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

Anonymous Author(s)
Affiliation
Address
email

Abstract

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

1 Introduction

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

Despite the powerful and versatile in-context learning ability of LLMs, the underlying mechanisms by which they operate within a given context still remain unclear. Previous works investigate which aspects of the given examples contribute to the final task performance, including ground-truth labels and example ordering [7–9]. Another line of recent work has focused on explaining and leveraging the in-context learning (ICL) mechanism [10–13]. However, the basic problem they have in common is that the in-context prompts they input are based on natural language queries to investigate the reasoning abilities of LLMs. However, according to the Dual Process Theory [14, 15], humans are capable of using symbolic reasoning with System II to solve complex logical reasoning problems. To fill the research gap, we systematically study the in-context reasoning ability of LLMs by decoupling the semantics from the language reasoning process. With extensive experiments, we aim to answer the following research question: *Are LLMs good in-context reasoners without semantics?*

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

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

2 Related Works

Reasoning in LLMs Reasoning is a fundamental cognitive process involving logical inferences and conclusions based on given information. Developing models with strong reasoning capabilities has attracted increasing attention and many researches have been conducted on this topic since early days in the NLP domain [17]. Since then, various benchmarks focusing on different aspects of reasoning have been proposed, including natural language inference (NLI) [18–20], commonsense reasoning [21, 22], multi-hop reasoning [23, 24] *etc*. In recent years, there has been growing interests in studying the reasoning abilities of LLMs. Researchers have explored various approaches to enable LLMs to perform better on reasoning tasks. For example, "chain-of-thought (CoT)" [5, 25] is proposed to facilitate models to generate a reasoning path that decomposes complex reasoning into multiple easier steps; LLMs are decent zero-shot reasoners by adding a simple prompt, "Let's think step by step", to facilitate step-by-step thinking before giving the final answer [6]. This significantly improves the performance on arithmetic [26], commonsense [21, 27], and symbolic reasoning [5] benchmarks. However, despite their impressive performance on various reasoning benchmarks, all the tasks evaluated are rich in semantics. Thus it is unclear where the reasoning abilities of LLMs come from. This motivates us to investigate LLMs' reasoning abilities when semantics are decoupled.

In-Context Learning LLMs' reasoning abilities are closely related to in-context learning (ICL). ICL refers to the ability of language models to adapt and learn from a few prompt examples during the inference process. In recent years, there has been a focus on exploring how to improve the performance of ICL. Specifically, some works select related demonstrations to the test instance using off-the-shelf unsupervised similarity metrics or train a prompt retriever to select examples [28–30]. Others incorporate task instructions or different task prompts [31, 32]. Despite the empirical success, the underlying mechanisms of ICL still remain unclear. A few studies have shown that the performance of ICL usually varies with the choice of in-context demonstrations [8, 33]. Specifically, the order of demonstrations may lead to large performance fluctuations [34, 9]. 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 [35] and input-label correspondence plays a more important role in contextual demonstration [36]. 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 [37]. 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 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 and cognitive science [38–40]. It involves manipulating symbols and applying logical rules to perform deduction [41], induction [39], and abduction [42]. Boole [43] introduced Boolean algebra, which laid the foundation for symbolic logic and provided a formal system for logical reasoning. McCarthy [44] introduced LISP programming language and the concept of symbolic computation, which boosted the development of sophisticated AI programs that could represent and manipulate complex ideas and relationships. Fuhr [45] introduced probabilistic Datalog, an extension of Datalog with probabilities, allowing for probabilistic reasoning in logic-based systems. Eiter et al. [46] introduced answer set programming (ASP), a logic-based programming paradigm that combines logic programming and non-monotonic reasoning. ASP has been used for various reasoning tasks, including planning, knowledge representation, and constraint solving. Yi et al. [47] proposed a neural-symbolic approach to visual question answering. It combines deep neural networks with symbolic rules to perform compositional and interpretable reasoning over visual and textual information. Shin et al. [48] explore using LLM-based models for program synthesis. They present an approach that leverages inferred execution traces to guide the generation of correct programs. Lample and Charton [49] 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. [50] 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.

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 [51–54]. 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.
- *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

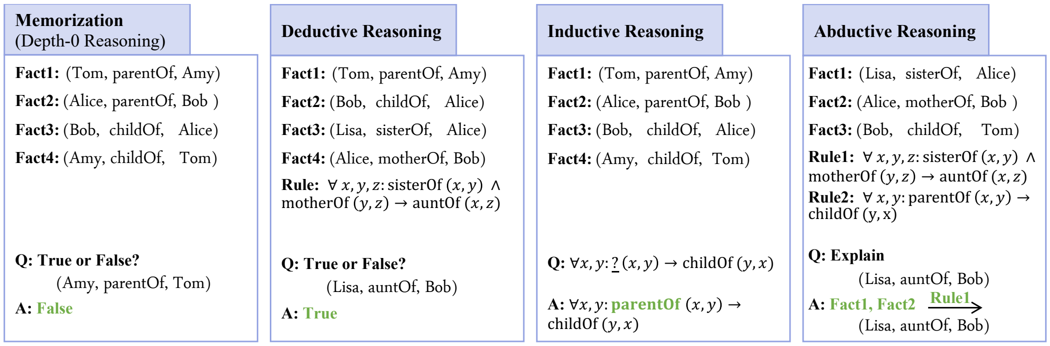

Figure 1: Task Definitions. **Memorization**: retrieving the predicted fact from in-context knowledge. **Deductive**: predicting the correctness of the predicted fact given rules and facts. **Inductive**: generating a rule based on multiple facts with similar patterns. **Abductive**: explaining the predicted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

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 [55]) 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 [56] 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** [57] 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.

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. Does it imply that the statement "The cow likes the squirrel." is True?	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 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 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).

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 "$\forall x, y$: parentOf$(x, y) \rightarrow$ 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 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

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

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 ambiguation problems and eliminating the influence of pre-existing knowledge on the memorization task. When testing, we follow the prompting of Taori et al. [58], 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+ [59], 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 the huge advantage of using KGs/external DBs to update knowledge rather than finetuning, aligning with the recent trend of retrieval-based LLMs.

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
Semantics	ChatGPT	Zero-Shot	66.1	36.4	2.94
		Zero-Shot-CoT	65.5	32.2	3.40
		Few-Shot-CoT	67.1	-	21.8
		Zero-Plus-Few-Shot-CoT	67.2	-	20.9
	GPT-4	Zero-Shot	79.2	52.5	27.3
		Zero-Shot-CoT	86.2	53.9	33.4
		Few-Shot-CoT	91.1	-	69.2
	Random	-	50.1	3.57	-
	Logic-based	-	100	57.1	100

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

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, 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 $\forall x, y$: parentOf$(x, y) \wedge$ female$(x) \rightarrow$ motherOf(x, y), we obtain $\forall x, y$: brotherOf$(x, y) \wedge$ male$(x) \rightarrow$ 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.

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

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

4.4 More analysis and discussions

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

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

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

(4) Zero-shot capabilities are approaching zero-shot-CoT capabilities: In *Symbols* setting, comparing zero-shot with zero-shot-CoT across deduction, induction, and abduction evaluations, we observe that zero-shot-CoT only marginally improves the performance compared to zero-shot learning. This finding suggests that the zero-shot capabilities of current LLMs are approaching their zero-shot-CoT learning abilities. One plausible explanation is that ChatGPT has already been trained on similar tasks with CoT and has memorized the instructions. Consequently, it implicitly follows these instructions when applied to the same queries, even without explicit CoT guidance [60].

(5) Utilizing internal knowledge outperforms external in-context knowledge: To explore the ability of LLMs to utilize internal and external knowledge, we conduct an additional experiment where we provide LLMs with only the relevant facts related to the predicted fact. We compare the performance of *Removing rules* (leveraging internal knowledge) with *Semantics* (providing external logical rules). Surprisingly, 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 this internal knowledge is more effective for reasoning than relying on external logical rules. Detailed results and case studies can be found in Appendix K.1.

5 Conclusion and Discussion

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

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

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

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

[26] Subhro Roy and Dan Roth. Solving general arithmetic word problems. *arXiv preprint arXiv:1608.01413*, 2016.

