demonstrate that, without advanced prompting techniques such as Chain-of-Thought (CoT), even GPT-40 performs almost at random. However, with CoT, GPT-40 shows a significant performance boost, while open-source models like Llama-3-8b continue to exhibit comparatively lower effectiveness in solving rule-based temporal logical reasoning problems.
 - We further conducted extensive analysis, including a detailed ablation study, natural language-based³ temporal logic reasoning experiments, error analysis, and human evaluation. Our analysis and key findings can serve as a foundation for LLM rule-based temporal logic reasoning benchmarking, and motivate further work to enhance these capabilities.

We will open source our developed TBEN to spur future research in improving the temporal logic reasoning ability of large language models.

152

117

118

122

133

134

135

136

137

138

139

140

141

142 143

144

145

146 147

148

DatalogMTL Brandt et al. (2018); Wałęga et al. (2019) is a powerful knowledge representation language, which extends Datalog Abiteboul et al. (1995), a widely used declarative logic programming language, with operators from metric temporal logic (MTL) Koymans (1990). Different Datalog designed to handle static facts and rules due to lack of built-in temporal constructs, DatalogMTL equipped with MTL operators is enabled to reasoning about properties of systems that evolve over time. These operators build upon the standard linear temporal logic (LTL) Huth & Ryan (2004) operators, such as \Leftrightarrow standing for "sometime in the past", \boxminus for "always in the past", and S for "since",

¹⁴⁹ 150 151

² DATALOGMTL

¹⁵⁹

 ³Manually converting each rule into natural language requires substantial human effort. Therefore, we propose an alternative approach: using customized templates to verbalize the temporal rules. The verbalized examples could be found in the appendix.

as well as their future counterparts \oplus for "sometime in the future", \boxplus for "always in the future", and *U* for "until". In MTL, however, these LTL operators are annotated with intervals; for instance, the expression $\oplus_{[1,2]} LiveIn(x, y)$ is true at time *t* if entity *x* lived in location *y* sometime between times t - 1 and t - 2. Similarly, $\boxminus_{[1,2]} LiveIn(x, y)$ holds at time *t* if *x* continuously lived in *y* throughout the aforementioned time interval. In this section, we recapitulate the syntax, semantics, and key temporal tereasoning problems in DatalogMTL.

Syntax We consider a *signature* consisting of pairwise disjoint countable sets of constants, variables, and predicates with non-negative integer arities. A term is either a constant or a variable. A *relational atom* is an expression of the form P(s), with P a predicate and s a tuple of terms whose length matches the arity of P. In this paper, we restrict ourselves to a fragment in which metric atoms are generated by the following grammar, where P(s) is a relational atom and ρ an interval including only non-negative numbers:

$$M ::= P(\mathbf{s}) \mid \Leftrightarrow_{\rho} M \mid \oplus_{\rho} M \mid \boxplus_{\rho} M \mid \boxplus_{\rho} M$$

A rule in this fragment is an expression of the form

 $P(\mathbf{s}) \leftarrow M_1 \wedge \dots \wedge M_n, \quad \text{for } n \ge 1,$ (1)

where the body atoms M_1, \ldots, M_n are metric atoms and the head atom P(s) is relational. A program is a finite set of rules.

Semantics An interpretation \mathfrak{I} is a function assigning truth values to ground relational atoms $P(\mathbf{c})$ and time points $t \in \mathbb{Z}$. It determines if $P(\mathbf{c})$ is satisfied at t, denoted as $\mathfrak{I}, t \models P(\mathbf{c})$, or not, denoted as $\mathfrak{I}, t \not\models P(\mathbf{c})$. This notion of truth assignment extends to other ground metric atoms in the considered fragment as follows:

$\mathfrak{I},t\models \Leftrightarrow_{\varrho}M$	iff	$\mathfrak{I}, t' \models M$ for some t' with $t - t' \in \varrho$,
$\mathfrak{I},t\models \oplus_{\varrho}M$	iff	$\mathfrak{I}, t' \models M$ for some t' with $t' - t \in \varrho$,
$\Im,t\models \exists_{\varrho}M$	iff	$\mathfrak{I}, t' \models M \text{ for all } t' \text{ with } t - t' \in \varrho,$
$\mathfrak{I}, t \models \boxplus_{\varrho} M$	iff	$\mathfrak{I}, t' \models M$ for all t' with $t' - t \in \varrho$.

For example, an interpretation making atom LiveIn(Ann, Paris) true everywhere within [10, 30] and false elsewhere makes $\exists_{[1,2]}LiveIn(Ann, Paris)$ true at the time point 31, but false at 32. An interpretation can be alternatively seen as the (possibly infinite) set of facts that it satisfies, which yields a natural meaning to containment and minimality of interpretations.

195 196 197

168

175 176

178

179

181 182

183

184

185

2.1 MAJOR TEMPORAL REASONING PROBLEMS

According to Brandt et al. (2018); Wałęga et al. (2019), temporal logic reasoning involves two major
 problems: consistency checking and fact entailment. Consistency checking is the task of determining
 whether a given program and dataset admit a common model Emerson (1990); Schnoebelen (2002).
 Fact entailment involves checking whether a program and dataset together entail a specific relational
 fact. Brandt et al. (2018) note that in *DatalogMTL*, consistency checking and fact entailment are
 complementary problems. Consequently, this paper focuses solely on the *fact entailment* problem to
 evaluate the temporal reasoning capabilities of large language models.

205 206

207

208

3 TBEN: A RULE-BASED TEMPORAL LOGIC REASONING BENCHMARK WITH DATALOGMTL

DatalogMTL is a powerful knowledge representation language that can characterize complex temporal conditions by defining various rules using combinations of different atoms and temporal operators (\diamond , \oplus , \equiv , \equiv) whose semantics has been described in Section 2. To some extent, the complexity of a *fact entailment* problem is largely determined by the complexity of associated temporal rules.

213

Challenges Although DatalogMTL is a powerful knowledge representation language, editing such
 rules requires considerable expertise and is very time-consuming. Currently, there are no suitable
 benchmarks in terms of size and rule diversity available for research. In the era of large language

models (LLMs), we are motivated to explore whether LLMs can assist in addressing temporal reasoning problems traditionally solved by symbolic reasoners. Therefore, developing data and rule generators that enable the automatic construction of benchmarks with flexible configurations for customized rule structures and task complexity is important and poses a substantial challenge.

Zero-shot Prompt Prefix

Given a dataset, temporal rules and a temporal fact, you need to apply the rules to the dataset and then judge whether the given fact is entailed by the dataset and rules. The rules are expressed as DatalogMTL, a language of temporal logic that extends Datalog with operators from metric temporal logic (MTL). The semantics of four MTL operators are given as follows: If $\Rightarrow_{[a,b]}A$ is true at the time t, it requires that A needs to be true at some time between t-b and t-a.

If $\exists_{[a,b]}A$ is true at the time t, it requires that A needs to be true continuously between t-b and t-a.

If $\oplus[a,b]A$ is true at the time t, it requires that A needs to be true at some point between t+a and t+b.

If $\boxplus[a,b]A$ is true at the time t, it requires that A needs to be true continuously between t+a and t+b.

Now, we have a data, some DatalogMTL rules and a fact entailment question. You should only output true or false, and please do not output other words.



Figure 2: Overview of the TBEN Benchmark, featuring six levels of temporal reasoning problems with varying complexity. We present an intuitive example representing each level, along with the corresponding rule, dataset, and fact entailment problem. A zero-shot-prompt prefix is also provided (see Appendix for additional prompt prefixes used in this paper). For better demonstration, we use the symbols \Leftrightarrow , \oplus , \Box , and Ξ , which are replaced by $\langle - \rangle$, $\langle + \rangle$, [-], and [+], respectively, in the actual prompts due to typing constraints.

246 247 248

249

265

220 221

222

224

225

226

227

228

229

230

231

232

233

234 235

237

238

239 240

241

242

243

244

245

3.1 LEVELING DATALOGMTL RULES

250 To address the aforementioned challenge and provide a more comprehensive evaluation of the temporal reasoning abilities of large language models, we aim to create a new synthetic benchmark with 251 flexible configurations for customizing rule structures and task complexity. We classify DatalogMTL rules into six classes (SingleAtom, ..., Recursive) based on their structural representations, consid-253 ering factors such as the number of body atoms, the number of temporal operators used, the number 254 of rules involved, and whether the rules are recursive. While we are unable to quantify the degree of complexity of each level, we assume that higher levels correspond to greater complexity. This as-256 sumption is based on the observation that more complex rule structures require additional temporal 257 reasoning steps when using a symbolic reasoner like MeTeoR Wang et al. (2022). 258

SingleAtom The most simplest form of a rule is $A \leftarrow \oslash_{[\rho]} B$, where \oslash could be one of the four metric temporal operators ($\Box, \Xi, \Leftrightarrow$ and \oplus). In particular, we ensure that A and B are two different atoms, so only one calculation operation. A **SingleAtom** example is given in Figure 2, where we can derive A@[5,7] based on the given dataset and the rule, entailing that A@6 is true. In particular, we consider DatalogMTL over the integer timeline, a fragment of DatalogMTL Wałęga et al. (2020) and use one type of MTL operator.

MultiAtoms In the **SingleAtom**, the body contains only one atom, so a single rule application is sufficient to complete the derivation. In **MultiAtoms**, we increase the number of atoms in the rule body, requiring not only the validation of each atom but also an intersection operation to obtain the final valid interval. As the example shown in Figure 2, the rule contains two atoms. First, we calculate the valid intervals for each atom. Based on the provided temporal facts, $\Box_{[3]}B$ holds only at the punctual time point [4, 4]], and $\exists_{[2,3]}A$ holds at the interval [4,5]. The intersection of these intervals, [4, 4] and [4, 5], is [4, 4]. Thus, we derive that A is true at the time point 4, so A@4 is entailed. As with **SingleAtom**, we consider DatalogMTL over the integer timeline Wałęga et al. (2020) and use only one type of MTL operator.

274

282

Rational Both SingleAtom and MultiAtoms focus solely on the integer timeline, which represents a relatively limited time space and simplifies reasoning due to the integer semantics Wałęga et al. (2020). In Rational, we build on top of MultiAtoms by expanding the timeline to include the rational numbers, incorporating decimal time points. Intuitively, rational-based numerical operations are more complex than their integer-based counterparts, and we aim to determine if large language models exhibit similar behavior. We continue to use only one type of MTL operator at this level.e
type of MTL operator in the level.

MixedOperators Using only one operator limits the expressiveness of DatalogMTL, preventing 283 the definition of complex temporal conditions. Thus, a natural expansion is to allow the use of MTL 284 operators. The four types of MTL operators can be used to define temporal conditions associated 285 with both the past and the future. A MixedOperators example is shown in Figure 2, which involves 286 two MTL operators (\Leftrightarrow and \boxplus). To complete the derivation, we first calculate the valid interval where 287 $\Leftrightarrow_{[1,2,4]}B$ with the past operator (\Leftrightarrow) holds, which is [2,3,4]. Then, we calculate $\boxplus_{[1,2]C}$, whose valid 288 interval is [1, 2]. After performing the interval intersection, we obtain that A holds at the time interval 289 [2, 2]. Thus, the temporal A@2.3 is not entailed. 290

291 **MultiRules** In the previous four levels, fact entailment is associated with only one temporal rule. 292 However, in more practical scenarios, multiple temporal rules may be required to express complex 293 temporal conditions. In this level, we consider a multi-rule temporal reasoning case, where fact 294 entailment involves multiple temporal rules and rule applications must be executed across these 295 rules to complete the derivation. As the example in Figure 2, to derive the target atom A, we need 296 to know both D and C. However, the dataset only provides the information about C. We can derive the D holds at 3 according to the first temporal rule $D \leftarrow \exists_{[2]} \land \exists_{[1,2]}C$; then, we can derive that A 297 holds at the interval [4.5, 5] according to the second rule. Hence, A@[4.5, 5] is entailed. 298

299

300 **Recursive** The fact entailment problem at this level is considered the hardest because it involves 301 recursion. Unlike static knowledge representation languages (e.g., Datalog), where all facts can be derived after a certain number of rule applications, some recursive rules in DatalogMTL may require 302 an infinite number of applications. Even for symbolic-based approaches, this presents a significant 303 challenge, and researchers have devoted considerable effort to addressing it Wałęga et al. (2021; 304 2023a). According to Wałęga et al. (2023a), in the recursive scenarios, periodic structures will ul-305 timately occur repeatedly, but calculating these periodic structures is challenging. From a human 306 perspective, however, identifying such periodic structures can be straightforward. For instance, con-307 sider a recursive rule \boxplus_{1year} Bday $(x) \leftarrow$ Bday(x), which states that anyone having their birthday at 308 a time point t will also be having their birthday at the same time the following year. If we know that 309 Ben has his birthday on Jun 8, 1991, it is easy to know that he will have his birthday on Jun 8, 1992, 310 Jun 8, 1993 and so on. However, this is difficult for traditional symbolic-based approaches to handle. 311 Therefore, we design fact entailment problems associated with recursive rules to test whether large language models can perform well in this setting. 312

Specifically, we use facts from both propositional logic Klement (2004) and first-order logic Barwise (1977). The former contains declarative statements that are either 'true' or 'false', while the latter includes expressions with one or more variables. For example, we allow both forms of temporal facts: Raining and Immune(x). The former states that an event (raining) is occurring, while the latter denotes that a property (immune) is associated with an entity, where x acts as a placeholder that can be instantiated to any entity, such as Immune(Ben), indicating that Ben is immune.

319

320 3.2 GENERATING TEMPORAL DATA AND RULES

321

The benchmark generation process can be mainly divided into the following three steps: 1) Graph construction, 2) Data generation, and 3) Rule generation. The pseudocode for this benchmark generation algorithm can be found in Part F of the appendix. Graph construction We employ a general-purpose random graph generator to generate a connected directed random graph. The nodes in the random graph represent predicates, such as A, B, and C. Each edge in this graph represents a body atom of a rule pointing to the corresponding head in the rule. In particular, a predicate can appear in bodies of multiple different rules.

Data generation After the construction of the graph, the program will traverse each nodes in the graph and randomly assign time points or time intervals to the chosen nodes. The time points or intervals are generated based on a given range.

Rule Generation Once the temporal data is generated, the rule generator traverses the edges of the graph, assigning random operators and intervals to the edges. To ensure the generated graph is non-trivial, a reasoning process is performed across the entire graph after completing this step to ensure new facts can be inferred. If multiple rules are required, the program repeats previous steps until a sufficient number of rules are generated.



Figure 3: An example of generating temporal data and rules. First, we randomly generate a graph. Next, our program selects specific nodes to assign time points. In our example, nodes B and D are chosen, resulting in two temporal facts: $\{B@[1,2], D@[2,3]\}$; Finally, we select a node as the head atom, with body atoms derived from the previous step. We then randomly assign temporal operators to these body atoms, resulting in the rule: $A \leftarrow \Leftrightarrow_{[1,2]} B \land \boxminus_{[1,1]} D$. The number of body atoms, the time range, and the temporal operators are specified as input parameters.

An example of generating temporal data and rules is shown in Figure 3. In particular, our program will have a post-processing operation to scan all the data and rules to ensure they have been utilized and removes any data and rules (in the ablation study, we will explore the impact of irrelevant data and rules) that are not participated in the the temporal reasoning process. We define the following flags for the samples to be generated based on their characteristics: rational number, multiple body atoms, recursive and mixed operators. These flags control the rule structures during the generation process.

	Prompt type	SingleAtom	MultiAtoms	Rational	MixedOperators	MultiRules	Recursive
	Zero-shot	45.8	43.2	37.1	57.3	53.3	37.7
GPT-40	Few-shot	40.4	38.0	27.2	51.6	36.7	32.2
	Zero-shot-CoT	85.6	85.1	85.7	90.3	74.0	58.0
	Zero-shot	40.7	44.0	43.9	60.5	39.1	8.7
Llama-3	Few-shot	38.4	44.3	44.4	47.1	36.1	30.2
	Zero-shot-CoT	59.9	58.4	68.2	64.1	59.0	48.5

Table 1: Model performance on the synthetic benchmarks across six rule structures, as defined in Section 3.1, shows that GPT-40 with chain-of-thought prompting significantly outperforms its counterparts—zero-shot and few-shot—by approximately 30% to 50. This indicates that it is crucial for LLMs to generate the reasoning process before arriving at the final conclusion. Specifically, we observe that the three prompting settings for Llama-3 achieve similar performance, suggesting that Llama-3 struggles with temporal reasoning in symbolic forms.



Figure 4: F1 score (%) on three benchmarks with *rule-based* and natural language (NL) based temporal logic reasoning and using Zero-shot-CoT.

4 EXPERIMENTS AND RESULTS

Based on our synthesized benchmarks, we aim to investigate *whether large language models can* solve temporal reasoning problems from the traditional logic-based AI domain by evaluating two large language models (GPT-40⁴ and Llama-3⁵) with three different kinds of prompting settings.

Baselines We adopt three basic prompting approaches as baselines. Specifically, zero-shot prompting (ZERO-SHOT), few-shot in-context learning (FEW-SHOT) (Brown et al., 2020), chain-of-thought prompting (COT) (Wei et al., 2022), and zero-shot chain-of-thought (0-COT) (Kojima et al., 2022) are leveraged to tackle various temporal reasoning problems in our constructed benchmark.

Benchmark statistics and experimental settings Unless otherwise specified, each benchmark level contains 200 samples selected from the facts derived using the chosen data and rule(s). For negative samples, a random interval is chosen, ensuring that these intervals do not overlap with those of the derived facts. Specifically, for all baselines, the temperature value is set to 0. For few-shot prompting techniques, the input prompt includes two manually constructed exemplars. In this paper, we use the F1 score as the evaluation metric. Details about the benchmark statistics, instructions, exemplars, and other relevant experimental information can be found in the the Appendix part.

