

---

# Large Language Models are In-Context Semantic Reasoners rather than Symbolic Reasoners

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1 The emergent few-shot reasoning capabilities of Large Language Models (LLMs)  
2 have excited the natural language and machine learning community over recent  
3 years. Despite of numerous successful applications, the underlying mechanism of  
4 such in-context capabilities still remains unclear. In this work, we hypothesize that  
5 the learned *semantics* of language tokens do the most heavy lifting during the reason-  
6 ing process. Different from human’s symbolic reasoning process, the semantic  
7 representations of LLMs could create strong connections among tokens, thus compos-  
8 ing a superficial logical chain. To test our hypothesis, we decouple semantics  
9 from the language reasoning process and evaluate three kinds of reasoning abilities,  
10 *i.e.*, deduction, induction and abduction. Our findings reveal that semantics play  
11 a vital role in LLMs’ in-context reasoning—LLMs perform significantly better  
12 when semantics are consistent with commonsense but struggle to solve symbolic  
13 or counter-commonsense reasoning tasks by leveraging in-context new knowledge.  
14 The surprising observations question whether modern LLMs have mastered the  
15 inductive, deductive and abductive reasoning abilities as in human intelligence, and  
16 motivate research on unveiling the magic existing within the black-box LLMs. On  
17 the whole, our analysis provides a novel perspective on the role of semantics in  
18 developing and evaluating language models’ reasoning abilities.

## 19 1 Introduction

20 In recent years, Large Language Models (LLMs) have achieved impressive performance on a variety  
21 of natural language tasks, including question answering, text summarization, machine translation,  
22 logic reasoning, *etc.* These successes have been largely attributed to the emergent ability of LLMs to  
23 utilize a “zero-shot” or “few-shot” learning approach without any gradient updates—a task description  
24 or a few examples are provided to guide their reasoning process [1–4]. One typical example is the  
25 “chain-of-thought (CoT)” approach, involving reasoning demonstrations or a simple prompt such as  
26 “Let’s think step by step” to perform complex reasoning tasks [5, 6].

27 Despite the powerful and versatile in-context learning ability of LLMs, the underlying mechanisms  
28 by which they operate within a given context still remain unclear. Previous works investigate which  
29 aspects of the given examples contribute to the final task performance, including ground-truth labels  
30 and example ordering [7–9]. Another line of recent work has focused on explaining and leveraging  
31 the in-context learning (ICL) mechanism [10–13]. However, the basic problem they have in common  
32 is that the in-context prompts they input are based on natural language queries to investigate the  
33 reasoning abilities of LLMs. According to the Dual Process Theory [14, 15], humans are capable of  
34 using symbolic reasoning with System II to solve complex logical reasoning problems. To fill the  
35 research gap, we systematically study the in-context reasoning ability of LLMs by decoupling the

36 semantics from the language reasoning process. With extensive experiments, we aim to answer the  
37 following research question: *Are LLMs good in-context reasoners without semantics?*

38 In this work, we hypothesize that the learned semantics of language tokens play an important role in  
39 the reasoning process, creating strong connections among tokens which help to compose a superficial  
40 logical chain (shortcut) instead of really performing the formal reasoning process. To test our  
41 hypothesis, given symbolic knowledge (facts and rules), we test three kinds of reasoning abilities  
42 (*i.e.*, deduction, induction, abduction) on a newly proposed synthetic dataset: Symbolic Tree dataset,  
43 which is composed of closed-world, noise-free, multi-hop symbolic reasoning data generated with  
44 logical rules. Besides, we also experiment with ProofWriter [16] task, containing questions whose  
45 answers require multi-hop reasoning. Our findings suggest that semantics indeed play a vital role  
46 in LLMs’ in-context reasoning: When semantics are consistent with commonsense, LLMs perform  
47 fairly well; when semantics are decoupled or in the counter-commonsense context, LLMs struggle  
48 to solve the reasoning tasks by leveraging in-context new knowledge. Moreover, we also study the  
49 memorization ability of LLMs to memorize new symbols and semantics information, allowing us to  
50 investigate the role of semantics on LLMs’ knowledge update ability.

51 To the best of our knowledge, this is the first study of the effect of semantics on LLMs’ in-context  
52 reasoning abilities. Our analysis underscores the importance of semantics in LLMs’ reasoning ability  
53 and questions whether modern LLMs have mastered the formal reasoning abilities as in human  
54 intelligence. We hope our findings can provide a novel perspective on the role of semantics in LLMs’  
55 in-context abilities, and inspire further research on unveiling the magic inside the black-box LLMs.

## 56 2 Related Works

57 **Reasoning in LLMs** Reasoning is a fundamental cognitive process involving logical inferences and  
58 conclusions based on given information. Developing models with strong reasoning capabilities has  
59 attracted increasing attention [17]. Since then, various benchmarks focusing on different aspects of  
60 reasoning have been proposed, including natural language inference (NLI) [18–20], commonsense  
61 reasoning [21, 22], multi-hop reasoning [23, 24], **logic reasoning [25–27] *etc.* These benchmarks**  
62 **are evaluated within semantics. However, we focus on decoupling semantics to test LLMs’ pure**  
63 **symbolic logic reasoning capabilities.** In recent years, there has been growing interests in studying  
64 the reasoning abilities of LLMs. For example, “chain-of-thought (CoT)” [5, 28, 6] is proposed to  
65 facilitate models to generate a reasoning path that decomposes complex reasoning into multiple  
66 easier steps. **Creswell et al. [29] solve multi-step reasoning tasks by interacting between selection**  
67 **and inference to generate immediate reasoning steps.** This significantly improves the performance  
68 on arithmetic [30], commonsense [21, 31], and symbolic reasoning [5] benchmarks. There has  
69 been also some work to explore the effect of semantics on reasoning. For example, Dasgupta et al.  
70 [32] evaluate three logical reasoning tasks, namely natural language inference (NLI), syllogisms  
71 and Wason selection based on whether the content of the problem is aligned with prior knowledge,  
72 concluding that LLMs show human-like content effects on reasoning. Schlegel et al. [33] also reach  
73 the similar conclusion. However, this is just a subordinate conclusion (LLMs are in-context semantic  
74 reasoners that rely on shallow semantic associations for prediction; consequently their predictions are  
75 naturally influenced by semantics). It cannot be used to extrapolate the conclusion that “LLMs are  
76 not symbolic reasoners. In contrast, our task focuses on pure symbolic reasoning. When we feed  
77 these to LLMs, they show significantly worse performance compared to when normal semantic words  
78 are fed. This phenomenon indicates that LLMs fail to invoke the basic formal reasoning abilities of  
79 humans but instead rely on shallow semantic associations for prediction.

80 **In-Context Learning** LLMs’ reasoning abilities are closely related to in-context learning (ICL).  
81 ICL refers to the ability of language models to adapt and learn from a few prompt examples during  
82 the inference process. In recent years, there has been a focus on exploring how to improve the  
83 performance of ICL [34–38]. Despite the empirical success, the underlying mechanisms of ICL still  
84 remain unclear. A few studies have shown how the chain-of-thought prompting effect the reasoning  
85 abilities of LLMs and which aspects of the prompting contribute to their performance [8, 39].  
86 Specifically, the order of demonstrations may lead to large performance fluctuations [40, 9]. **Wang**  
87 **et al. [41] explores factors such as the order of reasoning steps, whether the demonstrations consist of**  
88 **invalid reasoning steps or are related to the query. Turpin et al. [42] introduces bias features into CoT**  
89 **demonstrations, highlighting the potential for misleading explanations even within a chain of thought.**  
90 **Madaan and Yazdanbakhsh [43] aims to gain a deeper understanding of which aspects of CoT-based**

prompting affect LLMs’ reasoning performance by conducting counterfactual experiments. Recent works also explore the effect of ground-truth labels and question the necessity of ground-truth input-output mapping—using incorrect labels in the examples only marginally lowers the performance [44] and input-label correspondence plays a more important role in contextual demonstration [45]. Wei et al. [46] investigate the effects of semantic priors and input-label mapping on in-context learning using different-scale and instruction-tuned models. Wei et al. [47] propose symbol tuning, which involves fine-tuning a language model using input-label mapping unrelated to semantics prior in order to investigate whether LLMs can induce input-label patterns further improve reasoning abilities. To further understand why in-context learning works, some work provides theoretical analysis that in-context learning can be formalized as Bayesian inference [13] or some instances of ICL can be understood as implicit implementation of known learning algorithms [48]. However, the existing analyses of ICL are mainly based on natural language input with rich semantic information. We hypothesize that this might not be able to reflect their true level of symbolic reasoning abilities, including deduction, induction and abduction. Thus, this paper aims to decouple semantics in LLMs’ in-context reasoning abilities.

**Symbolic Reasoning** Symbolic reasoning has long been studied in the field of artificial intelligence [49–53] and cognitive science [54–58]. It involves manipulating symbols and applying logical rules to perform deduction [59], induction [55], and abduction [60]. Recently, there has been some work to explore LLMs’ ability of symbolic reasoning. Qian et al. [61] evaluate a set of simple symbolic manipulation (copy, reverse and addition) tasks to uncover the difficulty of the LLMs in handling OOD symbolic generalization, highlighting the limitation in arithmetic and symbolic induction. BIG-Bench [62] contain the symbol interpretation task aimed at reasoning and interpreting a simple scene consisting of some defined objects. Shin et al. [63] explore using LLM-based models for program synthesis by leveraging inferred execution traces to guide the generation of correct programs. Lamplé and Charton [64] focus on applying LLM-based models to mathematical reasoning, proposing a framework that combines deep learning with symbolic mathematics to perform algebraic reasoning, equation solving, and theorem proving. Pallagani et al. [65] use LLMs for automated planning—a branch of AI concerned with realizing action sequences (plans) to achieve certain goals, typically executed by intelligent agents, autonomous robots, and unmanned vehicles. Gao et al. [66] use LLMs to read natural language problems and generate programs as the intermediate reasoning steps proposes a new method using a symbolic module to address to improve LLMs’ symbolic reasoning abilities.

### 3 Decoupling Semantics from In-Context Reasoning

#### 3.1 Task Definitions

To begin, we first introduce the definitions of reasoning and memorization mechanisms and provide task descriptions for each. Examples of the tasks are shown in Figure 1.

**Reasoning** In the field of psychology, reasoning refers to the process of using logical operations to draw conclusions or make inferences based on available information [67–70]. As an abstract notion, it encompasses a variety of aspects. Traditionally, we can classify it into three categories:

- *Deductive reasoning* is a logical process in which a conclusion can be derived from given premises or principles, meaning predicting new facts based on existing facts and logical rules. For example, given the two facts (Lisa, sisterOf, Alice) and (Alice, motherOf, Bob) along with a logical rule  $\forall x, y, z : \text{sisterOf}(x, y) \wedge \text{motherOf}(y, z) \rightarrow \text{auntOf}(x, z)$ , the new fact (Lisa, auntOf, Bob) can be derived through deductive reasoning. The task is to predict the True/False of a predicted fact given facts and rules. The accuracy is the proportion of correct predictions.
- *Inductive reasoning* involves making generalizations based on specific observations or evidence. In other words, a logical rule can be induced from given facts. For instance, given a set of observations that person A is the parent of person B and person B is the child of person A, inductive reasoning is to conclude the logical rule  $\forall x, y : \text{parentOf}(x, y) \rightarrow \text{childOf}(y, x)$ . We perform the *rule generation* task. Given multiple facts with similar patterns and a rule template, the goal is to induce a rule that entails these facts. We test the generated rules against the ground truth rules. If the generated rule matches the ground truth rule exactly, we regard the rule as correct; otherwise, we regard the rule as incorrect. The precision is the proportion of correct rules. More details of the rule template and the ground-truth rules are provided in Appendix G.

Memorization (Depth-0 Reasoning)	Deductive Reasoning	Inductive Reasoning	Abductive Reasoning
<b>Fact1:</b> (Tom, parentOf, Amy) <b>Fact2:</b> (Alice, parentOf, Bob ) <b>Fact3:</b> (Bob, childOf, Alice) <b>Fact4:</b> (Amy, childOf, Tom)  <b>Q: True or False?</b> (Amy, parentOf, Tom) <b>A: False</b>	<b>Fact1:</b> (Tom, parentOf, Amy) <b>Fact2:</b> (Bob, childOf, Alice) <b>Fact3:</b> (Lisa, sisterOf, Alice) <b>Fact4:</b> (Alice, motherOf, Bob) <b>Rule:</b> $\forall x, y, z: \text{sisterOf}(x, y) \wedge \text{motherOf}(y, z) \rightarrow \text{auntOf}(x, z)$  <b>Q: True or False?</b> (Lisa, auntOf, Bob) <b>A: True</b>	<b>Fact1:</b> (Tom, parentOf, Amy) <b>Fact2:</b> (Alice, parentOf, Bob ) <b>Fact3:</b> (Bob, childOf, Alice) <b>Fact4:</b> (Amy, childOf, Tom)  <b>Q: <math>\forall x, y: ?(x, y) \rightarrow \text{childOf}(y, x)</math></b> <b>A: <math>\forall x, y: \text{parentOf}(x, y) \rightarrow \text{childOf}(y, x)</math></b>	<b>Fact1:</b> (Lisa, sisterOf, Alice) <b>Fact2:</b> (Alice, motherOf, Bob ) <b>Fact3:</b> (Bob, childOf, Tom) <b>Rule1:</b> $\forall x, y, z: \text{sisterOf}(x, y) \wedge \text{motherOf}(y, z) \rightarrow \text{auntOf}(x, z)$ <b>Rule2:</b> $\forall x, y: \text{parentOf}(x, y) \rightarrow \text{childOf}(y, x)$  <b>Q: Explain</b> (Lisa, auntOf, Bob) <b>A: Fact1, Fact2 <math>\xrightarrow{\text{Rule1}}</math> (Lisa, auntOf, Bob)</b>

Figure 1: Task Definitions. **Memorization**: retrieving the predicted fact from in-context knowledge. **Deductive**: predicting the correctness of the predicted fact given rules and facts. **Inductive**: generating a rule based on multiple facts with similar patterns. **Abductive**: explaining the predicted fact based on given rules and facts.

- *Abductive reasoning* is a logical process of seeking a hypothesis that best fits or explains a set of observations. For example, given a lot of facts including (Lisa, sisterOf, Alice) and (Alice, motherOf, Bob), along with a set of logical rules including  $\forall x, y, z : \text{sisterOf}(x, y) \wedge \text{motherOf}(y, z) \rightarrow \text{auntOf}(x, z)$ , if we observe Lisa is Bob’s aunt, one possible explanation is that Lisa is Alice’s sister and Alice is Bob’s mother. We use *explanation generation* to evaluate the abductive reasoning ability. Given a *theory* including facts and logical rules, the task is to select specific facts and a logical rule from the given theory to explain the *observation*. The *observation* is chosen from inferred facts. We use Proof Accuracy (PA) as an evaluation metric, *i.e.*, the fraction of examples where the generated proof matches exactly any of the gold proofs.

**Memorization** Memory plays a crucial role in reasoning, as it involves storing the in-context or parametric knowledge necessary for the reasoning process. In some sense, memory can be considered as Depth=0 reasoning, where the question is a known fact. The reasoning task involves retrieving the fact itself from in-context or knowledge within the parameters. However, the specific impact of semantics on memorization has not been extensively explored. Thus, in addition to decoupling semantics from reasoning, we also try to study the impact of semantics on memorization. Specifically, we use a new dataset to fine-tune a language model and test its *time*, *efficiency* and *forgetting ratio*: *time* is the fine-tuning time cost of adding/updating facts, *efficiency* is the filter MRR (the mean reciprocal of rank of the correct entity [71]) of the facts added/updated, and *forgetting ratio* is the filter MRR of the facts that should not be updated. When evaluating whether a fact has been successfully added or updated, we query LLM with a question about the tail entity and rank the probability of the true tail against all entities. The better LLM remembers a triplet, the higher the MRR gets.

### 3.2 Evaluation Datasets

Our goal is to decouple semantics from the in-context reasoning process and solely rely on the given (new) knowledge to perform reasoning tasks. To implement this, we use Symbolic Tree [72] and ProofWriter [16] datasets, which contain both relevant and irrelevant facts and LLMs need to infer the unknown facts after selecting relevant facts from memory.

The Symbolic Tree dataset is an artificially close-world and noise-free symbolic dataset generated with complex logical rules. The dataset consists of randomly sampled “*basic facts*”, which include gender information and “*parentOf*” relations among individuals. With the given logical rules, the dataset allows for reasoning about 28 different types of family relations, ranging from easy inferences (*e.g.*, fatherhood), to more elaborate ones (*e.g.*, a daughter of someone’s cousin). *Facts* consist of *basic facts* (in-context knowledge) and *inferred facts* (what to reason). Note that Symbolic Tree is a close-world dataset, which means that any facts not presented in the dataset are assumed to be false. Thus, we construct the false facts by replacing the head entity or tail entity with a random entity as negative examples in *inferred facts*. Considering the context window size limitation, we restrict each tree’s depth to 5 to generate the dataset. We experiment with 10 sampled Symbolic Trees; each has 30 kinds of relations (28 inferred relations, gender and parentOf relation), 26 entities, about 35 basic facts, 300 inferred facts and 300 false ones.

To decouple the semantics within the dataset, we replace the relation names (such as “parent”) with hand-crafted symbols (e.g., “r1”, “r2”, ...), so that LLMs cannot leverage the semantics of the predicates in reasoning but must resort to the given new knowledge (presented as in-context facts and rules). We also experiment with replacing entity names (such as “Alice”) with “e1”, “e2”, ..., but find that it has little impact on performance (more details are provided in Appendix Q). During the symbol generation process, we also try to randomly sample some letters as relation names (e.g., “lnqgv” instead of “r1”), but we observe that LLMs struggle to understand garbled characters, which may negatively affect performance (further discussion is provided in Appendix N).

ProofWriter [16] tasks provide artificial facts and rules expressed in natural language. For our experiments, we use a subset of the ProofWriter Open World Assumption (OWA) dataset with a depth of 1, 2, 3 and 5 (there is no depth 4 task), which contains many small rulebases of facts and rules, expressed in English and do not exist in LLMs’ knowledge base. Each rulebase has a set of questions (English statements) that can be proven true, false or “Unknown”. Note that if we want to prove something Unknown, it is necessary to enumerate all possible facts and check their true/false. Thus, we remove all the Unknowns and replace the subject and object with entity IDs. This dataset is simpler than Symbolic Tree. Considering most of the predicates in the sentences are unmeaningful verbs like “is” and “can”, we only replace the entities with entity IDs to decouple semantics. Take “Anne is kind.” as an example. We substitute subject (Anne) and object (kind) with “e1” and “e2”, respectively, resulting in “e1 is e2”. Figure 2 provides an illustrative example.

## 4 Experiment

**Models Selected for Evaluation** We primarily evaluate the performance of ChatGPT, GPT-4 and LLaMA. **ChatGPT** and **GPT-4** are advanced AI models developed by OpenAI and have demonstrated strong reasoning abilities across various tasks and benchmarks. **LLaMA** is an open-source large language model developed by Meta AI, with number of parameters ranging from 7B to 65B. Due to computational resource constraints, we could only fine-tune the LLaMA-7B version, which is used in our memorization test. **Note that in our study, we did attempt reasoning experiments using fine-tuned LLaMA-7b model. It performs significantly worse in reasoning tasks and even struggles to understand the instructions. Additionally, tasks requiring extensive facts and logical rules are hindered by the limited context window size. As a result, we did not conduct reasoning experiments with it.** Additionally, when comparing the reasoning abilities of LLMs, we also use some **logic-based** symbolic methods to conduct experiments as the baseline. To compare memorization, we use a popular graph database **Neo4j** [73] as the baseline. To ensure a relatively fair comparison, we configure Neo4j with a pre-stored knowledge base that has comparable disk space size to LLaMA. More introduction of Neo4j is represented in Appendix F.