[27] 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.

[28] 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`.

[29] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*, 2021.

[30] 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.

[31] 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.

[32] 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.

[33] Yao Lu, Max Bartolo, AlastairP. Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity, Apr 2021.

[34] 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`.

[35] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work?

[36] Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, Kang Min Yoo, and Taeuk Kim. Ground-truth labels matter: A deeper look into input-label demonstrations. *arXiv preprint arXiv:2205.12685*, 2022.

[37] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models, Nov 2022.

[38] 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.

[39] Nada Lavrac and Saso Dzeroski. Inductive logic programming. In *WLP*, pages 146–160. Springer, 1994.

[40] Allen Newell and Herbert A Simon. Computer science as empirical inquiry: Symbols and search. In *ACM Turing award lectures*, page 1975. 2007.

[41] Philip N Johnson-Laird. Deductive reasoning. *Annual review of psychology*, 50(1):109–135, 1999.

[42] Gyöngyi Kovács and Karen M Spens. Abductive reasoning in logistics research. *International journal of physical distribution & logistics management*, 2005.

[43] George Boole. *The mathematical analysis of logic*. Philosophical Library, 1847.

[44] John McCarthy. Recursive functions of symbolic expressions and their computation by machine, part i. *Communications of the ACM*, 3(4):184–195, 1960.

[45] Norbert Fuhr. Probabilistic datalog: Implementing logical information retrieval for advanced applications. *Journal of the American Society for Information Science*, 51(2):95–110, 2000.

[46] Thomas Eiter, Giovambattista Ianni, and Thomas Krennwallner. *Answer set programming: A primer*. Springer, 2009.

[47] 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.

[48] Eui Chul Shin, Illia Polosukhin, and Dawn Song. Improving neural program synthesis with inferred execution traces. *Advances in Neural Information Processing Systems*, 31, 2018.

[49] Guillaume Lample and François Charton. Deep learning for symbolic mathematics. *arXiv preprint arXiv:1912.01412*, 2019.

[50] 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.

[51] Peter Cathcart Wason and Philip Nicholas Johnson-Laird. *Psychology of reasoning: Structure and content*, volume 86. Harvard University Press, 1972.

[52] 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`.

[53] Ronald Fagin, Joseph Y Halpern, Yoram Moses, and Moshe Vardi. *Reasoning about knowledge*. MIT press, 2004.

[54] Douglas N Walton. What is reasoning? what is an argument? *The journal of Philosophy*, 87(8):399–419, 1990.

[55] 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.

[56] Patrick Hohenecker and Thomas Lukasiewicz. Ontology reasoning with deep neural networks. *Journal of Artificial Intelligence Research*, 68:503–540, 2020.

[57] Massimo Carro. Nosql databases. *arXiv preprint arXiv:1401.2101*, 2014.

[58] 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`, 2023.

[59] 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.

[60] Jiuhai 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

A Prompts

A.1 Deductive reasoning

A.1.1 Zero-Shot

```
system: You are a helpful assistant with deductive reasoning abilities.
user: I will provide a set of logical rules L1 to L{number of rules} and facts F1 to F{number
     of basic facts}. Please select one single logical rule from L1 to L{number of rules} and
     a few facts from F1 to F{number of basic facts} to predict True/False of the unknown fact
      using deductive reasoning.
Logical rules: {logical rules}
Facts: {basic facts}
Unknown fact: {statement}
The answer (True or False) is:
```

A.1.2 Zero-Shot-CoT

```
system: You are a helpful assistant with deductive reasoning abilities. Please select one
     single logical rule and a few facts to predict True/False of the following statement.
user: I will provide a set of logical rules L1 to L{number of rules} and facts F1 to F{number
     of basic facts}. Please select one single logical rule from L1 to L{number of rules} and
     a few facts from F1 to F{number of basic facts} to predict True/False of the following
     statement using deductive reasoning.
Logical rules: {logical rules}
```

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

A.1.3 Few-Shot-CoT

```
system: You are a helpful assistant with deductive reasoning abilities. Please select one
     single logical rule and a few facts to predict True/False of the following statement.
user:I will provide a set of logical rules L1 to L{number of rules} and facts F1 to F{number
     of basic facts}.
Logical rules: {logical rules}
Facts: {basic facts}
Please select one single logical rule from L1 to L{number of rules} and a few facts from F1 to
      F{number of basic facts} to predict True/False of the following statement using
     deductive reasoning.
{Demonstrations}
Statement: {statement}
Answer:
```

A.1.4 Examples of Few-Shot-CoT Demonstrations

Statement: $r14$(Amelie, Jonathan)
Answer: We can use L11: $\forall A, B, C, D, E : r3(B, A) \wedge r3(B, C) \wedge r3(C, D) \wedge r3(D, E) \wedge r2(A) \rightarrow r14(A, E)$ to make a deduction. From the facts, we know that $r3$(Amelie, Thomas) and $r3$(Thomas, Jonathan). Therefore, we can apply L11 with A = Amelie, B = Thomas, C = Jonathan, D and E as variables. This gives us: $r3$(Thomas, Jonathan) $\wedge$ $r3$(Amelie, Thomas) $\wedge$ $r3$(Thomas, D) $\wedge$ $r3$(D, E) $\wedge$ $r2$(Amelie) $\rightarrow$ $r14$(Amelie, E) Since we know that $r2$(Amelie) is true from F7, we can simplify the above statement to: $r3$(Thomas, Jonathan) $\wedge$ $r3$(Amelie, Thomas) $\wedge$ $r3$(Thomas, D) $\wedge$ $r3$(D, E) $\rightarrow$ $r14$(Amelie, E) Now, we can substitute the facts $r3$(Amelie, Thomas) and $r3$(Thomas, Jonathan) to get: $r3$(Thomas, Jonathan) $\wedge$ $r3$(Thomas, D) $\wedge$ $r3$(D, E) $\rightarrow$ $r14$(Amelie, E) We can see that this statement is true if we choose D = Alina and E = Jonathan, since we know that $r3$(Thomas, Alina) from F50 and $r3$(Alina, Jonathan) from F56. Therefore, the statement $r14$(Amelie, Jonathan) is true.

Statement: $r31$(Philipp, Nina)
Answer: Let's use L28: $\forall A, B, C, D : r3(B, A) \wedge r3(C, B) \wedge r3(C, D) \wedge r1(A) \rightarrow r31(A, D)$ to see if we can prove the statement $r31$(Philipp, Nina). From the facts, we know that $r3$(Luisa, Nina) (F52) and $r3$(Emilia, Philipp) (F62). We can use L28 with $A = Philipp$, $B = Emilia$, $C = Luisa$, and $D = Nina$ to get: $r3$(Emilia, Philipp) $\wedge$ $r3$(Luisa, Emilia) $\wedge$ $r3$(Luisa, Nina) $\wedge$ $r1$(Philipp) $\rightarrow$ $r31$(Philipp, Nina) Since all the conditions are met, we can conclude that $r31$(Philipp, Nina) is True. Therefore, the answer is True.

Statement: $r8$(Leonie, Nico)
Answer: We can use L5 to make a deduction about the statement $r8$(Leonie, Nico). L5 states that if there exist three individuals A, B, and C such that A is related to B, B is related to C, and A has a certain property (r2), then A is related to C in a certain way (r8). Using this rule, we can see that we have the following facts: - $r3$(Leonie, Emily)
- $r3$(Emily, Nico)
Therefore, we can conclude that $r8$(Leonie, Nico) is true.

Statement: $r7$(Patrick, Alina)
Answer: We can use L4: $\forall A, B : r3(A, B) \wedge r1(A) \rightarrow r7(A, B)$ to determine if $r7$(Patrick, Alina) is true or false. From the facts, we know that $r3$(Patrick, Alina) is true (F55). We also know that $r1$(Patrick) is true (F9). Therefore, we can apply L4 to conclude that $r7$(Patrick, Alina) is true. Answer: True.