408 409

410

386

387 388 389

390 391 392

393

394 395 396

397

398

399 400

4.1 MAIN RESULTS

We observe that for GPT-40, the *zero-shot* and *few-shot* prompting settings achieve extremely low 411 accuracy across the six benchmarks, with some experiencing more than a 50% accuracy drop com-412 pared to zero-shot-CoT. This indicates the necessity of using the chain-of-thought strategy with 413 LLMs to solve temporal reasoning problems effectively. Notably, on the three benchmarks Sin-414 gleAtom, MultiAtoms, and Rational, we observe similar accuracy with less than 2% variance 415 between them. Interestingly, we found that using multiple operators yields the best performance. 416 This is surprising because we assumed that employing more operators would require the model to 417 understand more semantics of DataloMTL, thus increasing the problem's complexity compared to 418 using only one type of MTL operator. However, considering the reasoning process using the sym-419 bolic reasoner, such results are also reasonable because if two rules contain the same number of body atoms, the number of calculating operators is the same. From this perspective, LLMs exhibit 420 similar behavior to symbolic reasoner-based approaches. 421

422 In the **MultiRules** and **Recursive** benchmarks, we notice a significant performance drop, with 423 **Recursive** achieving the lowest performance at 64.5%. These results suggest that recursive rules are 424 particularly challenging for models, as they require not only understanding the language semantics 425 and performing step-by-step reasoning but also possessing strong *inductive ability*. Unfortunately, 426 the results achieved by Llama-3 are almost random, indicating that the model lacks the capability to solve symbolic form temporal reasoning problems. One possible reason is the model size being 427 too small, preventing it from generalizing to unseen tasks. Additionally, the lack of relevant training 428 datasets during pre-training or fine-tuning stages could also contribute to this issue. 429

⁴https://openai.com/index/hello-gpt-40/

⁵https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

432 4.2 RULE BASED VS. NATURAL LANGUAGE (NL) BASED TEMPORAL LOGIC REASONING

In this paper, our primary focus is on benchmarking and testing the temporal logic reasoning ca-pabilities of LLMs by evaluating them in a rule-based format. However, it is also interesting to compare this with another setting: verbalizing the rules before presenting them to the LLMs, as explored in many previous similar works like CronQuestions Saxena et al. (2021); Ismayilzada et al. (2023). Given that manually converting each rule into its corresponding natural language expression is a labor-intensive process, we adopt a template-based approach to automate this verbalization. Al-though this method may result in expressions that are not entirely natural⁶, it provides a practical alternative to manual translation.

From Figure 4, we observe that both the rule-based and natural language-based settings achieve similar results, with the rule-based approach performing slightly better. The comparison indicates that LLMs are also capable of understanding the semantics of input expressed in rules, provided that each notation is clearly explained in the instructions. Notably, both settings struggle with the MultiRules and Recursive cases. One possible reason for this is that, while LLMs can understand the semantics of temporal logic language, they still face significant challenges in executing multi-ple deductions, retaining intermediate results, and recognizing repeated patterns-tasks that require delicate algorithms to accomplish effectively in the field of temporal logic Wałęga et al. (2023a).

4.3 ABLATION STUDY

To explore which component of the rule structure most significantly impact the reasoning complex-ity for LLMs, we designed four sets of ablation study experiments. These experiments explored the effects of the number of relevant rules, the number of operators considered, the percentage of irrele-vant data, and the percentage of irrelevant rules. From Figure 5 (a), we observe that as the lengths of dependent rules increase, the model's performance noticeably degrades. One possible reason is that when multiple rules are mutually dependent, the model needs to store intermediate results during the derivation process to complete subsequent steps that rely on previously derived outcomes. Unlike symbolic reasoners, which can explicitly store intermediate results, it may be challenging for large language models (LLMs) to retain such information in an auto-regressive manner. Additionally, Figure 5 (b) demonstrates that using more types of operators does not affect reasoning complexity, indicating that understanding the semantics of the temporal logic language is not a major issue for the model. Lastly, the results in Figures 5 (c) and (d) show that the model's performance is only minimally affected by irrelevant information, whether in the form of temporal data or rules, demon-strating its ability to correctly select relevant rules and remain resistant to distracting information.



Figure 5: Results of ablation study. (a) presents the benchmark results for temporal rules of varying lengths; (b) illustrates the benchmark results for different operators; (c) displays the benchmark results with varying percentages of irrelevant data relative to relevant data (e.g., if there are 10 temporal facts required for the temporal reasoning process, 100% denotes that an additional 10 irrelevant temporal facts are included in the input); (d) showcases the benchmark results with varying percentages of irrelevant rules relative to target rules (the meaning of percentages in the x-axis is similar to the setting of adding irrelevant temporal facts.

⁶Some examples of these conversions can be found in the Appendix.

486 5 RELATED WORKS

487 488

Symbolic approaches for temporal reasoning Knowledge representation languages, such as Lin-489 ear Temporal Logic (LTL) Huth & Ryan (2004) and DatalogMTL Brandt et al. (2018), have become 490 the de facto standard for specifying temporal properties in both formal verification and artificial in-491 telligence. Many temporal reasoning problems have proven to be PSPACE-complete Wałega et al. (2019); Fionda & Greco (2018); Bauland et al. (2009); Wałega et al. (2020). Satisfiability checking, 492 that is, the problem of deciding whether a given formula admits a satisfying model, is one of the most 493 important computational tasks associated with the logic, and one of the first that have been carefully 191 studied Sistla & Clarke (1985). Similarly, the main reasoning tasks considered in DatalogMTL are 495 fact entailment and consistency checking. These problems polynomially reduce to the complements 496 of each other Brandt et al. (2018). Despite this theoretically high computational complexity, numer-497 ous techniques and tools have been developed to solve different temporal reasoning problems, which 498 range from tableau systems Goré & Widmann (2009); Bertello et al. (2016) to reductions to model 499 checking Cavada et al. (2014), to automata techniques Li et al. (2014); Wang et al. (2022). 500

501 **Prompting LLMs for temporal reasoning** Although the aforementioned temporal reasoning 502 problems have been widely explored in the traditional logic-based AI domain, they remain underexplored in the regime of LLMs. There is a substantial body of research evaluating the reasoning 504 abilities of LLMs in an in-context learning setting, covering areas such as arithmetic reasoning, logical reasoning, and commonsense reasoning. Notably, simple math problem datasets like AQUA Ling 505 et al. (2017), GSM8K (Cobbe et al., 2021), and SVAMP (Patel et al., 2021) are frequently used to 506 assess arithmetic reasoning (Touvron et al., 2023; Shi et al., 2023). Welleck et al. (2021) developed 507 Natural Proofs, a multi-domain dataset for studying mathematical reasoning in natural language, 508 while Welleck et al. (2022) investigated LLMs' abilities to generate the next step in mathemati-509 cal proofs and complete full proofs. Additionally, LLMs have been evaluated on logical reasoning 510 tasks, including symbolic tasks like Coin Flip and Last Letter Concatenation (Wei et al., 2022), and 511 Logic Grid Puzzles on the BIG-BENCH (Srivastava et al., 2023). Commonsense reasoning datasets 512 (Talmor et al., 2019) have also been proposed for evaluating LLMs. Most relevant to our work are 513 various approaches to evaluating and enhancing the algorithmic reasoning abilities of LLMs (Zhou 514 et al., 2022). In this work, we focus on evaluating LLMs on temporal logic reasoning tasks.

515 516

517

6 CONCLUSION

518 We present TBEN, a new synthetic benchmark consisting of rule-based temporal logic reasoning 519 problems, designed to test the temporal reasoning abilities of LLMs. TBEN contains diverse tem-520 poral reasoning problems of varying complexities, providing flexible configurations for customising 521 generated benchmarks. Our experimental results reveal that models have difficulties in understand-522 ing symbolic data and performing temporal reasoning. Even though powerful LLMs like GPT-40 523 have been trained on extensive data from across the Internet, they struggle with some simple temporal reasoning tasks. By proposing this benchmark, we aim to draw attention to the data-based 524 temporal reasoning abilities of LLMs. Our goal is to facilitate the development of specialized meth-525 ods to enhance these abilities, such as training models in the agent reasoning style with data as input 526 and equipping models with more temporal reasoning and symbolic knowledge. 527

528 529

530

7 LIMITATION

531 Our experiments were constrained by the speed, computational resources, and financial costs asso-532 ciated with utilizing the closed-source GPT-40. For instance, although our generator allows for the 533 creation of benchmarks of temporal data and rules with arbitrary sizes, we obtained results across 534 multiple temporal reasoning datasets of varying complexities on a relatively small scale due to the 535 financial costs associated with GPT-40 API calls.

Another limitation of this preliminary exploration into testing the temporal reasoning abilities of
 LLMs is that we only present experimental results from three prompting settings, despite the avail ability of many other advanced prompting strategies. Furthermore, while we demonstrate that a
 chain-of-thought approach is necessary for solving temporal reasoning problems, we do not offer an
 effective method for enhancing the LLM's ability to handle temporal logic reasoning problems.

540 REFERENCES 541

546

547

558

570

575

581

582

542	Serge Abiteboul, Richard Hull, and Victor Vianu	. Foundations of Databases.	Addison-Wesley,
543	1995.		

- 544 Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. Large language models for mathematical reasoning: Progresses and challenges. arXiv preprint arXiv:2402.00157, 2024.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. arXiv preprint arXiv:2309.16609, 2023. 548
- 549 Jon Barwise. An introduction to first-order logic. In Studies in Logic and the Foundations of 550 Mathematics, volume 90, pp. 5-46. Elsevier, 1977. 551
- Michael Bauland, Martin Mundhenk, Thomas Schneider, Henning Schnoor, Ilka Schnoor, and 552 Heribert Vollmer. The tractability of model-checking for ltl: The good, the bad, and the ugly 553 fragments. Electronic Notes in Theoretical Computer Science, 231:277–292, 2009. 554
- 555 Matteo Bertello, Nicola Gigante, Angelo Montanari, Mark Reynolds, et al. Leviathan: A new Itl 556 satisfiability checking tool based on a one-pass tree-shaped tableau. In IJCAI-International Joint Conference on Artificial Intelligence, pp. 950–956. AAAI Press, 2016.
- Sebastian Brandt, Elem Gužel Kalaycı, Vladislav Ryzhikov, Guohui Xiao, and Michael Za-559 kharyaschev. Querying log data with metric temporal logic. Journal of Artificial Intelligence 560 Research, 62:829-877, 2018. 561
- 562 Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, 563 Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020. 564
- 565 Roberto Cavada, Alessandro Cimatti, Michele Dorigatti, Alberto Griggio, Alessandro Mariotti, An-566 drea Micheli, Sergio Mover, Marco Roveri, and Stefano Tonetta. The nuxmv symbolic model 567 checker. In Computer Aided Verification: 26th International Conference, CAV 2014, Held as Part 568 of the Vienna Summer of Logic, VSL 2014, Vienna, Austria, July 18-22, 2014. Proceedings 26, pp. 569 334-342. Springer, 2014.
- Jingde Cheng. The fundamental role of entailment in knowledge representation and reasoning. 571 Journal of Computing and Information, 2(1):853–873, 1996. 572
- 573 Liying Cheng, Xingxuan Li, and Lidong Bing. Is gpt-4 a good data analyst? In Findings of the 574 Association for Computational Linguistics: EMNLP 2023, pp. 9496–9514, 2023.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, 576 Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to 577 solve math word problems. arXiv preprint arXiv:2110.14168, 2021. 578
- 579 Alexandre Donzé. On signal temporal logic. In Runtime Verification: 4th International Conference, 580 RV 2013, Rennes, France, September 24-27, 2013. Proceedings 4, pp. 382–383. Springer, 2013.
 - E Allen Emerson. Temporal and modal logic. In Formal Models and Semantics, pp. 995–1072. Elsevier, 1990.
- 584 Daniel R. Feikin, Melissa M. Higdon, Laith J. Abu-Raddad, Nick Andrews, Rafael Araos, Yair 585 Goldberg, Michelle J. Groome, Amit Huppert, Katherine L. O'Brien, Peter G. Smith, Annelies Wilder-Smith, Scott Zeger, Maria Deloria Knoll, and Minal K. Patel. Duration of effectiveness 586 of vaccines against sars-cov-2 infection and covid-19 disease: Results of a systematic review and meta-regression. The Lancet, 399(10328):924-944, 2022. doi: 10.1016/S0140-6736(22)00152-0. 588
- 589 Valeria Fionda and Gianluigi Greco. Ltl on finite and process traces: Complexity results and a 590 practical reasoner. Journal of Artificial Intelligence Research, 63:557–623, 2018. 591
- Rajeev Goré and Florian Widmann. An optimal on-the-fly tableau-based decision procedure for 592 pdl-satisfiability. In International Conference on Automated Deduction, pp. 437–452. Springer, 2009.

617

618

619 620

621

622

628

638

639

640

641

- Michael Huth and Mark Ryan. Logic in Computer Science: Modelling and reasoning about systems. 595 Cambridge university press, 2004. 596 597 Mete Ismayilzada, Debjit Paul, Syrielle Montariol, Mor Geva, and Antoine Bosselut. Crow: Benchmarking commonsense reasoning in real-world tasks. arXiv preprint arXiv:2310.15239, 2023. 598 Kevin C Klement. Propositional logic. 2004. 600 601 Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large 602 language models are zero-shot reasoners. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, 603 and Kyunghyun Cho (eds.), Advances in Neural Information Processing Systems, 2022. URL 604 https://openreview.net/forum?id=e2TBb5y0yFf. 605 606 Ron Koymans. Specifying real-time properties with metric temporal logic. *Real-time Systems*, pp. 607 255-299, 1990. 608 Jianwen Li, Yinbo Yao, Geguang Pu, Lijun Zhang, and Jifeng He. Aalta: an ltl satisfiability checker 609 over infinite/finite traces. In Proceedings of the 22nd ACM SIGSOFT international symposium on 610 foundations of software engineering, pp. 731–734, 2014. 611 612 Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale gen-613
- wang Ellig, Dam Togatana, Chris Dyer, and Film Brunson. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 158–167, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1015. URL https://aclanthology.org/P17-1015.
 - Paula Maddigan and Teo Susnjak. Chat2vis: Generating data visualisations via natural language using chatgpt, codex and gpt-3 large language models. *Ieee Access*, 2023.
 - Gultekin Ozsoyoglu and Richard T Snodgrass. Temporal and real-time databases: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 7(4):513–532, 1995.
- Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple
 math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2080–
 2094, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.
 naacl-main.168. URL https://aclanthology.org/2021.naacl-main.168.
- Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste
 Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini
 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.
- Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. Question answering over temporal knowledge graphs. *arXiv preprint arXiv:2106.01515*, 2021.
- Philippe Schnoebelen. The complexity of temporal logic model checking. *Advances in modal logic*, 4(393-436):35, 2002.
 - Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pp. 31210–31227. PMLR, 2023.
- A Prasad Sistla and Edmund M Clarke. The complexity of propositional linear temporal logics. *Journal of the ACM (JACM)*, 32(3):733–749, 1985.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

648 649 650 651 652 653	Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In <i>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)</i> , pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421.
654 655 656 657	 Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. <i>arXiv preprint arXiv:2302.13971</i>, 2023.
658 659 660	Przemysław A Wałęga, Bernardo Cuenca Grau, Mark Kaminski, and Egor V Kostylev. Datalogmtl over the integer timeline. In <i>Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning</i> , volume 17, pp. 768–777, 2020.
661 662 663	Przemysław A Wałęga, Michał Zawidzki, and Bernardo Cuenca Grau. Finitely materialisable data- log programs with metric temporal operators. In <i>Proceedings of the International Conference on</i> <i>Principles of Knowledge Representation and Reasoning</i> , volume 18, pp. 619–628, 2021.
665 666 667	Przemysław A Wałęga, Michał Zawidzki, Dingmin Wang, and Bernardo Cuenca Grau. Materialisation-based reasoning in datalogmtl with bounded intervals. In <i>Proceedings of the AAAI</i> <i>Conference on Artificial Intelligence</i> , volume 37, pp. 6566–6574, 2023a.
668 669 670	Przemysław Andrzej Wałęga, Michał Zawidzki, and Christoph Haase. Computing all facts entailed by an ltl specification. In <i>Proceedings of the International Conference on Principles of Knowledge</i> <i>Representation and Reasoning</i> , volume 19, pp. 679–689, 2023b.
672 673 674	Dingmin Wang, Pan Hu, Przemysław Andrzej Wałęga, and Bernardo Cuenca Grau. Meteor: Practi- cal reasoning in datalog with metric temporal operators. In <i>Proceedings of the AAAI Conference</i> <i>on Artificial Intelligence</i> , volume 36, pp. 5906–5913, 2022.
675 676 677	Yuqing Wang and Yun Zhao. Gemini in reasoning: Unveiling commonsense in multimodal large language models. <i>arXiv preprint arXiv:2312.17661</i> , 2023a.
678 679	Yuqing Wang and Yun Zhao. Tram: Benchmarking temporal reasoning for large language models. <i>arXiv preprint arXiv:2310.00835</i> , 2023b.
680 681 682 683	Przemysław Andrzej Wałęga, B Cuenca Grau, Mark Kaminski, and Egor V Kostylev. DatalogMTL: Computational complexity and expressive power. In <i>Proceedings of the International Joint Con-</i> <i>ferences on Artificial Intelligence</i> , 2019.
684 685 686 687 688	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), <i>Ad-</i> <i>vances in Neural Information Processing Systems</i> , 2022. URL https://openreview.net/ forum?id=_VjQlMeSB_J.
689 690 691 692	Sean Welleck, Jiacheng Liu, Ronan Le Bras, Hannaneh Hajishirzi, Yejin Choi, and Kyunghyun Cho. Naturalproofs: Mathematical theorem proving in natural language. In <i>Thirty-fifth Conference on</i> <i>Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)</i> , 2021. URL https://openreview.net/forum?id=Jvxa8adr3iY.
693 694 695 696 697	Sean Welleck, Jiacheng Liu, Ximing Lu, Hannaneh Hajishirzi, and Yejin Choi. Naturalprover: Grounded mathematical proof generation with language models. In Alice H. Oh, Alekh Agar- wal, Danielle Belgrave, and Kyunghyun Cho (eds.), <i>Advances in Neural Information Processing</i> <i>Systems</i> , 2022. URL https://openreview.net/forum?id=rhdfTOiXBng.
698 699 700	Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. Large language models can learn temporal reasoning. <i>arXiv preprint arXiv:2401.06853</i> , 2024.

701 Zirui Zhao, Wee Sun Lee, and David Hsu. Large language models as commonsense knowledge for large-scale task planning. *Advances in Neural Information Processing Systems*, 36, 2024.

702 Hattie Zhou, Azade Nova, Hugo Larochelle, Aaron Courville, Behnam Neyshabur, and 703 Hanie Sedghi. Teaching algorithmic reasoning via in-context learning. arXiv preprint 704 arXiv:2211.09066, 2022. 705

PROMPTS USED IN THE BASIC EVALUATION Α

709 For all evaluations, we prepend a system message to introduce the syntax of DatalogMTL language 710 as below:

712 You are given a dataset and a temporal rule, and your task is to judge whether the given fact is entailed by the dataset and the rule. 713 The rules are expressed as DatalogMTL, a knowledge representation language that extends Dat-714 alog with operators from metric temporal logic (MTL). The semantics of four MTL operators are 715 given as follows: 716 If Diamondminus[a,b]A is true at the time t, it requires that A needs to be true at some time 717 between t-b and t-a. 718 If Boxminus[a,b]A is true at the time t, it requires that A needs to be true continuously between 719 t-b and t-a. 720 If Diamondplus[a,b]A is true at the time t, it requires that A needs to be true at some point 721 between t+a and t+b. 722

If Boxplus[a,b]A is true at the time t, it requires that A needs to be true continuously between t+a and t+b.