**Evaluation Setup** For reasoning, we use Symbolic Tree and ProofWriter as evaluation data. We refer to the raw data, where semantics are retained, as *Semantics*. When semantics are decoupled using symbols, we refer to it as *Symbols*. For the Symbolic Tree dataset, we experiment with 10 sampled trees and report the average results, where facts and rules can be represented as logical language and natural language text as the input of LLMs. For example, the fact “motherOf(Alice, Bob)” can be represented as “Alice is Bob’s mother”; the rule “ $\forall x, y : \text{parentOf}(x, y) \rightarrow \text{childOf}(y, x)$ ” can be represented as “If x is parent of y, then y is child of x.”. **Through numerous trials, we find that for the Symbols setting, LLMs tend to perform better when using logic language representations. Conversely, for the Semantics setting, LLMs tend to perform better when using natural language text.** We select the representation that yields better performance in LLMs’ reasoning. Additional results are presented

Given a set of rules and facts, you have to reason whether a statement is true or false. Here are some facts and rules:	Given a set of rules and facts, you have to reason whether a statement is true or false. Here are some facts and rules:
The bear likes the dog. The cow is round. The cow likes the bear. The cow needs the bear. The dog needs the squirrel. The dog sees the cow. The squirrel needs the dog. If someone is round then they like the squirrel. If the bear is round and the bear likes the squirrel then the squirrel needs the bear. If the cow needs the dog then the cow is cold.	The e4 likes the e5. The e14 is e2. The e14 likes the e4. The e14 needs the e4. The e5 needs the e26. The e5 sees the e14. The e26 needs the e5. If someone is e2 then they like the e26. If the e4 is e2 and the e4 likes the e26 then the e26 needs the e4. If the e14 needs the e5 then the e14 is e1.
Does it imply that the statement "The cow likes the squirrel." is True?	Does it imply that the statement "The e14 likes the e26." is True?

Figure 2: Decoupling semantics from the ProofWriter task. In the original ProofWriter task, entities are represented by their names (left). However, in our decoupled setting, we replace the entity names with unique entity IDs (right).



in Appendix M. We consider zero-shot, zero-shot CoT, few-shot CoT and zero-plus-few-shot-CoT as baselines. To generate explanations for few-shot CoT experiments, for deductive reasoning, we use zero-shot CoT (*i.e.*, Let’s think step by step) to generate explanations given the random questions; for abductive reasoning, we randomly select five examples and manually design their demonstrations. We provide all prompts and CoT demonstrations in Appendix A. **We use the accuracy of various tasks as the reasoning result, including deducing the correctness of a conclusion, inducing correct rules, or finding explanations for hypotheses.**

For memorization, we randomly selected 1,258 triplets from four sampled Symbolic Trees to fine-tune the LLaMA. **After adding these triplets, we perform a second fine-tuning step where we update half of the added triplets. To obtain the updated facts, we select the triplets in the first two trees and replace the tail entities with other random entities. Since these updates are chosen from two independent Symbolic Trees, they did not overlap with the remaining half of the facts. We then used the other two trees to evaluate the impact of updating knowledge on other knowledge, namely the forgetting ratio.** We still use *Symbols* and *Semantics* to denote different experiment settings. Both settings ensure that the new information provided does not overlap with the old knowledge base of LLMs, avoiding any ambiguity problems and eliminating the influence of pre-existing knowledge on the memorization task. When testing, we follow the prompting of Taori et al. [74], using the head entity and relation as instructions and providing all candidate tails as input. The detailed prompts are contained in Appendix A.

**Implementation Details** For ChatGPT and GPT-4, we use the chat completion API provided by OpenAI. We use a temperature of zero to generate output. Additionally, we set the frequency penalty to zero and top p to 1, which are the default values for these APIs.

For LLaMA-7B, we utilized 4 A100 80G GPUs with batch size 64 for finetuning. The training process involved 100 epochs, employing a cosine learning rate schedule with an initial learning rate of  $2e-5$ . We run these experiments three times and recorded their mean MRR and standard deviations. Please refer to Appendix H for more details.

For logic-based symbolic baseline, in the deductive reasoning setting, it enumerates paths between head  $h$  and tail  $t$  and uses activated rules to infer the answer; For inductive reasoning, we adopt AMIE+ [75], which first enumerates possible rules and then learns a scalar weight for each rule to encode its quality. For abductive reasoning, we locate the logical rule that reason about the relation of the fact and find all paths connecting the head and tail that can activate the rule. These path facts, along with the logical rule, serve as the explanations.

#### 4.1 Semantics Matter in LLMs’ memorizing

We first test the memorization ability of LLMs when new knowledge are presented in semantics/symbols forms. The results are reported in Table 1.

**Results** From Table 1, the *efficiency* of adding and updating semantic knowledge is higher compared to symbolic knowledge. This suggests that semantic knowledge is easier for LLMs to memorize than symbolic knowledge, similar to human’s memory capabilities (memorizing symbols is generally more challenging than memorizing words with semantic meanings). However, we also find that the *forgetting ratio* of *Semantics* setting is higher than the symbolic setting. This could be attributed to the fact that semantic knowledge has stronger correlation with each other than symbolic knowledge in LLMs. **In other words, LLMs may utilize shallow semantic associations for memorization. When a portion of knowledge is updated, it can inadvertently affect other knowledge that should remain unaffected. In contrast, symbolic LLMs rely on rote memorization, which makes them less susceptible to such inadvertent changes and forgetting.**

We also compare fine-tuned language models with the deterministic graph DB Neo4J to explore the memorization abilities of neural-based and symbolic-based methods (**More illustrations are included in Appendix E**). From the results shown in Table 4.1, we can see that knowledge update using Neo4j achieves 100% accuracy when inserting new triplets or editing existing triplets, regardless of whether the knowledge is symbolic or semantic. As expected, since the added or updated knowledge does not overlap with the existing knowledge base, there is no further influence on the existing knowledge in the database. Additionally, compared to the computational cost of fine-tuning LLMs, updating knowledge in a graph database with optimized storage mechanisms is significantly faster. **This affirms**

Table 1: Memorization abilities: LLaMA-7B and Neo4j. MRR are in %.

Method	Category	Adding Efficiency (MRR)	Updating Efficiency (MRR)	Forgetting (MRR↓)	Time/1k triplets
LLaMA-7B	Semantics	50.375 ± 1.27	51.34 ± 0.55	7.02 ± 1.55	41.5 min
	Symbols	48.91 ± 4.3	40.74 ± 2.26	2.2 ± 0.99	41.5 min
Neo4j	Semantics	100	100	0	19s
	Symbols	100	100	0	19s

Table 2: The reasoning results of Symbolic Tree. Results are in %.

Category	Model	Baseline	deduction	induction	abduction
Symbols	ChatGPT	Zero-Shot	52.6	6.10	1.50
		Zero-Shot-CoT	55.7	7.86	4.90
		Few-Shot-CoT	54.8	-	18.2
		Zero-Plus-Few-Shot-CoT	55.7	-	16.8
	GPT-4	Zero-Shot	68.8	9.28	25.0
		Zero-Shot-CoT	71.1	8.93	31.2
		Few-Shot-CoT	67.6	-	44.2
	Random	Zero-Shot	66.1	36.4	2.94
		Zero-Shot-CoT	65.5	32.2	3.40
Semantics	ChatGPT	Few-Shot-CoT	67.1	-	21.8
		Zero-Plus-Few-Shot-CoT	67.2	-	20.9
		Zero-Shot	79.2	52.5	27.3
		Zero-Shot-CoT	86.2	53.9	33.4
	GPT-4	Few-Shot-CoT	91.1	-	69.2
		Random	50.1	3.57	-
		Logic-based	100	57.1	100

the huge advantage of using KGs/external DBs to update knowledge rather than finetuning, aligning with the recent trend of retrieval-based LLMs.

## 4.2 Semantics Play a Vital Role in LLMs’ Reasoning

In this section, we evaluate the impact of decoupling semantics from LLMs’ in-context reasoning. In Table 2, we present the results of deductive, inductive, and abductive reasoning tasks on the Symbolic Tree datasets.

**Results** From Table 2, we observe that in all reasoning scenarios, *Semantics* setting significantly outperforms *Symbols* setting. Notably, in the inductive experiments, *Semantics* achieves approximately 30% higher absolute accuracy compared to *Symbols* setting. This indicates that preserving rich semantics in the reasoning process leads to better performance for LLMs.

Despite the improved in-context reasoning performance of LLMs with rich semantics, when compared to logic-based symbolic methods, LLMs still exhibit inferior performance in all reasoning tasks. This suggests that while LLMs possess a broad knowledge base and strong language understanding,

Table 3: The deduction results of ProofWriter tasks (ChatGPT). Results are in %.

Category	Baseline	depth-1	depth-2	depth-3	depth-5
Symbols	Zero-Shot	69.1	62.3	59.4	52.8
	Zero-Shot-CoT	56.2	49.4	45.2	38.6
	Few-Shot-CoT	65.8	58.1	57.8	45.9
Semantics	Zero-Shot	69.0	63.5	60.3	51.4
	Zero-Shot-CoT	51.5	45.8	40.3	30.9
	Few-Shot-CoT	62.5	56.7	56.9	47.8

symbolic reasoning is not their primary strength compared to methods specifically designed for symbolic reasoning. This also suggests the potential of future neural-symbolic AI systems.

### 4.3 More Fine-grained Analysis about Semantics

The aforementioned experiments offer initial evidence highlighting the significance of semantics in the reasoning of LLMs. To further investigate this observation, we examine the influence of commonsense knowledge stored within LLMs on their semantic reasoning performance. Specifically, we explore three aspects: First, **we examine the influence of commonsense knowledge stored within LLMs on their semantic reasoning performance. To achieve this, we remain the semantics (as semantics can encompass commonsense knowledge) and remove all given logical rules (in deduction) and facts (in induction). Please refer to Appendix A for prompts. This forces the LLMs to rely solely on their prior commonsense knowledge to infer the answers and allows us to assess the extent to which LLMs can leverage their internal knowledge to reason effectively without explicit in-context knowledge.** Second, we retain the semantics of the datasets but introduce counter-commonsense logical rules. This requires LLMs to leverage in-context new knowledge and navigate the reasoning process by strictly adhering to the new information conflicting with the old knowledge. We implement it by shuffling relations as new relation labels to construct a new counter-commonsense dataset. For instance, we replace “motherOf” with “sisterOf”, “parentOf” with “brotherOf”, and “female” with “male”. Consequently, for a rule such as  $\forall x, y : \text{parentOf}(x, y) \wedge \text{female}(x) \rightarrow \text{motherOf}(x, y)$ , we obtain  $\forall x, y : \text{brotherOf}(x, y) \wedge \text{male}(x) \rightarrow \text{sisterOf}(x, y)$ . Thirdly, we use a subset of the ProofWriter OWA datasets for depths 0, 1, 2, 3 and 5, which contains synthetic facts and rules despite written in natural language but irrelevant to commonsense (see Figure 2). These investigations allow us to gain deeper insights into the effect of semantics on the reasoning capabilities of LLMs.

**When semantics are consistent with commonsense** As shown in Table 4, in the deductive reasoning experiment, *Removing rules/facts* achieves comparable results to *Semantics*; in the inductive reasoning experiment, *Removing rules/facts* outperforms *Symbols*, achieving 35.7% in GPT-4. These findings suggest that LLMs can perform deductive reasoning comparably by leveraging their stored commonsense knowledge without using the provided semantic knowledge, and providing symbolic instead of semantic knowledge in induction might even hurt the performance. Besides, **GPT-4 significantly outperforms ChatGPT across all evaluation settings. The results may be attributed to the fact that the stored commonsense knowledge within GPT-4 is likely more extensive than that in ChatGPT or GPT-4 potentially possesses stronger reasoning capabilities. Additionally, there is a possibility of potential data contamination in the training process of GPT-4. For example, it has been trained on datasets, such as ProofWriter, which influenced the results.**

**When semantics are not consistent with commonsense** To investigate the impact of semantics that are not consistent with commonsense, we introduce counter-commonsense (Counter-CS) scenarios, which is also shown in table 4. In comparison to *Semantics* and *Symbols*, we find that *Counter-Commonsense* performs worse than *Semantics*, even *Symbols*.

These findings suggest that when the in-context new knowledge conflicts with commonsense, LLMs struggle to accurately reason and predict.

**When semantics are irrelevant to commonsense** We use the ProofWriter tasks to test whether unmeaningful semantics are still useful. The results are shown in table 3. The *Symbols* setting performs comparably to the *Semantics* setting in the zero-shot setting, suggesting that when semantics are irrelevant to commonsense, they have little effect on the reasoning abilities of LLMs. In other words, when the task does not require deep semantic understanding or relies minimally on commonsense knowledge, the presence or absence of semantics does not significantly impact the performance of LLMs. However, in the CoT settings, we observe that *Semantics* is significantly worse than *Symbols*. This might be because step-by-step reasoning magnifies the disturbing effect brought by weird semantics such as “The squirrel needs the dog”. Additionally, we observe that the CoT settings even perform worse than the zero-shot setting, with a higher frequency of the answer

Table 4: Semantics, removing rules/facts and counter-commonsense reasoning experiments (ChatGPT and GPT-4). Results are in %.

	deductive (Few-Shot-CoT)		inductive (Zero-Shot-CoT)	
	ChatGPT	GPT-4	ChatGPT	GPT-4
Semantics	71.8	90.0	25.0	53.6
Symbols	53.7	67.6	7.14	21.4
Remove R/F	70.1	90.4	7.14	35.7
Counter-CS	48.9	73.4	7.14	17.8



359 “Cannot be determined.”. Similar phenomenons are also observed in table 2, indicating that CoT may  
360 not be always helpful for reasoning tasks with in-context new knowledge.

#### 361 4.4 More analysis and discussions

362 **(1) Induction and abduction underperform deduction:** We compare the reasoning abilities of  
363 LLMs across induction and abduction tasks and find that they perform notably worse compared to  
364 deduction, regardless of whether semantics or symbols are used. When semantics are decoupled, the  
365 drop in performance is even more significant. These findings highlight the considerable room for  
366 improvement in LLMs’ reasoning abilities and suggest that relying solely on semantics to achieve  
367 symbolic reasoning is challenging.

368 **(2) Shorter in-context knowledge enhances reasoning performance:** To examine the influence  
369 of context length on reasoning, we conducted an abductive reasoning experiment using a smaller  
370 Symbolic Tree, containing approximately 12 entities and 100 facts. The results, provided in Ap-  
371 pendix P, show that abductive reasoning with a shorter context leads to better performance compared  
372 to a longer context. Besides, we also conduct deduction and induction experiments where LLMs are  
373 directly provided with the relevant facts related to the predicted fact or the predicted rule. The results  
374 are presented in Appendix K. This finding suggests that LLMs struggle with processing excessively  
375 long in-context information, particularly in reasoning tasks. The length of the context influences  
376 reasoning performance, as shorter contexts make it easier to select relevant and useful information  
377 while minimizing the impact of unrelated content.

378 **(3) Effectiveness of commonsense expressed in natural language:** We explore the representation  
379 of knowledge in natural language and logic language forms in our experiments. The results, presented  
380 in Appendix M, indicate that for tasks involving semantics, natural language descriptions are more  
381 effective than logical language representations. Conversely, for symbolic and counter-commonsense  
382 tasks, logic language performs better. This observation suggests that natural language represen-  
383 tations better stimulate the semantic understanding capabilities of LLMs, while logical language  
384 representations are more conducive to symbolic reasoning.

385 **(4) Zero-shot capabilities are approaching zero-shot-CoT capabilities:** In *Symbols* setting,  
386 comparing zero-shot with zero-shot-CoT across deduction, induction, and abduction evaluations,  
387 we observe that zero-shot-CoT only marginally improves the performance compared to zero-shot  
388 learning. This finding suggests that the zero-shot capabilities of current LLMs are approaching their  
389 zero-shot-CoT learning abilities. One plausible explanation is that ChatGPT has already been trained  
390 on similar tasks with CoT and has memorized the instructions. Consequently, it implicitly follows  
391 these instructions when applied to the same queries, even without explicit CoT guidance [76].

392 **(5) Utilizing internal knowledge outperforms external in-context knowledge:** To explore the  
393 ability of LLMs to utilize internal and external knowledge, we conduct an additional experiment  
394 where we provide LLMs with only the relevant facts related to the predicted fact. We compare the  
395 performance of *Removing rules* (leveraging internal knowledge) with *Semantics* (providing external  
396 logical rules). Surprisingly, we find that *Removing rules* performed better than *Semantics*. This  
397 suggests that LLMs possess the necessary internal knowledge to support answering questions and  
398 reasoning tasks, and leveraging this internal knowledge is more effective for reasoning than relying  
399 on external logical rules. Detailed results and case studies can be found in Appendix K.1.

## 400 5 Conclusion and Discussion

401 Our paper presents the first comprehensive investigation of the role of semantics in LLMs’ in-context  
402 reasoning abilities by decoupling semantics from in-context prompts. Experimental results suggest  
403 that: When semantics are consistent with commonsense, LLMs perform fairly well; when semantics  
404 are decoupled or counter-commonsense, LLMs struggle to solve the reasoning tasks by leveraging  
405 in-context new knowledge. These findings reveal the importance of semantics in LLMs’ reasoning  
406 abilities and inspire further research on unveiling the magic existing within the black-box LLMs. In  
407 light of the findings identified in our analysis, we point out several potential future directions for the  
408 development of large foundation models:

409 **More complex symbolic reasoning benchmark:** To improve LLMs’ in-context symbolic reasoning  
410 abilities, developing new datasets with decoupled semantics and more complex reasoning tasks is  
411 necessary. These benchmarks should challenge LLMs with diverse and intricate symbolic knowledge.

412 **Combination with external non-parametric knowledge base:** As our experimental results show,  
413 the memorization abilities of LLMs are not comparable to existing graph-based methods. This  
414 motivates integrating LLMs with external non-parametric knowledge bases, such as graph databases,  
415 to enhance their knowledge insertion and updating. This hybrid approach can leverage the strengths  
416 of LLMs’ language understanding and the comprehensive, accurate and up-to-date knowledge stored  
417 in non-parametric sources.

418 **Improving the ability of processing in-context knowledge:** We observe that LLMs perform better  
419 under shorter context (discussion 4.5 (2)) and when only provided with the relevant facts related to  
420 the question (discussion 4.5 (6)). The results indicate that LLMs probably struggle with processing  
421 excessively long in-context information. Moreover, discussion 4.5 (5) also suggests LLMs are  
422 more reliable to leverage internal knowledge. As a result, effectively utilizing external (in-context)  
423 knowledge to perform situated tasks remains an important challenge for LLMs. This includes  
424 developing mechanisms to better encode and retrieve relevant information from the in-context  
425 knowledge.

