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

Anonymous Author(s)

Affiliation

Address

email

Abstract

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

19 1 Introduction

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

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

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

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.

144 the fact itself from in-context or knowledge within the parameters. However, the specific impact
145 of semantics on memorization has not been extensively explored. Thus, in addition to decoupling
146 semantics from reasoning, we also try to study the impact of semantics on memorization. Specifically,
147 we use a new dataset to fine-tune a language model and test its *time*, *efficiency* and *forgetting ratio*:
148 *time* is the fine-tuning time cost of adding/updating facts, *efficiency* is the filter MRR (the mean
149 reciprocal of rank of the correct entity [55]) of the facts added/updated, and *forgetting ratio* is the filter
150 MRR of the facts that should not be updated. When evaluating whether a fact has been successfully
151 added or updated, we query LLM with a question about the tail entity and rank the probability of the
152 true tail against all entities. The better LLM remembers a triplet, the higher the MRR gets.

153 3.2 Evaluation Datasets

154 Our goal is to decouple semantics from the in-context reasoning process and solely rely on the given
155 (new) knowledge to perform reasoning tasks. To implement this, we use Symbolic Tree [56] and
156 ProofWriter [16] datasets, which contain both relevant and irrelevant facts and LLMs need to infer
157 the unknown facts after selecting relevant facts from memory.

158 The Symbolic Tree dataset is an artificially close-world and noise-free symbolic dataset generated
159 with complex logical rules. The dataset consists of randomly sampled “*basic facts*”, which include
160 gender information and “parentOf” relations among individuals. With the given logical rules, the
161 dataset allows for reasoning about 28 different types of family relations, ranging from easy inferences
162 (e.g., fatherhood), to more elaborate ones (e.g., a daughter of someone’s cousin). *Facts* consist of
163 *basic facts* (in-context knowledge) and *inferred facts* (what to reason). Note that Symbolic Tree is a
164 close-world dataset, which means that any facts not presented in the dataset are assumed to be false.
165 Thus, we construct the false facts by replacing the head entity or tail entity with a random entity as
166 negative examples in *inferred facts*. Considering the context window size limitation, we restrict each
167 tree’s depth to 5 to generate the dataset. We experiment with 10 sampled Symbolic Trees; each has
168 30 kinds of relations (28 inferred relations, gender and parentOf relation), 26 entities, about 35 basic
169 facts, 300 inferred facts and 300 false ones.

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

178 ProofWriter [16] tasks provide artificial facts and rules expressed in natural language. For our
179 experiments, we use a subset of the ProofWriter Open World Assumption (OWA) dataset with a
180 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.

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

<p>Given a set of rules and facts, you have to reason whether a statement is true or false. Here are some facts and rules:</p> <p>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.</p> <p>Does it imply that the statement "The cow likes the squirrel." is True?</p>	<p>Given a set of rules and facts, you have to reason whether a statement is true or false. Here are some facts and rules:</p> <p>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.</p> <p>Does it imply that the statement "The e14 likes the e26." is True?</p>
---	---

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).

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 \pm 1.27	51.34 \pm 0.55	7.02 \pm 1.55	41.5 min
		48.91 \pm 4.3	40.74 \pm 2.26	2.2 \pm 0.99	41.5 min
Neo4j	Semantics	100	100	0	19s
		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 ambiguity 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 $\forall x, y : \text{parentOf}(x, y) \wedge \text{female}(x) \rightarrow \text{motherOf}(x, y)$, we obtain $\forall x, y : \text{brotherOf}(x, y) \wedge \text{male}(x) \rightarrow \text{sisterOf}(x, y)$. Thirdly, we use a subset of the ProofWriter OWA datasets for depths 0, 1, 2, 3 and 5, which contains synthetic facts and rules despite written in natural language but irrelevant to commonsense (see Figure 2). These investigations allow us to gain deeper insights into the effect of semantics on the reasoning capabilities of LLMs.

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

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

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

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)

410 knowledge to perform situated tasks remains an important challenge for LLMs. This includes
411 developing mechanisms to better encode and retrieve relevant information from the in-context
412 knowledge.