727

728

729 730

731

706

707 708

711

Zero-shot For zero-shot evaluations, the system prompt we uses is the above general introduction plus the statement: You should not give any explanation and you should only output "true" or "false". We are using the statement Now we have some temporal data and some rules, data: {data} rule: *{rule}, Is {inquiry} true or not?* as the user prompt to evaluate LLM's reasoning ability.

Here is an example of the complete prompt we constructed to do zero-shot evaluation.

732		
700	System Prompt	You are given a dataset and a temporal rule, and your task is to judge whether
133		the given fact is entailed by the dataset and the rule.
734		The rules are expressed as DatalogMTL, a knowledge representation lan-
735		guage that extends Datalog with operators from metric temporal logic
736		(MTL). The semantics of four MTL operators are given as follows:
737		If Diamondminus[a,b]A is true at the time t, it requires that A needs to be
738		true at some time between t-b and t-a.
739		If Boxminus[a,b]A is true at the time t, it requires that A needs to be true
740		continuously between t-b and t-a.
741		If Diamondplus[a,b]A is true at the time t, it requires that A needs to be true
742		at some point between t+a and t+b.
743		If Boxplus[a,b]A is true at the time t, it requires that A needs to be true continuously between the and the
744		Vou should not give one evaluation and you should only output "true" or
745		"false"
746	User Prompt	Now we have some temporal data and some rules, data: B@[3,10]
747	1	rule: A:-Diamondplus[6,10]B
748		Is $A@[1,4]$ true or not?
749	LLM's output	false
750	Expected Answer	true

751 752

753 **Few-shot** For few-shot evaluations, just like the zero-shot case, the system prompt we uses is the above general introduction plus the statement: You should not give any explanation and you should 754 only output "true" or "false". However, in the user prompt, we are integrating some examples using 755 the following syntax:

To help you better understand the task, I will provide two examples.
Example 1: data: {pos data} rule: {pos rule} in this case you should output "true" for {pos inquiry}.
Example 2: data: {neg data} rule: {neg rule} in this case you should output "false" for {neg inquiry}.
Now we have some temporal data and some rules, data: {data} rule: {rule}
Is {inquiry} true or not?"

{pos data}, {pos rule} and *{pos inquiry}* are from a positive sample, *{neg data}, {neg rule}* and *{neg inquiry}* are from a negative sample. They are samples not in the testing set, but has the same type as the testing samples.

Here is an example of the complete prompt we constructed to do few-shot evaluation.

System Prompt User Prompt	You are given a dataset and a temporal rule, and your task is to judge whether the given fact is entailed by the dataset and the rule. The rules are expressed as DatalogMTL, a knowledge representation lan- guage that extends Datalog with operators from metric temporal logic (MTL). The semantics of four MTL operators are given as follows: If Diamondminus[a,b]A is true at the time t, it requires that A needs to be true at some time between t-b and t-a. If Boxminus[a,b]A is true at the time t, it requires that A needs to be true continuously between t-b and t-a. If Diamondplus[a,b]A is true at the time t, it requires that A needs to be true at some point between t+a and t+b. If Boxplus[a,b]A is true at the time t, it requires that A needs to be true continuously between t+a and t+b. If Boxplus[a,b]A is true at the time t, it requires that A needs to be true continuously between t+a and t+b. To help you better understand the task, I will provide two examples. Example 1: data: B@[5,7] rule: A:-Boxminus[10,12]B in this case you should output "true" for A@[17,17]			
	in this case you should output "true" for A@[17,17] Example 2: data: B@[1,9]			
	rule: A:-Diamondplus[3,3]B			
	In this case you should output faise for $A \ll [-25, -0]$ Now we have some temporal data and some rules data: $B @ [3, 10]$			
	rule: A:-Diamondplus[6.10]B			
	Is A@[1,4] true or not?"			
LLM's output	false			
Expected Answer	true			

Zero-shot-CoT For zero-shot-cot evaluations, the system prompt we uses is the above general introduction without the additional the statement that we used in zero-shot or few-shot to ask LLM's to answer directly. We are going to put some directions to let LLM think step by step in the user prompt below:

Now we have some temporal data and some rules, data: {data}	
rule: {rule}	
Is {inquiry} true or not? Do not answer directly, think step by step.	

After that, we will put the LLM's response in the context and then asks LLM again about the final answer by *Based on your previous response, without any explanation state your answers to the question. You should only output "true" or "false".*

Here is an example of the complete prompt we constructed to do zero-shot evaluation.

Q10		X7 1 1 1 1 1 1 1 1 1 1
010	System Prompt	You are given a dataset and a temporal rule, and your task is to judge whether
011		the given fact is entailed by the dataset and the rule.
812		The rules are expressed as DatalogMTL, a knowledge representation lan-
813		guage that extends Datalog with operators from metric temporal logic
814		(MTL). The semantics of four MTL operators are given as follows:
815		If Diamondminus[a,b]A is true at the time t, it requires that A needs to be
816		true at some time between t-b and t-a.
817		If Boxminus[a,b]A is true at the time t, it requires that A needs to be true
818		continuously between t-b and t-a.
819		If Diamondplus[a,b]A is true at the time t, it requires that A needs to be true
220		at some point between t+a and t+b.
020		If Boxplus[a,b]A is true at the time t, it requires that A needs to be true
821		continuously between t+a and t+b.
822	User Prompt	Now we have some temporal data and some rules, data: B@[3,10]
823		rule: A:-Diamondplus[6,10]B
824		Is A@[1,4] true or not? Do not answer directly, think step by step.
825	LLM's output	First, let's understand the rule: A:-Diamondplus[6,10]B. This rule means
826		that if B is true at some point between t+6 and t+10, then A is true at time t.
827		Now, let's look at the data: $B@[3,10]$. This means that B is true at some
828		point between time 3 and time 10.
829		We are asked to determine if $A@[1,4]$ is true or not. This means we are
830		looking to see if A is true at some point between time 1 and time 4.
831		Given the rule, for A to be true at time 1, B would need to be true at some
001		point between time 7 (1+6) and 11 (1+10). For A to be true at time 4, B
002		would need to be true at some point between time $10 (4+6)$ and $14 (4+10)$.
833		However, B is only true between time 3 and time 10. Therefore, B is not true
834		at any point between time 7 and 11 or between time 10 and 14.
835		So, A@[1,4] is not true.
836	User Prompt	Based on your previous response, without any explanation state your an-
837		swers to the question. You should only output "true" or "false"
838	LLM's output	false
839	Expected Answer	true

B PROMPTS USED IN THE NATURAL LANGUAGE EVALUATION

For the evaluations using natural language, we no longer need to introduce the syntax of DatalogMTL language. Therefore there is no general system prompt.

Zero-shot For zero-shot setting, we change the system prompt to a simple sentence to ensure that LLMs output in a desired format *You should not give any explanation and you should only output "true" or "false".* We are using the statement *Now we have some temporal data and some rules, data: {data} rule: {rule}, Is {inquiry} true or not?* as the user prompt to evaluate LLM's reasoning ability. {data}, {rule} and {inquiry} are all replaced by their verbalized representation.

Here is an example of the complete prompt we constructed to do zero-shot evaluation.

Syster	m Prompt	You should not give any explanation and you should only output "true" or			
		"false"			
User	Prompt	Now we have some temporal data and some rules, data:			
		A holds From 10.000 to 10.000			
		rule: B holds in each time such that A will hold sometime between 4.000			
		and 15.000 hours in the future			
		Is B holds From -5.000 to 1.000 true or not?			
LLM	's output	false			
Expect	ed Answer	true			

Few-shot For few-shot evaluations, just like the zero-shot case, the system prompt we uses is the same: *You should not give any explanation and you should only output "true" or "false"*. However, in the user prompt, we are integrating some examples using the following syntax:

To help you better understand the task, I will provide two examples. Example 1: data: {pos data} rule: {pos rule} in this case you should output "true" for {pos inquiry}. Example 2: data: {pog data} rule: {pog rule} in this case you should output "false" for {pog

Example 2: data: {neg data} rule: {neg rule} in this case you should output "false" for {neg inquiry}.

Now we have some temporal data and some rules, data: {data} rule: {rule}

{pos data}, {pos rule} and *{pos inquiry}* are verbalized representations from a positive sample, *{neg data}, {neg rule}* and *{neg inquiry}* are verbalized representations from a negative sample. They are samples not in the testing set, but has the same type as the testing samples.

Here is an example of the complete prompt we constructed to do few-shot evaluation.

System Prompt	You should not give any explanation and you should only output "true" or "false"		
User Prompt	To help you better understand the task, I will provide two examples.		
_	Example 1: data:		
	B holds From 5.000 to 7.000		
	rule: A holds in each time such that B did hold continuously for 10.000 and		
	12.000 hours before		
	in this case you should output "true" for A holds From 17.000 to 17.000		
	Example 2: data:		
	B holds From 1.000 to 9.000		
	rule: A holds in each time such that B will hold sometime between 3.000		
	and 3.000 hours in the future		
	in this case you should output "false" for A holds From -25.000 to -6.000		
	Now we have some temporal data and some rules, data:		
	B holds From 3.000 to 10.000		
	rule: A holds in each time such that B will hold sometime between 6.000		
	and 10.000 hours in the future		
	Is A holds From 1.000 to 4.000 true or not?		
LLM's output	true		
Expected Answer	true		

Zero-shot-CoT For verbalized zero-shot-cot evaluations, we no longer use system message since we don't need to introduce the synatx nor ask it to answer in a specific format. We organize our user prompts in the following format:

Now we have some temporal data and some rules, data: {data} rule: {rule} Is {inquiry} true or not? Do not answer directly, think step by step.

⁹¹³ where {data}, {rule} and {inquiry} refers to the verbalized rules.

After that, we will put the LLM's response in the context and then asks LLM again about the final answer by *Based on your previous response, without any explanation state your answers to the question. You should only output "true" or "false".*

Here is an example of the complete verbalized prompt we constructed to do zero-shot evaluation.

18	System Prompt	
19	User Prompt	Now we have some temporal data and some rules, data:
20		B holds From 3.000 to 10.000
21		rule: A holds in each time such that B will hold sometime between 6.000
22		and 10.000 hours in the future
23		Is A holds From 1.000 to 4.000 true or not? Do not answer directly, think
24		step by step.
25	LLM's output	To determine if A holds from 1.000 to 4.000 based on the given rule and
26		data, let's break it down step by step:
27		1. **Understand the Data:**
28		- B holds from 3.000 to 10.000.
20		2. **Understand the Rule:**
120		- A holds at each time such that B will hold sometime between 6.000 and
21		10.000 hours in the future.
20		3. ***Analyze the Time Intervals:*** We need to sheak if P holds in the interval [6,000, 10,000] hours in the
132		- we need to check if B holds in the interval [0.000, 10.000] hours in the future from any time within [1.000, 4.000]
133		1 uture noin any time within $[1.000, 4.000]$.
034		- For time 1 000:
135		- 6 000 to 10 000 hours in the future from 1 000 is [7 000, 11 000]
)36		- B holds from 3 000 to 10 000, which overlaps with [7 000, 10 000]
937		- Therefore B holds sometime in [7 000 11 000]
)38		- For time 2.000:
39		- 6.000 to 10.000 hours in the future from 2.000 is [8.000, 12.000].
40		- B holds from 3.000 to 10.000, which overlaps with [8.000, 10.000].
41		- Therefore, B holds sometime in [8.000, 12.000].
42		- For time 3.000:
43		- 6.000 to 10.000 hours in the future from 3.000 is [9.000, 13.000].
44		- B holds from 3.000 to 10.000, which overlaps with [9.000, 10.000].
45		- Therefore, B holds sometime in [9.000, 13.000].
46		- For time 4.000:
47		- 6.000 to 10.000 hours in the future from 4.000 is [10.000, 14.000].
48		- B holds from 3.000 to 10.000, which overlaps with [10.000, 10.000].
49		- Therefore, B holds sometime in [10.000, 14.000].
50		5. **Conclusion:** Γ_{1}
51		- For each time point in the interval [1.000, 4.000], B holds sometime in the
952		Therefore, according to the rule. A holds from 1 000 to 4 000
53		So based on the given rule and data A holds from 1 000 to 4 000.
54	User Prompt	Based on your previous response without any explanation state your an
55	0 set i tompt	swers to the question. You should only output "true" or "false"
56	LLM's output	frue
57	Expected Answer	true
57	Expected Answel	

С

We generated different number of samples based on the type of DatalogMTl, such as SingleAtom, MultiAtoms, Rational, MixedOperators, MultiRules and Recursive.

DETAILED INFORMATION OF OUR BENCHMARK

972	Type of Sample	# of Positive Samples	# of Negative Samples
973	SingleAtom	500	500
974	MultiAtoms	300	300
975	Rational	500	500
976	MixedOperators (with 2 operators)	1739	1739
977	MixedOperators (with 3 operators)	145	145
978	MixedOperators (with 3 operators)	126	126
979	MultiRules (with 2 rules)	250	250
020	MultiRules (with 4 rules)	250	250
900	MultiRules (with 6 rules)	150	150
901	Recursive	500	500

1000 1001 1002

1003 1004

1005

Table 2: The number of samples of different categories in our dataset

No matter how many samples are there in the our dataset, for each category, we only use the first 100 positive samples and the last 100 negative samples as the testing samples. For in-context/few-shot evaluations, we use one additional positive sample (the 101-th sample) and one additional negative sample (the last 101-th sample) as in-context learning sample to put into the context.

For MultiAtoms, we don't specify the number of operators it has in the rule nor evaluate them separately, while in general it follows the following distribution:

Type of Sample	# of Positive Samples	# of Negative Samples
MultiAtoms (with 2 atoms in the rule)	109	115
MultiAtoms (with 3 atoms in the rule)	79	79
MultiAtoms (with 4 atoms in the rule)	61	64
MultiAtoms (with 5 atoms in the rule)	51	42
Total	300	300

Table 3: The distribution of the number of atoms in our MultiAtoms subset of our dataset

D BAD CASE ANALYSIS

D.1 CASES THAT ARE INCORRECT IN SYMBOLIZED REPRESENTATION BUT ARE CORRECT IN VERBALIZED REPRESENTATION

Symbolic Verbalized 1008 1009 1010 1011 1011 1012 1013 1014 1015 1016 1017 1018 1019 1020 1021 1022 1022 1023 1024 1025	1006			
1008 1009 1010 1011 1012 1013 1014 1015 1016 1017 1018 1019 1020 1021 1022 1023 1024 1025	1007	Symbolic	Verbalized	
1009 1010 1011 1012 1013 1014 1015 1016 1017 1018 1019 1020 1021 1022 1023 1024 1025	1008			
1010 1011 1012 1013 1014 1015 1016 1017 1018 1019 1020 1021 1022 1023 1024 1025	1009			
1011 1012 1013 1014 1015 1016 1017 1018 1019 1020 1021 1022 1023 1024 1025	1010			
1012 1013 1014 1015 1016 1017 1018 1019 1020 1021 1022 1023 1024 1025	1011			
1013 1014 1015 1016 1017 1018 1019 1020 1021 1022 1023 1024 1025	1012			
1014 1015 1016 1017 1018 1019 1020 1021 1022 1023 1024 1025	1013			
1015 1016 1017 1018 1019 1020 1021 1022 1023 1024 1025	1014			
1016 1017 1018 1019 1020 1021 1022 1023 1023	1015			
1017 1018 1019 1020 1021 1022 1023 1024 1025	1016			
1018 1019 1020 1021 1022 1023 1024 1025	1017			
1019 1020 1021 1022 1023 1024 1025	1018			
1020 1021 1022 1023 1024 1025	1019			
1021 1022 1023 1024 1025	1020			
1022 1023 1024 1025	1021			
1023 1024 1025	1022			
1024 1025	1023			
1025	1024			
	1025			

1026			
1027	System Prompt	You are given a dataset and a tem-	(No introduction to the syntax is
1028	v <u>1</u>	poral rule, and your task is to judge	required for verbalized representa-
1029		whether the given fact is entailed by	tions)
1030		the dataset and the rule.	
1031		The rules are expressed as Data-	
1032		logMTL, a knowledge representa-	
1033		tion language that extends Datalog	
1034		with operators from metric temporal	
1035		logic (MTL). The semantics of four	
1036		MIL operators are given as follows:	
1030		If Diamondminus[a,b]A is true at	
1037		the time t, it requires that A needs	
1038		to be true at some time between t-b	
1039		If Royminus[a b] A is true at the	
1040		time t it requires that A needs to be	
1041		true continuously between t-b and t-	
1042		a	
1043		If Diamondplus[a b]A is true at the	
1044		time t it requires that A needs to be	
1045		true at some point between t+a and	
1046		t+b.	
1047		If Boxplus[a,b]A is true at the time	
1048		t, it requires that A needs to be true	
1049		continuously between t+a and t+b.	
1050			
1051	User Prompt	Now we have some temporal data	Now we have some temporal data
1052		and some rules, data: B@[3,10]	and some rules, data:
1053		rule: A:-Diamondplus[6,10]B	B holds From 3.000 to 10.000
1054		Is A@[1,4] true or not? Do not an-	rule: A holds in each time such that
1055		swer directly, think step by step.	B will hold sometime between 6.000
1056			and 10.000 hours in the future
1057			Is A holds From 1.000 to 4.000 true
1058			or not? Do not answer directly, think
1059			step by step.
1060			
1061			
1062			
1062			
1064			
1065			
1000			
1007			
1007			
1008			
1069			
1070			
1071			
1072			
1073			
1074			
1075			
1076			
1077			
1078			

1080			
1081	LLM's output	First, let's understand the rule:	To determine if A holds from 1.000
1082	1	A:-Diamondplus[6,10]B. This rule	to 4.000 based on the given rule and
1083		means that if B is true at some point	data, let's break it down step by step:
1084		between t+6 and t+10, then A is true	1. **Understand the Data:**
1085		at time t.	- B holds from 3.000 to 10.000.
1086		Now, let's look at the data:	2. **Understand the Rule:**
1087		B@[3,10]. This means that B is true	- A holds at each time such that B
1088		at some point between time 3 and	will hold sometime between 6.000
1089		time 10.	and 10.000 hours in the future.
1090		we are asked to determine if	3. **Analyze the 11me Intervals:**
1001		A@[1,4] is true of not. This means	- we need to check II B holds in the
1002		some point between time 1 and time	future from any time within [1 000
1002		4	4 0001
1095		Given the rule for A to be true at	4 **Check Each Time Point in
1094		time 1. B would need to be true at	[1.000, 4.000]:**
1095		some point between time 7 (1+6)	- For time 1.000:
1090		and $11(1+10)$. For A to be true at	- 6.000 to 10.000 hours in the future
1097		time 4, B would need to be true at	from 1.000 is [7.000, 11.000].
1098		some point between time 10 (4+6)	- B holds from 3.000 to 10.000,
1099		and 14 (4+10).	which overlaps with [7.000,
1100		However, B is only true between	10.000].
1101		time 3 and time 10. Therefore, B is	- Therefore, B holds sometime in
1102		not true at any point between time 7	[7.000, 11.000].
1103		and 11 or between time 10 and 14.	- For time 2.000 :
1104		So, $A@[1,4]$ is not true.	- 6.000 to 10.000 nours in the ruture
1105			$\begin{array}{c} \text{Irom 2.000 is [8.000, 12.000].} \\ \text{B holds from 3.000 to 10.000} \end{array}$
1106			- B holds from 5.000 to 10.000, which overlaps with [8,000
1107			10 0001
1108			- Therefore, B holds sometime in
1109			[8.000, 12.000].
1110			- For time 3.000:
1111			- 6.000 to 10.000 hours in the future
1112			from 3.000 is [9.000, 13.000].
1113			- B holds from 3.000 to 10.000,
1114			which overlaps with [9.000,
1115			10.000].
1116			- Therefore, B holds sometime in
1117			[9.000, 13.000].
1118			- FOI time 4.000. 6 000 to 10 000 hours in the future
1119			from 4 000 is [10 000 14 000]
1120			- B holds from 3 000 to 10 000
1121			which overlaps with [10.000,
1122			10.000].
1123			- Therefore, B holds sometime in
1124			[10.000, 14.000].
1125			5. **Conclusion:**
1126			- For each time point in the interval
1127			[1.000, 4.000], B holds sometime in
1128			the future interval [6.000, 10.000]
1129			hours from that time point.
1130			- Therefore, according to the rule, A
1131			noids from 1.000 to 4.000.
1132			So, based on the given rule and data, A holds from 1 000 to 4 000
1133			A notus nom 1.000 to 4.000.