Statement: $r27$(Jonathan, Leonie)
Answer: Let's use L4: $\forall A, B : r3(A, B) \wedge r1(A) \rightarrow r7(A, B)$ and F56: $r3$(Patrick, Jonathan) to infer that $r7$(Patrick, Jonathan) is true. Then, we can use L24: $\forall A, B, C : r3(B, A) \wedge r3(C, B) \wedge r1(A) \rightarrow r27(A, C)$ and F44: $r3$(Leonie, Emily) to infer that $r27$(Jonathan, Leonie) is false, since there is no fact that supports $r3$(Jonathan, Emily). Therefore, the answer is False.

A.1.5 Zero-Shot of removing rules setting

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

A.2 Inductive reasoning

A.2.1 Zero-Shot

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

A.2.2 Zero-Shot CoT

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

A.2.3 Zero-Shot of removing facts setting

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

A.3 Abductive reasoning

A.3.1 Zero-Shot

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

A.3.2 Zero-Shot-CoT

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

A.3.3 Few-Shot-CoT prompt

```
system: You are a helpful assistant with abductive reasoning abilities. Please select one
    single logical rule and a few facts to explain the following statement.
user:
   I will provide a set of logical rules L1 to L{number of rules} and facts F1 to F{number of
       basic facts}. Please select one single logical rule from L1 to L{number of rules} and
       a few facts from F1 to F{number of basic facts} to explain the following statement.
   Rules: {logical rules}
   Facts: {basic facts}
   {Demonstrations}
   Statement: {statement}
   Answer:
```

A.3.4 Examples of Few-Shot-CoT Demonstrations

Statement: $r4$(Laura, Fabian)
Answer: To explain the statement $r4$(Laura, Fabian), we can use the rule L3 and fact F1 and F37.
L3: $\forall A, B : r1(A, B) \wedge r44(A) \rightarrow r4(A, B)$
F1: $r1$(Laura,Fabian)
F37: $r44$(Laura)
According to L3, if there is a relationship r1 between A and B, and A has the property r44, then the relationship r4 holds between A and B. In this case, we have a relationship r1 between Laura and Fabian (F1) and Laura has the property r44 (F37). Therefore, the relationship r4 holds between Laura and Fabian, which is the given statement.
The selected rule and facts are: L3, F1, and F37.

Statement: $r5$(Stefan, Emily)
Answer: To explain the statement $r5$(Stefan, Emily), we can use the rule L4 and fact F17. L4: $\forall A, B : r1(A, B) \wedge r43(A) \rightarrow r5(A, B)$
F17: $r1$(Stefan,Emily)
According to rule L4, if there is a relation r1(A, B) and A has the property r43, then the relation r5(A, B) holds. In our case, we have the fact F17, which states that there is a relation r1(Stefan, Emily). Additionally, Stefan has the property r43 (F61). Therefore, we can conclude that the relation r5(Stefan, Emily) holds.
The selected rule and facts are: L4, F17, and F61.

Statement: $r23$(Felix, Elias)
Answer: To explain the statement $r23$(Felix, Elias), we can use the rule L22 and fact F5. L22: $\forall A, B : r1(B, A) \wedge r43(A) \rightarrow r23(A, B)$
F5: $r1$(Elias,Felix)
According to rule L22, if there exists a relationship $r1$(B, A) and A has the property $r43$, then the relationship $r23$(A, B) holds. In our case, we have the fact F5, which states that there is a relationship $r1$(Elias, Felix). Additionally, we know that Felix has the property $r43$ (F59). Therefore, applying rule L22, we can conclude that the relationship $r23$(Felix, Elias) holds.
The selected rule and facts are: L22, F5, and F59.

Statement: $r3$(Felix, Fabian)
Answer: To explain the statement $r3$(Felix, Fabian), we can use the rule L2 and facts F2, F1, and F59.
L2: $\forall A, B, C : r1(B, A) \wedge r1(B, C) \wedge r43(A) \rightarrow r3(A, C)$
F1: $r1$(Laura,Fabian)

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

Statement: $r7$(Fabian, Helga)
Answer: To explain the statement $r7$(Fabian, Helga), we can use the rule L6 and facts F13, F21, and F39.
L6: $\forall A, B, C : r1(A, B) \wedge r1(B, C) \wedge r43(A) \rightarrow r7(A, C)$
F13: $r1$(Fabian,Thomas)
F21: $r1$(Thomas,Helga)
F39: $r43$(Fabian)
According to L6, if there exist A, B, and C such that $r1(A, B)$, $r1(B, C)$, and $r43(A)$, then $r7(A, C)$ holds. In our case, A = Fabian, B = Thomas, and C = Helga. Since we have $r1$(Fabian,Thomas) from F13, $r1$(Thomas,Helga) from F21, and $r43$(Fabian) from F39, we can conclude that $r7$(Fabian, Helga) is true.
So, the selected rule and facts are: L6, F13, F21, and F39.

A.3.5 Few-Shot CoT of removing rules setting

```
system: You are a helpful assistant with abductive reasoning abilities. Please select a few
    facts to explain the following statement.
user: I will provide a set of facts F1 to F63. Please select a few facts from F1 to F63 to
    explain the following statement.
Facts: {facts}

Statement: Laura is mother of Felix.
Answer: To explain the statement "Laura is mother of Felix", we can use Facts:
Fact F2 states: Laura is parent of Felix.
Fact F37 states: Laura is female.
Using F2 and F37, we can conclude that "Laura is mother of Felix" holds.
Therefore, the selected rule and facts are F2, F37.

Statement: Samuel is brother of Alina.
Answer: To infer the statement "Samuel is brother of Alina", we have:
F27: Patrick is parent of Samuel.
F28: Patrick is parent of Alina.
F47: Samuel is male.
Based on these facts, we can infer "Samuel is brother of Alina":
Therefore, the selected rule and facts are F27, F28, F47.

Statement: Patrick is grandfather of David.
Answer: To explain the statement "Patrick is grandfather of David", we have:
F28: Patrick is parent of Alina.
F7: Alina is parent of David.
F45: Patrick is male.
Based on these facts, we can infer "Patrick is grandfather of David":
Therefore, the selected rule and facts are F28, F7, F45.

Statement: Amelie is daughter of Elena.
Answer: To explain the statement "Amelie is daughter of Elena", we have:
F20: Elena is parent of Amelie.
F43: Amelie is female.
Based on these facts, we can infer "Amelie is daughter of Elena".
Therefore, the selected rule and facts are F20, F43.

Statement: Claudia is sister of Felix
Answer: To prove the statement "Claudia is sister of Felix", we can use facts:
F3: Laura is parent of Claudia.
F2: Laura is parent of Felix.
F40: Claudia is female.
Based on these facts, we can infer "Claudia is sister of Felix".
Therefore, the selected rule and facts are F3, F2, F40.

Statement: Laura is mother of Fabian.
```

```
Answer:
```