## References

- [1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, 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.
- [2] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022.
- [3] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.
- [4] Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. Pretrained transformers as universal computation engines. *arXiv preprint arXiv:2103.05247*, 1, 2021.
- [5] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.
- [6] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022.
- [7] Sewon Min, Xinxu Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.
- [8] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*, 2021.
- [9] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR, 2021.
- [10] Timo Schick and Hinrich Schütze. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*, 2020.
- [11] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*, 2021.
- [12] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, 2022.
- [13] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.
- [14] Steven A Sloman. The empirical case for two systems of reasoning. *Psychological bulletin*, 119(1):3, 1996.
- [15] Thomas Gilovich, Dale Griffin, and Daniel Kahneman. *Heuristics and biases: The psychology of intuitive judgment*. Cambridge university press, 2002.
- [16] Oyvind Taffjord, Bhavana Dalvi Mishra, and Peter Clark. Proofwriter: Generating implications, proofs, and abductive statements over natural language. *arXiv preprint arXiv:2012.13048*, 2020.
- [17] Robin Cooper. The fracas consortium robin cooper, dick crouch, jan van eijck, chris fox, josef van genabith, jan jaspars, hans kamp, david milward, manfred pinkal, massimo poesio, steve pulman. 1996.
- [18] SamuelR. Bowman, Gabor Angeli, Christopher Potts, and ChristopherD. Manning. A large annotated corpus for learning natural language inference, Aug 2015.
- [19] Adina Williams, Nikita Nangia, and SamuelR. Bowman. A broad-coverage challenge corpus for sentence understanding through inference, Apr 2017.
- [20] Jean-Philippe Bernardy and Stergios Chatzikyriakidis. Improving the precision of natural textual entailment problem datasets., May 2020.
- [21] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge, Nov 2018.

- [22] Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. Explanations for commonsenseqa: New dataset and models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Jul 2021. doi: 10.18653/v1/2021.acl-long.238. URL <http://dx.doi.org/10.18653/v1/2021.acl-long.238>.
- [23] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs, Mar 2019.
- [24] Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and WilliamYang Wang. Hybridqa: A dataset of multi-hop question answering over tabular and textual data, Apr 2020.
- [25] Zonglin Yang, Li Dong, Xinya Du, Hao Cheng, Erik Cambria, Xiaodong Liu, Jianfeng Gao, and Furu Wei. Language models as inductive reasoners. *arXiv preprint arXiv:2212.10923*, 2022.
- [26] Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *arXiv preprint arXiv:2210.01240*, 2022.
- [27] Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, et al. Folio: Natural language reasoning with first-order logic. *arXiv preprint arXiv:2209.00840*, 2022.
- [28] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.
- [29] Antonia Creswell, Murray Shanahan, and Irina Higgins. Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712*, 2022.
- [30] Subhro Roy and Dan Roth. Solving general arithmetic word problems. *arXiv preprint arXiv:1608.01413*, 2016.
- [31] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021.
- [32] Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*, 2022.
- [33] Viktor Schlegel, Kamen V Pavlov, and Ian Pratt-Hartmann. Can transformers reason in fragments of natural language? *arXiv preprint arXiv:2211.05417*, 2022.
- [34] Rajarshi Das, Manzil Zaheer, Dung Thai, Ameya Godbole, Ethan Perez, Jay Yoon Lee, Lizhen Tan, Lazaros Polymenakos, and Andrew McCallum. Case-based reasoning for natural language queries over knowledge bases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9594–9611, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.755. URL <https://aclanthology.org/2021.emnlp-main.755>.
- [35] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*, 2021.
- [36] Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. Selective annotation makes language models better few-shot learners, 2022.
- [37] Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. Multitask prompted training enables zero-shot task generalization, 2022.
- [38] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners, 2022.

- [39] Yao Lu, Max Bartolo, Alastair P. Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity, Apr 2021.
- [40] Albert Webson and Ellie Pavlick. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.167. URL <https://aclanthology.org/2022.naacl-main.167>.
- [41] Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. Towards understanding chain-of-thought prompting: An empirical study of what matters. *arXiv preprint arXiv:2212.10001*, 2022.
- [42] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*, 2023.
- [43] Aman Madaan and Amir Yazdanbakhsh. Text and patterns: For effective chain of thought, it takes two to tango. *arXiv preprint arXiv:2209.07686*, 2022.
- [44] Sewon Min, Xinxu Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work?
- [45] Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, Kang Min Yoo, and Taek Kim. Ground-truth labels matter: A deeper look into input-label demonstrations. *arXiv preprint arXiv:2205.12685*, 2022.
- [46] Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023.
- [47] Jerry Wei, Le Hou, Andrew Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, et al. Symbol tuning improves in-context learning in language models. *arXiv preprint arXiv:2305.08298*, 2023.
- [48] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models, Nov 2022.
- [49] George Boole. *The mathematical analysis of logic*. Philosophical Library, 1847.
- [50] John McCarthy. Recursive functions of symbolic expressions and their computation by machine, part i. *Communications of the ACM*, 3(4):184–195, 1960.
- [51] Norbert Fuhr. Probabilistic datalog: Implementing logical information retrieval for advanced applications. *Journal of the American Society for Information Science*, 51(2):95–110, 2000.
- [52] Thomas Eiter, Giovambattista Ianni, and Thomas Krennwallner. *Answer set programming: A primer*. Springer, 2009.
- [53] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31, 2018.
- [54] John McCarthy and Patrick J Hayes. Some philosophical problems from the standpoint of artificial intelligence. In *Readings in artificial intelligence*, pages 431–450. Elsevier, 1981.
- [55] Nada Lavrac and Saso Dzeroski. Inductive logic programming. In *WLP*, pages 146–160. Springer, 1994.
- [56] Allen Newell and Herbert A Simon. Computer science as empirical inquiry: Symbols and search. In *ACM Turing award lectures*, page 1975. 2007.
- [57] David Landy, Colin Allen, and Carlos Zednik. A perceptual account of symbolic reasoning. *Frontiers in psychology*, 5:275, 2014.
- [58] Megan Fowler. A human-centric system for symbolic reasoning about code. 2021.
- [59] Philip N Johnson-Laird. Deductive reasoning. *Annual review of psychology*, 50(1):109–135, 1999.
- [60] Gyöngyi Kovács and Karen M Spens. Abductive reasoning in logistics research. *International journal of physical distribution & logistics management*, 2005.



- [61] Jing Qian, Hong Wang, Zekun Li, Shiyang Li, and Xifeng Yan. Limitations of language models in arithmetic and symbolic induction. *arXiv preprint arXiv:2208.05051*, 2022.
- [62] 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. *arXiv preprint arXiv:2206.04615*, 2022.
- [63] Eui Chul Shin, Illia Polosukhin, and Dawn Song. Improving neural program synthesis with inferred execution traces. *Advances in Neural Information Processing Systems*, 31, 2018.
- [64] Guillaume Lample and François Charton. Deep learning for symbolic mathematics. *arXiv preprint arXiv:1912.01412*, 2019.
- [65] Vishal Pallagani, Bharath Muppasani, Keerthiram Murugesan, Francesca Rossi, Lior Horesh, Biplav Srivastava, Francesco Fabiano, and Andrea Loreggia. Plansformer: Generating symbolic plans using transformers. *arXiv preprint arXiv:2212.08681*, 2022.
- [66] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR, 2023.
- [67] Peter Cathcart Wason and Philip Nicholas Johnson-Laird. *Psychology of reasoning: Structure and content*, volume 86. Harvard University Press, 1972.
- [68] P. C. Wason. Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, page 273–281, Jul 2007. doi: 10.1080/14640746808400161. URL <http://dx.doi.org/10.1080/14640746808400161>.
- [69] Ronald Fagin, Joseph Y Halpern, Yoram Moses, and Moshe Vardi. *Reasoning about knowledge*. MIT press, 2004.
- [70] Douglas N Walton. What is reasoning? what is an argument? *The journal of Philosophy*, 87(8):399–419, 1990.
- [71] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2013.
- [72] Patrick Hohenacker and Thomas Lukasiewicz. Ontology reasoning with deep neural networks. *Journal of Artificial Intelligence Research*, 68:503–540, 2020.
- [73] Massimo Carro. Nosql databases. *arXiv preprint arXiv:1401.2101*, 2014.
- [74] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. [https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca), 2023.
- [75] Luis Galárraga, Christina Teflioudi, Katja Hose, and Fabian M Suchanek. Fast rule mining in ontological knowledge bases with amie  
+  
+. *The VLDB Journal*, 24(6):707–730, 2015.
- [76] Jiu hai Chen, Lichang Chen, Heng Huang, and Tianyi Zhou. When do you need chain-of-thought prompting for chatgpt? *arXiv preprint arXiv:2304.03262*, 2023.

## Table of Contents

---

<b>A</b>	<b>Prompts</b>	<b>16</b>
A.1	Deductive reasoning . . . . .	16
A.1.1	Zero-Shot . . . . .	16
A.1.2	Zero-Shot-CoT . . . . .	16
A.1.3	Few-Shot-CoT . . . . .	17
A.1.4	Examples of Few-Shot-CoT Demonstrations . . . . .	17
A.1.5	Zero-Shot of removing rules setting . . . . .	17
A.2	Inductive reasoning . . . . .	18
A.2.1	Zero-Shot . . . . .	18
A.2.2	Zero-Shot CoT . . . . .	18
A.2.3	Zero-Shot of removing facts setting . . . . .	18
A.3	Abductive reasoning . . . . .	18
A.3.1	Zero-Shot . . . . .	18
A.3.2	Zero-Shot-CoT . . . . .	19
A.3.3	Few-Shot-CoT prompt . . . . .	19
A.3.4	Examples of Few-Shot-CoT Demonstrations . . . . .	19
A.3.5	Few-Shot CoT of removing rules setting . . . . .	20
A.4	Commonsense reasoning . . . . .	21
A.4.1	Zero-Shot prompt . . . . .	21
A.4.2	Few-Shot prompt . . . . .	21
A.4.3	Examples of Few-Shot Demonstrations . . . . .	21
A.5	LLaMA Fine-tuning Prompt . . . . .	21
<b>B</b>	<b>Deduction examples of Symbolic Tree datasets</b>	<b>21</b>
B.1	Semantics . . . . .	22
B.1.1	Logic language representations . . . . .	22
B.1.2	Natural language representations . . . . .	23
B.2	Symbolization . . . . .	25
B.2.1	Logic language representations . . . . .	25
B.2.2	Natural language representations: . . . . .	27
B.3	Semantics of removing rule setting . . . . .	28
<b>C</b>	<b>Examples of ProofWriter</b>	<b>29</b>
C.1	Semantics . . . . .	29
C.2	Symbols . . . . .	30
<b>D</b>	<b>Different Zero-Shot prompting</b>	<b>31</b>
<b>E</b>	<b>Comparison of memorization abilities of neural-based and symbolic-based methods</b>	<b>31</b>

645	<b>F Introduction of Neo4j</b>	<b>32</b>
646	<b>G Task definitions</b>	<b>32</b>
647	<b>H Implementation of memorization</b>	<b>33</b>
648	<b>I Consistency of knowledge base</b>	<b>33</b>
649	<b>J Different Few-Shot-CoT examples</b>	<b>33</b>
650	J.1 Few-Shot-CoT examples with Auto-Filter . . . . .	33
651	J.2 Few-Shot-CoT examples with Manually-Designed . . . . .	34
652	<b>K Reasoning results of after human selection</b>	<b>36</b>
653	K.1 Comparison of Removing rules and Semantics . . . . .	36
654	K.2 Reasoning results after human selection . . . . .	36
655	<b>L Reasoning with natural language</b>	<b>37</b>
656	<b>M Reasoning results of two representations</b>	<b>37</b>
657	<b>N Reasoning with garbled symbols</b>	<b>37</b>
658	<b>O More reasoning results of Symbolic Tree</b>	<b>39</b>
659	<b>P Abductive reasoning on smaller datasets</b>	<b>39</b>
660	<b>Q Replacing entity labels</b>	<b>40</b>
661	<b>R Multi-short rules</b>	<b>40</b>
662	<b>A Prompts</b>	
663	<b>A.1 Deductive reasoning</b>	
664	<b>A.1.1 Zero-Shot</b>	
665	<b>system:</b> You are a helpful assistant with deductive reasoning abilities.	
666	<b>user:</b> I will provide a set of logical rules L1 to L{number of rules} and facts F1 to F{number	
667	of basic facts}. Please select one single logical rule from L1 to L{number of rules} and	
668	a few facts from F1 to F{number of basic facts} to predict True/False of the unknown fact	
669	using deductive reasoning.	
670	Logical rules: {logical rules}	
671	Facts: {basic facts}	
672	Unknown fact: {statement}	
673	The answer (True or False) is:	
674	<b>A.1.2 Zero-Shot-CoT</b>	
675	<b>system:</b> You are a helpful assistant with deductive reasoning abilities. Please select one	
676	single logical rule and a few facts to predict True/False of the following statement.	
677	<b>user:</b> I will provide a set of logical rules L1 to L{number of rules} and facts F1 to F{number	
678	of basic facts}. Please select one single logical rule from L1 to L{number of rules} and	
679	a few facts from F1 to F{number of basic facts} to predict True/False of the following	
680	statement using deductive reasoning.	
681	Logical rules: {logical rules}	

```

682 Facts: {basic facts}
683 Statement: {statement}
684 Answer with True or False? Let's think step by step.

```

### 685 A.1.3 Few-Shot-CoT

```

686 system: You are a helpful assistant with deductive reasoning abilities. Please select one
687       single logical rule and a few facts to predict True/False of the following statement.
688 user: I will provide a set of logical rules L1 to L{number of rules} and facts F1 to F{number
689       of basic facts}.
690 Logical rules: {logical rules}
691 Facts: {basic facts}
692 Please select one single logical rule from L1 to L{number of rules} and a few facts from F1 to
693       F{number of basic facts} to predict True/False of the following statement using
694       deductive reasoning.
695 {Demonstrations}
696 Statement: {statement}
697 Answer:

```

### 698 A.1.4 Examples of Few-Shot-CoT Demonstrations

699 Statement:  $r_{14}(\text{Amelie}, \text{Jonathan})$   
700 Answer: We can use L11:  $\forall A, B, C, D, E : r_3(B, A) \wedge r_3(B, C) \wedge r_3(C, D) \wedge r_3(D, E) \wedge r_2(A) \rightarrow$   
701  $r_{14}(A, E)$  to make a deduction. From the facts, we know that  $r_3(\text{Amelie}, \text{Thomas})$  and  $r_3(\text{Thomas},$   
702  $\text{Jonathan})$ . Therefore, we can apply L11 with  $A = \text{Amelie}$ ,  $B = \text{Thomas}$ ,  $C = \text{Jonathan}$ ,  $D$  and  $E$  as  
703 variables. This gives us:  $r_3(\text{Thomas}, \text{Jonathan}) \wedge r_3(\text{Amelie}, \text{Thomas}) \wedge r_3(\text{Thomas}, D) \wedge r_3(D, E)$   
704  $\wedge r_2(\text{Amelie}) \rightarrow r_{14}(\text{Amelie}, E)$  Since we know that  $r_2(\text{Amelie})$  is true from F7, we can simplify  
705 the above statement to:  $r_3(\text{Thomas}, \text{Jonathan}) \wedge r_3(\text{Amelie}, \text{Thomas}) \wedge r_3(\text{Thomas}, D) \wedge r_3(D, E)$   
706  $\rightarrow r_{14}(\text{Amelie}, E)$  Now, we can substitute the facts  $r_3(\text{Amelie}, \text{Thomas})$  and  $r_3(\text{Thomas}, \text{Jonathan})$   
707 to get:  $r_3(\text{Thomas}, \text{Jonathan}) \wedge r_3(\text{Thomas}, D) \wedge r_3(D, E) \rightarrow r_{14}(\text{Amelie}, E)$  We can see that this  
708 statement is true if we choose  $D = \text{Alina}$  and  $E = \text{Jonathan}$ , since we know that  $r_3(\text{Thomas}, \text{Alina})$   
709 from F50 and  $r_3(\text{Alina}, \text{Jonathan})$  from F56. Therefore, the statement  $r_{14}(\text{Amelie}, \text{Jonathan})$  is true.

710 Statement:  $r_{31}(\text{Philipp}, \text{Nina})$   
711 Answer: Let's use L28:  $\forall A, B, C, D : r_3(B, A) \wedge r_3(C, B) \wedge r_3(C, D) \wedge r_1(A) \rightarrow r_{31}(A, D)$  to  
712 see if we can prove the statement  $r_{31}(\text{Philipp}, \text{Nina})$ . From the facts, we know that  $r_3(\text{Luisa}, \text{Nina})$   
713 (F52) and  $r_3(\text{Emilia}, \text{Philipp})$  (F62). We can use L28 with  $A = \text{Philipp}$ ,  $B = \text{Emilia}$ ,  $C = \text{Luisa}$ ,  
714 and  $D = \text{Nina}$  to get:  $r_3(\text{Emilia}, \text{Philipp}) \wedge r_3(\text{Luisa}, \text{Emilia}) \wedge r_3(\text{Luisa}, \text{Nina}) \wedge r_1(\text{Philipp}) \rightarrow$   
715  $r_{31}(\text{Philipp}, \text{Nina})$  Since all the conditions are met, we can conclude that  $r_{31}(\text{Philipp}, \text{Nina})$  is True.  
716 Therefore, the answer is True.

717 Statement:  $r_8(\text{Leonie}, \text{Nico})$   
718 Answer: We can use L5 to make a deduction about the statement  $r_8(\text{Leonie}, \text{Nico})$ . L5 states that if  
719 there exist three individuals  $A$ ,  $B$ , and  $C$  such that  $A$  is related to  $B$ ,  $B$  is related to  $C$ , and  $A$  has a  
720 certain property ( $r_2$ ), then  $A$  is related to  $C$  in a certain way ( $r_8$ ). Using this rule, we can see that we  
721 have the following facts:  $- r_3(\text{Leonie}, \text{Emily})$   
722  $- r_3(\text{Emily}, \text{Nico})$   
723 Therefore, we can conclude that  $r_8(\text{Leonie}, \text{Nico})$  is true.

724 Statement:  $r_7(\text{Patrick}, \text{Alina})$   
725 Answer: We can use L4:  $\forall A, B : r_3(A, B) \wedge r_1(A) \rightarrow r_7(A, B)$  to determine if  $r_7(\text{Patrick}, \text{Alina})$   
726 is true or false. From the facts, we know that  $r_3(\text{Patrick}, \text{Alina})$  is true (F55). We also know that  
727  $r_1(\text{Patrick})$  is true (F9). Therefore, we can apply L4 to conclude that  $r_7(\text{Patrick}, \text{Alina})$  is true.  
728 Answer: True.

729 Statement:  $r_{27}(\text{Jonathan}, \text{Leonie})$   
730 Answer: Let's use L4:  $\forall A, B : r_3(A, B) \wedge r_1(A) \rightarrow r_7(A, B)$  and F56:  $r_3(\text{Patrick}, \text{Jonathan})$  to infer  
731 that  $r_7(\text{Patrick}, \text{Jonathan})$  is true. Then, we can use L24:  $\forall A, B, C : r_3(B, A) \wedge r_3(C, B) \wedge r_1(A) \rightarrow$   
732  $r_{27}(A, C)$  and F44:  $r_3(\text{Leonie}, \text{Emily})$  to infer that  $r_{27}(\text{Jonathan}, \text{Leonie})$  is false, since there is no  
733 fact that supports  $r_3(\text{Jonathan}, \text{Emily})$ . Therefore, the answer is False.