1135	User Prompt	Based on your previous response, without any explanation state your an-	
1136	_	swers to the question. You should only	y output "true" or "false"
1137	LLM's output	false	true
1138	Expected Answer	true	

D.2 CASES THAT ARE INCORRECT IN ZERO-SHOT BY ARE CORRECT WHEN CHAIN OF THOUGHT IS APPLIED

Example 1:

	Without CoT	With CoT
System Prompt	You are given a dataset and a tem-	You are given a dataset and a t
	poral rule, and your task is to judge	poral rule, and your task is to ju
	whether the given fact is entailed by	whether the given fact is entailed
	the dataset and the rule.	the dataset and the rule.
	The rules are expressed as Data-	The rules are expressed as D
	logMTL, a knowledge representa-	logMTL, a knowledge represe
	tion language that extends Datalog	tion language that extends Dat
	with operators from metric temporal	with operators from metric temp
	logic (MTL). The semantics of four	logic (MTL). The semantics of
	MTL operators are given as follows:	MTL operators are given as follo
	If Diamondminus[a,b]A is true at	If Diamondminus[a,b]A is tru
	the time t, it requires that A needs	the time t, it requires that A no
	to be true at some time between t-b	to be true at some time between
	and t-a.	and t-a.
	If Boxminus[a,b]A is true at the time	If Boxminus[a,b]A is true at the
	t, it requires that A needs to be true	t, it requires that A needs to be
	continuously between t-b and t-a.	continuously between t-b and t-a
	II Diamondplus[a,b]A is true at the	II Diamondplus[a,b]A is true at time to it requires that A people
	time t, it requires that A needs to be	time t, it requires that A needs t
	tute at some point between t+a and	t b
	If $Boxplus[a b]A$ is true at the time	If Boxplus[2 b] Λ is true at the t
	t it requires that A needs to be true	t it requires that A needs to be
	continuously between t+a and t+b	continuously between $t+a$ and t-
	You should not give any explanation	continuously between the and t
	and you should only output "true" or	
	"false"	
User Prompt	Now we have some temporal data	Now we have some temporal
1	and some rules, data:	and some rules, data:
	B@[6,9]	B@[6,9]
	rule: A:-Diamondminus[6,15]B	rule: A:-Diamondminus[6,15]B
	Is A@[12,18] true or not?	Is A@[12,18] true or not? Do
		answer directly, think step by ste

1189	LLM's output		First, let's understand the rule: A:-
1190			Diamondminus[6,15]B. This rule
1191			means that if B is true at some point
1192			between t-15 and t-6, then A is true
1193			at time t.
1194			Now, let's consider the fact
1195			A@[12,18]. This means we are
1196			checking if A is true at some point
1197			between time 12 and 18.
1198			is true at some point between (12
1199			15) and $(12-6)$ i.e. between time -3
1200			and 6, or between $(18-15)$ and $(18-15)$
1201			6), i.e., between time 3 and 12.
1202			From the given data, we know that
1203			B is true between time 6 and 9. This
1204			falls within the required time range
1205			for A to be true.
1206			Therefore, A@[12,18] is true.
1207	User Prompt		Based on your previous response,
1208			without any explanation state your
1200			answers to the question. You should
1210	TTM2	6.1.	only output "true" or "false"
1011	LLM's output	false	true
1211	Expected Answer	true	

1213	
1214	Example 2:

215			
216		Without CoT	With CoT
210	System Prompt	You are given a dataset and a tem-	You are given a dataset and a tem-
217 218		poral rule, and your task is to judge	poral rule, and your task is to judge
210		whether the given fact is entailed by	whether the given fact is entailed by
219		the dataset and the rule.	the dataset and the rule.
220		The rules are expressed as Data-	The rules are expressed as Data-
221		logMTL, a knowledge representa-	logMTL, a knowledge representa-
222		tion language that extends Datalog	tion language that extends Datalog
223		with operators from metric temporal	with operators from metric temporal
224		logic (MTL). The semantics of four	logic (MTL). The semantics of four
225		MTL operators are given as follows:	MTL operators are given as follows:
226		If Diamondminus[a,b]A is true at	If Diamondminus[a,b]A is true at
227		the time t, it requires that A needs	the time t, it requires that A needs
228		to be true at some time between t-b	to be true at some time between t-b
220		and t-a.	and t-a.
229		If Boxminus[a,b]A is true at the time	If Boxminus[a,b]A is true at the time
230		t, it requires that A needs to be true	t, it requires that A needs to be true
231		continuously between t-b and t-a.	continuously between t-b and t-a.
232		If Diamondplus[a,b]A is true at the	If Diamondplus[a,b]A is true at the
233		time t, it requires that A needs to be	time t, it requires that A needs to be
234		true at some point between t+a and	true at some point between t+a and
235		t+b.	t+b.
236		If Boxplus[a,b]A is true at the time	If Boxplus[a,b]A is true at the time
237		t, it requires that A needs to be true	t, it requires that A needs to be true
238		continuously between t+a and t+b.	continuously between t+a and t+b.
239		You should not give any explanation	
240		and you should only output "true" or	
2/1		"talse"	
24 I			

Now we have some temporal data

Now we have some temporal data

rule: B:-Diamondminus[10,12]A

swer directly, think step by step.

Is B@[21,24] true or not? Do not an-

First, let's understand the rule: B:-

Diamondminus[10,12]A. This rule

means that if A is true at some point

between t-12 and t-10, then B is true

Now, let's look at the data: A@[8,13]. This means that A

is true continuously from time 8 to

Now, we need to check if

B@[21,24] is true or not. For

B to be true at any time between 21

and 24, A needs to be true at some

Let's check for each time t in

- For t=21, A needs to be true at

some point between 21-12=9 and 21-10=11. A is true at these times

- For t=22, A needs to be true at

some point between 22-12=10 and

22-10=12. A is true at these times

- For t=23, A needs to be true at

some point between 23-12=11 and

23-10=13. A is true at these times

- For t=24, A needs to be true at

some point between 24-12=12 and 24-10=14. A is true at time 12, but

not at time 14. However, the rule re-

quires A to be true at some point be-

tween these times, not at all times,

Based on your previous response,

without any explanation state your

answers to the question. You should

so this condition is satisfied.

only output "true" or "false"

true

Therefore, B@[21,24] is true.

point between (t-12) and (t-10).

according to the data.

according to the data.

according to the data.

and some rules, data:

A@[8,13]

at time t.

time 13.

[21,24]:

	and some rules, data: A@[8,13] rule: B:-Diamondminus[10,12]A Is B@[21,24] true or not?
LLM's output	

1242 1243 1244

1245

1246

1247

1248 1249

1250

1251

1252

1253

1254

1255

1256

1257

1258

1259

1260

1261

1262

1263

1264

1265

1266

1267

1268

1269

1270

1271

1272

1273

1274

1275

1276

1277

1278

1279

1280

1281

1282

1283

1284

1285

1286

1287

User Prompt

User Prompt LLLM's output false Expected Answer true

1288 1289 1290

1291

D.3 SINGLEATOM ERROR CASES

False Negative Example:

1292 1293

1296		
1297	System Prompt	You are given a dataset and a temporal rule, and your task is to judge whether
1298	, 1	the given fact is entailed by the dataset and the rule. The rules are expressed
1299		as DatalogMTL, a knowledge representation language that extends Datalog
1300		with operators from metric temporal logic (MTL). The semantics of four
1301		MTL operators are given as follows: If Diamondminus[a,b]A is true at the
1302		time t, it requires that A needs to be true at some time between t-b and
1303		t-a.If Boxminus[a,b]A is true at the time t, it requires that A needs to be
1304		true continuously between t-b and t-a.If Diamondplus[a,b]A is true at the
1305		time t, it requires that A needs to be true at some point between t+a and
1306		t+b.If Boxplus[a,b]A is true at the time t, it requires that A needs to be true
1307	Llass Dramat	continuously between t+a and t+b.
1202	User Prompt	Now we have some temporal data and some rules, data: $A \ll [7,15]$
1200		Iuic. DDOAIIIIIIus[5,5]A Is R@[8 30] true or not? Do not answer directly think sten by sten
1210	LLM's output	To determine whether ($B@[8,30]$) is true given the data and the rule, we
1310	LLW Soutput	need to follow these steps:
1311		1 **Understand the Rule**' The rule is (B leftarrow Boxminus[3 5]A)
1312		This means that (B) is true at time (t) if (A) is true continuously in the
1313		interval ([t-5, t-3]).
1314		2. **Check the Interval for (B@[8,30])**: We need to check if (B) is
1315		true for every time point in the interval ([8, 30]). For each time (t) in this
1316		interval, (A) must be true continuously in the interval ([t-5, t-3]).
1317		3. **Analyze the Data**: The data given is (A@[7,15]), which means (A
1318) is true continuously from time 7 to time 15.
1319		4. **Check Each Time Point in ([8, 30])**:
1320		- For $(t = 8)$: The interval $([t-5, t-3])$ is $([3, 5])$. (A) is not true in this
1321		interval because (A) is only true from 7 to 15. Eq. $(1 - 0)$ The interval $(1 - 5 + 2)$ is $(1 - 6)$ (A) is set to a in this
1322		- For $(t = 9)$: The interval $([t-5, t-3])$ is $([4, 6])$. (A) is not true in this
1323		Interval. For $(t = 10)$: The interval ([t 5 t 3]) is ([5 7]) (A) is not true continu
1324		- FOI ($t = 10$). The interval ($[t-3, t-3]$) is ($[3, 7]$). (A) is not true continu-
1325		- For $(t = 11)$: The interval ([t-5, t-3]) is ([6, 8]) (A) is true continuously
1326		in this interval
1327		- For ($t = 12$): The interval ([t-5, t-3]) is ([7, 9]). (A) is true continuously
1328		in this interval.
1329		- For (t = 13): The interval ([t-5, t-3]) is ([8, 10]). (A) is true continuously
1330		in this interval.
1331		- For ($t = 14$): The interval ([t-5, t-3]) is ([9, 11]). (A) is true continuously
1332		in this interval.
1333		- For ($t = 15$): The interval ([t-5, t-3]) is ([10, 12]). (A) is true continuously
1334		In this interval. $\mathbf{T} = \{ f \in \mathcal{T} : f \in \mathcal{O} \} : (f + 1, 2) \in (A, b) \in (A, b) $
1335		- For ($t = 16$): The interval ([t-5, t-3]) is ([11, 13]). (A) is true continuously
1336		In this interval. Eq. $(t - 17)$: The interval ([t 5, t 2]) is ([12, 14]) (A) is true continuously.
1337		- FOI $(1 = 17)$. The interval $([1-3, 1-3])$ is $([12, 14])$. (A) is true continuously in this interval
1338		For $(t - 18)$: The interval ([t-5, t-3]) is ([13, 15]) (A) is true continuously
1339		in this interval
1340		- For $(t = 19)$: The interval ([t-5 t-3]) is ([14 16]) (A) is not true in this
1341		interval (it ends at 15).
1342		- For ($t = 20$) to ($t = 30$): The intervals ([t-5, t-3]) will be ([15, 17]) to
1343		([25, 27]). (A) is not true in any of these intervals.
1344		5. **Conclusion**: (B) is not true for all time points in the interval ([8,
1345		30]). Specifically, (B) is not true at ($t = 8, 9, 10, \overline{19}$,) and any (t) from 20
1346		to 30. Therefore, (B@[8,30]) is not true.
1347	User Prompt	Based on your previous response, without any explanation state your an-
1348		swers to the question. You should only output "true" or "false"
1349	LLM's output	talse
	Expected Answer	true

False Positive Example:

1050		
1352	System Prompt	You are given a dataset and a temporal rule, and your task is to judge whether
1353		the given fact is entailed by the dataset and the rule. The rules are expressed
1354		as DatalogMTL, a knowledge representation language that extends Datalog
1355		with operators from metric temporal logic (MTL). The semantics of four
1356		MTL operators are given as follows: If Diamondminus[a,b]A is true at the
1357		time t, it requires that A needs to be true at some time between t-b and
1358		t-a.If Boxminus[a,b]A is true at the time t, it requires that A needs to be
1359		true continuously between t-b and t-a.If Diamondplus[a,b]A is true at the
1360		time t, it requires that A needs to be true at some point between t+a and
1361		t+b.If Boxplus[a,b]A is true at the time t, it requires that A needs to be true
1362		continuously between t+a and t+b.
1363	User Prompt	Now we have some temporal data and some rules, data: $A@[0,3]$
1264		rule: B:-Diamondminus[2,4]A
1304		Is $B@[-34,8]$ true or not? Do not answer directly, think step by step.
1365	LLM's output	To determine whether ($B@[-34,8]$) is true given the data and the rule, we
1366		need to follow these steps:
1367		1. **Understand the Rule**: The rule is (B leftarrow Diamondminus[2,4]A
1368). This means that (B) is true at time (t) if (A) is true at some time between $(t, t) = 1$
1369		(t-4) and $(t-2)$.
1370		2. **Determine the line interval for $(B@[-34,8])^{**}$: We need to check if
1371		(B) is true at any time within the interval $([-34, 8])$.
1372		3. **Apply the Rule to Each Time Point in the Interval**: - For each time (4) in the interval ([24, 8]) are used to should if (A) is true at some time
1373		(t) in the interval ([-34, 8]), we need to check if (A) is true at some time
1374		between (1-4) and (1-2). $4 \times Charle the Determined the data since is (A @[0.2]) which means (A) is$
1375		4. **Check the Data**: The data given is ($A@[0,5]$), which means (A) is
1376		true from time 0 to time 5. 5 **Evaluate the Dule for Each Time Doint**: For $(t - 24)$: We need (
1377		5. Evaluate the Kule for Each finite form For $(1 = -54)$, we need (4) to be true between ([0, 3])
1070		A) to be true in this interval. For $(t = 33)$: We need (Λ) to be true
1070		(A) is not true in this interval For $(t = -55)$. We need (A) to be true between ([37, 35]) Again (A) is not true in this interval. Continue this
1379		process up to $(t - 8)$: For $(t - 0)$: We need (A) to be true between $([-4]$
1380		(1-4) (A) is not true in this interval - For $(1-4)$. We need (A) to be true
1381		between ([-3, -1]) (A) is not true in this interval. For $(t = 1)$, we need (A) to be true between ([-3, -1]) (A) is not true in this interval.
1382		A) to be true between $([-2, 0])$ (A) is not true in this interval. For $(t = 3)$.
1383). We need (A) to be true between $([-1, 1])$ (A) is true at time 0 which is
1384		within this interval For $(t = 4)$: We need (A) to be true between ([0, 2]).
1385		(A) is true in this interval For ($t = 5$): We need (A) to be true between
1386		([1, 3]). (A) is true in this interval For ($t = 6$): We need (A) to be true
1387		between ([2, 4]). (A) is true at time 2 and 3, which is within this interval
1388		For $(t = 7)$: We need (A) to be true between ([3, 5]). (A) is true at time 3,
1389		which is within this interval For $(t = 8)$: We need (A) to be true between
1390		([4, 6]). (A) is not true in this interval.
1391		6. **Conclusion**: Since (A) is true at some time between (t-4) and (t-2
1392) for ($t = 3, 4, 5, 6$,) and (7), (B) is true at these times. Therefore, (B) is
1393		true at some points within the interval ([-34, 8]).
1304		Thus, (B@[-34,8]) is true.
1205	User Prompt	Based on your previous response, without any explanation state your an-
1000		swers to the question. You should only output "true" or "false"
1390	LLM's output	true
1397	Expected Answer	false
1398	-	

D.4 RECURSIVE ERROR CASES

1401False Negative Example:

1403

1399

System PromptYou are given a dataset and a temporal rule, and your task is to judge whether the given fact is entailed by the dataset and the rule. The rules are expressed as DatalogMTL, a knowledge representation language that extends Datalog with operators are given as follows:If Diamondminus[ab]A is true at the time t, it requires that A needs to be true at some time between t-b and t-a.4f Boxminus[ab]A is true at the time t, it requires that A needs to be true continuously between t-b and t-a.1f Diamondminus[ab]A is itrue at the time t, it requires that A needs to be true at some point between t+a and t-b.ff Boxplus[ab]A is true at the time t, it requires that A needs to be true continuously between t4a and t-b.User PromptNow we have some temporal data and some rules, data: A@[10,12] rule: A:-Diamondplus[4,13]A Is a @[-469,-221] true or not? Do not answer directly, think step by step.LLM's outputTo determine whether (A@[-469,-221]) is true based on the given data and rule, we need to follow a step-by-step reasoning process. Let's break it down:### Step 1: Understand the Data The data given is: [A@[10,12]] This means that (A) is true continuously from time 10 to time 12.### Step 2: Understand the Rule The rule provided is: [A leftarrow Dia- mondplus[4,13]A] This rule states that (A) is irue at time (t) if (A) is true at some point between (t+4) and (t+13).### Step 3: Apply the Rule We need to check if (A@[-469,-221]) can be derived using the rule and progragate (A) backward or forward in time to cover the interval ([10, 12]) where (A) is forward Let's start from the known in- terval (10, 12]) and see if the cule provada (L+4) and (L+6), -221]).### Step 3:: Check the Time Interval The interval ([-469, -221]).### Step 3:: Chock the Time true at some point between (<th>1404</th> <th></th> <th></th>	1404		
1006 1007 1008 1009the given fact is entailed by the dataset and the rule. The rules are expressed 	1405	System Prompt	You are given a dataset and a temporal rule, and your task is to judge whether
as DatalogMTL, a knowledge representation language that extends Datalog with operators from metric temporal logic (MTL). The semantics of four MTL operators are given as follows:If Diamondplus[a,b]A is true at the time t, it requires that A needs to be true at some pint between t-b and t-a.If Boxminus[a,b]A is true at the time t, it requires that A needs to be true continuously between t-b and t-a.If Diamondplus[a,b]A is true at the time t, it requires that A needs to be true at some point between t-b and t-b.If Boxplus[a,b]A is true at the time t, it requires that A needs to be true continuously between t-a and t-b. Now we have some temporal data and some rules, data: A@[10,12] rule: A:-Diamondplus[4,13]A to A@[-409,-22] It ure on rol? Do not answer directly, think step by step. Is A@[-409,-22] true on rol? Do not answer directly, think step by step. LLM's output to determine whether (A@[-469,-22]]) is true based on the given data and rule, we need to follow a step-by-step reasoning process. Let's break it down: ### Step 3: Understand the Data The data given is: [A@[10,12]] This means that (A) is true continuously from time 10 to time 12. ### Step 2: Understand the Rule The rule provided is: [A [A [Farrow Dia- mondplus[4,13]A] This rule states that (A) is true at time (t) if (A) is true at some point between (t+4) and (t+13). ### Step 3: Apply the Rule We need to cheak if (A@[-469,-221]) can be derived using the rule and the given data. #### Step 3: Propagate (A) backward or forward in time to cover the interval [(10, 12]) where (A) is known to be true. However, we need to see if the rule can propagate (A) backward or forward in time to cover the interval [(-40, -221]). #### Step 3:? Propagate the Rule Forward Let's start from the known in terval (110, 12]) and see if we can propagate (A) forward using the rule. - For (A) to be true at time (t), (A) is true at some point between (t+4) and (t+13). Starting from (t = 10): - (A) is true at (t = 10) if (A) is true at some point between (12+4) and (10+13	1406		the given fact is entailed by the dataset and the rule. The rules are expressed
with operators from metric temporal logic (MTL). The semantics of four MTL operators are given as follows: IT Diamondminus(ab)A is true at the time t, it requires that A needs to be true at some time between t-b and t-a.IT Boxminus(ab)A is true at the time t, it requires that A needs to be true continuously between t-b and t-a.IT Diamondpinus(ab)A is true at the time t, it requires that A needs to be true at some point between t-a and t-b.IT Boxpinus(ab)A is true at the time t, it requires that A needs to be true continuously between t-a and t-tb.User PromptNow we have some temporal data and some rules, data: A@[10,12] rule: A:-Diamondpinus(A)[A]A Is A@[-469,-221] true or not? Do not answer directly, think step by step.LLM's outputTo determine whether (A@[-469,-221]) is true based on the given data and rule, we need to follow a step-by-step reasoning process. Let's break it down:down:### Step 1: Understand the Data The data given is: [A@[10,12]] This means that (A) is true continuously from time I0 to time 12.### Step 2: Understand the Rule The rule provided is: [A leftarrow Dia- mondplus[4,13]A] This rule states that (A) is true at time (t) if (A) is true at some point between (t+4) and (t+13).### Step 3: Check the Time Interval The interval [[-469, -221]) can be derived using the rule and the given data.### Step 3: Check the Time Interval The interval [[-469, -221]) is quite far from the interval (10, 12]) where (A) is known to be true. However, we need to see if the rule can propagate (A) backward or forward in time to cover the interval ([-469, -221]).### Step 3: Check the Give A can at a some point between (t+4) and (t+13).#### Step 3: Check the Give A can at some point between (t+4) and (1+13), i.e., between (14) and (23).	1407		as DatalogMTL, a knowledge representation language that extends Datalog
$\begin{array}{c c c c c c c c c c c c c c c c c c c $	1408		with operators from metric temporal logic (MTL). The semantics of four
1410time t, it requires that A needs to be true at some time between t-b and t-a.If Boxninus[a,b]A is true at the time t, it requires that A needs to be true continuously between t-b and t-a.If Diamondplus[a,b]A is true at the time t, it requires that A needs to be true at some point between t-a and t+b.If Boxplus[a,b]A is true at the time t, it requires that A needs to be true continuously between t-a and t-b.1413User PromptNow we have some temporal data and some rules, data: A@[10,12] rule: A:-Diamondplus[4,13]A Is A@[-469,-221] true or not? Do not answer directly, think step by step.1414It of determine whether (A@[-469,-221]) is true based on the given data and rule, we need to follow a step-by-step reasoning process. Let's break it down:1420### Step 1: Understand the Data The data given is: [A@[10,12]] This means that (A) is true continuously from time 10 to time 12. ### Step 2: Understand the Rule The rule provided is: [A leftarrow Dia- mondplus[4,13]A] This rule states that (A) is true at time (1) if (A) is true at some point between (1+4) and (1+13).1422### Step 3: Check the Time Interval The interval [(-469, -221]) can be derived using the rule and the given data. #### Step 3: Check the Time Interval The interval ([-469, -221]) is quite far from the interval ([10, 12]) and see if we can progate (A) backward or forward in time to cover the interval ([10, 12]). #### Step 3: Check the Time Interval The interval ([-469, -221]) is quite far from the interval ([-10, 12]) make if we can progate (A) forward using the rule. - For (A) to be true at time (1), (A) is true at some point between (1+4) and (1+10, 12]) where (A) is knew of (A) is true at some point between (1+2) and (1+3), i.e., between (14) and (23). Starting from (t = 10): - (A) is true at (10, 12]. #### Step 3:3: S	1409		MTL operators are given as follows: If Diamondminus[a,b]A is true at the
1411Ica.II Boxminusla,b A is true at the time t, it requires that A needs to be true continuously between t-b and t-a.II Diamondplus[a,b]A is true at the time t, it requires that A needs to be true at some point between t+a and t+b.IF Boxplus[a,b]A is true at the time t, it requires that A needs to be true continuously between t+a and t+b.1414User PromptNow we have some temporal data and some rules, data: A@[10,12] rule: A:-Diamondplus[4,13]A1415Is A@[-469,-221] true or not? Do not answer directly, think step by step.1416It Me soutputTo determine whether (A@[-469,-221]) is true based on the given data and rule, we need to follow a step-by-step reasoning process. Let's break it down:1421### Step 1: Understand the Data The data given is: [A@[10,12]] This means that (A) is true continuously from time 10 to time 12.1422### Step 2: Understand the Rule The rule provided is: [A @[10,12]] can be derived using the rule and the given data.1423mondplus[4,13]A] This rule states that (A) is true at time (t) if (A) is true at some point between (1+4) and (t+13).1424### Step 3.1: Check the Time Interval The interval (1(469,-221]) can be derived using the rule and rule gar propagate (A) backward or forward in time to cover the interval (140, 12] where (A) is known to be true. Howveer, we need to see if the rule can propagate (A) backward or forward in time to cover the interval (140, 12), i.e., between (14) and (23).1434Starting from (t = 10): - (A) is true at some point between (t+4) and (12+13), i.e., between (14) and (23).1435Starting from (t = 12): - (A) is true at tome point between (t+4) and (12+13), i.e., between (14) and (23).1436Starting from (t	1410		time t, it requires that A needs to be true at some time between t-b and
1412Irree continuously between 1-b and 1-a.II Damondpulus[a,b]A is true at the time t, it requires that A needs to be true at some point between 1+a and 1+b.If Boxplus[a,b]A is true at the time t, it requires that A needs to be true continuously between 1+a and 1+b.1414User PromptNow we have some temporal data and some rules, data: A@[10,12] rule: A:-Diamondplus[4,13]A1415Is A@[-469,-221] two er not? Do not answer directly, think step by step.1416Is A@[-469,-221] two er not? Do not answer directly, think step by step.1417To determine whether (A@[-469,-221]) is true based on the given data and rule, we need to follow a step-by-step reasoning process. Let's break it down:1420### Step 1: Understand the Data The data given is: [A@[10,12]] This means that (A) is true continuously from time 10 to time 12. ### Step 2: Understand the Rule The rule provided is: [A leftarrow Dia- mondplus[4,13]A] This rule states that (A) is true at time (1) if (A) is true at some point between (1+4) and (1+3). ### Step 3: Apply the Rule We need to check if (A@[-469,-221]) can be derived using the rule and the given data. #### Step 3.1: Check the Time Interval The interval [[-469, -221]) is quite far from the interval (10, 12]) where (A) is known to be true. However, terval (10, 12]) where (A) is known to be true. However, we need to see if the rule can propagate (A) backward or forward in time to cover the interval ([-469, -221]).1433#### Step 3.2: Propagate the Rule Forward Let's start from the known in- terval ([10, 12]) and see if we can propagate (A) forward using the rule. - For (A) to be true at time (1). (A) must be true at some point between (124) and (124). Starting from (1 = 10): - (A) is true at (1 = 10) if (A) is true at some point between (124) and	1411		t-a.If Boxminus[a,b]A is true at the time t, it requires that A needs to be
1413 14141111 14141111 14141414 14141411 14111411 14111415 1416User PromptNow we have some temporal data and some rules, data: A@[10,12] rule: A:-Diamondplus[4,13]A 14171416 14171412 14181418 141915 15 14181419 141915 15 14191410 141115 15 14111411 141215 15 14121412 141416 15 14121413 141416 15 14121414 1414171415 142017 14141416 14211417 1422 14221418 1421 14221421 1422 14221422 1422 14231424 1424 14241425 1424 1424 14241426 1424 1424 1424 14241427 1424 1424 1424 1425 1425 1426 1427 1428 1428 1429 1428 1429 1429 1420 1421 1421 1421 1421 1422 1422 1422 1422 1423 1424 1433 1434 1434 1434 1434 1434 14	1412		true continuously between t-b and t-a.lf Diamondplus[a,b]A is true at the
1414 141515 Obsplastable is the at the time t, if requires that A freeds to be the continuously between t+a and t+b.1416 1417User PromptNow we have some temporal data and some rules, data: A@[10,12] rule: A:-Diamondplus[4,13]A 15 A@[-469,-221] true or not? Do not answer directly, think step by step.1418 1419LLM's outputTo determine whether (A@[-469,-221]) is true based on the given data and rule, we need to follow a step-by-step reasoning process. Let's break it down: ### Step 1: Understand the Data The data given is: [A@[10,12]] This means that (A) is true continuously from time 10 to time 12. ### Step 2: Understand the Rule The rule provided is: [A leftarrow Dia- mondplus[4,13]A] This rule states that (A) is true at time (t 1) if (A) is true at some point between (t+4) and (t+13). ### Step 3: Dheck the Time Interval [1469,-221]) can be derived using the rule and the given data. #### Step 3: Check the Time Interval [1469,-221]) is quite far from the interval ([10, 12]) where (A) is known to be true. However, we need to see if the rule can propagate (A) backward or forward in time to cover the interval ([10, 12]) where (A) forward using the rule. 41301422 1423#### Step 3.2: Propagate the Rule Forward Let's start from the known in- terval ([10, 12]) and see if we can propagate (A) hockward or forward in time to cover the interval ([-469, -221]).1431 1432 1433#### Step 3.2: Propagate the Rule Forward Let's start from the known in- terval ([10, 12]) and see if we can propagate (A) forward using the rule. 4. Yo to be true at time (t), (A) must be true at some point between (10+4) and (10+13) i.e., between (14) and (23). Starting from (t = 10): - (A) is true at (t = 10) if (A) is true at some point between (10+4) and (10+13) i.e., between (16) and (25).	1413		the It is the state of the time to be true at some point between $t+a$ and the If Boxplus $[a, b] A$ is true at the time to it requires that A needs to be true
1415User PromptNow we have some temporal data and vero.1416User PromptNow we have some temporal data and some rules, data: A@[10,12] rule: A:-Diamondplus[4,13]A1417Is A@[-469,-221] true or not? Do not answer directly, think step by step.1418LLM's outputTo determine whether (A@[-469,-221]) is true based on the given data and rule, we need to follow a step-by-step reasoning process. Let's break it down:1420### Step 1: Understand the Data The data given is: [A@[10,12]] This means that (A) is true continuously from time 10 to time 12. ### Step 2: Understand the Rule The rule provided is: [A leftarrow Dia- mondplus[4,13]A] This rule states that (A) is true at time (t) if (A) is true at some point between (1+4) and (1+13). ### Step 3: Apply the Rule We need to check if (A@[-469,-221]) can be derived using the rule and the given data. #### Step 3: Check the Time Interval [[-469, -221]). true at some point between (1-469, -221]). we need to see if the rule can propagate (A) backward or forward in time to cover the interval ([-469, -221]).1431#### Step 3: 2: Propagate the Rule Forward Let's start from the known in- terval ([10, 12]) and see if we can propagate (A) backward or forward in time to cover the interval (1-44) and (10+13), i.e., between (14) and (23). Starting from (t = 10): - (A) is true at tome point between (1+4) and (24). Starting from (t = 10): - (A) is true at t (= 10) if (A) is true at some point between (10+4) and (10+13), i.e., between (14) and (25). Since we do not have any data indicating that (A) is true at some point between (10+4) and (10+13), i.e., between (16) and (25). Since we do not have any data indicating that (A) is true at some point between (10+4) and (10+13), i.e., between (16) and (1414		continuously between t+a and t+b
14161500 fr 10mpt141715 A @[-469,-221] true or not? Do not answer directly, think step by step.141815 A @[-469,-221] true or not? Do not answer directly, think step by step.141916 determine whether (A @[-469,-221]) is true based on the given data and rule, we need to follow a step-by-step reasoning process. Let's break it down:1421### Step 1: Understand the Data The data given is: [A @[10,12]] This means that (A) is true continuously from time 10 to time 12.1421### Step 2: Understand the Rule The rule provided is: [A leftarrow Dia- mondplus[4,13] A] This rule states that (A) is true at time (t) if (A) is true at some point between (1+4) and (1+13).1425### Step 3: Apply the Rule We need to check if (A@[-469,-221]) can be derived using the rule and the given data.1426### Step 3: Apply the Rule We need to check if (A@[-469,-221]) is quite far from the interval ([10, 12]) where (A) is known to be true. However, we need to see if the rule can propagate (A) backward or forward in time to cover the interval ([10, 12]) and see if we can propagate (A) forward using the rule. - For (A) to be true at time (t), (A) must be true at some point between (14+4) and (1+13).1434### Step 3.2: Propagate the Rule Forward Let's start from the known in- terval ([10, 12]) and see if we can propagate (A) is true at some point between (10+4) and (10+13), i.e., between (14) and (23).1434Starting from (t = 10): - (A) is true at (t = 10) if (A) is true at some point between (10+4) and (10+13), i.e., between (14) and (25).1436Starting from (t = 12): - (A) is true at these new intervals, we cannot propagate (A) sust and the stome propagate (A) is at some point between (10+4) and (10+13),	1415	User Prompt	Now we have some temporal data and some rules data: A@[10.12]
1417Is A@[-469,-221] true or not? Do not answer directly, think step by step.1418LLM's outputTo determine whether (A@[-469,-221]) is true based on the given data and rule, we need to follow a step-by-step reasoning process. Let's break it down:1421### Step 1: Understand the Data The data given is: [A@[10,12]] This means that (A) is true contunously from time 10 to time 12.1422### Step 2: Understand the Rule The rule provided is: [A leftarrow Dia- mondplus[4,13]A] This rule states that (A) is true at time (t) if (A) is true at some point between (1+4) and (1+13).1425### Step 3: Apply the Rule We need to check if (A@[-469,-221]) can be derived using the rule and the given data.1426derived using the rule and the given data.1427### Step 3.: Check the Time Interval The interval ([-469, -221]) is quite far from the interval ([10, 12]) where (A) is known to be true. However, we need to see if the rule can propagate (A) backward or forward in time to cover the interval ([-469, -221]).1430#### Step 3.: Propagate the Rule Forward Let's start from the known in- terval ([10, 12]) and see if we can propagate (A) backward or forward in time to cover the interval ([-461 -3,).c., between (14) and (23).1431Starting from (t = 10): - (A) is true at (t = 10) if (A) is true at some point between (12+4) and (1+13), i.e., between (16) and (23).1436Starting from (t = 12): - (A) is true at some point between (t+4) and (t+13).1437Starting from (t = 12): - (A) is true at some point between (t+4) and (t+13).1438Starting from (t = 12): - (A) is true at (t = 12) if (A) is true at some point between (12+4) and (1+13)	1416	eser riompt	rule: A:-Diamondplus[4,13]A
1418LLM's outputTo determine whether (A@[-469,-221]) is true based on the given data and rule, we need to follow a step-by-step reasoning process. Let's break it down:1420and rule, we need to follow a step-by-step reasoning process. Let's break it down:1421### Step 1: Understand the Data The data given is: [A@[10,12]] This means that (A) is true continuously from time 10 to time 12.1422### Step 2: Understand the Rule The rule provided is: [A leftarrow Dia- mondplus[4,13]A] This rule states that (A) is true at time (t) if (A) is true at some point between (1+4) and (1+13).1424### Step 3: Apply the Rule We need to check if (A@[-469,-221]) can be derived using the rule and the given data. #### Step 3: Di propagate the Rule Forward Let's start from the known in- terval ([10, 12]) where (A) is known to be true. However, we need to see if the rule can propagate (A) backward or forward in time to cover the interval ([-469, -221]).1434### Step 3: Di propagate the Rule Forward Let's start from the known in- terval ([10, 12]) and see if we can propagate (A) forward using the rule. - For (A) to be true at time (t), (A) must be true at some point between (144) and (12+13).1436Starting from (t = 10): - (A) is true at (t = 10) if (A) is true at some point between (10+4) and (10+13), i.e., between (16) and (23). Since we do not have any data indicating that (A) is true at some point between (12+4) and (12+13).1436Since we do not have any data indicating that (A) is true at some point between (12+4) and (12+13).1437Since we do not have any data indicating that (A) is true at some point between (12+4) and (10, 12).1438We do not have any data	1417		Is A@ $[-469,-221]$ true or not? Do not answer directly, think step by step.
1419and rule, we need to follow a step-by-step reasoning process. Let's break it down:1420### Step 1: Understand the Data The data given is: [A@[10,12]] This means that (A) is true continuously from time 10 to time 12.1421### Step 1: Understand the Rule The rule provided is: [A leftarrow Dia- mondplus[4,13]A] This rule states that (A) is true at time (t) if (A) is true at some point between (t+4) and (t+13).1424### Step 3: Apply the Rule We need to check if (A@[-469,-221]) can be derived using the rule and the given data.1425### Step 3: Check the Time Interval The interval ([-469, -221]) is quite far from the interval ([10, 12]) where (A) is known to be true. However, we need to see if the rule can propagate (A) backward or forward in time to cover the interval ([-469, -221]).1430### Step 3.2: Propagate the Rule Forward Let's start from the known in- terval ([10, 12]) and see if we can propagate (A) forward using the rule. - For (A) to be true at time (t), (A) must be true at some point between (10+4) and (10+13), i.e., between (14) and (25).1434Starting from (t = 10): - (A) is true at (t = 12) if (A) is true at some point between (12+4) and (12+13), i.e., between (16) and (25).1435Since we do not have any data indicating that (A) is true in these new intervals, we cannot propagate (A) forward from the interval ([10, 12]).1444For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13), i.e., between (16) and (25).1436Since we do not have any data indicating that (A) is true in these new intervals, we cannot propagate (A) forward from the interval ([10, 12]).1440#### Step 3.3: Propagate the Rule Backward Now, let's consider if we can 	1418	LLM's output	To determine whether (A@[-469221]) is true based on the given data
down:the form the form	1419	. .	and rule, we need to follow a step-by-step reasoning process. Let's break it
#### Step 1: Understand the Data The data given is: [A@[10,12]] This means that (A) is true continuously from time 10 to time 12.### Step 1: Understand the Rule The rule provided is: [A leftarrow Dia- mondplus[4,13]A] This rule states that (A) is true at time (t) if (A) is true at some point between (1+4) and (1+13).### Step 3: Apply the Rule We need to check if (A@[-469,-221]) can be derived using the rule and the given data.#### Step 3: Apply the Rule We need to check if (A@[-469,-221]) is quite far from the interval ([10, 12]) where (A) is known to be true. However, we need to see if the rule can propagate (A) backward or forward in time to cover the interval ([469, -221]).#### Step 3: Apply and see if we can propagate (A) backward or forward in time to cover the interval ([469, -221]).#### Step 3: Apply and see if we can propagate (A) is true at some point between (t+4) and (10, 12]) and see if we can propagate (A) is true at some point between (10+4) and (10+13), i.e., between (14) and (23).Starting from (t = 10): - (A) is true at (t = 12) if (A) is true at some point between (10+4) and (12+13), i.e., between (14) and (23).Starting from (t = 12): - (A) is true at (t = 12) if (A) is true at some point between (12+4) and (12+13), i.e., between (14) and (23).Starting from (t = 12): - (A) is true at (t = 12) if (A) is true at some point between (12+4) and (12+13), i.e., between (14) and (23).Starting from (t = 10): - (A) is true at some point between (14) and (12+13), i.e., between (14) and (23).Starting from (t = 10): - (A) is true at the can propagate (A) is true in these new intervals, we cannot propagate (A) forward from the interval (10, 12]).######## Step 3.3: Propagate the Rule Backward Now, let's consider if we can <td>1420</td> <td></td> <td>down:</td>	1420		down:
nearmeans that (A) is true continuously from time 10 to time 12.1422### Step 2: Understand the Rule The rule provided is: [A leftarrow Dia- mondplus[4,1]A] This rule states that (A) is true at time (t) if (A) is true at some point between (t+4) and (t+13).1424### Step 3: Apply the Rule We need to check if (A@[-469,-221]) can be derived using the rule and the given data.1427### Step 3: Check the Time Interval The interval [(-469, -221]) is quite 	1421		### Step 1: Understand the Data The data given is: [A@[10,12]] This
### Step 2: Understand the Rule The rule provided is: [A leftarrow Dia- mondplus[4,13]A] This rule states that (A) is true at time (t) if (A) is true at some point between (t+4) and (t+13).##25### Step 3: Apply the Rule We need to check if (A@[-469,-221]) can be derived using the rule and the given data.### Step 3: I. Check the Time Interval The interval ([-469, -221]) is quite far from the interval ([10, 12]) where (A) is known to be true. However, we need to see if the rule can propagate (A) backward or forward in time to cover the interval ([-469, -221]).### Step 3: Propagate the Rule Forward Let's start from the known in- terval ([10, 12]) and see if we can propagate (A) forward using the rule. - For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).#434Starting from (t = 10): - (A) is true at (t = 10) if (A) is true at some point between (10+4) and (10+13), i.e., between (14) and (23).#335Starting from (t = 12): - (A) is true at (t = 12) if (A) is true at some point between (12+4) and (12+13), i.e., between (14) and (23).#346Starting from (t = 12): - (A) is true at (t = 12) if (A) is true at some point between (12+4) and (12+13), i.e., between (14) and (23).#347Since we do not have any data indicating that (A) is true in these new intervals, we cannot propagate (A) forward from the interval ([10, 12]).#444H## Step 3.3: Propagate the Rule Backward Now, let's consider if we can propagate (A) backward to cover the interval ([-469, -221]),#444For (A to be true at time (t), (A) must be true at some point between (t+4) and (t+13).#444H# (t+13).#444We conclusion Given the data (A@[10,12]).#444H# (t+13)	1422		means that (A) is true continuously from time 10 to time 12.
mondplus[4,13]AThis rule states that (A) is true at time (t) if (A) is true at some point between (t+4) and (t+13).1425### Step 3: Apply the Rule We need to check if (A@[-469,-221]) can be derived using the rule and the given data.1427#### Step 3.1: Check the Time Interval ([-469, -221]) is quite far from the interval ([10, 12]) where (A) is known to be true. However, we need to see if the rule can propagate (A) backward or forward in time to cover the interval ([-469, -221]).1430#### Step 3.2: Propagate the Rule Forward Let's start from the known in- terval ([10, 12]) and see if we can propagate (A) forward using the rule. - For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1434Starting from (t = 10): - (A) is true at (t = 10) if (A) is true at some point between (10+4) and (10+13), i.e., between (14) and (23).1435Starting from (t = 12): - (A) is true at (t = 12) if (A) is true at some point between (12+4) and (12+13), i.e., between (14) and (25).1438Since we do not have any data indicating that (A) is true in these new intervals, we cannot propagate (A) forward from the interval [[10, 12]).1440#### Step 3.3: Propagate the Rule Backward Now, let's consider if we can propagate (A) backward to cover the interval [[-469, -221]).1441For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1442- For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1444that (t+4) it (t+13).1445Hift (t+4) it (t+13).1446the interval [[-469, -221]].1447#### Conclusion Given the data (A@[10,12]).<	1423		### Step 2: Understand the Rule The rule provided is: [A leftarrow Dia-