A.4 Commonsense reasoning

A.4.1 Zero-Shot prompt

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

A.4.2 Few-Shot prompt

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

A.4.3 Examples of Few-Shot Demonstrations

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

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

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

A.5 LLaMA Fine-tuning Prompt

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

B Deduction examples of Symbolic Tree datasets

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

B.1 Semantics

B.1.1 Logic language representations

```
Logical rules:
L1: $\forall A,B,C: parentOf(B, A) \land parentOf(B, C) \land female(A) \rightarrow sisterOf(A,
     C)$
L2: $\forall A,B,C: parentOf(B, A) \land parentOf(B, C) \land male(A) \rightarrow brotherOf(A,
     C)$
L3: $\forall A,B: parentOf(A, B) \land female(A) \rightarrow motherOf(A,B)$
L4: $\forall A,B: parentOf(A, B) \land male(A) \rightarrow fatherOf(A,B)$
L5: $\forall A,B,C: parentOf(A, B) \land parentOf(B, C) \land female(A) \rightarrow
     grandmotherOf(A,C)$
L6: $\forall A,B,C: parentOf(A, B) \land parentOf(B, C) \land male(A) \rightarrow
     grandfatherOf(A,C)$
L7: $\forall A,B,C,D: parentOf(A, B) \land parentOf(B, C) \land parentOf(C, D) \land female(A)
      \rightarrow greatGrandmotherOf(A,D)$
L8: $\forall A,B,C,D: parentOf(A, B) \land parentOf(B, C) \land parentOf(C, D) \land male(A) \
     rightarrow greatGrandfatherOf(A,D)$
L9: $\forall A,B,C,D: parentOf(B, A) \land parentOf(B, C) \land parentOf(C, D) \land female(A)
      \rightarrow auntOf(A,D)$
L10: $\forall A,B,C,D: parentOf(B, A) \land parentOf(B, C) \land parentOf(C, D) \land male(A)
     \rightarrow uncleOf(A,D)$
L11: $\forall A,B,C,D,E: parentOf(B, A) \land parentOf(B, C) \land parentOf(C, D) \land
     parentOf(D, E) \land female(A) \rightarrow greatAuntOf(A,E)$
L12: $\forall A,B,C,D,E: parentOf(B, A) \land parentOf(B, C) \land parentOf(C, D) \land
     parentOf(D, E) \land male(A) \rightarrow greatUncleOf(A,E)$
L13: $\forall A,B,C,D,E,F: parentOf(B, A) \land parentOf(C, B) \land parentOf(C, D) \land
     parentOf(D, E) \land parentOf(E, F) \land female(A) \rightarrow secondAuntOf(A,F)$
L14: $\forall A,B,C,D,E,F: parentOf(B, A) \land parentOf(C, B) \land parentOf(C, D) \land
     parentOf(D, E) \land parentOf(E, F) \land male(A) \rightarrow secondUncleOf(A,F)$
L15: $\forall A,B,C,D,E: parentOf(B, A) \land parentOf(C, B) \land parentOf(C, D) \land
     parentOf(D, E) \land female(A) \rightarrow girlCousinOf(A,E)$
L16: $\forall A,B,C,D,E: parentOf(B, A) \land parentOf(C, B) \land parentOf(C, D) \land
     parentOf(D, E) \land male(A) \rightarrow boyCousinOf(A,E)$
L17: $\forall A,B,C,D,E,F,G: parentOf(B, A) \land parentOf(C, B) \land parentOf(D, C) \land
     parentOf(D, E) \land parentOf(E, F) \land parentOf(F, G) \land female(A) \rightarrow
     girlSecondCousinOf(A,G)$
L18: $\forall A,B,C,D,E,F,G: parentOf(B, A) \land parentOf(C, B) \land parentOf(D, C) \land
     parentOf(D, E) \land parentOf(E, F) \land parentOf(F, G) \land male(A) \rightarrow
     boySecondCousinOf(A,G)$
L19: $\forall A,B,C,D,E,F: parentOf(B, A) \land parentOf(C, B) \land parentOf(D, C) \land
     parentOf(D, E) \land parentOf(E, F) \land female(A) \rightarrow
     girlFirstCousinOnceRemovedOf(A,F)$
L20: $\forall A,B,C,D,E,F: parentOf(B, A) \land parentOf(C, B) \land parentOf(D, C) \land
     parentOf(D, E) \land parentOf(E, F) \land male(A) \rightarrow boyFirstCousinOnceRemovedOf
     (A,F)$
L21: $\forall A,B: parentOf(B, A) \land female(A) \rightarrow daughterOf(A,B)$
L22: $\forall A,B: parentOf(B, A) \land male(A) \rightarrow sonOf(A,B)$
L23: $\forall A,B,C: parentOf(B, A) \land parentOf(C, B) \land female(A) \rightarrow
     granddaughterOf(A,C)$
L24: $\forall A,B,C: parentOf(B, A) \land parentOf(C, B) \land male(A) \rightarrow grandsonOf(
     A,C)$
L25: $\forall A,B,C,D: parentOf(B, A) \land parentOf(C, B) \land parentOf(D, C) \land female(A
     ) \rightarrow greatGranddaughterOf(A,D)$
L26: $\forall A,B,C,D: parentOf(B, A) \land parentOf(C, B) \land parentOf(D, C) \land male(A)
     \rightarrow greatGrandsonOf(A,D)$
L27: $\forall A,B,C,D: parentOf(B, A) \land parentOf(C, B) \land parentOf(C, D) \land female(A
     ) \rightarrow nieceOf(A,D)$
L28: $\forall A,B,C,D: parentOf(B, A) \land parentOf(C, B) \land parentOf(C, D) \land male(A)
     \rightarrow nephewOf(A,D)$

Facts:
F1: female(Laura)
F2: male(Elias)
F3: male(Fabian)
F4: female(Claudia)
F5: female(Elena)
F6: male(Thomas)
F7: female(Amelie)
F8: female(Luisa)
F9: male(Patrick)
```

```
F10: female(Emilia)
F11: male(Samuel)
F12: female(Alina)
F13: male(Jonathan)
F14: male(Philipp)
F15: male(Nico)
F16: male(David)
F17: female(Emily)
F18: male(Konstantin)
F19: male(Florian)
F20: female(Helga)
F21: female(Nina)
F22: female(Lea)
F23: male(Felix)
F24: female(Leonie)
F25: male(Stefan)
F26: male(Gabriel)
F27: male(Tobias)
F28: parentOf(Laura, Fabian)
F29: parentOf(Laura, Felix)
F30: parentOf(Laura, Claudia)
F31: parentOf(Elias, Fabian)
F32: parentOf(Elias, Felix)
F33: parentOf(Elias, Claudia)
F34: parentOf(Alina, David)
F35: parentOf(Alina, Lea)
F36: parentOf(Nico, David)
F37: parentOf(Nico, Lea)
F38: parentOf(Emily, Nico)
F39: parentOf(Konstantin, Nico)
F40: parentOf(Fabian, Thomas)
F41: parentOf(Fabian, Amelie)
F42: parentOf(Nina, Tobias)
F43: parentOf(Leonie, Emily)
F44: parentOf(Stefan, Emily)
F45: parentOf(Gabriel, Tobias)
F46: parentOf(Elena, Thomas)
F47: parentOf(Elena, Amelie)
F48: parentOf(Thomas, Helga)
F49: parentOf(Thomas, Nina)
F50: parentOf(Thomas, Patrick)
F51: parentOf(Luisa, Helga)
F52: parentOf(Luisa, Nina)
F53: parentOf(Luisa, Patrick)
F54: parentOf(Patrick, Samuel)
F55: parentOf(Patrick, Alina)
F56: parentOf(Patrick, Jonathan)
F57: parentOf(Patrick, Philipp)
F58: parentOf(Patrick, Florian)
F59: parentOf(Emilia, Samuel)
F60: parentOf(Emilia, Alina)
F61: parentOf(Emilia, Jonathan)
F62: parentOf(Emilia, Philipp)
F63: parentOf(Emilia, Florian)

Unknown fact: boyCousinOf(Tobias, David)
```