### 734 A.1.5 Zero-Shot of removing rules setting

735 **system:** Please answer the question only with True or False.  
 736 **user:** I will provide a set of facts. Please predict True/False of the unknown fact based on  
 737 given facts.  
 738 Facts: {facts}  
 739 Unknown fact: {statement}  
 740 The answer (True or False) is:

## 741 A.2 Inductive reasoning

### 742 A.2.1 Zero-Shot

743 **system:** You are a helpful assistant with inductive reasoning abilities. Please generate one  
 744 single rule to match the template and logically entail the facts. Note that the symbol  
 745 '##' in the template should be filled with either 'r1' or 'r45', while the symbol '++'  
 746 should be filled with either 'r43' or 'r44'.  
 747 **user:** I will give you a set of facts F1 to F{number of basic facts}, facts G1 to G{number of  
 748 inferred fact} and a template for a logical rule. Please generate one single rule to  
 749 match the template and logically entail the facts G1 to G{number of inferred fact} based  
 750 on facts F1 to F{number of basic facts}.  
 751 Facts: {facts}  
 752 Template: {rule template}  
 753 Note that the symbol '##' in the template should be filled with either 'r1' or 'r45', while  
 754 the symbol '++' should be filled with either 'r43' or 'r44'.  
 755 After filling in the template, the generated rule is:

### 756 A.2.2 Zero-Shot CoT

757 **system:** You are a helpful assistant with inductive reasoning abilities. Please generate one  
 758 single rule to match the template and logically entail the facts. Note that the symbol  
 759 '##' in the template should be filled with either 'r1' or 'r45', while the symbol '++'  
 760 should be filled with either 'r43' or 'r44'.  
 761 **user:** I will give you a set of facts F1 to F{number of basic facts}, facts G1 to G{number of  
 762 inferred fact} and a template for a logical rule. Please generate one single rule to  
 763 match the template and logically entail the facts G1 to G{number of inferred fact} based  
 764 on facts F1 to F{number of basic facts}.  
 765 Facts: {facts}  
 766 Template: {rule template}  
 767 Note that the symbol '##' in the template should be filled with either 'r1' or 'r45', while  
 768 the symbol '++' should be filled with either 'r43' or 'r44'.  
 769 After filling in the template, the generated rule is: Let's think step by step.

### 770 A.2.3 Zero-Shot of removing facts setting

771 **system:** Please generate one single rule to match the template. Note that the symbol '##' in  
 772 the template should be filled with either 'parent' or 'child', while the symbol '++'  
 773 should be filled with either 'male' or 'female'.  
 774 **user:** I will give you a template for a logical rule. Please generate one single rule to match  
 775 the template and logically infer the relation sister  
 776 Template: If A is ## of B and B is ## of C and A is ++, then A is sister of C.  
 777 Note that the symbol '##' in the template should be filled with either 'parent' or 'child',  
 778 while the symbol '++' should be filled with either 'male' or 'female'.  
 779 After filling in the template, the generated rule is:

## 780 A.3 Abductive reasoning

### 781 A.3.1 Zero-Shot

782 **system:** You are a helpful assistant with abductive reasoning abilities. Please select one  
 783 single logical rule and a few facts to explain the following statement.  
 784 **user:** I will provide a set of logical rules L1 to L{number of rules} and facts F1 to F{number  
 785 of basic facts}. Please select one single logical rule from L1 to L{number of rules} and  
 786 a few facts from F1 to F{number of basic facts} to explain the following statement.  
 787 Rules: {logical rules}  
 788 Facts: {basic facts}  
 789 Statement: {statement}  
 790 Answer with the numbers of the selected rule and facts. The selected rule and facts are:



### 791 A.3.2 Zero-Shot-CoT

792 **system:** You are a helpful assistant with abductive reasoning abilities. Please select one  
 793 single logical rule and a few facts to explain the following statement.  
 794 **user:** I will provide a set of logical rules L1 to L{number of rules} and facts F1 to F{number  
 795 of basic facts}. Please select one single logical rule from L1 to L{number of rules} and  
 796 a few facts from F1 to F{number of basic facts} to explain the following statement.  
 797 Rules: {logical rules}  
 798 Facts: {basic facts}  
 799 Statement: {statement}  
 800 Answer with the numbers of the selected rule and facts. The selected rule and facts are: Let's  
 801 think step by step.

### 802 A.3.3 Few-Shot-CoT prompt

803 **system:** You are a helpful assistant with abductive reasoning abilities. Please select one  
 804 single logical rule and a few facts to explain the following statement.  
 805 **user:**  
 806 I will provide a set of logical rules L1 to L{number of rules} and facts F1 to F{number of  
 807 basic facts}. Please select one single logical rule from L1 to L{number of rules} and  
 808 a few facts from F1 to F{number of basic facts} to explain the following statement.  
 809 Rules: {logical rules}  
 810 Facts: {basic facts}  
 811 {Demonstrations}  
 812 Statement: {statement}  
 813 Answer:

### 814 A.3.4 Examples of Few-Shot-CoT Demonstrations

815 Statement:  $r4(\text{Laura}, \text{Fabian})$   
 816 Answer: To explain the statement  $r4(\text{Laura}, \text{Fabian})$ , we can use the rule L3 and fact F1 and F37.  
 817 L3:  $\forall A, B : r1(A, B) \wedge r44(A) \rightarrow r4(A, B)$   
 818 F1:  $r1(\text{Laura}, \text{Fabian})$   
 819 F37:  $r44(\text{Laura})$   
 820 According to L3, if there is a relationship  $r1$  between A and B, and A has the property  $r44$ , then the  
 821 relationship  $r4$  holds between A and B. In this case, we have a relationship  $r1$  between Laura and  
 822 Fabian (F1) and Laura has the property  $r44$  (F37). Therefore, the relationship  $r4$  holds between Laura  
 823 and Fabian, which is the given statement.  
 824 The selected rule and facts are: L3, F1, and F37.  
 825 Statement:  $r5(\text{Stefan}, \text{Emily})$   
 826 Answer: To explain the statement  $r5(\text{Stefan}, \text{Emily})$ , we can use the rule L4 and fact F17. L4:  
 827  $\forall A, B : r1(A, B) \wedge r43(A) \rightarrow r5(A, B)$   
 828 F17:  $r1(\text{Stefan}, \text{Emily})$   
 829 According to rule L4, if there is a relation  $r1(A, B)$  and A has the property  $r43$ , then the relation  
 830  $r5(A, B)$  holds. In our case, we have the fact F17, which states that there is a relation  $r1(\text{Stefan},$   
 831  $\text{Emily})$ . Additionally, Stefan has the property  $r43$  (F61). Therefore, we can conclude that the relation  
 832  $r5(\text{Stefan}, \text{Emily})$  holds.  
 833 The selected rule and facts are: L4, F17, and F61.  
 834 Statement:  $r23(\text{Felix}, \text{Elias})$   
 835 Answer: To explain the statement  $r23(\text{Felix}, \text{Elias})$ , we can use the rule L22 and fact F5. L22:  
 836  $\forall A, B : r1(B, A) \wedge r43(A) \rightarrow r23(A, B)$   
 837 F5:  $r1(\text{Elias}, \text{Felix})$   
 838 According to rule L22, if there exists a relationship  $r1(B, A)$  and A has the property  $r43$ , then the  
 839 relationship  $r23(A, B)$  holds. In our case, we have the fact F5, which states that there is a relationship  
 840  $r1(\text{Elias}, \text{Felix})$ . Additionally, we know that Felix has the property  $r43$  (F59). Therefore, applying  
 841 rule L22, we can conclude that the relationship  $r23(\text{Felix}, \text{Elias})$  holds.  
 842 The selected rule and facts are: L22, F5, and F59.  
 843 Statement:  $r3(\text{Felix}, \text{Fabian})$   
 844 Answer: To explain the statement  $r3(\text{Felix}, \text{Fabian})$ , we can use the rule L2 and facts F2, F1, and F59.  
 845 L2:  $\forall A, B, C : r1(B, A) \wedge r1(B, C) \wedge r43(A) \rightarrow r3(A, C)$   
 846 F1:  $r1(\text{Laura}, \text{Fabian})$

847 F2:  $r1(\text{Laura}, \text{Felix})$   
848 F59:  $r43(\text{Felix})$   
849 By substituting the values into L2, we get:  
850  $r1(\text{Laura}, \text{Felix}) \wedge r1(\text{Laura}, \text{Fabian}) \wedge r43(\text{Felix}) \rightarrow r3(\text{Felix}, \text{Fabian})$   
851 Since all the premises are true according to the given facts, the conclusion  $r3(\text{Felix}, \text{Fabian})$  is also  
852 true.  
853 The selected rule and facts are: L2, F1, F2, and F59.

854 Statement:  $r7(\text{Fabian}, \text{Helga})$   
855 Answer: To explain the statement  $r7(\text{Fabian}, \text{Helga})$ , we can use the rule L6 and facts F13, F21, and  
856 F39.  
857 L6:  $\forall A, B, C : r1(A, B) \wedge r1(B, C) \wedge r43(A) \rightarrow r7(A, C)$   
858 F13:  $r1(\text{Fabian}, \text{Thomas})$   
859 F21:  $r1(\text{Thomas}, \text{Helga})$   
860 F39:  $r43(\text{Fabian})$   
861 According to L6, if there exist A, B, and C such that  $r1(A, B)$ ,  $r1(B, C)$ , and  $r43(A)$ , then  $r7(A, C)$   
862 holds. In our case, A = Fabian, B = Thomas, and C = Helga. Since we have  $r1(\text{Fabian}, \text{Thomas})$   
863 from F13,  $r1(\text{Thomas}, \text{Helga})$  from F21, and  $r43(\text{Fabian})$  from F39, we can conclude that  $r7(\text{Fabian},$   
864  $\text{Helga})$  is true.  
865 So, the selected rule and facts are: L6, F13, F21, and F39.

### 866 A.3.5 Few-Shot CoT of removing rules setting

867 **system:** You are a helpful assistant with abductive reasoning abilities. Please select a few  
868 facts to explain the following statement.  
869 **user:** I will provide a set of facts F1 to F63. Please select a few facts from F1 to F63 to  
870 explain the following statement.  
871 Facts: {facts}  
872  
873 Statement: Laura is mother of Felix.  
874 Answer: To explain the statement "Laura is mother of Felix", we can use Facts:  
875 Fact F2 states: Laura is parent of Felix.  
876 Fact F37 states: Laura is female.  
877 Using F2 and F37, we can conclude that "Laura is mother of Felix" holds.  
878 Therefore, the selected rule and facts are F2, F37.  
879  
880 Statement: Samuel is brother of Alina.  
881 Answer: To infer the statement "Samuel is brother of Alina", we have:  
882 F27: Patrick is parent of Samuel.  
883 F28: Patrick is parent of Alina.  
884 F47: Samuel is male.  
885 Based on these facts, we can infer "Samuel is brother of Alina":  
886 Therefore, the selected rule and facts are F27, F28, F47.  
887  
888 Statement: Patrick is grandfather of David.  
889 Answer: To explain the statement "Patrick is grandfather of David", we have:  
890 F28: Patrick is parent of Alina.  
891 F7: Alina is parent of David.  
892 F45: Patrick is male.  
893 Based on these facts, we can infer "Patrick is grandfather of David":  
894 Therefore, the selected rule and facts are F28, F7, F45.  
895  
896 Statement: Amelie is daughter of Elena.  
897 Answer: To explain the statement "Amelie is daughter of Elena", we have:  
898 F20: Elena is parent of Amelie.  
899 F43: Amelie is female.  
900 Based on these facts, we can infer "Amelie is daughter of Elena".  
901 Therefore, the selected rule and facts are F20, F43.  
902  
903 Statement: Claudia is sister of Felix  
904 Answer: To prove the statement "Claudia is sister of Felix", we can use facts:  
905 F3: Laura is parent of Claudia.  
906 F2: Laura is parent of Felix.  
907 F40: Claudia is female.  
908 Based on these facts, we can infer "Claudia is sister of Felix".  
909 Therefore, the selected rule and facts are F3, F2, F40.  
910  
911 Statement: Laura is mother of Fabian.

912 Answer:

## 913 A.4 Commonsense reasoning

### 914 A.4.1 Zero-Shot prompt

915 **system:** Please choose answer from [A, B, C, D, E] and output one capital letter only.  
916 **user:** Question: {Question}  
917 {Multiple choices}  
918 Answer:

### 919 A.4.2 Few-Shot prompt

920 **system:** Please choose answer from [A, B, C, D, E] and output one capital letter only.  
921 **user:** {Demonstrations:}  
922 Question: {Question}  
923 {Multiple choices}  
924 Answer:

### 925 A.4.3 Examples of Few-Shot Demonstrations

#### 926 **Standard:**

927 Question: When is hotter?  
928 Choices: A.desert B.night C.morning D.noon  
929 Answer: A.

#### 930 **Predict-explain:**

931 Question: When is hotter?  
932 Choices: A.desert B.night C.morning D.noon  
933 Answer: A. The desert is hotter than the other options listed because it receives direct sunlight and  
934 has little to no vegetation to provide shade or cool the air. During the day, the sun beats down on the  
935 desert, causing temperatures to rise significantly. At night, the lack of vegetation and moisture in the  
936 air means that heat is quickly lost, causing temperatures to drop rapidly. In the morning and at noon,  
937 the sun is still high in the sky, but temperatures may be slightly cooler due to the presence of some  
938 shade and cooler air from the night before.

#### 939 **Expain-predict:**

940 Question: When is hotter?  
941 Choices: A.desert B.night C.morning D.noon  
942 The desert is hotter than the other options listed because it receives direct sunlight and has little to no  
943 vegetation to provide shade or cool the air. During the day, the sun beats down on the desert, causing  
944 temperatures to rise significantly. At night, the lack of vegetation and moisture in the air means that  
945 heat is quickly lost, causing temperatures to drop rapidly. In the morning and at noon, the sun is still  
946 high in the sky, but temperatures may be slightly cooler due to the presence of some shade and cooler  
947 air from the night before. Answer: A.

## 948 A.5 LLaMA Fine-tuning Prompt

949 Below is an instruction that describes a task, paired with an input that provides further  
950 context.  
951 Write a response that appropriately completes the request.  
952 Instruction: {Head} is the {Relation} of {Tail}  
953 Input: {input}  
954 Response:

## 955 B Deduction examples of Symbolic Tree datasets

956 In this section, we provide examples of deduction experiments conducted on the Symbolic Tree  
957 datasets. We present examples for both the *Semantics* and *Symbols* settings, represented in both  
958 natural language text and logic language

## 959 B.1 Semantics

### 960 B.1.1 Logic language representations

```
961 Logical rules:
962 L1:  $\forall A,B,C: \text{parentOf}(B, A) \wedge \text{parentOf}(B, C) \wedge \text{female}(A) \rightarrow \text{sisterOf}(A,$ 
963  $C)$ 
964 L2:  $\forall A,B,C: \text{parentOf}(B, A) \wedge \text{parentOf}(B, C) \wedge \text{male}(A) \rightarrow \text{brotherOf}(A,$ 
965  $C)$ 
966 L3:  $\forall A,B: \text{parentOf}(A, B) \wedge \text{female}(A) \rightarrow \text{motherOf}(A,B)$ 
967 L4:  $\forall A,B: \text{parentOf}(A, B) \wedge \text{male}(A) \rightarrow \text{fatherOf}(A,B)$ 
968 L5:  $\forall A,B,C: \text{parentOf}(A, B) \wedge \text{parentOf}(B, C) \wedge \text{female}(A) \rightarrow$ 
969  $\text{grandmotherOf}(A,C)$ 
970 L6:  $\forall A,B,C: \text{parentOf}(A, B) \wedge \text{parentOf}(B, C) \wedge \text{male}(A) \rightarrow$ 
971  $\text{grandfatherOf}(A,C)$ 
972 L7:  $\forall A,B,C,D: \text{parentOf}(A, B) \wedge \text{parentOf}(B, C) \wedge \text{parentOf}(C, D) \wedge \text{female}(A)$ 
973  $\rightarrow \text{greatGrandmotherOf}(A,D)$ 
974 L8:  $\forall A,B,C,D: \text{parentOf}(A, B) \wedge \text{parentOf}(B, C) \wedge \text{parentOf}(C, D) \wedge \text{male}(A) \rightarrow$ 
975  $\text{greatGrandfatherOf}(A,D)$ 
976 L9:  $\forall A,B,C,D: \text{parentOf}(B, A) \wedge \text{parentOf}(B, C) \wedge \text{parentOf}(C, D) \wedge \text{female}(A)$ 
977  $\rightarrow \text{auntOf}(A,D)$ 
978 L10:  $\forall A,B,C,D: \text{parentOf}(B, A) \wedge \text{parentOf}(B, C) \wedge \text{parentOf}(C, D) \wedge \text{male}(A)$ 
979  $\rightarrow \text{uncleOf}(A,D)$ 
980 L11:  $\forall A,B,C,D,E: \text{parentOf}(B, A) \wedge \text{parentOf}(B, C) \wedge \text{parentOf}(C, D) \wedge$ 
981  $\text{parentOf}(D, E) \wedge \text{female}(A) \rightarrow \text{greatAuntOf}(A,E)$ 
982 L12:  $\forall A,B,C,D,E: \text{parentOf}(B, A) \wedge \text{parentOf}(B, C) \wedge \text{parentOf}(C, D) \wedge$ 
983  $\text{parentOf}(D, E) \wedge \text{male}(A) \rightarrow \text{greatUncleOf}(A,E)$ 
984 L13:  $\forall A,B,C,D,E,F: \text{parentOf}(B, A) \wedge \text{parentOf}(C, B) \wedge \text{parentOf}(C, D) \wedge$ 
985  $\text{parentOf}(D, E) \wedge \text{parentOf}(E, F) \wedge \text{female}(A) \rightarrow \text{secondAuntOf}(A,F)$ 
986 L14:  $\forall A,B,C,D,E,F: \text{parentOf}(B, A) \wedge \text{parentOf}(C, B) \wedge \text{parentOf}(C, D) \wedge$ 
987  $\text{parentOf}(D, E) \wedge \text{parentOf}(E, F) \wedge \text{male}(A) \rightarrow \text{secondUncleOf}(A,F)$ 
988 L15:  $\forall A,B,C,D,E: \text{parentOf}(B, A) \wedge \text{parentOf}(C, B) \wedge \text{parentOf}(C, D) \wedge$ 
989  $\text{parentOf}(D, E) \wedge \text{female}(A) \rightarrow \text{girlCousinOf}(A,E)$ 
990 L16:  $\forall A,B,C,D,E: \text{parentOf}(B, A) \wedge \text{parentOf}(C, B) \wedge \text{parentOf}(C, D) \wedge$ 
991  $\text{parentOf}(D, E) \wedge \text{male}(A) \rightarrow \text{boyCousinOf}(A,E)$ 
992 L17:  $\forall A,B,C,D,E,F,G: \text{parentOf}(B, A) \wedge \text{parentOf}(C, B) \wedge \text{parentOf}(D, C) \wedge$ 
993  $\text{parentOf}(D, E) \wedge \text{parentOf}(E, F) \wedge \text{parentOf}(F, G) \wedge \text{female}(A) \rightarrow$ 
994  $\text{girlSecondCousinOf}(A,G)$ 
995 L18:  $\forall A,B,C,D,E,F,G: \text{parentOf}(B, A) \wedge \text{parentOf}(C, B) \wedge \text{parentOf}(D, C) \wedge$ 
996  $\text{parentOf}(D, E) \wedge \text{parentOf}(E, F) \wedge \text{parentOf}(F, G) \wedge \text{male}(A) \rightarrow$ 
997  $\text{boySecondCousinOf}(A,G)$ 
998 L19:  $\forall A,B,C,D,E,F: \text{parentOf}(B, A) \wedge \text{parentOf}(C, B) \wedge \text{parentOf}(D, C) \wedge$ 
999  $\text{parentOf}(D, E) \wedge \text{parentOf}(E, F) \wedge \text{female}(A) \rightarrow$ 
1000  $\text{girlFirstCousinOnceRemovedOf}(A,F)$ 
1001 L20:  $\forall A,B,C,D,E,F: \text{parentOf}(B, A) \wedge \text{parentOf}(C, B) \wedge \text{parentOf}(D, C) \wedge$ 
1002  $\text{parentOf}(D, E) \wedge \text{parentOf}(E, F) \wedge \text{male}(A) \rightarrow \text{boyFirstCousinOnceRemovedOf}$ 
1003  $(A,F)$ 
1004 L21:  $\forall A,B: \text{parentOf}(B, A) \wedge \text{female}(A) \rightarrow \text{daughterOf}(A,B)$ 
1005 L22:  $\forall A,B: \text{parentOf}(B, A) \wedge \text{male}(A) \rightarrow \text{sonOf}(A,B)$ 
1006 L23:  $\forall A,B,C: \text{parentOf}(B, A) \wedge \text{parentOf}(C, B) \wedge \text{female}(A) \rightarrow$ 
1007  $\text{granddaughterOf}(A,C)$ 
1008 L24:  $\forall A,B,C: \text{parentOf}(B, A) \wedge \text{parentOf}(C, B) \wedge \text{male}(A) \rightarrow \text{grandsonOf}($ 
1009  $A,C)$ 
1010 L25:  $\forall A,B,C,D: \text{parentOf}(B, A) \wedge \text{parentOf}(C, B) \wedge \text{parentOf}(D, C) \wedge \text{female}(A)$ 
1011  $\rightarrow \text{greatGranddaughterOf}(A,D)$ 
1012 L26:  $\forall A,B,C,D: \text{parentOf}(B, A) \wedge \text{parentOf}(C, B) \wedge \text{parentOf}(D, C) \wedge \text{male}(A)$ 
1013  $\rightarrow \text{greatGrandsonOf}(A,D)$ 
1014 L27:  $\forall A,B,C,D: \text{parentOf}(B, A) \wedge \text{parentOf}(C, B) \wedge \text{parentOf}(C, D) \wedge \text{female}(A)$ 
1015  $\rightarrow \text{nieceOf}(A,D)$ 
1016 L28:  $\forall A,B,C,D: \text{parentOf}(B, A) \wedge \text{parentOf}(C, B) \wedge \text{parentOf}(C, D) \wedge \text{male}(A)$ 
1017  $\rightarrow \text{nephewOf}(A,D)$ 
1018
1019 Facts:
1020 F1:  $\text{female}(\text{Laura})$ 
1021 F2:  $\text{male}(\text{Elias})$ 
1022 F3:  $\text{male}(\text{Fabian})$ 
1023 F4:  $\text{female}(\text{Claudia})$ 
1024 F5:  $\text{female}(\text{Elena})$ 
1025 F6:  $\text{male}(\text{Thomas})$ 
1026 F7:  $\text{female}(\text{Amelie})$ 
1027 F8:  $\text{female}(\text{Luisa})$ 
1028 F9:  $\text{male}(\text{Patrick})$ 
```

```

1029 F10: female(Emilia)
1030 F11: male(Samuel)
1031 F12: female(Alina)
1032 F13: male(Jonathan)
1033 F14: male(Philipp)
1034 F15: male(Nico)
1035 F16: male(David)
1036 F17: female(Emily)
1037 F18: male(Konstantin)
1038 F19: male(Florian)
1039 F20: female(Helga)
1040 F21: female(Nina)
1041 F22: female(Lea)
1042 F23: male(Felix)
1043 F24: female(Leonie)
1044 F25: male(Stefan)
1045 F26: male(Gabriel)
1046 F27: male(Tobias)
1047 F28: parentOf(Laura, Fabian)
1048 F29: parentOf(Laura, Felix)
1049 F30: parentOf(Laura, Claudia)
1050 F31: parentOf(Elias, Fabian)
1051 F32: parentOf(Elias, Felix)
1052 F33: parentOf(Elias, Claudia)
1053 F34: parentOf(Alina, David)
1054 F35: parentOf(Alina, Lea)
1055 F36: parentOf(Nico, David)
1056 F37: parentOf(Nico, Lea)
1057 F38: parentOf(Emily, Nico)
1058 F39: parentOf(Konstantin, Nico)
1059 F40: parentOf(Fabian, Thomas)
1060 F41: parentOf(Fabian, Amelie)
1061 F42: parentOf(Nina, Tobias)
1062 F43: parentOf(Leonie, Emily)
1063 F44: parentOf(Stefan, Emily)
1064 F45: parentOf(Gabriel, Tobias)
1065 F46: parentOf(Elena, Thomas)
1066 F47: parentOf(Elena, Amelie)
1067 F48: parentOf(Thomas, Helga)
1068 F49: parentOf(Thomas, Nina)
1069 F50: parentOf(Thomas, Patrick)
1070 F51: parentOf(Luisa, Helga)
1071 F52: parentOf(Luisa, Nina)
1072 F53: parentOf(Luisa, Patrick)
1073 F54: parentOf(Patrick, Samuel)
1074 F55: parentOf(Patrick, Alina)
1075 F56: parentOf(Patrick, Jonathan)
1076 F57: parentOf(Patrick, Philipp)
1077 F58: parentOf(Patrick, Florian)
1078 F59: parentOf(Emilia, Samuel)
1079 F60: parentOf(Emilia, Alina)
1080 F61: parentOf(Emilia, Jonathan)
1081 F62: parentOf(Emilia, Philipp)
1082 F63: parentOf(Emilia, Florian)
1083
1084 Unknown fact: boyCousinOf(Tobias, David)