The at some point between (1+4) and (1+13).1425### Step 3: Apply the Rule We need to check if (A@[-469,-221]) can be derived using the rule and the given data.1426derived using the rule and the given data.1427#### Step 3.1: Check the Time Interval The interval ([-469, -221]) is quite far from the interval ([10, 12]) where (A) is known to be true. However, we need to see if the rule can propagate (A) backward or forward in time to cover the interval ([-469, -221]).1430#### Step 3.2: Propagate the Rule Forward Let's start from the known in- terval ([10, 12]) and see if we can propagate (A) forward using the rule. - For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1434Starting from (t = 10): - (A) is true at (t = 10) if (A) is true at some point between (10+4) and (10+13), i.e., between (14) and (23). Starting from (t = 12): - (A) is true at (t = 12) if (A) is true at some point between (12+4) and (12+13), i.e., between (16) and (25).1436Starting from (t = 12): - (A) is true at (Te 12) if (A) is true at some point between (12+4) and (12+13), i.e., between (16) and (25).1438Since we do not have any data indicating that (A) is true in these new intervals, we cannot propagate (A) forward from the interval ([10, 12]).1440#### Step 3.3: Propagate the Rule Backward Now, let's consider if we can propagate (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1442- For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1444true therval ([-469, -221]), we would need to find a time (t) such that (t+4) to (t+13) overlaps with ([10, 12]).1441Propagate	1494		mondplus[4,13]A] This rule states that (A) is true at time (t) if (A) is
1426### Step 3: Apply the Rule We need to check if ($A@[-469,-221]$) can be derived using the rule and the given data.1427#### Step 3:1: Check the Time Interval The interval ([-469, -221]) is quite far from the interval ([10, 12]) where (A) is known to be true. However, we need to see if the rule can propagate (A) backward or forward in time to cover the interval ([-469, -221]).1430#### Step 3:2: Propagate the Rule Forward Let's start from the known in- terval ([10, 12]) and see if we can propagate (A) forward using the rule. - For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t=13).1431#### Step 3:2: Propagate the Rule Forward Let's start from the known in- terval ([10, 12]) and see if we can propagate (A) forward using the rule. - For (A) to be true at time (t), (A) must be true at some point between (t+4) and (10+13), i.e., between (14) and (23).1435Starting from (t = 10): - (A) is true at (t = 10) if (A) is true at some point between (12+4) and (10+13), i.e., between (16) and (25).1438Since we do not have any data indicating that (A) is true in these new intervals, we cannot propagate (A) forward from the interval [[10, 12]).1440#### Step 3:Propagate the Rule Backward Now, let's consider if we can propagate (A) backward to cover the interval [[-469, -221]).1441propagate (A) backward to cover the interval [[-469, -221]).1442- For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1441propagate (A) backward to cover the interval [[-469, -221]).1442- For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1443To cover the interval [[-4	1425		true at some point between $(t+4)$ and $(t+13)$.
1427derived using the rule and the given data.1427#### Step 3.1: Check the Time Interval The interval ([-469, -221]) is quite far from the interval ([10, 12]) where (A) is known to be true. However, we need to see if the rule can propagate (A) backward or forward in time to cover the interval ([-469, -221]).1430#### Step 3.2: Propagate the Rule Forward Let's start from the known in- terval ([10, 12]) and see if we can propagate (A) forward using the rule. - For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1434Starting from (t = 10): - (A) is true at (t = 10) if (A) is true at some point between (10+4) and (10+13), i.e., between (14) and (23).1435Starting from (t = 12): - (A) is true at (t = 12) if (A) is true at some point between (12+4) and (12+13), i.e., between (16) and (25).1438Since we do not have any data indicating that (A) is true in these new intervals, we cannot propagate (A) forward how, let's consider if we can propagate (A) backward to cover the interval ([-469, -221]).1441Propagate (A) backward to cover the interval ([-469, -221]).1442- For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1444To cover the interval [[-469, -221]), we would need to find a time (t) such that (t+4) to (t+13) overlaps with ([10, 12]).1444However, since ([10, 12]) is far in the future compared to ([-469, -221]), there is no way to find such at (t) that satisfies the rule.1444User PromptBased on your previous response, without any explanation state your an- swers to the question. You should only output "true" or "false"1445LLM's outputfalse <td>1/126</td> <td></td> <td>### Step 3: Apply the Rule we need to check if ($A@[-469,-221]$) can be</td>	1/126		### Step 3: Apply the Rule we need to check if ($A@[-469,-221]$) can be
1427#### Step 5.1: Check the lifter lifter the interval ([-409, -221]) is quite far from the interval ([10, 12]) where (A) is known to be true. However, we need to see if the rule can propagate (A) backward or forward in time to cover the interval ([-469, -221]).1430#### Step 3.2: Propagate the Rule Forward Let's start from the known in- terval ([10, 12]) and see if we can propagate (A) forward using the rule. - For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1434t+4) and (t+13).1435Starting from (t = 10): - (A) is true at (t = 10) if (A) is true at some point between (10+4) and (10+13), i.e., between (14) and (23).1436Starting from (t = 12): - (A) is true at (t = 12) if (A) is true at some point between (12+4) and (12+13), i.e., between (16) and (25).1438Since we do not have any data indicating that (A) is true in these new intervals, we cannot propagate (A) forward from the interval ([10, 12]).1440#### Step 3.3: Propagate the Rule Backward Now, let's consider if we can propagate (A) backward to cover the interval ([-469, -221]).1441- For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1442- For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1444that (t+4) to (t+13) overlaps with ([10, 12]).1445However, since ([10, 12]) is far in the future compared to ([-469, -221]), there is no way to find such a (t) that satisfies the rule.1444that (t+4) to (t+13) overlaps with ([10, 12]) and the rule (A leftarrow Diamondplus[4,13]A), we cannot derive (A@[-469,-221]). Therefore, (A@[-469,-221]) is **not** true.1445	1/107		derived using the rule and the given data.
1429141 find the interval (10, 12) which (A) is which to be due. Howevel, we need to see if the rule can propagate (A) backward or forward in time to cover the interval ([-469, -221]).1431#### Step 3.2: Propagate the Rule Forward Let's start from the known in- terval (10, 12) and see if we can propagate (A) forward using the rule. - For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1434Starting from (t = 10): - (A) is true at (t = 10) if (A) is true at some point between (10+4) and (10+13), i.e., between (14) and (23).1435Starting from (t = 12): - (A) is true at (t = 12) if (A) is true at some point between (12+4) and (12+13), i.e., between (16) and (25).1438Since we do not have any data indicating that (A) is true in these new intervals, we cannot propagate (A) forward from the interval [[10, 12]).1440#### Step 3.3: Propagate the Rule Backward Now, let's consider if we can propagate (A) backward to cover the interval [[-469, -221]).1441- For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1442- For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1444To cover the interval ([-469, -221]), we would need to find a time (t) such that (t+4) to (t+13) overlaps with ([10, 12]).1444However, since ([10, 12]) is far in the future compared to ([-469, -221]), there is no way to find such a (t) that satisfies the rule.1445User Prompt1446Mar (14, 13]A), we cannot derive (A@[-469, -221]). Therefore, (A@[-469, -221]) is **not** true.1450User Prompt1451Expected Answer1452 </td <td>1/22</td> <td></td> <td>far from the interval ([10, 12]) where (Λ) is known to be true. However</td>	1/22		far from the interval ([10, 12]) where (Λ) is known to be true. However
1429In the latter in the tart of product of the back while of rot while in the tart1430to cover the interval ([-469, -221]).1431#### Step 3.2: Propagate the Rule Forward Let's start from the known in- terval ([10, 12]) and see if we can propagate (A) forward using the rule. - For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1434Starting from (t = 10): - (A) is true at (t = 10) if (A) is true at some point between (10+4) and (10+13), i.e., between (14) and (23).1435Starting from (t = 12): - (A) is true at (t = 12) if (A) is true at some point between (12+4) and (12+13), i.e., between (16) and (25).1438Since we do not have any data indicating that (A) is true in these new intervals, we cannot propagate (A) forward from the interval ([10, 12]).1440#### Step 3.3: Propagate the Rule Backward Now, let's consider if we can propagate (A) backward to cover the interval ([-469, -221]).1441- For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1442- For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1444To cover the interval ([-469, -221]), we would need to find a time (t) such that (t+4) to (t+13) overlaps with ([10, 12]).1444However, since ([10, 12]) is far in the future compared to ([-469, -221]), there is no way to find such a (t) that satisfies the rule.1445User PromptBased on your previous response, without any explanation state your an- swers to the question. You should only output "true" or "false"1450LLM's outputfalse1453Expected Answertrue	1420		we need to see if the rule can propagate (A) backward or forward in time
1431#### Step 3.2: Propagate the Rule Forward Let's start from the known in- terval ([10, 12]) and see if we can propagate (A) forward using the rule. - For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1434Starting from (t = 10): - (A) is true at (t = 10) if (A) is true at some point between (10+4) and (10+13), i.e., between (14) and (23).1436Starting from (t = 12): - (A) is true at (t = 12) if (A) is true at some point between (12+4) and (12+13), i.e., between (16) and (25).1438Since we do not have any data indicating that (A) is true in these new intervals, we cannot propagate (A) forward from the interval ([10, 12]).1440#### Step 3.3: Propagate the Rule Backward Now, let's consider if we can propagate (A) backward to cover the interval ([-469, -221]). - For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1444To cover the interval ([-469, -221]), we would need to find a time (t) such that (t+4) to (t+13) overlaps with ([10, 12]).1444However, since ([10, 12]) is far in the future compared to ([-469, -221]), there is no way to find such a (t) that satisfies the rule.1447### Conclusion Given the data (A@[10,12]) and the rule (A leftarrow Diamondplus[4,13]A), we cannot derive (A@[-469,-221]). Therefore, (A@[-469,-221]) is **not** true.1450User PromptBased on your previous response, without any explanation state your an- swers to the question. You should only output "true" or "false"1451LLM's outputfalse1452Expected Answertrue	1429		to cover the interval ([-469, -221]).
1432terval ([10, 12]) and see if we can propagate (A) forward using the rule. - For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1434Starting from (t = 10): - (A) is true at (t = 10) if (A) is true at some point between (10+4) and (10+13), i.e., between (14) and (23).1435Starting from (t = 12): - (A) is true at (t = 12) if (A) is true at some point between (12+4) and (12+13), i.e., between (16) and (25).1438Since we do not have any data indicating that (A) is true in these new intervals, we cannot propagate (A) forward from the interval ([10, 12]).1440#### Step 3.3: Propagate the Rule Backward Now, let's consider if we can propagate (A) backward to cover the interval ([-469, -221]).1441- For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1442To cover the interval ([-469, -221]), we would need to find a time (t) such that (t+4) to (t+13) overlaps with ([10, 12]).1444However, since ([10, 12]) is far in the future compared to ([-469, -221]), there is no way to find such a (t) that satisfies the rule.1447### Conclusion Given the data (A@[10,12]) and the rule (A leftarrow Diamondplus[4,13]A), we cannot derive (A@[-469,-221]). Therefore, (A@[-469,-221]) is **not** true.1450User PromptBased on your previous response, without any explanation state your an- swers to the question. You should only output "true" or "false"1451LLM's outputfalse1452LLM's outputfalse	1/21		#### Step 3.2: Propagate the Rule Forward Let's start from the known in-
1432 1433- For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13). Starting from (t = 10): - (A) is true at (t = 10) if (A) is true at some point between (10+4) and (10+13), i.e., between (14) and (23). Starting from (t = 12): - (A) is true at (t = 12) if (A) is true at some point between (12+4) and (12+13), i.e., between (16) and (25). Since we do not have any data indicating that (A) is true in these new intervals, we cannot propagate (A) forward from the interval ([10, 12]). #### Step 3.3: Propagate the Rule Backward Now, let's consider if we can propagate (A) backward to cover the interval ([-469, -221]).1440 1441 1442 1444- For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13). To cover the interval ([-469, -221]), we would need to find a time (t) such that (t+4) to (t+13) overlaps with ([10, 12]). However, since ([10, 12]) is far in the future compared to ([-469, -221]), there is no way to find such a (t) that satisfies the rule. ### Conclusion Given the data (A@[10,12]) and the rule (A leftarrow Diamondplus[4,13]A), we cannot derive (A@[-469,-221]). Therefore, (A@[-469,-221]) is **not** true.1450 1450User PromptBased on your previous response, without any explanation state your an- swers to the question. You should only output "true" or "false"1451 1452LLM's outputfalse	1401		terval ([10, 12]) and see if we can propagate (A) forward using the rule.
1433t+4) and (t+13).1434Starting from (t = 10): - (A) is true at (t = 10) if (A) is true at some point between (10+4) and (10+13), i.e., between (14) and (23).1436Starting from (t = 12): - (A) is true at (t = 12) if (A) is true at some point between (12+4) and (12+13), i.e., between (16) and (25).1438Since we do not have any data indicating that (A) is true in these new intervals, we cannot propagate (A) forward from the interval ([10, 12]).1440#### Step 3.3: Propagate the Rule Backward Now, let's consider if we can propagate (A) backward to cover the interval ([-469, -221]).1441- For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1442- For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1444To cover the interval ([-469, -221]), we would need to find a time (t) such that (t+4) to (t+13) overlaps with ([10, 12]).1446However, since ([10, 12]) is far in the future compared to ([-469, -221]), there is no way to find such a (t) that satisfies the rule.1447### Conclusion Given the data (A@[10,12]) and the rule (A leftarrow Diamondplus[4,13]A), we cannot derive (A@[-469,-221]). Therefore, (A@[-469,-221]) is **not** true.1450User PromptBased on your previous response, without any explanation state your an- swers to the question. You should only output "true" or "false"1451Expected Answertrue	1402		- For (A) to be true at time (t), (A) must be true at some point between (
1435Starting from (t = 10): - (A) is true at (t = 10) if (A) is true at some point between (10+4) and (10+13), i.e., between (14) and (23).1436Starting from (t = 12): - (A) is true at (t = 12) if (A) is true at some point between (12+4) and (12+13), i.e., between (16) and (25).1438Since we do not have any data indicating that (A) is true in these new intervals, we cannot propagate (A) forward from the interval ([10, 12]).1440#### Step 3.3: Propagate the Rule Backward Now, let's consider if we can propagate (A) backward to cover the interval ([-469, -221]).1442- For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1444To cover the interval ([-469, -221]), we would need to find a time (t) such that (t+4) to (t+13) overlaps with ([10, 12]).1446However, since ([10, 12]) is far in the future compared to ([-469, -221]), there is no way to find such a (t) that satisfies the rule.1447### Conclusion Given the data (A@[10,12]) and the rule (A leftarrow Diamondplus[4,13]A), we cannot derive (A@[-469,-221]). Therefore, (A@[-469,-221]) is **not** true.1450User PromptBased on your previous response, without any explanation state your an- swers to the question. You should only output "true" or "false"1451Expected Answertrue	1400		t+4) and (t+13).
1435point between $(10+4)$ and $(10+13)$, i.e., between (14) and (23) .1436Starting from $(t = 12): - (A)$ is true at $(t = 12)$ if (A) is true at some1437point between $(12+4)$ and $(12+13)$, i.e., between (16) and (25) .1438Since we do not have any data indicating that (A) is true in these new1439intervals, we cannot propagate (A) forward from the interval $([10, 12])$.1440#### Step 3.3: Propagate the Rule Backward Now, let's consider if we can1441propagate (A) backward to cover the interval $([-469, -221])$.1442- For (A) to be true at time (t) , (A) must be true at some point between $(t+4)$ and $(t+13)$.1443To cover the interval $([-469, -221])$, we would need to find a time (t) such that $(t+4)$ to $(t+13)$ overlaps with $([10, 12])$.1444However, since $([10, 12])$ is far in the future compared to $([-469, -221])$, there is no way to find such a (t) that satisfies the rule.1447### Conclusion Given the data $(A@[10,12])$ and the rule $(A leftarrow Diamondplus[4,13]A)$, we cannot derive $(A@[-469,-221])$. Therefore, $(A@[-469,-221])$ is **not** true.1450User PromptBased on your previous response, without any explanation state your answers to the question. You should only output "true" or "false"1451LLM's outputfalse1452Expected Answertrue	1434		Starting from ($t = 10$): - (A) is true at ($t = 10$) if (A) is true at some
1436Starting from (t = 12): - (A) is true at (t = 12) if (A) is true at some point between (12+4) and (12+13), i.e., between (16) and (25).1438Since we do not have any data indicating that (A) is true in these new intervals, we cannot propagate (A) forward from the interval ([10, 12]).1440#### Step 3.3: Propagate the Rule Backward Now, let's consider if we can propagate (A) backward to cover the interval ([-469, -221]).1442- For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1443To cover the interval ([-469, -221]), we would need to find a time (t) such that (t+4) to (t+13) overlaps with ([10, 12]).1444Here is no way to find such a (t) that satisfies the rule.1445### Conclusion Given the data (A@[10,12]) and the rule (A leftarrow Diamondplus[4,13]A), we cannot derive (A@[-469,-221]). Therefore, (A@[-469,-221]) is **not** true.1450User PromptBased on your previous response, without any explanation state your an- swers to the question. You should only output "true" or "false"1451LLM's outputfalse	1433		point between (10+4) and (10+13), i.e., between (14) and (23).
1437point between (12+4) and (12+13), i.e., between (16) and (25).1438Since we do not have any data indicating that (A) is true in these new intervals, we cannot propagate (A) forward from the interval ([10, 12]).1440#### Step 3.3: Propagate the Rule Backward Now, let's consider if we can propagate (A) backward to cover the interval ([-469, -221]).1441- For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1442- For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1443To cover the interval ([-469, -221]), we would need to find a time (t) such that (t+4) to (t+13) overlaps with ([10, 12]).1444However, since ([10, 12]) is far in the future compared to ([-469, -221]), there is no way to find such a (t) that satisfies the rule.1447### Conclusion Given the data (A@[10,12]) and the rule (A leftarrow Diamondplus[4,13]A), we cannot derive (A@[-469,-221]). Therefore, (A@[-469,-221]) is **not** true.1450User PromptBased on your previous response, without any explanation state your an- swers to the question. You should only output "true" or "false"1451LLM's outputfalse1453Expected Answertrue	1430		Starting from ($t = 12$): - (A) is true at ($t = 12$) if (A) is true at some
1430Since we do not have any data indicating that (A) is true in these new intervals, we cannot propagate (A) forward from the interval ([10, 12]). #### Step 3.3: Propagate the Rule Backward Now, let's consider if we can propagate (A) backward to cover the interval ([-469, -221]). - For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1442 1443 1444 1444 1444 1445 1446 1446 1446 1447- For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13). To cover the interval ([-469, -221]), we would need to find a time (t) such that (t+4) to (t+13) overlaps with ([10, 12]). However, since ([10, 12]) is far in the future compared to ([-469, -221]), there is no way to find such a (t) that satisfies the rule. ### Conclusion Given the data (A@[10,12]) and the rule (A leftarrow Diamondplus[4,13]A), we cannot derive (A@[-469,-221]). Therefore, (A@[-469,-221]) is **not** true.1450 1450User PromptBased on your previous response, without any explanation state your an- swers to the question. You should only output "true" or "false"1451 1453Expected Answertrue	1437		point between $(12+4)$ and $(12+13)$, i.e., between (16) and (25) .
1439Intervals, we cannot propagate (A) forward from the interval ([10, 12]).1440#### Step 3.3: Propagate the Rule Backward Now, let's consider if we can propagate (A) backward to cover the interval ([-469, -221]). - For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1443- For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1444To cover the interval ([-469, -221]), we would need to find a time (t) such that (t+4) to (t+13) overlaps with ([10, 12]).1445However, since ([10, 12]) is far in the future compared to ([-469, -221]), there is no way to find such a (t) that satisfies the rule.1446### Conclusion Given the data (A@[10,12]) and the rule (A leftarrow Diamondplus[4,13]A), we cannot derive (A@[-469,-221]). Therefore, (A@[-469,-221]) is **not** true.1450User PromptBased on your previous response, without any explanation state your an- swers to the question. You should only output "true" or "false"1451LLM's outputfalse1453Expected Answertrue	1430		Since we do not have any data indicating that (A) is true in these new intervals we connect account (A) forward form the interval ([10, 12])
1440#### Step 5.5. Propagate the Rule Backward Now, let's consider If we can propagate (A) backward to cover the interval ([-469, -221]). - For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13).1442 1443 1444 1444- For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13). To cover the interval ([-469, -221]), we would need to find a time (t) such that (t+4) to (t+13) overlaps with ([10, 12]).1445 1446 1447 1446 1447However, since ([10, 12]) is far in the future compared to ([-469, -221]), there is no way to find such a (t) that satisfies the rule. ### Conclusion Given the data (A@[10,12]) and the rule (A leftarrow Diamondplus[4,13]A), we cannot derive (A@[-469,-221]). Therefore, (A@[-469,-221]) is **not** true.1450User PromptBased on your previous response, without any explanation state your an- swers to the question. You should only output "true" or "false"1451LLM's outputfalse1453Expected Answertrue	1439		##### Step 3.3: Propagate the Rule Backward Now let's consider if we can
 For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13). For (A) to be true at time (t), (A) must be true at some point between (t+4) and (t+13). To cover the interval ([-469, -221]), we would need to find a time (t) such that (t+4) to (t+13) overlaps with ([10, 12]). However, since ([10, 12]) is far in the future compared to ([-469, -221]), there is no way to find such a (t) that satisfies the rule. ### Conclusion Given the data (A@[10,12]) and the rule (A leftarrow Diamondplus[4,13]A), we cannot derive (A@[-469,-221]). Therefore, (A@[-469,-221]) is **not** true. User Prompt Based on your previous response, without any explanation state your answers to the question. You should only output "true" or "false" LLM's output false Expected Answer true 	1440		propagate (A) backward to cover the interval ([-469 -221])
144214431443t+4) and (t+13).1444To cover the interval ([-469, -221]), we would need to find a time (t) such that (t+4) to (t+13) overlaps with ([10, 12]).1445However, since ([10, 12]) is far in the future compared to ([-469, -221]), there is no way to find such a (t) that satisfies the rule.1446### Conclusion Given the data (A@[10,12]) and the rule (A leftarrow Diamondplus[4,13]A), we cannot derive (A@[-469,-221]). Therefore, (A@[-469,-221]) is **not** true.1450User PromptBased on your previous response, without any explanation state your an- swers to the question. You should only output "true" or "false"1451LLM's outputfalse1453Expected Answertrue	1441		- For (A) to be true at time (t), (A) must be true at some point between (
1443To cover the interval ([-469, -221]), we would need to find a time (t) such that (t+4) to (t+13) overlaps with ([10, 12]).1445However, since ([10, 12]) is far in the future compared to ([-469, -221]), there is no way to find such a (t) that satisfies the rule.1447### Conclusion Given the data (A@[10,12]) and the rule (A leftarrow Diamondplus[4,13]A), we cannot derive (A@[-469,-221]). Therefore, (A@[-469,-221]) is **not** true.1450User PromptBased on your previous response, without any explanation state your an- swers to the question. You should only output "true" or "false"1451LLM's outputfalse1453Expected Answertrue	1442		t+4) and ($t+13$).
1444 1445 1446 1446that (t+4) to (t+13) overlaps with ([10, 12]). However, since ([10, 12]) is far in the future compared to ([-469, -221]), there is no way to find such a (t) that satisfies the rule. ### Conclusion Given the data (A@[10,12]) and the rule (A leftarrow Diamondplus[4,13]A), we cannot derive (A@[-469,-221]). Therefore, (A@[-469,-221]) is **not** true.1449 1450 1450 1451 1452 1452 1453User Prompt LLM's output false	1443		To cover the interval ($[-469, -221]$), we would need to find a time (t) such
1445 1446 1447However, since ([10, 12]) is far in the future compared to ([-469, -221]), there is no way to find such a (t) that satisfies the rule. ### Conclusion Given the data (A@[10,12]) and the rule (A leftarrow Diamondplus[4,13]A), we cannot derive (A@[-469,-221]). Therefore, (A@[-469,-221]) is **not** true.1449User PromptBased on your previous response, without any explanation state your an- swers to the question. You should only output "true" or "false"1450LLM's outputfalse1453Expected Answertrue	1444		that (t+4) to (t+13) overlaps with ([10, 12]).
1446 1447 1448there is no way to find such a (t) that satisfies the rule. ### Conclusion Given the data (A@[10,12]) and the rule (A leftarrow Diamondplus[4,13]A), we cannot derive (A@[-469,-221]). Therefore, (A@[-469,-221]) is **not** true.1450User PromptBased on your previous response, without any explanation state your an- swers to the question. You should only output "true" or "false"1451LLM's outputfalse1453Expected Answertrue	1445		However, since ([10, 12]) is far in the future compared to ([-469, -221]),
1447### Conclusion Given the data (A@[10,12]) and the rule (A leftarrow Diamondplus[4,13]A), we cannot derive (A@[-469,-221]). Therefore, (A@[-469,-221]) is **not** true.1450User PromptBased on your previous response, without any explanation state your an- swers to the question. You should only output "true" or "false"1451LLM's outputfalse1453Expected Answertrue	1440		there is no way to find such a (t) that satisfies the rule.
1440 Diamondplus[4,13]A), we cannot derive (A@[-469,-221]). Therefore, (A@[-469,-221]) is **not** true. 1449 User Prompt Based on your previous response, without any explanation state your answers to the question. You should only output "true" or "false" 1450 LLM's output false 1453 Expected Answer true	1447		### Conclusion Given the data (A@[10,12]) and the rule (A leftarrow
1449 A@[-409,-221]) is **not** true. 1450 User Prompt Based on your previous response, without any explanation state your an- swers to the question. You should only output "true" or "false" 1451 LLM's output false 1453 Expected Answer true	1440		Diamondplus[4,13]A), we cannot derive ($A@[-469,-221]$). Therefore, (
Instruction Instruction Based on your previous response, without any explanation state your answers to the question. You should only output "true" or "false" Instruction Instruction Instruction	1449	Lloon Durant	A@[-409,-221]) is **not** true.
1451 Swers to the question. You should only output true or faise 1452 LLM's output 1453 Expected Answer	1450	User Prompt	based on your previous response, without any explanation state your an-
1452 Expected Answer true	1451	LIM's output	sweis to the question. Tou should only output true of Talse
1400 Expected / Inswer ute	1452	Expected Answer	true
1454	1403	Expected miswel	