B.1.2 Natural language representations

```
Logical rules:
L1: If B is parent of A and B is parent of C and A is female, then A is sister of D.
L2: If B is parent of A and B is parent of C and A is male, then A is brother of D.
L3: If A is parent of B and A is female, then A is mother of C.
L4: If A is parent of B and A is male, then A is father of C.
L5: If A is parent of B and B is parent of C and A is female, then A is grandmother of D.
L6: If A is parent of B and B is parent of C and A is male, then A is grandfather of D.
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
    greatGrandmother of E.
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
    greatGrandfather of E.
```

```
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
     aunt of E.
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
     uncle of E.
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
     is female, then A is greatAunt of F.
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
     is male, then A is greatUncle of F.
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
     is parent of F and A is female, then A is secondAunt of G.
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
     is parent of F and A is male, then A is secondUncle of G.
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
     is female, then A is girlCousin of F.
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
     is male, then A is boyCousin of F.
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
     is parent of F and F is parent of G and A is female, then A is girlSecondCousin of H.
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
     is parent of F and F is parent of G and A is male, then A is boySecondCousin of H.
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
     is parent of F and A is female, then A is girlFirstCousinOnceRemoved of G.
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
     is parent of F and A is male, then A is boyFirstCousinOnceRemoved of G.
L21: If B is parent of A and A is female, then A is daughter of C.
L22: If B is parent of A and A is male, then A is son of C.
L23: If B is parent of A and C is parent of B and A is female, then A is granddaughter of D.
L24: If B is parent of A and C is parent of B and A is male, then A is grandson of D.
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
     greatGranddaughter of E.
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
     greatGrandson of E.
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
     niece of E.
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
     nephew of E.

Facts:
F1: Laura is female.
F2: Elias is male.
F3: Fabian is male.
F4: Claudia is female.
F5: Elena is female.
F6: Thomas is male.
F7: Amelie is female.
F8: Luisa is female.
F9: Patrick is male.
F10: Emilia is female.
F11: Samuel is male.
F12: Alina is female.
F13: Jonathan is male.
F14: Philipp is male.
F15: Nico is male.
F16: David is male.
F17: Emily is female.
F18: Konstantin is male.
F19: Florian is male.
F20: Helga is female.
F21: Nina is female.
F22: Lea is female.
F23: Felix is male.
F24: Leonie is female.
F25: Stefan is male.
F26: Gabriel is male.
F27: Tobias is male.
F28: Laura is parent of Fabian.
F29: Laura is parent of Felix.
F30: Laura is parent of Claudia.
F31: Elias is parent of Fabian.
F32: Elias is parent of Felix.
F33: Elias is parent of Claudia.
F34: Alina is parent of David.
```

```
F35: Alina is parent of Lea.
F36: Nico is parent of David.
F37: Nico is parent of Lea.
F38: Emily is parent of Nico.
F39: Konstantin is parent of Nico.
F40: Fabian is parent of Thomas.
F41: Fabian is parent of Amelie.
F42: Nina is parent of Tobias.
F43: Leonie is parent of Emily.
F44: Stefan is parent of Emily.
F45: Gabriel is parent of Tobias.
F46: Elena is parent of Thomas.
F47: Elena is parent of Amelie.
F48: Thomas is parent of Helga.
F49: Thomas is parent of Nina.
F50: Thomas is parent of Patrick.
F51: Luisa is parent of Helga.
F52: Luisa is parent of Nina.
F53: Luisa is parent of Patrick.
F54: Patrick is parent of Samuel.
F55: Patrick is parent of Alina.
F56: Patrick is parent of Jonathan.
F57: Patrick is parent of Philipp.
F58: Patrick is parent of Florian.
F59: Emilia is parent of Samuel.
F60: Emilia is parent of Alina.
F61: Emilia is parent of Jonathan.
F62: Emilia is parent of Philipp.
F63: Emilia is parent of Florian.

Unknown fact: Gabriel is uncle of Lea.
```

B.2 Symbolization

B.2.1 Logic language representations

```
Logical rules:
L1: $\forall A,B,C: r3(B, A) \land r3(B, C) \land r2(A) \rightarrow r4(A, C)$
L2: $\forall A,B,C: r3(B, A) \land r3(B, C) \land r1(A) \rightarrow r5(A, C)$
L3: $\forall A,B: r3(A, B) \land r2(A) \rightarrow r6(A, B)$
L4: $\forall A,B: r3(A, B) \land r1(A) \rightarrow r7(A, B)$
L5: $\forall A,B,C: r3(A, B) \land r3(B, C) \land r2(A) \rightarrow r8(A, C)$
L6: $\forall A,B,C: r3(A, B) \land r3(B, C) \land r1(A) \rightarrow r9(A, C)$
L7: $\forall A,B,C,D: r3(A, B) \land r3(B, C) \land r3(C, D) \land r2(A) \rightarrow r10(A, D)$
L8: $\forall A,B,C,D: r3(A, B) \land r3(B, C) \land r3(C, D) \land r1(A) \rightarrow r11(A, D)$
L9: $\forall A,B,C,D: r3(B, A) \land r3(B, C) \land r3(C, D) \land r2(A) \rightarrow r12(A, D)$
L10: $\forall A,B,C,D: r3(B, A) \land r3(B, C) \land r3(C, D) \land r1(A) \rightarrow r13(A, D)$
L11: $\forall A,B,C,D,E: r3(B, A) \land r3(B, C) \land r3(C, D) \land r3(D, E) \land r2(A) \rightarrow r14(A, E)$
L12: $\forall A,B,C,D,E: r3(B, A) \land r3(B, C) \land r3(C, D) \land r3(D, E) \land r1(A) \rightarrow r15(A, E)$
L13: $\forall A,B,C,D,E,F: r3(B, A) \land r3(C, B) \land r3(C, D) \land r3(D, E) \land r3(E, F) \land r2(A) \rightarrow r16(A, F)$
L14: $\forall A,B,C,D,E,F: r3(B, A) \land r3(C, B) \land r3(C, D) \land r3(D, E) \land r3(E, F) \land r1(A) \rightarrow r17(A, F)$
L15: $\forall A,B,C,D,E: r3(B, A) \land r3(C, B) \land r3(C, D) \land r3(D, E) \land r2(A) \rightarrow r18(A, E)$
L16: $\forall A,B,C,D,E: r3(B, A) \land r3(C, B) \land r3(C, D) \land r3(D, E) \land r1(A) \rightarrow r19(A, E)$
L17: $\forall A,B,C,D,E,F,G: r3(B, A) \land r3(C, B) \land r3(D, C) \land r3(D, E) \land r3(E, F) \land r3(F, G) \land r2(A) \rightarrow r20(A, G)$
L18: $\forall A,B,C,D,E,F,G: r3(B, A) \land r3(C, B) \land r3(D, C) \land r3(D, E) \land r3(E, F) \land r3(F, G) \land r1(A) \rightarrow r21(A, G)$
L19: $\forall A,B,C,D,E,F: r3(B, A) \land r3(C, B) \land r3(D, C) \land r3(D, E) \land r3(E, F) \land r2(A) \rightarrow r22(A, F)$
L20: $\forall A,B,C,D,E,F: r3(B, A) \land r3(C, B) \land r3(D, C) \land r3(D, E) \land r3(E, F) \land r1(A) \rightarrow r23(A, F)$
```

```
L21: $\forall A,B: r3(B, A) \land r2(A) \rightarrow r24(A, B)$
L22: $\forall A,B: r3(B, A) \land r1(A) \rightarrow r25(A, B)$
L23: $\forall A,B,C: r3(B, A) \land r3(C, B) \land r2(A) \rightarrow r26(A, C)$
L24: $\forall A,B,C: r3(B, A) \land r3(C, B) \land r1(A) \rightarrow r27(A, C)$
L25: $\forall A,B,C,D: r3(B, A) \land r3(C, B) \land r3(D, C) \land r2(A) \rightarrow r28(A, D)$
L26: $\forall A,B,C,D: r3(B, A) \land r3(C, B) \land r3(D, C) \land r1(A) \rightarrow r29(A, D)$
L27: $\forall A,B,C,D: r3(B, A) \land r3(C, B) \land r3(C, D) \land r2(A) \rightarrow r30(A, D)$
L28: $\forall A,B,C,D: r3(B, A) \land r3(C, B) \land r3(C, D) \land r1(A) \rightarrow r31(A, D)$

Facts:
F1: $r2$(Laura)
F2: $r1$(Elias)
F3: $r1$(Fabian)
F4: $r2$(Claudia)
F5: $r2$(Elena)
F6: $r1$(Thomas)
F7: $r2$(Amelie)
F8: $r2$(Luisa)
F9: $r1$(Patrick)
F10: $r2$(Emilia)
F11: $r1$(Samuel)
F12: $r2$(Alina)
F13: $r1$(Jonathan)
F14: $r1$(Philipp)
F15: $r1$(Nico)
F16: $r1$(David)
F17: $r2$(Emily)
F18: $r1$(Konstantin)
F19: $r1$(Florian)
F20: $r2$(Helga)
F21: $r2$(Nina)
F22: $r2$(Lea)
F23: $r1$(Felix)
F24: $r2$(Leonie)
F25: $r1$(Stefan)
F26: $r1$(Gabriel)
F27: $r1$(Tobias)
F28: $r3$(Laura, Fabian)
F29: $r3$(Laura, Felix)
F30: $r3$(Laura, Claudia)
F31: $r3$(Elias, Fabian)
F32: $r3$(Elias, Felix)
F33: $r3$(Elias, Claudia)
F34: $r3$(Alina, David)
F35: $r3$(Alina, Lea)
F36: $r3$(Nico, David)
F37: $r3$(Nico, Lea)
F38: $r3$(Emily, Nico)
F39: $r3$(Konstantin, Nico)
F40: $r3$(Fabian, Thomas)
F41: $r3$(Fabian, Amelie)
F42: $r3$(Nina, Tobias)
F43: $r3$(Leonie, Emily)
F44: $r3$(Stefan, Emily)
F45: $r3$(Gabriel, Tobias)
F46: $r3$(Elena, Thomas)
F47: $r3$(Elena, Amelie)
F48: $r3$(Thomas, Helga)
F49: $r3$(Thomas, Nina)
F50: $r3$(Thomas, Patrick)
F51: $r3$(Luisa, Helga)
F52: $r3$(Luisa, Nina)
F53: $r3$(Luisa, Patrick)
F54: $r3$(Patrick, Samuel)
F55: $r3$(Patrick, Alina)
F56: $r3$(Patrick, Jonathan)
F57: $r3$(Patrick, Philipp)
F58: $r3$(Patrick, Florian)
```

```
F59: $r3$(Emilia, Samuel)
F60: $r3$(Emilia, Alina)
F61: $r3$(Emilia, Jonathan)
F62: $r3$(Emilia, Philipp)
F63: $r3$(Emilia, Florian)

Unknown fact: $r9$(Thomas, Claudia)
```