```

## 1085 B.1.2 Natural language representations

```

1086 Logical rules:
1087 L1: If B is parent of A and B is parent of C and A is female, then A is sister of D.
1088 L2: If B is parent of A and B is parent of C and A is male, then A is brother of D.
1089 L3: If A is parent of B and A is female, then A is mother of C.
1090 L4: If A is parent of B and A is male, then A is father of C.
1091 L5: If A is parent of B and B is parent of C and A is female, then A is grandmother of D.
1092 L6: If A is parent of B and B is parent of C and A is male, then A is grandfather of D.
1093 L7: If A is parent of B and B is parent of C and C is parent of D and A is female, then A is
1094 greatGrandmother of E.
1095 L8: If A is parent of B and B is parent of C and C is parent of D and A is male, then A is
1096 greatGrandfather of E.

```



1097 L9: If B is parent of A and B is parent of C and C is parent of D and A is female, then A is  
 1098 aunt of E.  
 1099 L10: If B is parent of A and B is parent of C and C is parent of D and A is male, then A is  
 1100 uncle of E.  
 1101 L11: If B is parent of A and B is parent of C and C is parent of D and D is parent of E and A  
 1102 is female, then A is greatAunt of F.  
 1103 L12: If B is parent of A and B is parent of C and C is parent of D and D is parent of E and A  
 1104 is male, then A is greatUncle of F.  
 1105 L13: If B is parent of A and C is parent of B and C is parent of D and D is parent of E and E  
 1106 is parent of F and A is female, then A is secondAunt of G.  
 1107 L14: If B is parent of A and C is parent of B and C is parent of D and D is parent of E and E  
 1108 is parent of F and A is male, then A is secondUncle of G.  
 1109 L15: If B is parent of A and C is parent of B and C is parent of D and D is parent of E and A  
 1110 is female, then A is girlCousin of F.  
 1111 L16: If B is parent of A and C is parent of B and C is parent of D and D is parent of E and A  
 1112 is male, then A is boyCousin of F.  
 1113 L17: If B is parent of A and C is parent of B and D is parent of C and D is parent of E and E  
 1114 is parent of F and F is parent of G and A is female, then A is girlSecondCousin of H.  
 1115 L18: If B is parent of A and C is parent of B and D is parent of C and D is parent of E and E  
 1116 is parent of F and F is parent of G and A is male, then A is boySecondCousin of H.  
 1117 L19: If B is parent of A and C is parent of B and D is parent of C and D is parent of E and E  
 1118 is parent of F and A is female, then A is girlFirstCousinOnceRemoved of G.  
 1119 L20: If B is parent of A and C is parent of B and D is parent of C and D is parent of E and E  
 1120 is parent of F and A is male, then A is boyFirstCousinOnceRemoved of G.  
 1121 L21: If B is parent of A and A is female, then A is daughter of C.  
 1122 L22: If B is parent of A and A is male, then A is son of C.  
 1123 L23: If B is parent of A and C is parent of B and A is female, then A is granddaughter of D.  
 1124 L24: If B is parent of A and C is parent of B and A is male, then A is grandson of D.  
 1125 L25: If B is parent of A and C is parent of B and D is parent of C and A is female, then A is  
 1126 greatGranddaughter of E.  
 1127 L26: If B is parent of A and C is parent of B and D is parent of C and A is male, then A is  
 1128 greatGrandson of E.  
 1129 L27: If B is parent of A and C is parent of B and C is parent of D and A is female, then A is  
 1130 niece of E.  
 1131 L28: If B is parent of A and C is parent of B and C is parent of D and A is male, then A is  
 1132 nephew of E.  
 1133  
 1134 Facts:  
 1135 F1: Laura is female.  
 1136 F2: Elias is male.  
 1137 F3: Fabian is male.  
 1138 F4: Claudia is female.  
 1139 F5: Elena is female.  
 1140 F6: Thomas is male.  
 1141 F7: Amelie is female.  
 1142 F8: Luisa is female.  
 1143 F9: Patrick is male.  
 1144 F10: Emilia is female.  
 1145 F11: Samuel is male.  
 1146 F12: Alina is female.  
 1147 F13: Jonathan is male.  
 1148 F14: Philipp is male.  
 1149 F15: Nico is male.  
 1150 F16: David is male.  
 1151 F17: Emily is female.  
 1152 F18: Konstantin is male.  
 1153 F19: Florian is male.  
 1154 F20: Helga is female.  
 1155 F21: Nina is female.  
 1156 F22: Lea is female.  
 1157 F23: Felix is male.  
 1158 F24: Leonie is female.  
 1159 F25: Stefan is male.  
 1160 F26: Gabriel is male.  
 1161 F27: Tobias is male.  
 1162 F28: Laura is parent of Fabian.  
 1163 F29: Laura is parent of Felix.  
 1164 F30: Laura is parent of Claudia.  
 1165 F31: Elias is parent of Fabian.  
 1166 F32: Elias is parent of Felix.  
 1167 F33: Elias is parent of Claudia.  
 1168 F34: Alina is parent of David.

1169 F35: Alina is parent of Lea.  
 1170 F36: Nico is parent of David.  
 1171 F37: Nico is parent of Lea.  
 1172 F38: Emily is parent of Nico.  
 1173 F39: Konstantin is parent of Nico.  
 1174 F40: Fabian is parent of Thomas.  
 1175 F41: Fabian is parent of Amelie.  
 1176 F42: Nina is parent of Tobias.  
 1177 F43: Leonie is parent of Emily.  
 1178 F44: Stefan is parent of Emily.  
 1179 F45: Gabriel is parent of Tobias.  
 1180 F46: Elena is parent of Thomas.  
 1181 F47: Elena is parent of Amelie.  
 1182 F48: Thomas is parent of Helga.  
 1183 F49: Thomas is parent of Nina.  
 1184 F50: Thomas is parent of Patrick.  
 1185 F51: Luisa is parent of Helga.  
 1186 F52: Luisa is parent of Nina.  
 1187 F53: Luisa is parent of Patrick.  
 1188 F54: Patrick is parent of Samuel.  
 1189 F55: Patrick is parent of Alina.  
 1190 F56: Patrick is parent of Jonathan.  
 1191 F57: Patrick is parent of Philipp.  
 1192 F58: Patrick is parent of Florian.  
 1193 F59: Emilia is parent of Samuel.  
 1194 F60: Emilia is parent of Alina.  
 1195 F61: Emilia is parent of Jonathan.  
 1196 F62: Emilia is parent of Philipp.  
 1197 F63: Emilia is parent of Florian.  
 1198  
 1199 Unknown fact: Gabriel is uncle of Lea.

## 1200 B.2 Symbolization

### 1201 B.2.1 Logic language representations

1202 Logical rules:  
 1203 L1:  $\forall A,B,C: r3(B, A) \wedge r3(B, C) \wedge r2(A) \rightarrow r4(A, C)$   
 1204 L2:  $\forall A,B,C: r3(B, A) \wedge r3(B, C) \wedge r1(A) \rightarrow r5(A, C)$   
 1205 L3:  $\forall A,B: r3(A, B) \wedge r2(A) \rightarrow r6(A, B)$   
 1206 L4:  $\forall A,B: r3(A, B) \wedge r1(A) \rightarrow r7(A, B)$   
 1207 L5:  $\forall A,B,C: r3(A, B) \wedge r3(B, C) \wedge r2(A) \rightarrow r8(A, C)$   
 1208 L6:  $\forall A,B,C: r3(A, B) \wedge r3(B, C) \wedge r1(A) \rightarrow r9(A, C)$   
 1209 L7:  $\forall A,B,C,D: r3(A, B) \wedge r3(B, C) \wedge r3(C, D) \wedge r2(A) \rightarrow r10(A, D)$   
 1210 \$  
 1211 L8:  $\forall A,B,C,D: r3(A, B) \wedge r3(B, C) \wedge r3(C, D) \wedge r1(A) \rightarrow r11(A, D)$   
 1212 \$  
 1213 L9:  $\forall A,B,C,D: r3(B, A) \wedge r3(B, C) \wedge r3(C, D) \wedge r2(A) \rightarrow r12(A, D)$   
 1214 \$  
 1215 L10:  $\forall A,B,C,D: r3(B, A) \wedge r3(B, C) \wedge r3(C, D) \wedge r1(A) \rightarrow r13(A, D)$   
 1216 )\$  
 1217 L11:  $\forall A,B,C,D,E: r3(B, A) \wedge r3(B, C) \wedge r3(C, D) \wedge r3(D, E) \wedge r2(A) \rightarrow r14(A, E)$   
 1218 \$  
 1219 L12:  $\forall A,B,C,D,E: r3(B, A) \wedge r3(B, C) \wedge r3(C, D) \wedge r3(D, E) \wedge r1(A) \rightarrow r15(A, E)$   
 1220 \$  
 1221 L13:  $\forall A,B,C,D,E,F: r3(B, A) \wedge r3(C, B) \wedge r3(C, D) \wedge r3(D, E) \wedge r3(E, F) \wedge r2(A) \rightarrow r16(A, F)$   
 1222 \$  
 1223 L14:  $\forall A,B,C,D,E,F: r3(B, A) \wedge r3(C, B) \wedge r3(C, D) \wedge r3(D, E) \wedge r3(E, F) \wedge r1(A) \rightarrow r17(A, F)$   
 1224 \$  
 1225 L15:  $\forall A,B,C,D,E: r3(B, A) \wedge r3(C, B) \wedge r3(C, D) \wedge r3(D, E) \wedge r2(A) \rightarrow r18(A, E)$   
 1226 \$  
 1227 L16:  $\forall A,B,C,D,E: r3(B, A) \wedge r3(C, B) \wedge r3(C, D) \wedge r3(D, E) \wedge r1(A) \rightarrow r19(A, E)$   
 1228 \$  
 1229 L17:  $\forall A,B,C,D,E,F,G: r3(B, A) \wedge r3(C, B) \wedge r3(D, C) \wedge r3(D, E) \wedge r3(E, F) \wedge r3(F, G) \wedge r2(A) \rightarrow r20(A, G)$   
 1230 \$  
 1231 L18:  $\forall A,B,C,D,E,F,G: r3(B, A) \wedge r3(C, B) \wedge r3(D, C) \wedge r3(D, E) \wedge r3(E, F) \wedge r3(F, G) \wedge r1(A) \rightarrow r21(A, G)$   
 1232 \$  
 1233 L19:  $\forall A,B,C,D,E,F: r3(B, A) \wedge r3(C, B) \wedge r3(D, C) \wedge r3(D, E) \wedge r3(E, F) \wedge r2(A) \rightarrow r22(A, F)$   
 1234 \$  
 1235 L20:  $\forall A,B,C,D,E,F: r3(B, A) \wedge r3(C, B) \wedge r3(D, C) \wedge r3(D, E) \wedge r3(E, F) \wedge r1(A) \rightarrow r23(A, F)$   
 1236 \$

```

1237 L21:  $\forall A, B: r3(B, A) \wedge r2(A) \rightarrow r24(A, B)$ 
1238 L22:  $\forall A, B: r3(B, A) \wedge r1(A) \rightarrow r25(A, B)$ 
1239 L23:  $\forall A, B, C: r3(B, A) \wedge r3(C, B) \wedge r2(A) \rightarrow r26(A, C)$ 
1240 L24:  $\forall A, B, C: r3(B, A) \wedge r3(C, B) \wedge r1(A) \rightarrow r27(A, C)$ 
1241 L25:  $\forall A, B, C, D: r3(B, A) \wedge r3(C, B) \wedge r3(D, C) \wedge r2(A) \rightarrow r28(A, D)$ 
1242  $)$ 
1243 L26:  $\forall A, B, C, D: r3(B, A) \wedge r3(C, B) \wedge r3(D, C) \wedge r1(A) \rightarrow r29(A, D)$ 
1244  $)$ 
1245 L27:  $\forall A, B, C, D: r3(B, A) \wedge r3(C, B) \wedge r3(C, D) \wedge r2(A) \rightarrow r30(A, D)$ 
1246  $)$ 
1247 L28:  $\forall A, B, C, D: r3(B, A) \wedge r3(C, B) \wedge r3(C, D) \wedge r1(A) \rightarrow r31(A, D)$ 
1248  $)$ 
1249
1250 Facts:
1251 F1:  $r2(Laura)$ 
1252 F2:  $r1(Elias)$ 
1253 F3:  $r1(Fabian)$ 
1254 F4:  $r2(Claudia)$ 
1255 F5:  $r2(Elena)$ 
1256 F6:  $r1(Thomas)$ 
1257 F7:  $r2(Amelie)$ 
1258 F8:  $r2(Luisa)$ 
1259 F9:  $r1(Patrick)$ 
1260 F10:  $r2(Emilia)$ 
1261 F11:  $r1(Samuel)$ 
1262 F12:  $r2(Alina)$ 
1263 F13:  $r1(Jonathan)$ 
1264 F14:  $r1(Philipp)$ 
1265 F15:  $r1(Nico)$ 
1266 F16:  $r1(David)$ 
1267 F17:  $r2(Emily)$ 
1268 F18:  $r1(Konstantin)$ 
1269 F19:  $r1(Florian)$ 
1270 F20:  $r2(Helga)$ 
1271 F21:  $r2(Nina)$ 
1272 F22:  $r2(Lea)$ 
1273 F23:  $r1(Felix)$ 
1274 F24:  $r2(Leonie)$ 
1275 F25:  $r1(Stefan)$ 
1276 F26:  $r1(Gabriel)$ 
1277 F27:  $r1(Tobias)$ 
1278 F28:  $r3(Laura, Fabian)$ 
1279 F29:  $r3(Laura, Felix)$ 
1280 F30:  $r3(Laura, Claudia)$ 
1281 F31:  $r3(Elias, Fabian)$ 
1282 F32:  $r3(Elias, Felix)$ 
1283 F33:  $r3(Elias, Claudia)$ 
1284 F34:  $r3(Alina, David)$ 
1285 F35:  $r3(Alina, Lea)$ 
1286 F36:  $r3(Nico, David)$ 
1287 F37:  $r3(Nico, Lea)$ 
1288 F38:  $r3(Emily, Nico)$ 
1289 F39:  $r3(Konstantin, Nico)$ 
1290 F40:  $r3(Fabian, Thomas)$ 
1291 F41:  $r3(Fabian, Amelie)$ 
1292 F42:  $r3(Nina, Tobias)$ 
1293 F43:  $r3(Leonie, Emily)$ 
1294 F44:  $r3(Stefan, Emily)$ 
1295 F45:  $r3(Gabriel, Tobias)$ 
1296 F46:  $r3(Elena, Thomas)$ 
1297 F47:  $r3(Elena, Amelie)$ 
1298 F48:  $r3(Thomas, Helga)$ 
1299 F49:  $r3(Thomas, Nina)$ 
1300 F50:  $r3(Thomas, Patrick)$ 
1301 F51:  $r3(Luisa, Helga)$ 
1302 F52:  $r3(Luisa, Nina)$ 
1303 F53:  $r3(Luisa, Patrick)$ 
1304 F54:  $r3(Patrick, Samuel)$ 
1305 F55:  $r3(Patrick, Alina)$ 
1306 F56:  $r3(Patrick, Jonathan)$ 
1307 F57:  $r3(Patrick, Philipp)$ 
1308 F58:  $r3(Patrick, Florian)$ 