False Positive Example:

1458		
1459	System Prompt	You are given a dataset and a temporal rule, and your task is to judge whether
1460		the given fact is entailed by the dataset and the rule. The rules are expressed
1461		as DatalogMTL, a knowledge representation language that extends Datalog
1462		with operators from metric temporal logic (MTL). The semantics of four
1463		MTL operators are given as follows: If Diamondminus[a,b]A is true at the
1464		time t, it requires that A needs to be true at some time between t-b and
1465		t-a.If Boxminus[a,b]A is true at the time t, it requires that A needs to be
1466		true continuously between t-b and t-a.lf Diamondplus[a,b]A is true at the
1467		time t, it requires that A needs to be true at some point between $t+a$ and the If Daymburghe heat A is true at the time t it requires that A needs to be true
1468		t+0.11 Boxplus[a,b]A is true at the time t, it requires that A needs to be true
1469	Llear Dromnt	Continuously between t+a and t+b. Now we have some temporal data and some rules, data: $A \otimes [2,2]$
1470	User Prompt	Now we have some temporal data and some rules, data: $A \otimes [2,2]$
1/171		Is $A @[-24.25]$ true or not? Do not answer directly think step by step
1/170	LLM's output	To determine whether $(A \otimes [-24 25])$ is true given the data and the rule, we
1472	EEM 5 Output	need to analyze the rule and the data step by step.
1473		### Step-by-Step Analysis
1474		1. **Understanding the Rule:** - The rule is (A leftarrow Diamondmi-
1473		nus[4,15]A) This means that (A) is true at time (t) if (A) was true at
1470		some time between $(t-15)$ and $(t-4)$.
1477		2. **Given Data:** - (A@[2,2]) means (A) is true at time 2.
1478		3. **Determine the Time Intervals:** - We need to check if (A) can be true
1479		at any time (t) within the interval ([-24, 25]).
1480		4. **Applying the Rule:** - For (A) to be true at any time (t) within ([-24,
1481		25]), there must be some time (t') within ([t-15, t-4]) where (A) is true.
1482		5. **Checking the Data:** - We know (A) is true at time 2 We need to
1483		find if there exists any (t) in $([-24, 25])$ such that (2) lies within $([t-15, 41)$
1484		[-4]). (**Einding the Valid (t):** Ear (2) to be within ([t 15 t 4]): [t 15 lag
1485		0. Finding the value (1), $r = For(2)$ to be writing ([t-15, t-4]). [t-15 leq 2 leg t 4]. Solving these inequalities: [t 15 leg 2 implies t leg 17] [2 leg
1486		t_{-4} implies t geq 6] - Therefore (t) must be in the interval ([6, 17])
1487		7. **Checking the Interval ([-24, 25]):** - The interval ([6, 17]) is within
1488		the interval ($[-24, 25]$).
1489		### Conclusion Since there exists a time (t) within ([-24, 25]) (specifically
1490		within ([6, 17])) where the rule can be satisfied by the given data, ($A@[-$
1491		24,25]) is true.
1492	User Prompt	Based on your previous response, without any explanation state your an-
1493		swers to the question. You should only output "true" or "false"
1494	LLM's output	true
1495	Expected Answer	false

1498

1499 1500

E CAN LLM UNDERSTAND SYMBOLIC REPRESENTATION?

1501 We ask LLM to translate generated natural language representations of logic rules to symbolic repre-1502 sentations and them compare the translated symbolic representations with the ground truth symbolic representation to verify if LLM has the ability to understand symbolic representations. Specifically, 1503 we passed the same prompt that used in our symbolic evaluations, "The rules are expressed as Data-1504 logMTL, a.....If Boxplus[a,b]A is true at the time t, it requires that A needs to be true continuously 1505 between t+a and t+b.", into LLMs, along with few examples telling LLM the output format, then 1506 ask LLM to translate verbalized samples into symbolic ones. If the translated symbolic rule from 1507 the verbalized rule is exactly the same as the original symbolic rule, then we consider LLM has the 1508 ability to understand both the symbolic rule and the verbalized rule. 1509

We passed 50 samples selected from MultiRules subset, which is considered the most challenging, into the LLMs. LLM accurately translated 96% of testing samples from verbalized representations to symbolic representations, so we consider that LLM can understand the symbolic representation.