B.2.2 Natural language representations:

```
Logical rules:
L1: If B is $r3$ of A and B is $r3$ of C and A is $r2$, then A is $r4$ of D.
L2: If B is $r3$ of A and B is $r3$ of C and A is $r1$, then A is $r5$ of D.
L3: If A is $r3$ of B and A is $r2$, then A is $r6$ of C.
L4: If A is $r3$ of B and A is $r1$, then A is $r7$ of C.
L5: If A is $r3$ of B and B is $r3$ of C and A is $r2$, then A is $r8$ of D.
L6: If A is $r3$ of B and B is $r3$ of C and A is $r1$, then A is $r9$ of D.
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
     E.
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
     E.
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
     E.
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
      E.
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 $r2$,
      then A is $r14$ of F.
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 $r1$,
      then A is $r15$ of F.
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 $r3$
     of F and A is $r2$, then A is $r16$ of G.
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 $r3$
     of F and A is $r1$, then A is $r17$ of G.
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 $r2$,
      then A is $r18$ of F.
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 $r1$,
      then A is $r19$ of F.
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 $r3$
     of F and F is $r3$ of G and A is $r2$, then A is $r20$ of H.
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 $r3$
     of F and F is $r3$ of G and A is $r1$, then A is $r21$ of H.
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 $r3$
     of F and A is $r2$, then A is $r22$ of G.
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 $r3$
     of F and A is $r1$, then A is $r23$ of G.
L21: If B is $r3$ of A and A is $r2$, then A is $r24$ of C.
L22: If B is $r3$ of A and A is $r1$, then A is $r25$ of C.
L23: If B is $r3$ of A and C is $r3$ of B and A is $r2$, then A is $r26$ of D.
L24: If B is $r3$ of A and C is $r3$ of B and A is $r1$, then A is $r27$ of D.
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
      E.
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
      E.
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
      E.
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
      E.

Facts:
F1: Laura is $r2$.
F2: Elias is $r1$.
F3: Fabian is $r1$.
F4: Claudia is $r2$.
F5: Elena is $r2$.
F6: Thomas is $r1$.
F7: Amelie is $r2$.
F8: Luisa is $r2$.
F9: Patrick is $r1$.
F10: Emilia is $r2$.
F11: Samuel is $r1$.
F12: Alina is $r2$.
```

```
F13: Jonathan is $r1$.
F14: Philipp is $r1$.
F15: Nico is $r1$.
F16: David is $r1$.
F17: Emily is $r2$.
F18: Konstantin is $r1$.
F19: Florian is $r1$.
F20: Helga is $r2$.
F21: Nina is $r2$.
F22: Lea is $r2$.
F23: Felix is $r1$.
F24: Leonie is $r2$.
F25: Stefan is $r1$.
F26: Gabriel is $r1$.
F27: Tobias is $r1$.
F28: Laura is $r3$ of Fabian.
F29: Laura is $r3$ of Felix.
F30: Laura is $r3$ of Claudia.
F31: Elias is $r3$ of Fabian.
F32: Elias is $r3$ of Felix.
F33: Elias is $r3$ of Claudia.
F34: Alina is $r3$ of David.
F35: Alina is $r3$ of Lea.
F36: Nico is $r3$ of David.
F37: Nico is $r3$ of Lea.
F38: Emily is $r3$ of Nico.
F39: Konstantin is $r3$ of Nico.
F40: Fabian is $r3$ of Thomas.
F41: Fabian is $r3$ of Amelie.
F42: Nina is $r3$ of Tobias.
F43: Leonie is $r3$ of Emily.
F44: Stefan is $r3$ of Emily.
F45: Gabriel is $r3$ of Tobias.
F46: Elena is $r3$ of Thomas.
F47: Elena is $r3$ of Amelie.
F48: Thomas is $r3$ of Helga.
F49: Thomas is $r3$ of Nina.
F50: Thomas is $r3$ of Patrick.
F51: Luisa is $r3$ of Helga.
F52: Luisa is $r3$ of Nina.
F53: Luisa is $r3$ of Patrick.
F54: Patrick is $r3$ of Samuel.
F55: Patrick is $r3$ of Alina.
F56: Patrick is $r3$ of Jonathan.
F57: Patrick is $r3$ of Philipp.
F58: Patrick is $r3$ of Florian.
F59: Emilia is $r3$ of Samuel.
F60: Emilia is $r3$ of Alina.
F61: Emilia is $r3$ of Jonathan.
F62: Emilia is $r3$ of Philipp.
F63: Emilia is $r3$ of Florian.

Unknown fact: Nico is $r27$ of Stefan.
```

B.3 Semantics of removing rule setting

```
I will provide a set of facts. Please predict True/False of the unknown fact based on given
    facts.
Facts:
F1: Laura is female.
F2: Elias is male.
F3: Fabian is male.
F4: Claudia is female.
F5: Elena is female.
F6: Thomas is male.
F7: Amelie is female.
F8: Luisa is female.
F9: Patrick is male.
F10: Emilia is female.
F11: Samuel is male.
F12: Alina is female.
```

```
F13: Jonathan is male.
F14: Philipp is male.
F15: Nico is male.
F16: David is male.
F17: Emily is female.
F18: Konstantin is male.
F19: Florian is male.
F20: Helga is female.
F21: Nina is female.
F22: Lea is female.
F23: Felix is male.
F24: Leonie is female.
F25: Stefan is male.
F26: Gabriel is male.
F27: Tobias is male.
F28: Laura is parent of Fabian.
F29: Laura is parent of Felix.
F30: Laura is parent of Claudia.
F31: Elias is parent of Fabian.
F32: Elias is parent of Felix.
F33: Elias is parent of Claudia.
F34: Alina is parent of David.
F35: Alina is parent of Lea.
F36: Nico is parent of David.
F37: Nico is parent of Lea.
F38: Emily is parent of Nico.
F39: Konstantin is parent of Nico.
F40: Fabian is parent of Thomas.
F41: Fabian is parent of Amelie.
F42: Nina is parent of Tobias.
F43: Leonie is parent of Emily.
F44: Stefan is parent of Emily.
F45: Gabriel is parent of Tobias.
F46: Elena is parent of Thomas.
F47: Elena is parent of Amelie.
F48: Thomas is parent of Helga.
F49: Thomas is parent of Nina.
F50: Thomas is parent of Patrick.
F51: Luisa is parent of Helga.
F52: Luisa is parent of Nina.
F53: Luisa is parent of Patrick.
F54: Patrick is parent of Samuel.
F55: Patrick is parent of Alina.
F56: Patrick is parent of Jonathan.
F57: Patrick is parent of Philipp.
F58: Patrick is parent of Florian.
F59: Emilia is parent of Samuel.
F60: Emilia is parent of Alina.
F61: Emilia is parent of Jonathan.
F62: Emilia is parent of Philipp.
F63: Emilia is parent of Florian.

Unknown fact: Jonathan is aunt of Thomas.
The answer (True or False) is:
```