References

- [1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [2] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022.
- [3] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.
- [4] Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. Pretrained transformers as universal computation engines. *arXiv preprint arXiv:2103.05247*, 1, 2021.
- [5] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.
- [6] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022.
- [7] Sewon Min, Xinxu Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.
- [8] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*, 2021.
- [9] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR, 2021.
- [10] Timo Schick and Hinrich Schütze. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*, 2020.
- [11] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*, 2021.
- [12] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, 2022.
- [13] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.
- [14] Steven A Sloman. The empirical case for two systems of reasoning. *Psychological bulletin*, 119(1):3, 1996.
- [15] Thomas Gilovich, Dale Griffin, and Daniel Kahneman. *Heuristics and biases: The psychology of intuitive judgment*. Cambridge university press, 2002.
- [16] Oyvind Taffjord, Bhavana Dalvi Mishra, and Peter Clark. Proofwriter: Generating implications, proofs, and abductive statements over natural language. *arXiv preprint arXiv:2012.13048*, 2020.
- [17] Robin Cooper. The fracas consortium robin cooper, dick crouch, jan van eijck, chris fox, josef van genabith, jan jaspars, hans kamp, david milward, manfred pinkal, massimo poesio, steve pulman. 1996.
- [18] SamuelR. Bowman, Gabor Angeli, Christopher Potts, and ChristopherD. Manning. A large annotated corpus for learning natural language inference, Aug 2015.
- [19] Adina Williams, Nikita Nangia, and SamuelR. Bowman. A broad-coverage challenge corpus for sentence understanding through inference, Apr 2017.
- [20] Jean-Philippe Bernardy and Stergios Chatzikyriakidis. Improving the precision of natural textual entailment problem datasets., May 2020.
- [21] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge, Nov 2018.

- [22] Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. Explanations for commonsenseqa: New dataset and models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Jul 2021. doi: 10.18653/v1/2021.acl-long.238. URL <http://dx.doi.org/10.18653/v1/2021.acl-long.238>.
- [23] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs, Mar 2019.
- [24] Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and WilliamYang Wang. Hybridqa: A dataset of multi-hop question answering over tabular and textual data, Apr 2020.
- [25] 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, Deba-jyoti 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, Xinxin 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 Hohenacker 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] Jiu-hai Chen, Lichang Chen, Heng Huang, and Tianyi Zhou. When do you need chain-of-thought prompting for chatgpt? *arXiv preprint arXiv:2304.03262*, 2023.

Table of Contents

558	A Prompts	15
559	A.1 Deductive reasoning	15
560	A.1.1 Zero-Shot	15
561	A.1.2 Zero-Shot-CoT	15
562	A.1.3 Few-Shot-CoT	16
563	A.1.4 Examples of Few-Shot-CoT Demonstrations	16
564	A.1.5 Zero-Shot of removing rules setting	16
565	A.2 Inductive reasoning	17
566	A.2.1 Zero-Shot	17
567	A.2.2 Zero-Shot CoT	17
568	A.2.3 Zero-Shot of removing facts setting	17
569	A.3 Abductive reasoning	17
570	A.3.1 Zero-Shot	17
571	A.3.2 Zero-Shot-CoT	18
572	A.3.3 Few-Shot-CoT prompt	18
573	A.3.4 Examples of Few-Shot-CoT Demonstrations	18
574	A.3.5 Few-Shot CoT of removing rules setting	19
575	A.4 Commonsense reasoning	20
576	A.4.1 Zero-Shot prompt	20
577	A.4.2 Few-Shot prompt	20
578	A.4.3 Examples of Few-Shot Demonstrations	20
579	A.5 LLaMA Fine-tuning Prompt	20
580	B Deduction examples of Symbolic Tree datasets	20
581	B.1 Semantics	21
582	B.1.1 Logic language representations	21
583	B.1.2 Natural language representations	22
584	B.2 Symbolization	24
585	B.2.1 Logic language representations	24
586	B.2.2 Natural language representations:	26
587	B.3 Semantics of removing rule setting	27
588	C Examples of ProofWriter	28
589	C.1 Semantics	28
590	C.2 Symbols	29
591	D Different Zero-Shot prompting	30
592	E Comparison of memorization abilities of neural-based and symbolic-based methods	30