```

1309 F59: \$r3\$(Emilia, Samuel)  
 1310 F60: \$r3\$(Emilia, Alina)  
 1311 F61: \$r3\$(Emilia, Jonathan)  
 1312 F62: \$r3\$(Emilia, Philipp)  
 1313 F63: \$r3\$(Emilia, Florian)  
 1314  
 1315 Unknown fact: \$r9\$(Thomas, Claudia)

## 1316 B.2.2 Natural language representations:

1317 Logical rules:  
 1318 L1: If B is \$r3\$ of A and B is \$r3\$ of C and A is \$r2\$, then A is \$r4\$ of D.  
 1319 L2: If B is \$r3\$ of A and B is \$r3\$ of C and A is \$r1\$, then A is \$r5\$ of D.  
 1320 L3: If A is \$r3\$ of B and A is \$r2\$, then A is \$r6\$ of C.  
 1321 L4: If A is \$r3\$ of B and A is \$r1\$, then A is \$r7\$ of C.  
 1322 L5: If A is \$r3\$ of B and B is \$r3\$ of C and A is \$r2\$, then A is \$r8\$ of D.  
 1323 L6: If A is \$r3\$ of B and B is \$r3\$ of C and A is \$r1\$, then A is \$r9\$ of D.  
 1324 L7: If A is \$r3\$ of B and B is \$r3\$ of C and C is \$r3\$ of D and A is \$r2\$, then A is \$r10\$ of  
 1325 E.  
 1326 L8: If A is \$r3\$ of B and B is \$r3\$ of C and C is \$r3\$ of D and A is \$r1\$, then A is \$r11\$ of  
 1327 E.  
 1328 L9: If B is \$r3\$ of A and B is \$r3\$ of C and C is \$r3\$ of D and A is \$r2\$, then A is \$r12\$ of  
 1329 E.  
 1330 L10: If B is \$r3\$ of A and B is \$r3\$ of C and C is \$r3\$ of D and A is \$r1\$, then A is \$r13\$ of  
 1331 E.  
 1332 L11: If B is \$r3\$ of A and B is \$r3\$ of C and C is \$r3\$ of D and D is \$r3\$ of E and A is \$r2\$,  
 1333 then A is \$r14\$ of F.  
 1334 L12: If B is \$r3\$ of A and B is \$r3\$ of C and C is \$r3\$ of D and D is \$r3\$ of E and A is \$r1\$,  
 1335 then A is \$r15\$ of F.  
 1336 L13: If B is \$r3\$ of A and C is \$r3\$ of B and C is \$r3\$ of D and D is \$r3\$ of E and E is \$r3\$  
 1337 of F and A is \$r2\$, then A is \$r16\$ of G.  
 1338 L14: If B is \$r3\$ of A and C is \$r3\$ of B and C is \$r3\$ of D and D is \$r3\$ of E and E is \$r3\$  
 1339 of F and A is \$r1\$, then A is \$r17\$ of G.  
 1340 L15: If B is \$r3\$ of A and C is \$r3\$ of B and C is \$r3\$ of D and D is \$r3\$ of E and A is \$r2\$,  
 1341 then A is \$r18\$ of F.  
 1342 L16: If B is \$r3\$ of A and C is \$r3\$ of B and C is \$r3\$ of D and D is \$r3\$ of E and A is \$r1\$,  
 1343 then A is \$r19\$ of F.  
 1344 L17: If B is \$r3\$ of A and C is \$r3\$ of B and D is \$r3\$ of C and D is \$r3\$ of E and E is \$r3\$  
 1345 of F and F is \$r3\$ of G and A is \$r2\$, then A is \$r20\$ of H.  
 1346 L18: If B is \$r3\$ of A and C is \$r3\$ of B and D is \$r3\$ of C and D is \$r3\$ of E and E is \$r3\$  
 1347 of F and F is \$r3\$ of G and A is \$r1\$, then A is \$r21\$ of H.  
 1348 L19: If B is \$r3\$ of A and C is \$r3\$ of B and D is \$r3\$ of C and D is \$r3\$ of E and E is \$r3\$  
 1349 of F and A is \$r2\$, then A is \$r22\$ of G.  
 1350 L20: If B is \$r3\$ of A and C is \$r3\$ of B and D is \$r3\$ of C and D is \$r3\$ of E and E is \$r3\$  
 1351 of F and A is \$r1\$, then A is \$r23\$ of G.  
 1352 L21: If B is \$r3\$ of A and A is \$r2\$, then A is \$r24\$ of C.  
 1353 L22: If B is \$r3\$ of A and A is \$r1\$, then A is \$r25\$ of C.  
 1354 L23: If B is \$r3\$ of A and C is \$r3\$ of B and A is \$r2\$, then A is \$r26\$ of D.  
 1355 L24: If B is \$r3\$ of A and C is \$r3\$ of B and A is \$r1\$, then A is \$r27\$ of D.  
 1356 L25: If B is \$r3\$ of A and C is \$r3\$ of B and D is \$r3\$ of C and A is \$r2\$, then A is \$r28\$ of  
 1357 E.  
 1358 L26: If B is \$r3\$ of A and C is \$r3\$ of B and D is \$r3\$ of C and A is \$r1\$, then A is \$r29\$ of  
 1359 E.  
 1360 L27: If B is \$r3\$ of A and C is \$r3\$ of B and C is \$r3\$ of D and A is \$r2\$, then A is \$r30\$ of  
 1361 E.  
 1362 L28: If B is \$r3\$ of A and C is \$r3\$ of B and C is \$r3\$ of D and A is \$r1\$, then A is \$r31\$ of  
 1363 E.  
 1364  
 1365 Facts:  
 1366 F1: Laura is \$r2\$.  
 1367 F2: Elias is \$r1\$.  
 1368 F3: Fabian is \$r1\$.  
 1369 F4: Claudia is \$r2\$.  
 1370 F5: Elena is \$r2\$.  
 1371 F6: Thomas is \$r1\$.  
 1372 F7: Amelie is \$r2\$.  
 1373 F8: Luisa is \$r2\$.  
 1374 F9: Patrick is \$r1\$.  
 1375 F10: Emilia is \$r2\$.  
 1376 F11: Samuel is \$r1\$.  
 1377 F12: Alina is \$r2\$.

1378 F13: Jonathan is \$r1\$.  
 1379 F14: Philipp is \$r1\$.  
 1380 F15: Nico is \$r1\$.  
 1381 F16: David is \$r1\$.  
 1382 F17: Emily is \$r2\$.  
 1383 F18: Konstantin is \$r1\$.  
 1384 F19: Florian is \$r1\$.  
 1385 F20: Helga is \$r2\$.  
 1386 F21: Nina is \$r2\$.  
 1387 F22: Lea is \$r2\$.  
 1388 F23: Felix is \$r1\$.  
 1389 F24: Leonie is \$r2\$.  
 1390 F25: Stefan is \$r1\$.  
 1391 F26: Gabriel is \$r1\$.  
 1392 F27: Tobias is \$r1\$.  
 1393 F28: Laura is \$r3\$ of Fabian.  
 1394 F29: Laura is \$r3\$ of Felix.  
 1395 F30: Laura is \$r3\$ of Claudia.  
 1396 F31: Elias is \$r3\$ of Fabian.  
 1397 F32: Elias is \$r3\$ of Felix.  
 1398 F33: Elias is \$r3\$ of Claudia.  
 1399 F34: Alina is \$r3\$ of David.  
 1400 F35: Alina is \$r3\$ of Lea.  
 1401 F36: Nico is \$r3\$ of David.  
 1402 F37: Nico is \$r3\$ of Lea.  
 1403 F38: Emily is \$r3\$ of Nico.  
 1404 F39: Konstantin is \$r3\$ of Nico.  
 1405 F40: Fabian is \$r3\$ of Thomas.  
 1406 F41: Fabian is \$r3\$ of Amelie.  
 1407 F42: Nina is \$r3\$ of Tobias.  
 1408 F43: Leonie is \$r3\$ of Emily.  
 1409 F44: Stefan is \$r3\$ of Emily.  
 1410 F45: Gabriel is \$r3\$ of Tobias.  
 1411 F46: Elena is \$r3\$ of Thomas.  
 1412 F47: Elena is \$r3\$ of Amelie.  
 1413 F48: Thomas is \$r3\$ of Helga.  
 1414 F49: Thomas is \$r3\$ of Nina.  
 1415 F50: Thomas is \$r3\$ of Patrick.  
 1416 F51: Luisa is \$r3\$ of Helga.  
 1417 F52: Luisa is \$r3\$ of Nina.  
 1418 F53: Luisa is \$r3\$ of Patrick.  
 1419 F54: Patrick is \$r3\$ of Samuel.  
 1420 F55: Patrick is \$r3\$ of Alina.  
 1421 F56: Patrick is \$r3\$ of Jonathan.  
 1422 F57: Patrick is \$r3\$ of Philipp.  
 1423 F58: Patrick is \$r3\$ of Florian.  
 1424 F59: Emilia is \$r3\$ of Samuel.  
 1425 F60: Emilia is \$r3\$ of Alina.  
 1426 F61: Emilia is \$r3\$ of Jonathan.  
 1427 F62: Emilia is \$r3\$ of Philipp.  
 1428 F63: Emilia is \$r3\$ of Florian.  
 1429  
 1430 Unknown fact: Nico is \$r27\$ of Stefan.

### 1431 B.3 Semantics of removing rule setting

1432 I will provide a set of facts. Please predict True/False of the unknown fact based on given  
 1433 facts.  
 1434 Facts:  
 1435 F1: Laura is female.  
 1436 F2: Elias is male.  
 1437 F3: Fabian is male.  
 1438 F4: Claudia is female.  
 1439 F5: Elena is female.  
 1440 F6: Thomas is male.  
 1441 F7: Amelie is female.  
 1442 F8: Luisa is female.  
 1443 F9: Patrick is male.  
 1444 F10: Emilia is female.  
 1445 F11: Samuel is male.  
 1446 F12: Alina is female.



1447 F13: Jonathan is male.  
 1448 F14: Philipp is male.  
 1449 F15: Nico is male.  
 1450 F16: David is male.  
 1451 F17: Emily is female.  
 1452 F18: Konstantin is male.  
 1453 F19: Florian is male.  
 1454 F20: Helga is female.  
 1455 F21: Nina is female.  
 1456 F22: Lea is female.  
 1457 F23: Felix is male.  
 1458 F24: Leonie is female.  
 1459 F25: Stefan is male.  
 1460 F26: Gabriel is male.  
 1461 F27: Tobias is male.  
 1462 F28: Laura is parent of Fabian.  
 1463 F29: Laura is parent of Felix.  
 1464 F30: Laura is parent of Claudia.  
 1465 F31: Elias is parent of Fabian.  
 1466 F32: Elias is parent of Felix.  
 1467 F33: Elias is parent of Claudia.  
 1468 F34: Alina is parent of David.  
 1469 F35: Alina is parent of Lea.  
 1470 F36: Nico is parent of David.  
 1471 F37: Nico is parent of Lea.  
 1472 F38: Emily is parent of Nico.  
 1473 F39: Konstantin is parent of Nico.  
 1474 F40: Fabian is parent of Thomas.  
 1475 F41: Fabian is parent of Amelie.  
 1476 F42: Nina is parent of Tobias.  
 1477 F43: Leonie is parent of Emily.  
 1478 F44: Stefan is parent of Emily.  
 1479 F45: Gabriel is parent of Tobias.  
 1480 F46: Elena is parent of Thomas.  
 1481 F47: Elena is parent of Amelie.  
 1482 F48: Thomas is parent of Helga.  
 1483 F49: Thomas is parent of Nina.  
 1484 F50: Thomas is parent of Patrick.  
 1485 F51: Luisa is parent of Helga.  
 1486 F52: Luisa is parent of Nina.  
 1487 F53: Luisa is parent of Patrick.  
 1488 F54: Patrick is parent of Samuel.  
 1489 F55: Patrick is parent of Alina.  
 1490 F56: Patrick is parent of Jonathan.  
 1491 F57: Patrick is parent of Philipp.  
 1492 F58: Patrick is parent of Florian.  
 1493 F59: Emilia is parent of Samuel.  
 1494 F60: Emilia is parent of Alina.  
 1495 F61: Emilia is parent of Jonathan.  
 1496 F62: Emilia is parent of Philipp.  
 1497 F63: Emilia is parent of Florian.  
 1498  
 1499 Unknown fact: Jonathan is aunt of Thomas.  
 1500 The answer (True or False) is:

## 1501 C Examples of ProofWriter

1502 In this section, we provide examples of deduction experiments conducted on the ProofWriter Depth-1  
 1503 dataset. We present examples for both the *Semantics* and *Symbols* settings.

### 1504 C.1 Semantics

1505 The bear likes the dog.  
 1506 The cow is round.  
 1507 The cow likes the bear.  
 1508 The cow needs the bear.  
 1509 The dog needs the squirrel.  
 1510 The dog sees the cow.  
 1511 The squirrel needs the dog.

```

1512 If someone is round then they like the squirrel.
1513 If the bear is round and the bear likes the squirrel then the squirrel needs the bear.
1514 If the cow needs the dog then the cow is cold.
1515 Does it imply that the statement "The cow likes the squirrel." is True?

```

```

1516 The bear likes the dog.
1517 The cow is round.
1518 The cow likes the bear.
1519 The cow needs the bear.
1520 The dog needs the squirrel.
1521 The dog sees the cow.
1522 The squirrel needs the dog.
1523 If someone is round then they like the squirrel.
1524 If the bear is round and the bear likes the squirrel then the squirrel needs the bear.
1525 If the cow needs the dog then the cow is cold.
1526 Does it imply that the statement "The cow does not like the squirrel." is True?

```

```

1527 Bob is blue.
1528 Erin is quiet.
1529 Fiona is cold.
1530 Harry is cold.
1531 All quiet things are blue.
1532 If Harry is blue then Harry is not young.
1533 Blue things are young.
1534 Blue, round things are cold.
1535 If something is blue and not red then it is round.
1536 If something is young then it is white.
1537 If Erin is red and Erin is not round then Erin is young.
1538 If Erin is red and Erin is not cold then Erin is white.
1539 Does it imply that the statement "Erin is white" is True?
1540 Answer with only True or False. The answer is:

```

```

1541 The bear likes the dog.
1542 The cow is round.
1543 The cow likes the bear.
1544 The cow needs the bear.
1545 The dog needs the squirrel.
1546 The dog sees the cow.
1547 The squirrel needs the dog.
1548 If someone is round then they like the squirrel.
1549 If the bear is round and the bear likes the squirrel then the squirrel needs the bear.
1550 If the cow needs the dog then the cow is cold.
1551 Does it imply that the statement "The cow likes the squirrel." is True?

```

## 1552 C.2 Symbols

```

1553 The e4 likes the e5.
1554 The e14 is e2.
1555 The e14 likes the e4.
1556 The e14 needs the e4.
1557 The e5 needs the e26.
1558 The e5 sees the e14.
1559 The e26 needs the e5.
1560 If someone is e2 then they like the e26.
1561 If the e4 is e2 and the e4 likes the e26 then the e26 needs the e4.
1562 If the e14 needs the e5 then the e14 is e1.
1563 Does it imply that the statement "The e14 likes the e26." is True?

```

```

1564 The e27 is e7.
1565 The e27 is e15.
1566 The e30 does not chase the e27.
1567 The e30 eats the e27.
1568 The e30 is e1.
1569 The e30 is e15.
1570 The e30 visits the e27.
1571 If something visits the e27 then the e27 does not visit the e30.
1572 If something is e1 and e15 then it visits the e30.
1573 Does it imply that the statement "The e30 visits the e30." is True?

```

1574 The e27 is e7.  
 1575 The e27 is e15.  
 1576 The e30 does not chase the e27.  
 1577 The e30 eats the e27.  
 1578 The e30 is e1.  
 1579 The e30 is e15.  
 1580 The e30 visits the e27.  
 1581 If something visits the e27 then the e27 does not visit the e30.  
 1582 If something is e1 and e15 then it visits the e30.  
 1583 Does it imply that the statement "The e30 visits the e30." is True?

## 1584 D Different Zero-Shot prompting

1585 We try different Zero-Shot prompts:

1586 (1)

1587 I will provide a set of logical rules L1 to L{number of rules} and facts F1 to F{number of  
 1588 basic facts}. Please select one single logical rule from L1 to L{number of rules} and a  
 1589 few facts from F1 to F{number of basic facts} to predict True/False of the unknown fact  
 1590 using deductive reasoning.  
 1591 Logical rules: {rules}  
 1592 Facts: {basic facts}  
 1593 Unknown fact: {unknown fact}  
 1594 The answer (True or False) is:

1595 (2)

1596 I will provide a set of logical rules L1 to L{number of rules} and facts F1 to F{number of  
 1597 basic facts}. Please predict True/False of the unknown fact using deductive reasoning.  
 1598 Logical rules: {rules}  
 1599 Facts: {basic facts}  
 1600 Unknown fact: {unknown fact}  
 1601 The answer (True or False) is:

1602 (3)

1603 Given a set of rules and facts, you have to reason whether a statement is True or False.  
 1604 Here are some rules: {rules}  
 1605 Here are some facts: {basic facts}  
 1606 Does it imply that the statement "{unknown fact}" is True?  
 1607 The answer (YES or NO) is:

1608 The results of the three prompts in the Zero-Shot setting are presented in Table 5. Among the three  
 1609 prompts, we select the one that achieves the best performance as our Zero-Shot prompt.

Table 5: Different Zero-Shot Prompts of deductive reasoning. Results are in %.

	prompt1	prompt2	prompt3
KG <sub>1</sub>	54.5	51.5	53.8

## 1610 E Comparison of memorization abilities of neural-based and symbolic-based 1611 methods

1612 We compare fine-tuned language models with the deterministic graph DB Neo4J to explore the  
 1613 memorization abilities of neural-based and symbolic-based methods. Language models can implicitly  
 1614 store and retrieve facts as "knowledge bases" within their neural parameters. They are trained on a  
 1615 snapshot of data and may not have access to the latest or most accurate information. In order to update  
 1616 or add facts, specific model parameters need to be modified, or the model needs to be fine-tuned  
 1617 with new data. In contrast, symbolic knowledge graphs can directly add or update individual triplets,  
 1618 making it easier to incorporate new information. Our comparison affirms the huge advantage of  
 1619 using KGs/external DBs to update knowledge rather than finetuning, aligning with the recent trend of  
 1620 retrieval-based LLM.

## F Introduction of Neo4j

The Symbolic Tree is also a knowledge graph dataset. We conduct a comparison between the memorization abilities of a popular graph database, **Neo4j**, and LLMs, **LLaMA-7B**. Neo4j is a widely used graph database system that provides convenient operations such as querying, inserting, deleting, and revising knowledge graphs. For our comparison, we deployed Neo4j on a high-performance server equipped with 2 Intel(R) Xeon(R) Platinum 8380 CPUs, each with 40 cores and 80 threads. The server has 512GB of memory and 4x1.8T NVME SSD disks.

To ensure a fair comparison, we configured Neo4j with a pre-stored knowledge base that has a comparable disk space size to the LLaMA language model. Specifically, we used the Freebase dataset for Neo4j, which occupies approximately 30GB of disk space after preprocessing. For the language model, we used LLaMA-7B, which requires about 14GB of disk space. By comparing the performance of Neo4j and LLaMA-7B in terms of their memorization abilities, we can gain insights into the advantages and limitations of graph databases and language models for storing and retrieving knowledge.