C Examples of ProofWriter

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

C.1 Semantics

```
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.
Does it imply that the statement "The cow likes the squirrel." is True?
```

```
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.
Does it imply that the statement "The cow does not like the squirrel." is True?
```

```
Bob is blue.
Erin is quiet.
Fiona is cold.
Harry is cold.
All quiet things are blue.
If Harry is blue then Harry is not young.
Blue things are young.
Blue, round things are cold.
If something is blue and not red then it is round.
If something is young then it is white.
If Erin is red and Erin is not round then Erin is young.
If Erin is red and Erin is not cold then Erin is white.
Does it imply that the statement "Erin is white" is True?
Answer with only True or False. The answer is:
```

```
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.
Does it imply that the statement "The cow likes the squirrel." is True?
```

C.2 Symbols

```
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 e14 likes the e26." is True?
```

```
The e27 is e7.
The e27 is e15.
The e30 does not chase the e27.
The e30 eats the e27.
The e30 is e1.
The e30 is e15.
The e30 visits the e27.
If something visits the e27 then the e27 does not visit the e30.
If something is e1 and e15 then it visits the e30.
Does it imply that the statement "The e30 visits the e30." is True?
```

```
The e27 is e7.
The e27 is e15.
The e30 does not chase the e27.
The e30 eats the e27.
The e30 is e1.
The e30 is e15.
The e30 visits the e27.
If something visits the e27 then the e27 does not visit the e30.
If something is e1 and e15 then it visits the e30.
Does it imply that the statement "The e30 visits the e30." is True?
```

D Different Zero-Shot prompting

We try different Zero-Shot prompts:

(1)

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

(2)

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

(3)

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

The results of the three prompts in the Zero-Shot setting are presented in Table 5. Among the three 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_1	54.5	51.5	53.8

E Comparison of memorization abilities of neural-based and symbolic-based methods

We compare fine-tuned language models with the deterministic graph DB Neo4J to explore the memorization abilities of neural-based and symbolic-based methods. Language models can implicitly store and retrieve facts as "knowledge bases" within their neural parameters. They are trained on a snapshot of data and may not have access to the latest or most accurate information. In order to update or add facts, specific model parameters need to be modified, or the model needs to be fine-tuned with new data. In contrast, symbolic knowledge graphs can directly add or update individual triplets, making it easier to incorporate new information. Our comparison affirms the huge advantage of using KGs/external DBs to update knowledge rather than finetuning, aligning with the recent trend of 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 $\forall x,y,z$: ##$(x,y) \wedge$ ##$(y,z) \wedge$ ++$(x) \rightarrow auntOf(x,z)$ 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 [55]) of facts added/updated, and *forgetting* is the 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 all tokens between all tail entities. The better LLM remembers a triplet, the higher the MRR gets. Note that, there may be more than one entity for each (head, relation) pair. We only consider the rank one of them.

H Implementation of memorization

We selected 1258 triplets from 4 Symbolic Trees to evaluate the effectiveness of adding knowledge. Following the prompting of Taori et al. [58], we use the head entity and relation as instructions and provide all candidate tails as input. The model's training objective is to autoregress toward the true tail entities. The detailed prompting is contained in Appendix A.5. In the updating step, we fine-tune the model on all 620 triplets from the first two trees whose tail entities are randomly flipped to false ones. Besides the effectiveness of updating, we evaluated the forgetting ratio using the remaining 638 triplets of the least two trees. These triplets have been remembered in the first step and haven't been updated in the second. Noting that, within each tree, the relationships between entities are independent, and the entities are distinct. Therefore, we propose that LLM should retain its memory of the previously remembered triplets when updating based on the first two trees. 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 and standard MRR.

I Consistency of knowledge base

In the context of updating a knowledge base, it is important to ensure the consistency of the knowledge base. When revising a fact, it is necessary to update other related facts accordingly to maintain a coherent and accurate knowledge base. For example, if we have the facts: Alice is Bob's mother, Amy is Alice's sister, Momo is Lily's sister, and Amy is Bob's aunt, and we update Bob's mother as Lily, it follows that Bob's aunt should be updated to Momo instead of Amy to maintain consistency.

However, in our experiments, we did not consider the situation of updating related facts in order to simplify the evaluation process, although it may lead to inconsistencies within the knowledge base. In real-world scenarios, ensuring consistency is an essential aspect of maintaining the accuracy and reliability of a knowledge base when updating or revising facts.

J Different Few-Shot-CoT examples

We use three different kinds of demonstrations generation:

(1) leverage LLMs with the "Let's think step by step" prompt to generate reasoning chains for demonstrations.

(2) manually designed each demonstration with a question and a reasoning chain. A reasoning chain is composed of a rationale (a series of intermediate reasoning steps) and an expected answer.

(3) leverage LLMs with the "Let's think step by step" prompt to generate reasoning chains and filter with those generated chains coming with mistakes for demonstrations.

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

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

Statement: $r8$(Elena, Nina)
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 deduce whether the statement $r8$(Elena, Nina) is true or false. From the facts, we know that $r3$(Elena, Thomas) and $r3$(Thomas, Nina). Therefore, we can use L5 with A =Elena, B =Thomas, and C =Nina to see if $r8$(Elena, Nina) is true or false. L5 states that if $r3$(A, B) and $r3$(B, C) and $r2$(A)

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

Statement: $r29$(Philipp, Elena)
Answer: We can use logical rule L26: $\forall A, B, C, D : r3(B, A) \wedge r3(C, B) \wedge r3(D, C) \wedge r1(A) \rightarrow r29(A, D)$ to infer that if $r3(B, A)$, $r3(C, B)$, $r3(D, C)$, and $r1(A)$ are true, then $r29(A, D)$ is also true.
From the given facts, we know that $r3$(Elena, Thomas) (F46) and $r3$(Thomas, Amelie) (F47). Therefore, we can use L26 with $A =$ Philipp, $B =$ Thomas, $C =$ Amelie, and $D =$ Elena to get:
$r3$(Thomas, Philipp) (from F62 and $r3(B, A)$)
$r3$(Amelie, Thomas) (from F47 and $r3(C, B)$)
$r3$(Elena, Amelie) (from F47 and $r3(D, C)$)
$r1$(Philipp) (from F14 and $r1(A)$)
Therefore, we can conclude that $r29$(Philipp, Elena) is True.

Statement: $r6$(Luisa, Helga)
Answer: We can use logical rule L3: $\forall A, B : r3(A, B) \wedge r2(A) \rightarrow r6(A, B)$ to deduce the truth value of the statement $r6$(Luisa, Helga).
From the facts, we know that $r3$(Luisa, Helga) and $r2$(Luisa) are true. Therefore, we can apply L3 to conclude that $r6$(Luisa, Helga) is true.
Thus, the answer is True.

Statement: $r14$(Amelie, Alina)
Answer: We can use logical rule L11: $\forall A, B, C, D, E : r3(B, A) \wedge r3(B, C) \wedge r3(C, D) \wedge r3(D, E) \wedge r2(A) \rightarrow r14(A, E)$ to infer $r14$(Amelie, Alina) if we can find facts that satisfy the antecedent of the rule.
From the facts, we can see that $r3$(Amelie, Thomas) and $r3$(Thomas, Helga) and $r3$(Helga, David) and $r3$(David, Alina) are true. Therefore, we can apply L11 with $A =$Amelie, $B =$Thomas, $C =$Helga, $D =$David, and $E =$Alina to get $r14$(Amelie, Alina) is True.
Therefore, the answer is True.

Statement: $r26$(Alina, Luisa)
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 determine the truth value of the statement $r26$(Alina, Luisa).
From the facts, we know that $r3$(Alina, David) and $r3$(David, Luisa) (F34 and F52). Therefore, we can substitute A = Alina, B = David, and C = Luisa into L23 to get:
$r3$(David, Alina) $\wedge$ $r3$(Luisa, David) $\wedge$ $r2$(Alina) $\rightarrow$ $r26$(Alina, Luisa)
$r3$(Alina, David) is true (F34) and $r2$(Alina) is also true (F12). Therefore, we can simplify the above expression to:
$r3$(Luisa, David) $\rightarrow$ $r26$(Alina, Luisa)
$r3$(Luisa, David) is true (F52), so we can conclude that $r26$(Alina, Luisa) is true.
Therefore, the answer is True.