1512 F DETAILED BENCHMARK CONSTRUCTION PSEUDO CODE 1513 1514

ır	dataset generation algorithm is driven by generating rules. In a high level view, it generate rules
ne l	by one in a same context, while the generation process for each rule contains the context check
nsu	y one in a sume context, while the generation process for each full contains the context check,
	ring the generated rules are non-trivial.
	ring the generated rules are non-trivial.
lgo ara ara	The set of features Enabled meters: <i>f</i> : The set of rules
Algc Para Para Para Para Outj $G \leftarrow$ whill • •	ring the generated rules are non-trivial. prithm 1: Generate imeters: f : The set of features Enabled imeters: N : The number of rules imeters: V : A boolean flag to control if the program should generate a positive sample or a negative sample put: A problem instance I containing a set of rules, a set of data, a query and a boolean value representing whether the query is valid or not. EmptyGraph(); e i in IN do lo $G \leftarrow GenerateGraph(G)$; while n in $G.nodes$ do Assign node with random values end $G \leftarrow GenerateRules(G)$ while New Info can be Inferred from I ; $as Data \leftarrow Extract Rules associted with G:$
lgc ara ara ara hil (hil () hil () hil () ()	ring the generated rules are non-trivial. prithm 1: Generate uneters: f : The set of features Enabled uneters: N : The number of rules uneters: V : A boolean flag to control if the program should generate a positive sample or a negative sample put: A problem instance I containing a set of rules, a set of data, a query and a boolean value representing whether the query is valid or not. EmptyGraph(); e i in $1N$ do lo lo lo lo l $G \leftarrow GenerateGraph(G)$; while n in $G.nodes$ do l Assign node with random values end $G \leftarrow GenerateRules(G)$ while New Info can be Inferred from I ; $es, Data \leftarrow Extract Rules$ associted with G ; $eaNew \leftarrow Facts Inferred From Graph G$;
lgo ara ara ara hil (hil (ula ula ula ula ula ula ula ula ula	ring the generated rules are non-trivial. prithm 1: Generate imeters: f : The set of features Enabled imeters: N : The number of rules imeters: N : The number of rules imeters: V : A boolean flag to control if the program should generate a positive sample or a negative sample put: A problem instance I containing a set of rules, a set of data, a query and a boolean value representing whether the query is valid or not. EmptyGraph(); e i in 1 N do lo l $G \leftarrow GenerateGraph(G)$; while n in G .nodes do i Assign node with random values end $G \leftarrow GenerateRules(G)$ while New Info can be Inferred from I ; es , Data \leftarrow Extract Rules associted with G ; $aNew \leftarrow$ Facts Inferred From Graph G ; $ryEntity$, Interval \leftarrow Randomly Select From DeltaNew;
$\frac{\log \alpha}{\log \alpha}$	ring the generated rules are non-trivial. prithm 1: Generate imeters: f : The set of features Enabled imeters: N : The number of rules imeters: N : The number of rules imeters: N : The number of rules imeters: V : A boolean flag to control if the program should generate a positive sample or a negative sample put: A problem instance I containing a set of rules, a set of data, a query and a boolean value representing whether the query is valid or not. EmptyGraph(); e <i>i in</i> $1N$ do lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo
$\frac{ \mathbf{g}_{0} }{ \mathbf{g}_{0} }$	ring the generated rules are non-trivial. prithm 1: Generate imeters: f: The set of features Enabled imeters: N: The number of rules imeters: N: A boolean flag to control if the program should generate a positive sample or a negative sample put: A problem instance I containing a set of rules, a set of data, a query and a boolean value representing whether the query is valid or not. • EmptyGraph(); e i in 1N do lo l $G \leftarrow GenerateGraph(G);$ while n in G.nodes do i Assign node with random values end $G \leftarrow GenerateRules(G)$ while New Info can be Inferred from I; es, Data \leftarrow Extract Rules associted with G; raNew \leftarrow Facts Inferred From Graph G; ryEntity, Interval \leftarrow Randomly Select From DeltaNew; then QueryInterval \leftarrow A random sub-interval from Interval;
lige Para Para Para Para Put] C ← Vhil C ← Vhil C ← C ← C ← C ← C ← C ← C ← C ← C ← C ←	ring the generated rules are non-trivial. prithm 1: Generate imeters: f: The set of features Enabled imeters: N: The number of rules imeters: N: A boolean flag to control if the program should generate a positive sample or a negative sample put: A problem instance I containing a set of rules, a set of data, a query and a boolean value representing whether the query is valid or not. • EmptyGraph(); e i in 1N do lo l $G \leftarrow GenerateGraph(G);$ while n in G.nodes do Assign node with random values end $G \leftarrow GenerateRules(G)$ while New Info can be Inferred from I; es, Data \leftarrow Extract Rules associted with G; raNew \leftarrow Facts Inferred From Graph G; ryEntity, Interval \leftarrow Randomly Select From DeltaNew; then QueryInterval \leftarrow A random sub-interval from Interval; Duese Unterval \leftarrow A random sub-interval that is not in Interval;
Algo Para Para Para Para Para Vhil (vhil (vhil (vhil 2 ulo 2 V (lse (nd	ring the generated rules are non-trivial. prithm 1: Generate umeters: f: The set of features Enabled umeters: N: The number of rules umeters: N: A boolean flag to control if the program should generate a positive sample or a negative sample put: A problem instance I containing a set of rules, a set of data, a query and a boolean value representing whether the query is valid or not. • EmptyGraph(); e i in 1N do lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo

cates such as A, B and C. We are going to attach details information about predicates and rules into
 the corresponding nodes and edges of the graph, but at this time we only need the structure of the
 graph, i.e. nodes and edges don't have special information attached.

Input: (Paramet	
Paramet	7: The existing graph
	rers: f: The set of features Enabled
Jutput:	G: The generated graph (including the old information in the existing graph) L = t[V]. The list of new nodes, representing and lists in the new result.
output:	List[V]: The list of new nodes, representing predicates, in the new graph
utput:	V_0 : The output node which depends on the some other nodes (in case that recursive is not enabled in f) in List[V]
Now No	not enabled in f) in $List[V]$
Vewivo Jetermii	$ue \leftarrow []$
nodes t	and r based on f
$V \leftarrow ra$	ndom(l r)
vhile <i>i i</i>	$n 1 \dots N do$
$p \leftarrow p$	A randomly assigned predicate:
G.A	ddNode(p):
Neu	Node.Push(p);
nd	
DutNod	$le \leftarrow RandomSelect(NewNode);$
vhile <i>p</i> i	n NewNode do
if "re	cursive" not in f and $p == OutNode$ then
	ontinue;
end	
end G.A	ddEdge(p,OutNode)
end G.A	ddEdge(p,OutNode)
end G.A. nd eturn C fter the f the gra- beneration we will s	ddEdge(p, OutNode) <i>G. NewNode, OutNode</i> e structure of the graph is generated, we are going to attach rule information to each edge aph using the Rule Generation algorithm 3. Since we are doing Graph Generation and Rule on alternately, in the rule generation we only care about edges that don't already has a rule kip the edges that already has a rule associated with that.
end G.A. end After the of the gra Generati ve will s	 ddEdge(p, OutNode) <i>G. NewNode, OutNode</i> e structure of the graph is generated, we are going to attach rule information to each edge aph using the Rule Generation algorithm 3. Since we are doing Graph Generation and Rule on alternately, in the rule generation we only care about edges that don't already has a rule kip the edges that already has a rule associated with that. m 3: Rule Generation
end G.A. end return C After the of the gra Generati we will s Algorith Input: C	 ddEdge(p, OutNode) <i>G. NewNode, OutNode</i> e structure of the graph is generated, we are going to attach rule information to each edge aph using the Rule Generation algorithm 3. Since we are doing Graph Generation and Rule on alternately, in the rule generation we only care about edges that don't already has a rule kip the edges that already has a rule associated with that. m 3: Rule Generation <i>G. The existing graph</i>
end G.A. end return C After the of the gradient Generation we will s Algorith Input: C Paramet	ddEdge(p, OutNode) G. NewNode, OutNode e structure of the graph is generated, we are going to attach rule information to each edge aph using the Rule Generation algorithm 3. Since we are doing Graph Generation and Rule on alternately, in the rule generation we only care about edges that don't already has a rule kip the edges that already has a rule associated with that. m 3: Rule Generation c: The existing graph cers: f: The set of features Enabled
end G.A. end return C After the of the gr. Generati we will s Algorith Input: C Paramet Output:	ddEdge(p, OutNode) G. NewNode, OutNode e structure of the graph is generated, we are going to attach rule information to each edge aph using the Rule Generation algorithm 3. Since we are doing Graph Generation and Rule on alternately, in the rule generation we only care about edges that don't already has a rule kip the edges that already has a rule associated with that. m 3: Rule Generation <i>G</i> : The existing graph <i>G</i> : The generated graph (including the old information in the existing graph)
end G.A. end return C After the of the gra- Generati we will s Algorith Input: C Paramet Output: Selected	$\frac{ddEdge(p, OutNode)}{G, NewNode, OutNode}$ e structure of the graph is generated, we are going to attach rule information to each edge aph using the Rule Generation algorithm 3. Since we are doing Graph Generation and Rule on alternately, in the rule generation we only care about edges that don't already has a rule kip the edges that already has a rule associated with that. m 3: Rule Generation <i>G</i> : The existing graph <i>G</i> : The set of features Enabled <i>G</i> : The generated graph (including the old information in the existing graph) $\frac{Op \leftarrow Set()}{Op \leftarrow Set()}$
end G.A. end return C After the of the gra Generati we will s Algorith Input: C Paramet Output: Selected Selected	$\frac{ddEdge(p, OutNode)}{dege(p, OutNode)}$ is structure of the graph is generated, we are going to attach rule information to each edge aph using the Rule Generation algorithm 3. Since we are doing Graph Generation and Rule on alternately, in the rule generation we only care about edges that don't already has a rule kip the edges that already has a rule associated with that. m 3: Rule Generation $\frac{1}{2}$: The existing graph ters: f: The set of features Enabled G: The generated graph (including the old information in the existing graph) $\frac{100p \leftarrow Set()}{100p.add(RandomSelect(Boxminus, Boxplus, Diamondplusm, Diamondminus))}$
end G.A. end return C After the of the gra Generati we will s Algorith Input: C Paramet Output: Selected if "mixa	$\frac{ddEdge(p, OutNode)}{G, NewNode, OutNode}$ e structure of the graph is generated, we are going to attach rule information to each edge aph using the Rule Generation algorithm 3. Since we are doing Graph Generation and Rule on alternately, in the rule generation we only care about edges that don't already has a rule kip the edges that already has a rule associated with that. m 3: Rule Generation F: The existing graph ters: f: The set of features Enabled G: The generated graph (including the old information in the existing graph) $\frac{Op \leftarrow Set()}{Op.add(RandomSelect(Boxminus, Boxplus, Diamondplusm, Diamondminus)))}{Doperators'' in f then here a set of a dd more operators to SelectedOp: \frac{Op \leftarrow Set()}{Op}$
After the of the gra Generati we will s Algorith Input: C Paramet Output: Selected if "mixed and	$\frac{ddEdge(p, OutNode)}{G, NewNode, OutNode}$ e structure of the graph is generated, we are going to attach rule information to each edge aph using the Rule Generation algorithm 3. Since we are doing Graph Generation and Rule on alternately, in the rule generation we only care about edges that don't already has a rule kip the edges that already has a rule associated with that. m 3: Rule Generation W: The existing graph G: The generated graph (including the old information in the existing graph) $Op \leftarrow Set()$ Op.add(RandomSelect(Boxminus, Boxplus, Diamondplusm, Diamondminus))) $d_operators'' in f$ then tomly select and add more operators to SelectedOp;
end G.A. end return C After the of the gra Generati we will s Algorith Input: C Paramet Output: Selected if "mixe end while Ed	$\frac{ddEdge(p, OutNode)}{ds, NewNode, OutNode}$ a structure of the graph is generated, we are going to attach rule information to each edge aph using the Rule Generation algorithm 3. Since we are doing Graph Generation and Rule on alternately, in the rule generation we only care about edges that don't already has a rule kip the edges that already has a rule associated with that. m 3: Rule Generation C: The existing graph ers: f: The set of features Enabled G: The generated graph (including the old information in the existing graph) $\frac{dOp \leftarrow Set()}{Dp.add(RandomSelect(Boxminus, Boxplus, Diamondplusm, Diamondminus)))}{Dp d_Operators'' in f then loomly select and add more operators to SelectedOp; date in G do$
end G.A. end return C After the of the gra Generati we will s Algorith Input: C Paramet Output: Selected Selected if "mixe end while Ed while Ed	$\frac{ddEdge(p, OutNode)}{G, NewNode, OutNode}$ a structure of the graph is generated, we are going to attach rule information to each edge aph using the Rule Generation algorithm 3. Since we are doing Graph Generation and Rule on alternately, in the rule generation we only care about edges that don't already has a rule kip the edges that already has a rule associated with that. m 3: Rule Generation C: The existing graph ers: f: The set of features Enabled G: The generated graph (including the old information in the existing graph) $\frac{Op \leftarrow Set()}{Op.add(RandomSelect(Boxminus, Boxplus, Diamondplusm, Diamondminus)))}{d_operators'' in f then lomly select and add more operators to SelectedOp; \frac{dge in G \mathbf{do}}{a \leftarrow G}$
end G.A. end return C After the of the gra Generati we will s Algorith Input: C Paramet Output: Selected if "mixe end while Ed u, v, if Na	$\frac{ddEdge(p, OutNode)}{G, NewNode, OutNode}$ e structure of the graph is generated, we are going to attach rule information to each edge aph using the Rule Generation algorithm 3. Since we are doing Graph Generation and Rule on alternately, in the rule generation we only care about edges that don't already has a rule kip the edges that already has a rule associated with that. m 3: Rule Generation C: The existing graph cers: f: The set of features Enabled G: The generated graph (including the old information in the existing graph) $\frac{Op}{Op.add(RandomSelect(Boxminus, Boxplus, Diamondplusm, Diamondminus)))}{Dog_operators'' in f them tomly select and add more operators to SelectedOp; \frac{dge in G \text{ do}}{a \leftarrow G;}rule is associated with Edge then$
end G.A. end return C After the of the gradement Generati we will s Algorith Input: C Paramet Output: Selected if "mixe while Ed u, v, if No if No	$\frac{ddEdge(p, OutNode)}{dege(p, OutNode)}$ is structure of the graph is generated, we are going to attach rule information to each edge aph using the Rule Generation algorithm 3. Since we are doing Graph Generation and Rule on alternately, in the rule generation we only care about edges that don't already has a rule kip the edges that already has a rule associated with that. m 3: Rule Generation c: The existing graph c: The set of features Enabled <i>G</i> : The generated graph (including the old information in the existing graph) $\frac{dOp \leftarrow Set()}{dOp.add(RandomSelect(Boxminus, Boxplus, Diamondplusm, Diamondminus))}{dogerators'' in f then} loomly select and add more operators to SelectedOp; \frac{dge in G \mathbf{do}}{a \leftarrow G;}rule is associated with Edge thenDp \leftarrow Randomly select an operator from SelectedOp;$
end $G.Ad$ return CAfter theof the gradeGeneratiwe will sAlgorithInput: CParametOutput:Selectedif "mixedendwhile Ed $u, v,$ if Not $u, v, v,$	$\frac{ddEdge(p, OutNode)}{ddEdge(p, OutNode)}$ is structure of the graph is generated, we are going to attach rule information to each edge aph using the Rule Generation algorithm 3. Since we are doing Graph Generation and Rule on alternately, in the rule generation we only care about edges that don't already has a rule kip the edges that already has a rule associated with that. m 3: Rule Generation $\frac{1}{2}$: The existing graph $\frac{1}{2}$: The set of features Enabled G: The generated graph (including the old information in the existing graph) $\frac{1}{2}Op \leftarrow Set()$ $\frac{1}{2}Op.add(RandomSelect(Boxminus, Boxplus, Diamondplusm, Diamondminus)))$ $\frac{1}{2}Operators'' in f$ then tomly select and add more operators to SelectedOp; $\frac{1}{2}dge in G$ do $a \leftarrow G$; rule is associated with Edge then $Dp \leftarrow Randomly select an operator from SelectedOp;$ $nterval \leftarrow Randomly create an interval;$
end G.A. return C After the of the grading Generati we will s Algorith Input: C Paramet Output: Selected if "mixed while Ed u, v, if No Input: C	$\frac{ddEdge(p, OutNode)}{ddEdge(p, OutNode)}$ structure of the graph is generated, we are going to attach rule information to each edge aph using the Rule Generation algorithm 3. Since we are doing Graph Generation and Rule on alternately, in the rule generation we only care about edges that don't already has a rule kip the edges that already has a rule associated with that. m 3: Rule Generation $\frac{1}{2}$: The existing graph $\frac{1}{2}$: The set of features Enabled G: The generated graph (including the old information in the existing graph) $\frac{1}{2}Op \leftarrow Set()$ $\frac{1}{2}Op.add(RandomSelect(Boxminus, Boxplus, Diamondplusm, Diamondminus)))$ $\frac{1}{2}Operators'' in f then$ tomly select and add more operators to SelectedOp; $\frac{1}{2}dge in G do$ $a \leftarrow G;$ rule is associated with Edge then $Dp \leftarrow Randomly select an operator from SelectedOp;$ $nterval \leftarrow Randomly create an interval;$ Create an item literal with Op and Interval and associated that with Edge;
end G.A. return C After the of the grading Generati we will s Algorith Input: C Paramet Output: Selected if "mixed end while Ed u, v, if No 0	ddEdge(p, OutNode) $f, NewNode, OutNode$ estructure of the graph is generated, we are going to attach rule information to each edge aph using the Rule Generation algorithm 3. Since we are doing Graph Generation and Rule on alternately, in the rule generation we only care about edges that don't already has a rule kip the edges that already has a rule associated with that. m 3: Rule Generation $f: The existing graph ers: f: The set of features Enabled G: The generated graph (including the old information in the existing graph) Dp \leftarrow Set() Dp.add(RandomSelect(Boxminus, Boxplus, Diamondplusm, Diamondminus))) d_operators'' in f then tomly select and add more operators to SelectedOp; dge in G \text{ do} a \leftarrow G; rule is associated with Edge then Dp \leftarrow Randomly select an operator from SelectedOp; nterval \leftarrow Randomly create an interval; Create an item literal with Op and Interval and associated that with Edge;$
end G.A. return C After the of the gra Generati we will s Algorith Input: C Paramet Output: Selected if "mixe end while Ed u, v, if No 0 if No 0 end end end end	$\frac{ddEdge(p, OutNode)}{G, NewNode, OutNode}$ e structure of the graph is generated, we are going to attach rule information to each edge aph using the Rule Generation algorithm 3. Since we are doing Graph Generation and Rule on alternately, in the rule generation we only care about edges that don't already has a rule kip the edges that already has a rule associated with that. m 3: Rule Generation $\frac{1}{G}$: The existing graph ers: f: The set of features Enabled G: The generated graph (including the old information in the existing graph) $\frac{1}{Op} \leftarrow Set()$ $\frac{1}{Op.add}(RandomSelect(Boxminus, Boxplus, Diamondplusm, Diamondminus)))$ $\frac{1}{d_operators'' in f then}$ tomly select and add more operators to SelectedOp; $\frac{1}{d_e} in G \text{ do}$ $a \leftarrow G;$ rule is associated with Edge then $Dp \leftarrow Randomly create an interval;$ Create an item literal with Op and Interval and associated that with $Edge;$