## G Task definitions

We define a few tasks to evaluate LLMs' abilities of three kinds of reasoning and memorization.

- *deductive reasoning*: we use *hypothesis classification*, i.e., predict the *correctness* of the *hypothesis* given the *theory* where *theory* consists of basic facts and logical rules, *correctness* can be true or false, and *hypothesis* is a predicted fact, which is one of the inferred facts or negative samples. The accuracy is the proportion of correct predictions.
- *inductive reasoning*: we perform the *rule generation* task. Given multiple facts with similar patterns and a rule template, the goal is to induce a rule that entails these facts. Specifically, for each relation  $r$ , we use basic facts and those inferred facts that contain only relation  $r$  as provided facts. The induced rule is generated after filling in the rule template. We test the generated rules against the ground truth rules. If the generated rule matches the ground truth rule exactly, we predict the rule to be correct; otherwise, we predict the rule to be incorrect. The precision is the proportion of correct predictions. Note that considering logical rules maybe not all chain rules (e.g.,  $r_1(y, x) \wedge r_2(y, z) \rightarrow r_3(x, z)$ ), we add inverse relation for each relation in order to transform them into chain rules and simplify the rule template (e.g.,  $r_1^{-1}(x, y) \wedge r_2(y, z) \rightarrow r_3(x, z)$ ). Furthermore, we provide a rule template for each relation. Take *auntOf* as example, its rule template can be  $\forall x, y, z : \#(x, y) \wedge \#(y, z) \wedge ++(x) \rightarrow \text{auntOf}(x, z)$  or "If  $x$  is  $\#$  of  $y$  and  $y$  is  $\#$  of  $z$  and  $x$  is  $++$ , then  $x$  is aunt of  $z$ .", where  $\#$  can be *parent* or *inverse\_parent*,  $++$  can be *female* or *male*.  
Besides, a single rule can be equivalent to multiple rules. For example, the rule  $\forall x, y, z : \text{parentOf}(x, y) \wedge \text{parentOf}(y, z) \wedge \text{gender}(x, \text{female}) \rightarrow \text{GrandmotherOf}(x, z)$  can be represented as  $\forall x, y, z : \text{parentOf}(x, y) \wedge \text{parentOf}(y, z) \rightarrow \text{GrandparentOf}(x, z), \text{GrandparentOf}(x, z) \wedge \text{gender}(x, \text{female}) \rightarrow \text{GrandmotherOf}(x, z)$ . We conduct the experiments with both rule representations and find single-longer rules perform better than multiple-short rules. Results are presented in Appendix R. Based on these observations and considering the simplicity of induction evaluation, we rewrite all logical rules by including only the *parentOf* and *gender* relations in the rule body. This also ensures that each inferred relation is implied by a single logical rule, referred to as *grounding truth rule*.
- *abductive reasoning*: We use *explanation generation* to evaluate abductive reasoning abilities. Given a *theory* including basic facts and all logical rules, the task is to select specific facts and a logical rule to explain the *observation*. The *observation* is chosen from inferred facts. We use Proof Accuracy (PA) as an evaluation metric, i.e., the fraction of examples where the generated proof matches exactly any of the gold proofs.
- *memorization*: We use a subset of Symbolic Trees to fine-tune the language model. For the symbolic setting, we use  $r_1, r_2, r_3$  to replace the original relations in the semantic setting. Note that the new dataset does not overlap with the old knowledge base of LLMs, ensuring no disambiguation problem and the influence of pre-existing knowledge. When memorizing, we use *time*, *efficiency* and *forgetting* as metrics: *time* is the cost time of adding/updating facts, *efficiency* is the MRR (mean reciprocal rank [71]) of facts added/updated, and *forgetting* is the MRR of

1674 facts that should not be updated. When evaluating whether a fact has been successfully added or  
 1675 updated, we query LLM with a question about the tail entity and rank the probability of all tokens  
 1676 between all tail entities. The better LLM remembers a triplet, the higher the MRR gets. Note that,  
 1677 there may be more than one entity for each (head, relation) pair. We only consider the rank one of  
 1678 them.

## 1679 H Implementation of memorization

1680 We selected 1258 triplets from 4 Symbolic Trees to evaluate the effectiveness of adding knowledge.  
 1681 Following the prompting of Taori et al. [74], we use the head entity and relation as instructions and  
 1682 provide all candidate tails as input. The model’s training objective is to autoregress toward the true  
 1683 tail entities. The detailed prompting is contained in Appendix A.5. In the updating step, we fine-tune  
 1684 the model on all 620 triplets from the first two trees whose tail entities are randomly flipped to false  
 1685 ones. Besides the effectiveness of updating, we evaluated the forgetting ratio using the remaining  
 1686 638 triplets of the least two trees. These triplets have been remembered in the first step and haven’t  
 1687 been updated in the second. Noting that, within each tree, the relationships between entities are  
 1688 independent, and the entities are distinct. Therefore, we propose that LLM should retain its memory  
 1689 of the previously remembered triplets when updating based on the first two trees. We utilized 4 A100  
 1690 80G GPUs with batch size 64 for finetuning. The training process involved 100 epochs, employing a  
 1691 cosine learning rate schedule with an initial learning rate of  $2e-5$ . We run these experiments three  
 1692 times and recorded their mean and standard MRR.

## 1693 I Consistency of knowledge base

1694 In the context of updating a knowledge base, it is important to ensure the consistency of the knowledge  
 1695 base. When revising a fact, it is necessary to update other related facts accordingly to maintain a  
 1696 coherent and accurate knowledge base. For example, if we have the facts: Alice is Bob’s mother,  
 1697 Amy is Alice’s sister, Momo is Lily’s sister, and Amy is Bob’s aunt, and we update Bob’s mother as  
 1698 Lily, it follows that Bob’s aunt should be updated to Momo instead of Amy to maintain consistency.  
 1699 However, in our experiments, we did not consider the situation of updating related facts in order to  
 1700 simplify the evaluation process, although it may lead to inconsistencies within the knowledge base.  
 1701 In real-world scenarios, ensuring consistency is an essential aspect of maintaining the accuracy and  
 1702 reliability of a knowledge base when updating or revising facts.

## 1703 J Different Few-Shot-CoT examples

1704 We use three different kinds of demonstrations generation:

- 1705 (1) leverage LLMs with the “Let’s think step by step” prompt to generate reasoning chains for  
 1706 demonstrations.
- 1707 (2) manually designed each demonstration with a question and a reasoning chain. A reasoning chain  
 1708 is composed of a rationale (a series of intermediate reasoning steps) and an expected answer.
- 1709 (3) leverage LLMs with the “Let’s think step by step” prompt to generate reasoning chains and filter  
 1710 with those generated chains coming with mistakes for demonstrations.

1711 The results of these approaches are provided in Table 6. From the results, we observe that the  
 1712 *Auto-Randomly* approach outperforms both the *Auto-Filter* and *Manually-Designed* approaches. This  
 1713 suggests that randomly generating reasoning chains is more suitable for LLMs and can generate a  
 1714 diverse range of solutions, resulting in better reasoning performance.

### 1715 J.1 Few-Shot-CoT examples with Auto-Filter

1716 Statement:  $r8(\text{Elena}, \text{Nina})$   
 1717 Answer: We can use logical rule L5:  $\forall A, B, C : r3(A, B) \wedge r3(B, C) \wedge r2(A) \rightarrow r8(A, C)$  to  
 1718 deduce whether the statement  $r8(\text{Elena}, \text{Nina})$  is true or false. From the facts, we know that  $r3(\text{Elena},$   
 1719  $\text{Thomas})$  and  $r3(\text{Thomas}, \text{Nina})$ . Therefore, we can use L5 with  $A = \text{Elena}$ ,  $B = \text{Thomas}$ , and  
 1720  $C = \text{Nina}$  to see if  $r8(\text{Elena}, \text{Nina})$  is true or false. L5 states that if  $r3(A, B)$  and  $r3(B, C)$  and  $r2(A)$

1721 are true, then  $r8(A, C)$  is true. In this case, we have  $r3(\text{Elena}, \text{Thomas})$  and  $r3(\text{Thomas}, \text{Nina})$  as  
 1722 true facts, and we also know that  $r2(\text{Elena})$  is true from F5. Therefore, we can apply L5 to get:  
 1723  $r8(\text{Elena}, \text{Nina})$  is true.  
 1724 Therefore, the answer is True.

1725 Statement:  $r29(\text{Philipp}, \text{Elena})$

1726 Answer: We can use logical rule L26:  $\forall A, B, C, D : r3(B, A) \wedge r3(C, B) \wedge r3(D, C) \wedge r1(A) \rightarrow$   
 1727  $r29(A, D)$  to infer that if  $r3(B, A)$ ,  $r3(C, B)$ ,  $r3(D, C)$ , and  $r1(A)$  are true, then  $r29(A, D)$  is also  
 1728 true.

1729 From the given facts, we know that  $r3(\text{Elena}, \text{Thomas})$  (F46) and  $r3(\text{Thomas}, \text{Amelie})$  (F47).  
 1730 Therefore, we can use L26 with  $A = \text{Philipp}$ ,  $B = \text{Thomas}$ ,  $C = \text{Amelie}$ , and  $D = \text{Elena}$  to get:

1731  $r3(\text{Thomas}, \text{Philipp})$  (from F62 and  $r3(B, A)$ )

1732  $r3(\text{Amelie}, \text{Thomas})$  (from F47 and  $r3(C, B)$ )

1733  $r3(\text{Elena}, \text{Amelie})$  (from F47 and  $r3(D, C)$ )

1734  $r1(\text{Philipp})$  (from F14 and  $r1(A)$ )

1735 Therefore, we can conclude that  $r29(\text{Philipp}, \text{Elena})$  is True.

1736 Statement:  $r6(\text{Luisa}, \text{Helga})$

1737 Answer: We can use logical rule L3:  $\forall A, B : r3(A, B) \wedge r2(A) \rightarrow r6(A, B)$  to deduce the truth  
 1738 value of the statement  $r6(\text{Luisa}, \text{Helga})$ .

1739 From the facts, we know that  $r3(\text{Luisa}, \text{Helga})$  and  $r2(\text{Luisa})$  are true. Therefore, we can apply L3 to  
 1740 conclude that  $r6(\text{Luisa}, \text{Helga})$  is true.

1741 Thus, the answer is True.

1742 Statement:  $r14(\text{Amelie}, \text{Alina})$

1743 Answer: We can use logical rule L11:  $\forall A, B, C, D, E : r3(B, A) \wedge r3(B, C) \wedge r3(C, D) \wedge$   
 1744  $r3(D, E) \wedge r2(A) \rightarrow r14(A, E)$  to infer  $r14(\text{Amelie}, \text{Alina})$  if we can find facts that satisfy the  
 1745 antecedent of the rule.

1746 From the facts, we can see that  $r3(\text{Amelie}, \text{Thomas})$  and  $r3(\text{Thomas}, \text{Helga})$  and  $r3(\text{Helga}, \text{David})$  and  
 1747  $r3(\text{David}, \text{Alina})$  are true. Therefore, we can apply L11 with  $A = \text{Amelie}$ ,  $B = \text{Thomas}$ ,  $C = \text{Helga}$ ,  
 1748  $D = \text{David}$ , and  $E = \text{Alina}$  to get  $r14(\text{Amelie}, \text{Alina})$  is True.

1749 Therefore, the answer is True.

1750 Statement:  $r26(\text{Alina}, \text{Luisa})$

1751 Answer: We can use logical rule L23:  $\forall A, B, C : r3(B, A) \wedge r3(C, B) \wedge r2(A) \rightarrow r26(A, C)$  to  
 1752 determine the truth value of the statement  $r26(\text{Alina}, \text{Luisa})$ .

1753 From the facts, we know that  $r3(\text{Alina}, \text{David})$  and  $r3(\text{David}, \text{Luisa})$  (F34 and F52). Therefore, we  
 1754 can substitute  $A = \text{Alina}$ ,  $B = \text{David}$ , and  $C = \text{Luisa}$  into L23 to get:

1755  $r3(\text{David}, \text{Alina}) \wedge r3(\text{Luisa}, \text{David}) \wedge r2(\text{Alina}) \rightarrow r26(\text{Alina}, \text{Luisa})$

1756  $r3(\text{Alina}, \text{David})$  is true (F34) and  $r2(\text{Alina})$  is also true (F12). Therefore, we can simplify the above  
 1757 expression to:

1758  $r3(\text{Luisa}, \text{David}) \rightarrow r26(\text{Alina}, \text{Luisa})$

1759  $r3(\text{Luisa}, \text{David})$  is true (F52), so we can conclude that  $r26(\text{Alina}, \text{Luisa})$  is true.

1760 Therefore, the answer is True.

## 1761 J.2 Few-Shot-CoT examples with Manually-Designed

1762 Statement:  $r14(\text{Amelie}, \text{Jonathan})$

1763 Answer: We can use logical rule L11:  $\forall A, B, C, D, E : r3(B, A) \wedge r3(B, C) \wedge r3(C, D) \wedge$   
 1764  $r3(D, E) \wedge r2(A) \rightarrow r14(A, E)$ . We can substitute  $A = \text{Amelie}$  and  $E = \text{Jonathan}$   
 1765 to get:  $r3(B, \text{Amelie}) \wedge r3(B, C) \wedge r3(C, D) \wedge r3(D, \text{Jonathan}) \wedge r2(\text{Amelie}) \rightarrow$   
 1766  $r14(\text{Amelie}, \text{Jonathan})$ . Now, we need to find all facts that match the conditions  $r3(B, \text{Amelie})$ ,  
 1767  $r3(B, C)$ ,  $r3(C, D)$ ,  $r3(D, \text{Jonathan})$  and  $r2(\text{Amelie})$ . We have:

1768 F41:  $r3(\text{Fabian}, \text{Amelie})$

1769 F47:  $r3(\text{Elena}, \text{Amelie})$

1770 F56:  $r3(\text{Patrick}, \text{Jonathan})$

1771 F61:  $r3(\text{Emilia}, \text{Jonathan})$

1772 F7:  $r2(\text{Amelie})$

1773 We can see that F41 and F47 match the condition  $r3(B, \text{Amelie})$  with  $B$  being Fabian and Elena,  
 1774 respectively. F56 and F61 match the condition  $r3(D, \text{Jonathan})$  with  $D$  being Patrick and Emilia, re-  
 1775 spectively. F7 matches the condition  $r2(\text{Amelie})$ . Therefore, we have two possible ways of assigning

Table 6: The deductive reasoning results of one Symbolic Tree KG with ChatGPT. Results are in %.

	Auto-Randomly	Auto-Filter	Manually-Designed
KG <sub>1</sub>	55.6	53.6	40.7

1776 values to B, C, and D:  
1777 (1) B = Fabian, D = Patrick.  
1778 (2) B = Elena, D = Emilia.  
1779 (3) B = Fabian, D = Emilia.  
1780 (4) B = Elena, D = Patrick.  
1781 Using the values of B and D from the possible assignments above, we can find the value of C that  
1782 satisfies the conditions  $r3(B, C)$  and  $r3(C, D)$ . For assignment (1), we have:  
1783 F40:  $r3(\text{Fabian}, \text{Thomas})$   
1784 F50:  $r3(\text{Thomas}, \text{Patrick})$   
1785 Since all the conditions are satisfied, we can conclude that the statement  $r14(\text{Amelie}, \text{Jonathan})$  is  
1786 True if we assign B = Fabian and C = Thomas and D = Patrick. Thus, the output is True.

1787 Statement:  $r31(\text{Philipp}, \text{Nina})$   
1788 Answer: We can use logical rule L28:  $\forall A, B, C, D : r3(B, A) \wedge r3(C, B) \wedge r3(C, D) \wedge r1(A) \rightarrow$   
1789  $r31(A, D)$ . We can substitute A = Philipp and D = Nina to get:  $r3(B, \text{Philipp}) \wedge r3(C, B) \wedge$   
1790  $r3(C, \text{Nina}) \wedge r1(\text{Philipp}) \rightarrow r31(\text{Philipp}, \text{Nina})$ . Now, we need to find all facts that match the  
1791 conditions  $r3(B, \text{Philipp})$ ,  $r3(C, B)$ ,  $r3(C, \text{Nina})$ , and  $r1(\text{Philipp})$ . We have:  
1792 F57:  $r3(\text{Patrick}, \text{Philipp})$   
1793 F62:  $r3(\text{Emilia}, \text{Philipp})$   
1794 F49:  $r3(\text{Thomas}, \text{Nina})$   
1795 F52:  $r3(\text{Luisa}, \text{Nina})$   
1796 We can see that F57 and F62 match the condition  $r3(B, \text{Philipp})$  with B being Patrick and Emilia,  
1797 respectively. F49 and F52 match the condition  $r3(C, \text{Nina})$  with C being Thomas and Luisa, respec-  
1798 tively. Therefore, there is only one single way of assigning values to B and C.  
1799 B = Patrick and C = Thomas. We can substitute B and C to get:  $r3(\text{Patrick}, \text{Philipp}) \wedge$   
1800  $r3(\text{Thomas}, \text{Patrick}) \wedge r3(\text{Thomas}, \text{Nina}) \wedge r1(\text{Philipp}) \rightarrow r31(\text{Philipp}, \text{Nina})$ . Since all  
1801 the conditions are satisfied, we can conclude that the statement  $r31(\text{Philipp}, \text{Nina})$  is True if we  
1802 assign B = Patrick and C = Thomas. Thus, the output is True.

1803 Statement:  $r8(\text{Leonie}, \text{Nico})$   
1804 Answer: We can use logical rule L5:  $\forall A, B, C : r3(A, B) \wedge r3(B, C) \wedge r2(A) \rightarrow r8(A, C)$ . We  
1805 can substitute A = Leonie and C = Nico to get:  $r3(\text{Leonie}, B) \wedge r3(B, \text{Nico}) \wedge r2(\text{Leonie}) \rightarrow$   
1806  $r8(\text{Leonie}, \text{Nico})$ . Now, we need to find all facts that match the conditions  $r3(\text{Leonie}, B)$ ,  $r3(B,$   
1807  $\text{Nico})$ , and  $r2(\text{Leonie})$ . We have:  
1808 F43:  $r3(\text{Leonie}, \text{Emily})$   
1809 F38:  $r3(\text{Emily}, \text{Nico})$   
1810 F39:  $r3(\text{Konstantin}, \text{Nico})$   
1811 F24:  $r2(\text{Leonie})$   
1812 We can see that F43 matches the condition  $r3(\text{Leonie}, B)$  with B being Emily. and F38 matches the  
1813 condition  $r3(B, \text{Nico})$  with B being Emily. Therefore, we can assign B = Emily. Also, F24 matches  
1814 the condition  $r2(\text{Leonie})$ . Therefore, we can conclude that the statement  $r8(\text{Leonie}, \text{Nico})$  is True if  
1815 we assign B = Emily. Therefore, the output is True.

1816 Statement:  $r7(\text{Patrick}, \text{Alina})$   
1817 Answer: We can use logical rule L4:  $\forall A, B : r3(A, B) \wedge r1(A) \rightarrow r7(A, B)$ . We can substitute A  
1818 = Patrick and B = Alina to get:  $r3(\text{Patrick}, \text{Alina}) \wedge r1(\text{Patrick}) \rightarrow r7(\text{Patrick}, \text{Alina})$ . Now,  
1819 we need to find all facts that match the conditions  $r3(\text{Patrick}, \text{Alina})$  and  $r1(\text{Patrick})$ . We have:  
1820 F55:  $r3(\text{Patrick}, \text{Alina})$   
1821 F9:  $r1(\text{Patrick})$   
1822 We can see that F55 matches the condition  $r3(\text{Patrick}, \text{Alina})$ , and F9 matches the condition  
1823  $r1(\text{Patrick})$ . Therefore, we can conclude that the statement  $r7(\text{Patrick}, \text{Alina})$  is True based on  
1824 L4. Thus, the output is True.