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

Statement: $r14$(Amelie, Jonathan)
Answer: We can use logical rule L11: $\forall A, B, C, D, E : r3(B, A) \wedge r3(B, C) \wedge r3(C, D) \wedge r3(D, E) \wedge r2(A) \rightarrow r14(A, E)$. We can substitute A = Amelie and E = Jonathan to get: $r3(B, Amelie) \wedge r3(B, C) \wedge r3(C, D) \wedge r3(D, Jonathan) \wedge r2(Amelie) \rightarrow r14(Amelie, Jonathan)$. Now, we need to find all facts that match the conditions $r3$(B, Amelie), $r3$(B, C), $r3$(C, D), $r3$(D, Jonathan) and $r2$(Amelie). We have:
F41: $r3$(Fabian, Amelie)
F47: $r3$(Elena, Amelie)
F56: $r3$(Patrick, Jonathan)
F61: $r3$(Emilia, Jonathan)
F7: $r2$(Amelie)
We can see that F41 and F47 match the condition $r3$(B, Amelie) with B being Fabian and Elena, respectively. F56 and F61 match the condition $r3$(D, Jonathan) with D being Patrick and Emilia, respectively. F7 matches the condition $r2$(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_1	55.6	53.6	40.7

values to B, C, and D:
(1) B = Fabian, D = Patrick.
(2) B = Elena, D = Emilia.
(3) B = Fabian, D = Emilia.
(4) B = Elena, D = Patrick.
Using the values of B and D from the possible assignments above, we can find the value of C that satisfies the conditions $r3$(B, C) and $r3$(C, D). For assignment (1), we have:
F40: $r3$(Fabian, Thomas)
F50: $r3$(Thomas, Patrick)
Since all the conditions are satisfied, we can conclude that the statement $r14$(Amelie, Jonathan) is True if we assign B = Fabian and C = Thomas and D = Patrick. Thus, the output is True.

Statement: $r31$(Philipp, Nina)
Answer: We can use logical rule L28: $\forall A, B, C, D : r3(B, A) \wedge r3(C, B) \wedge r3(C, D) \wedge r1(A) \rightarrow r31(A, D)$. We can substitute A = Philipp and D = Nina to get: $r3(B, Philipp) \wedge r3(C, B) \wedge r3(C, Nina) \wedge r1(Philipp) \rightarrow r31(Philipp, Nina)$. Now, we need to find all facts that match the conditions $r3$(B, Philipp), $r3$(C, B), $r3$(C, Nina), and $r1$(Philipp). We have:
F57: $r3$(Patrick, Philipp)
F62: $r3$(Emilia, Philipp)
F49: $r3$(Thomas, Nina)
F52: $r3$(Luisa, Nina)
We can see that F57 and F62 match the condition $r3$(B, Philipp) with B being Patrick and Emilia, respectively. F49 and F52 match the condition $r3$(C, Nina) with C being Thomas and Luisa, respectively. Therefore, there is only one single way of assigning values to B and C.
B = Patrick and C = Thomas. We can substitute B and C to get: $r3(Patrick, Philipp) \wedge r3(Thomas, Patrick) \wedge r3(Thomas, Nina) \wedge r1(Philipp) \rightarrow r31(Philipp, Nina)$. Since all the conditions are satisfied, we can conclude that the statement $r31$ (Philipp, Nina) is True if we assign B = Patrick and C = Thomas. Thus, the output is True.

Statement: $r8$(Leonie, Nico)
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 can substitute A = Leonie and C = Nico to get: $r3(Leonie, B) \wedge r3(B, Nico) \wedge r2(Leonie) \rightarrow r8(Leonie, Nico)$. Now, we need to find all facts that match the conditions $r3$(Leonie, B), $r3$(B, Nico), and $r2$(Leonie). We have:
F43: $r3$(Leonie, Emily)
F38: $r3$(Emily, Nico)
F39: $r3$(Konstantin, Nico)
F24: $r2$(Leonie)
We can see that F43 matches the condition $r3$(Leonie, B) with B being Emily. and F38 matches the condition $r3$(B, Nico) with B being Emily. Therefore, we can assign B = Emily. Also, F24 matches the condition $r2$(Leonie). Therefore, we can conclude that the statement $r8$(Leonie, Nico) is True if we assign B = Emily. Therefore, the output is True.

Statement: $r7$(Patrick, Alina)
Answer: We can use logical rule L4: $\forall A, B : r3(A, B) \wedge r1(A) \rightarrow r7(A, B)$. We can substitute A = Patrick and B = Alina to get: $r3(Patrick, Alina) \wedge r1(Patrick) \rightarrow r7(Patrick, Alina)$. Now, we need to find all facts that match the conditions $r3$(Patrick, Alina) and $r1$(Patrick). We have:
F55: $r3$(Patrick, Alina)
F9: $r1$(Patrick)
We can see that F55 matches the condition $r3$(Patrick, Alina), and F9 matches the condition $r1$(Patrick). Therefore, we can conclude that the statement $r7$(Patrick, Alina) is True based on 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
Deductive	standard	52.6	56.1
	removing irr	55.7	63.0
Inductive	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
Logic language	Zero-Shot	52.6	7.14	1.95
	Zero-Shot-CoT	56.1	7.14	3.57
	Few-Shot-CoT	53.7	-	13.3
Natural language	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 "$\forall x, y$: parentOf(x, y) $\rightarrow$ 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
Symbols	logic	52.6	56.1
	natural language	49.0	51.1
Semantics	logic	61.4	61.9
	natural language	69.3	64.3
Counter-CS	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_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	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
	ChatGPT	Zero-Shot	52.6	50.6	50.5	49.5	55.2	53.1	50.0	53.4	56.6	54.0	52.6
		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	100

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

Category	Model	Baseline	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	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	10.7	7.14	7.14	10.7	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

Statement: $icers$(Jonathan, David)
Answer: We can use logical rule L10: $\forall A, B, C, D : inely(B, A) \land inely(B, C) \land inely(C, D) \land atis(A) \rightarrow icers(A, D)$, which states that if A is an artist and B, C, and D are in a line with B in the middle, then A influences D.
We can use the following facts: - $inely$(Jonathan, Alina) (F55)
- $inely$(Alina, David) (F34)
From these facts, we can conclude that Jonathan, Alina, and David are in a line with Alina in the middle. Also, we know that Alina is an artist. Therefore, we can apply L10 to conclude that Alina influences David.
Since Alina influences David, we can use L4: $\forall A, B : inely(A, B) \land atis(A) \rightarrow lesai(A, B)$, which states that if A influences B and A is an artist, then B is a musician. Therefore, we can conclude that David is a musician.
Therefore, the statement $icers$(Jonathan, David) is False, since David is not an ice skater.

Statement: $indr$(Florian, Nina)
Answer: We can use logical rule L24: $\forall A, B : inely(B, A) \land atis(A) \rightarrow indr(A, B)$, which states 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. From the given facts, we know that $atis$(Florian) and $inely$(Emilia, Nina). Therefore, we can conclude that Nina is indirectly a fan of Florian, since Emilia is a fan of Florian and Nina is a fan of Emilia.
Thus, the answer is True.

O More reasoning results of Symbolic Tree

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

P Abductive reasoning on smaller datasets

We use smaller Symbolic Tree datasets to conduct the abductive reasoning experiment, which contains about 12 entities and 100 facts. The results are provided in Table 15. We compare *Symbols* and *Semantics* and find that the *Semantics* setting still outperforms the *Symbols* setting. This reinforces 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_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	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 $\forall x, y, z$: parentOf$(x, y) \wedge$ parentOf$(y, z) \wedge$ gender$(x,$ female$) \rightarrow$ GrandmotherOf(x, z) can be represented as $\forall x, y, z$: parentOf$(x, y) \wedge$ parentOf$(y, z) \rightarrow$ GrandparentOf(x, z), GrandparentOf$(x, z) \wedge$ gender$(x,$ female$) \rightarrow$ 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.