## K Reasoning results of after human selection

In this section, we perform all experiments after human selection, providing the LLMs with only the relevant facts related to the predicted fact.

For example, we have facts:

F1: Lea is female.  
F2: Philipp is male.  
F3: Leonie is female.  
F4: Patrick is parent of Alina.  
F5: Patrick is parent of Philipp.  
F6: Nico is parent of David.  
F7: Nico is parent of Lea.  
F8: Laura is parent of Fabian.  
F9: Laura is parent of Felix.

If we predict the unknown fact "Philipp is the brother of Alina", we filter out the irrelevant facts and retain only the related facts:

F1: Philipp is male.  
F2: Patrick is parent of Alina.  
F3: Patrick is parent of Philipp.

### K.1 Comparison of Removing rules and Semantics

To explore the ability of LLMs to utilize internal and external knowledge, we conduct an additional experiment where we provided LLMs with only the relevant facts related to the predicted fact. We compared the performance of *Removing rules* (leveraging internal knowledge) with *Semantics* (providing external logical rules). Surprisingly, from Table 7, we find that *Removing rules* performed better than *Semantics*. This suggests that LLMs possess the necessary internal knowledge to support answering questions and reasoning tasks, and leveraging internal knowledge is more effective for reasoning than relying on external logical rules.

Table 7: Comparison of Removing and providing rules of deductive reasoning. Results are in %.

	Remove rules	Provide rules
Zero-Shot	80.5	77.1

LLMs’ output examples are provided below:

**Semantics:** According to the given logical rule, for A to be a granddaughter of D, B must be the parent of A and C must be the parent of B. However, there is no direct connection between Emilia and Lea in the given facts. Although Emilia is the parent of Alina, who is the parent of Lea, Emilia is not the parent of Lea’s parent. Therefore, we cannot conclude that Lea is Emilia’s granddaughter based on the given facts and logical rule.

**Removing rules:** Based on the given facts, Emilia is the parent of Alina, who is the parent of Lea. Therefore, Emilia is the grandmother of Lea, making Lea Emilia’s granddaughter.

From the example, we can observe that when relying on external logical rules, LLMs need to strictly adhere to the reasoning process, which can be more challenging for LLMs to predict unknown answers compared to utilizing the commonsense knowledge already contained within LLMs. This suggests that leveraging the internal knowledge of LLMs can be more effective for reasoning tasks.

### K.2 Reasoning results after human selection

We conduct deductive and inductive reasoning experiments to examine the performance of LLMs when only provided with the relevant facts related to the predicted fact. The results are presented in Table 8. They demonstrate that after selecting useful information, LLMs perform reasoning tasks more effectively. This finding suggests that LLMs face challenges when processing excessively long in-context information. Selecting relevant facts helps to reduce the memorization load on LLMs and enables them to focus on the most relevant information for reasoning, leading to improved performance.



Table 8: Reasoning results after removing irrelevant information. Results are %.

		Zero-Shot	Zero-Shot-CoT
<b>Deductive</b>	standard	52.6	56.1
	removing irr	55.7	63.0
<b>Inductive</b>	standard	7.14	7.14
	removing irr	67.9	67.9

Table 9: The reasoning results of Symbolic Tree (ChatGPT). Results are in %.

Category	Baseline	deduction	induction	abduction
<b>Logic language</b>	Zero-Shot	52.6	7.14	1.95
	Zero-Shot-CoT	56.1	7.14	3.57
	Few-Shot-CoT	53.7	-	13.3
<b>Natural language</b>	Zero-Shot	50.6	3.57	3.90
	Zero-Shot-CoT	50.2	7.14	1.95
	Few-Shot-CoT	51.9	-	8.13

## L Reasoning with natural language

In this section, we conducted experiments using the *Symbols* setting with deduction, induction, and abduction on a Symbolic Tree dataset expressed in natural language. The results are presented in Table 9. We observed that, in general, LLMs performed better when using logical language compared to natural language.

## M Reasoning results of two representations

For the Symbolic Tree dataset, facts and rules can be represented as logic language and natural language text as the input of LLMs. For example, the fact “motherOf(Alice, Bob)” can be represented as “Alice is Bob’s mother”; the fact “r1(Alice, Bob)” can be represented as “Alice is r1 of Bob”; the rule “ $\forall x, y : \text{parentOf}(x, y) \rightarrow \text{childOf}(y, x)$ ” can be represented as “If x is parent of y, then y is parent of x.”. Through numerous trials, we find that for the *Symbols* or *Counter-CS* setting, LLMs tend to perform better when using logic language representations. Conversely, for the *Semantics* setting, LLMs tend to perform better when using natural language text. The results are presented in Table 10. These observations suggest that natural language representations better stimulate the semantic understanding capabilities of LLMs, while logical language representations are more conducive to symbolic reasoning.

Table 10: Deductive reasoning results in different representations. Results are %.

		Zero-Shot	Zero-Shot-CoT
<b>Symbols</b>	logic	52.6	56.1
	natural language	49.0	51.1
<b>Semantics</b>	logic	61.4	61.9
	natural language	69.3	64.3
<b>Counter-CS</b>	logic	52.6	54.4
	natural language	48.7	48.3

## N Reasoning with garbled symbols

In this section, we randomly sample 4-8 letters to construct a garbled symbols word as each relation label. However, because LLMs process text by tokens, common sequences of characters found in the text, this random label is often tokenized as a single letter by LLMs. This leads to difficulty for LLMs to comprehend the word-level meaning of the label, harming the performance of reasoning.

Table 11: The deductive reasoning results of one Symbolic Tree KG with ChatGPT. Results are in %.

	Zero-Shot	Zero-Shot-CoT	Few-Shot-CoT	Zero-Plus-Few-Shot-CoT
garbled symbols	52.4	54.4	55.5	52.9
single-token symbols	52.5	53.2	53.9	52.3
ID symbols	54.5	57.0	57.2	52.3
short rules	52.1	51.1	52.9	54.6
long rules	54.5	57.0	57.2	52.3

We also attempted to use single-token garbled symbols as relation labels, but similar challenges were encountered, with the labels often being interpreted as typos by LLMs, further hindering their reasoning abilities. Our experiments and case studies also demonstrate this point.

As shown in Table 11, using ID symbols (e.g., “r1”) as relation labels performs better than using garbled symbols or single-token symbols.

Some examples with garbled symbols:

Statement: *ckxovsaq*(Alina, Philipp)

Answer: We can use logical rule L1 to make a deduction about Alina and Philipp. L1 states that if there is a person B who is a friend of both A and C, and A is not outgoing, then A trusts C. We know that Alina is not outgoing (F12: *ntoea*(Alina)) and that Philipp is a friend of Emilia (F62: *yufevh*(Emilia, Philipp)). However, we do not have any information about Alina and Emilia’s friendship. Therefore, we cannot use L1 to make a deduction about *ckxovsaq*(Alina, Philipp). The answer is False.

Statement: *nyheg*(Emily, Nico)

Answer: We can use logical rule L3:  $\forall A, B : yufevh(A, B) \wedge ntoea(A) \rightarrow nyheg(A, B)$  to determine if the statement is true or false. This rule states that if A is not outgoing and there is a friendship relation between A and B, then A is jealous of B. From the facts, we know that Emily is not outgoing (F17: *ntoea*(Emily)) and she is friends with Nico (F38: *yufevh*(Emily, Nico)). Therefore, we can apply L3 to conclude that *nyheg*(Emily, Nico) is true. Therefore, the answer is True.

Statement: *phnd*(Luisa, Jonathan)

Answer: We can use logical rule L5:  $\forall A, B, C : yufevh(A, B) \wedge yufevh(B, C) \wedge ntoea(A) \rightarrow phnd(A, C)$  to determine if the statement is true or false. This rule states that if A is not outgoing and there is a path from A to C through B, then A is phoned by C. From the given facts, we know that Luisa is not outgoing (F8: *ntoea*(Luisa)). We also know that there is a path from Luisa to Jonathan through Patrick (F53: *yufevh*(Luisa, Patrick) and F56: *yufevh*(Patrick, Jonathan)). Therefore, we can apply L5 with A = Luisa, B = Patrick, and C = Jonathan to get: *phnd*(Luisa, Jonathan) is true. Therefore, the answer is True.

Some examples with single-token symbols:

Statement: *iance*(Patrick, Lea)

Answer: We can use logical rule L4:  $\forall A, B : inely(A, B) \wedge atis(A) \rightarrow lesai(A, B)$ , which states that if A is an artist and B is in A’s inner circle, then B is a muse of A. We can also use fact F53: *inely*(Luisa, Patrick), which means that Luisa is in Patrick’s inner circle. Therefore, we can conclude that *lesai*(Patrick, Luisa) is true.

Next, we can use logical rule L6:  $\forall A, B, C : inely(A, B) \wedge inely(B, C) \wedge atis(A) \rightarrow iance(A, C)$ , which states that if A is an artist, B is in A’s inner circle, and C is in B’s inner circle, then A and C have an artistic alliance. We can use fact F35: *inely*(Alina, Lea), which means that Alina is in Lea’s inner circle. Therefore, we can conclude that *iance*(Alina, Lea) is true.

Finally, we can use logical rule L1:  $\forall A, B, C : inely(B, A) \wedge inely(B, C) \wedge uitka(A) \rightarrow reib(A, C)$ , which states that if A is a writer, B is in A’s inner circle, and B is also in C’s inner circle, then A and C have a literary rivalry. We can use fact F54: *inely*(Patrick, Samuel), which means that Samuel is in Patrick’s inner circle. We can also use fact F35: *inely*(Alina, Lea), which means that Alina is in Lea’s inner circle. Therefore, we can conclude that *reib*(Patrick, Lea) is true. Since Patrick has a literary rivalry with Lea, but not an artistic alliance, we can conclude that the statement *iance*(Patrick, Lea) is False.

Table 12: The deductive reasoning results of Symbolic Tree datasets. Results are in %.

Category	Model	Baseline	S <sub>1</sub>	S <sub>2</sub>	S <sub>3</sub>	S <sub>4</sub>	S <sub>5</sub>	S <sub>6</sub>	S <sub>7</sub>	S <sub>8</sub>	S <sub>9</sub>	S <sub>10</sub>	Avg.
Symbols	Random	-	52.4	50.8	51.3	50.2	49.3	49.1	48.1	52.3	48.4	49.0	50.1
		Zero-Shot	52.6	50.6	50.5	49.5	55.2	53.1	50.0	53.4	56.6	54.0	52.6
	ChatGPT	Zero-Shot-CoT	56.1	57.0	55.4	57.0	54.5	56.1	55.5	56.9	50.0	58.0	55.7
		Few-Shot-CoT	53.7	56.9	55.2	54.4	55.1	52.0	54.0	55.8	56.8	54.5	54.8
		Zero-Plus-Few-Shot-CoT	53.7	53.6	55.4	51.4	54.0	50.9	54.0	54.2	58.4	54.5	54.0
Semantics	ChatGPT	Zero-Shot	70.0	64.8	70.4	65.8	61.4	63.8	65.8	67.4	63.0	68.9	66.1
		Zero-Shot-CoT	66.7	64.8	64.6	64.1	64.4	67.2	66.5	66.7	64.6	65.4	65.5
		Few-Shot-CoT	71.8	70.4	63.9	69.2	66.7	59.3	68.7	68.3	67.9	64.4	67.1
		Zero-Plus-Few-Shot-CoT	71.3	67.8	66.6	69.5	65.7	60.9	68.4	68.3	66.5	66.8	67.2
		Logic-based	-	100	100	100	100	100	100	100	100	100	100

Table 13: The inductive reasoning results of Symbolic Tree datasets. Results are in %.

Category	Model	Baseline	S <sub>1</sub>	S <sub>2</sub>	S <sub>3</sub>	S <sub>4</sub>	S <sub>5</sub>	S <sub>6</sub>	S <sub>7</sub>	S <sub>8</sub>	S <sub>9</sub>	S <sub>10</sub>	Avg.
Symbols	ChatGPT	Zero-Shot	7.14	9.09	3.57	7.14	4.54	14.3	4.54	7.14	3.57	0.0	6.10
		Zero-Shot-CoT	7.14	7.14	3.57	14.3	14.3	7.14	3.57	0.0	14.3	7.14	7.86
	GPT-4	Zero-Shot	14.3	10.7	10.7	7.14	7.14	10.7	7.14	7.14	7.14	7.14	9.28
		Zero-Shot-CoT	21.4	7.14	17.9	7.14	3.57	7.14	7.14	7.14	7.14	3.57	8.93
Semantics	ChatGPT	Zero-Shot	25.0	32.1	39.3	39.3	42.9	39.3	35.7	32.1	35.7	42.9	36.4
		Zero-Shot-CoT	25.0	28.6	35.7	28.6	35.7	35.7	28.6	35.7	39.3	28.6	32.2
	GPT-4	Zero-Shot	53.6	53.6	50.0	53.6	50.0	53.6	50.0	57.1	53.6	50.0	52.5
		Zero-Shot-CoT	53.6	57.1	53.6	53.6	57.1	53.6	50.0	53.6	57.1	50.0	53.9
Rule-based		-	64.3	60.7	60.7	46.4	67.9	50.0	64.3	57.1	53.6	46.4	57.1

1936 Statement: *icers*(Jonathan, David)  
1937 Answer: We can use logical rule L10:  $\forall A, B, C, D : inely(B, A) \wedge inely(B, C) \wedge inely(C, D) \wedge$   
1938  $atis(A) \rightarrow icers(A, D)$ , which states that if A is an artist and B, C, and D are in a line with B in the  
1939 middle, then A influences D.  
1940 We can use the following facts: - *inely*(Jonathan, Alina) (F55)  
1941 - *inely*(Alina, David) (F34)  
1942 From these facts, we can conclude that Jonathan, Alina, and David are in a line with Alina in the  
1943 middle. Also, we know that Alina is an artist. Therefore, we can apply L10 to conclude that Alina  
1944 influences David.  
1945 Since Alina influences David, we can use L4:  $\forall A, B : inely(A, B) \wedge atis(A) \rightarrow lesai(A, B)$ , which  
1946 states that if A influences B and A is an artist, then B is a musician. Therefore, we can conclude that  
1947 David is a musician.  
1948 Therefore, the statement *icers*(Jonathan, David) is False, since David is not an ice skater.  
1949 Statement: *indr*(Florian, Nina)  
1950 Answer: We can use logical rule L24:  $\forall A, B : inely(B, A) \wedge atis(A) \rightarrow indr(A, B)$ , which states  
1951 that if A is an artist and B is a fan of A, then B is indirectly a fan of all the artists that A is a fan of.  
1952 From the given facts, we know that *atis*(Florian) and *inely*(Emilia, Nina). Therefore, we can  
1953 conclude that Nina is indirectly a fan of Florian, since Emilia is a fan of Florian and Nina is a fan of  
1954 Emilia.  
1955 Thus, the answer is True.

## 1956 O More reasoning results of Symbolic Tree

1957 We experiment with 10 sampled trees and report the average results in the main body. In this section,  
1958 we provide the reasoning results of each sampled Symbolic Tree, presented in Table Tables 12 to 14.

## 1959 P Abductive reasoning on smaller datasets

1960 We use smaller Symbolic Tree datasets to conduct the abductive reasoning experiment, which contains  
1961 about 12 entities and 100 facts. The results are provided in Table 15. We compare *Symbols* and  
1962 *Semantics* and find that the *Semantics* setting still outperforms the *Symbols* setting. This reinforces  
1963 the hypothesis that preserving semantics enhances the reasoning capabilities of LLMs.

Table 14: The abductive reasoning results of Symbolic Tree KGs. Results are in %.

Category	Model	Baseline	S <sub>1</sub>	S <sub>2</sub>	S <sub>3</sub>	S <sub>4</sub>	S <sub>5</sub>	S <sub>6</sub>	S <sub>7</sub>	S <sub>8</sub>	S <sub>9</sub>	S <sub>10</sub>	Avg.
Symbols	ChatGPT	Zero-Shot	1.95	0.31	1.07	1.52	2.36	1.45	1.06	0.75	3.1	1.39	1.50
		Zero-Shot-CoT	3.57	4.08	5.00	3.03	3.70	3.77	5.28	7.55	7.78	5.21	4.90
		Few-Shot-CoT	13.3	7.70	8.39	7.42	10.8	8.55	10.7	14.3	8.95	7.99	9.81
		Zero-Plus-Few-Shot-CoT	22.7	16.7	15.0	11.5	19.9	12.6	12.7	25.3	15.2	16.3	16.8
Semantics	ChatGPT	Zero-Shot	1.95	3.14	3.57	1.52	2.69	2.32	3.87	3.02	3.89	3.47	2.94
		Zero-Shot-CoT	4.22	5.34	4.64	3.63	2.69	2.90	4.23	1.89	3.11	1.39	3.40
		Few-Shot-CoT	9.90	13.2	10.9	7.42	8.59	0.97	11.3	13.0	11.3	11.1	9.77
		Zero-Plus-Few-Shot-CoT	17.5	25.2	22.1	16.7	16.5	18.0	22.2	27.2	22.6	21.5	20.9
Rule-based		-	100	100	100	100	100	100	100	100	100	100	100

Additionally, abductive reasoning in a shorter context yielded better performance compared to a longer context. This suggests that the length of the context has an impact on reasoning performance. Shorter contexts make selecting relevant and useful information easier while minimizing the influence of unrelated content.

Table 15: The abductive reasoning results of a smaller Symbolic Tree. Results are in %.

Category	Baseline	short context	long context
Symbols	ChatGPT: Zero-Shot-CoT	9.78	3.57
	GPT-4: Zero-Shot-CoT	46.7	32.1
Semantics	ChatGPT: Zero-Shot-CoT	5.43	4.22
	GPT-4: Zero-Shot-CoT	59.8	31.8

## Q Replacing entity labels

In this section, we conducted experiments to investigate the effects of replacing entity names (such as “Alice”) with entity IDs (*e.g.*, “e1”) in the context of reasoning tasks. The results are provided in Table 16. Comparing the performance of replacing relation names with replacing both entity and relation names, we observe that replacing entity names after replacing relation names had little impact on the overall performance.

Furthermore, we consider the scenario of only replacing entity names. Compared to the case of not replacing any labels, the results indicate that although replacing entity labels retains some level of semantics, it has a detrimental effect on reasoning performance. Additionally, we observed that the negative impact of decoupling the semantics of relations was more significant than that of decoupling the semantics of entities. These findings indicate a substantial portion of the semantic information is concentrated in the relation names.

Table 16: Comparison of replacing entity labels in deductive reasoning experiment (ChatGPT). Results are in %.

	Zero-Shot	Zero-Shot-CoT
replacing none	69.3	66.1
replacing ent	63.6	58.9
replacing rel	54.5	54.5
replacing ent & rel	57.5	55.6

## R Multi-short rules

Besides, a single rule can be equivalent to multiple rules. For example, the rule  $\forall x, y, z : \text{parentOf}(x, y) \wedge \text{parentOf}(y, z) \wedge \text{gender}(x, \text{female}) \rightarrow \text{GrandmotherOf}(x, z)$  can be represented as  $\forall x, y, z : \text{parentOf}(x, y) \wedge \text{parentOf}(y, z) \rightarrow \text{GrandparentOf}(x, z), \text{GrandparentOf}(x, z) \wedge \text{gender}(x, \text{female}) \rightarrow \text{GrandmotherOf}(x, z)$ . We conduct the experiments with both rule representations and find single-longer rules perform better than multiple-short rules. Results are presented in Table 11.