593	F Introduction of Neo4j	31
594	G Task definitions	31
595	H Implementation of memorization	32
596	I Consistency of knowledge base	32
597	J Different Few-Shot-CoT examples	32
598	J.1 Few-Shot-CoT examples with Auto-Filter	32
599	J.2 Few-Shot-CoT examples with Manually-Designed	33
600	K Reasoning results of after human selection	35
601	K.1 Comparison of Removing rules and Semantics	35
602	K.2 Reasoning results after human selection	35
603	L Reasoning with natural language	36
604	M Reasoning results of two representations	36
605	N Reasoning with garbled symbols	36
606	O More reasoning results of Symbolic Tree	38
607	P Abductive reasoning on smaller datasets	38
608	Q Replacing entity labels	39
609	R Multi-short rules	39
610	A Prompts	
611	A.1 Deductive reasoning	
612	A.1.1 Zero-Shot	
613	system: You are a helpful assistant with deductive reasoning abilities.	
614	user: I will provide a set of logical rules L1 to L{number of rules} and facts F1 to F{number	
615	of basic facts}. Please select one single logical rule from L1 to L{number of rules} and	
616	a few facts from F1 to F{number of basic facts} to predict True/False of the unknown fact	
617	using deductive reasoning.	
618	Logical rules: {logical rules}	
619	Facts: {basic facts}	
620	Unknown fact: {statement}	
621	The answer (True or False) is:	
622	A.1.2 Zero-Shot-CoT	
623	system: You are a helpful assistant with deductive reasoning abilities. Please select one	
624	single logical rule and a few facts to predict True/False of the following statement.	
625	user: I will provide a set of logical rules L1 to L{number of rules} and facts F1 to F{number	
626	of basic facts}. Please select one single logical rule from L1 to L{number of rules} and	
627	a few facts from F1 to F{number of basic facts} to predict True/False of the following	
628	statement using deductive reasoning.	
629	Logical rules: {logical rules}	

```

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

```

633 A.1.3 Few-Shot-CoT

```

634 system: You are a helpful assistant with deductive reasoning abilities. Please select one
635       single logical rule and a few facts to predict True/False of the following statement.
636 user: I will provide a set of logical rules L1 to L{number of rules} and facts F1 to F{number
637       of basic facts}.
638 Logical rules: {logical rules}
639 Facts: {basic facts}
640 Please select one single logical rule from L1 to L{number of rules} and a few facts from F1 to
641       F{number of basic facts} to predict True/False of the following statement using
642       deductive reasoning.
643 {Demonstrations}
644 Statement: {statement}
645 Answer:

```

646 A.1.4 Examples of Few-Shot-CoT Demonstrations

647 Statement: $r_{14}(\text{Amelie}, \text{Jonathan})$
648 Answer: We can use L11: $\forall A, B, C, D, E : r_3(B, A) \wedge r_3(B, C) \wedge r_3(C, D) \wedge r_3(D, E) \wedge r_2(A) \rightarrow$
649 $r_{14}(A, E)$ to make a deduction. From the facts, we know that $r_3(\text{Amelie}, \text{Thomas})$ and $r_3(\text{Thomas},$
650 $\text{Jonathan})$. Therefore, we can apply L11 with $A = \text{Amelie}$, $B = \text{Thomas}$, $C = \text{Jonathan}$, D and E as
651 variables. This gives us: $r_3(\text{Thomas}, \text{Jonathan}) \wedge r_3(\text{Amelie}, \text{Thomas}) \wedge r_3(\text{Thomas}, D) \wedge r_3(D, E)$
652 $\wedge r_2(\text{Amelie}) \rightarrow r_{14}(\text{Amelie}, E)$ Since we know that $r_2(\text{Amelie})$ is true from F7, we can simplify
653 the above statement to: $r_3(\text{Thomas}, \text{Jonathan}) \wedge r_3(\text{Amelie}, \text{Thomas}) \wedge r_3(\text{Thomas}, D) \wedge r_3(D, E)$
654 $\rightarrow r_{14}(\text{Amelie}, E)$ Now, we can substitute the facts $r_3(\text{Amelie}, \text{Thomas})$ and $r_3(\text{Thomas}, \text{Jonathan})$
655 to get: $r_3(\text{Thomas}, \text{Jonathan}) \wedge r_3(\text{Thomas}, D) \wedge r_3(D, E) \rightarrow r_{14}(\text{Amelie}, E)$ We can see that this
656 statement is true if we choose $D = \text{Alina}$ and $E = \text{Jonathan}$, since we know that $r_3(\text{Thomas}, \text{Alina})$
657 from F50 and $r_3(\text{Alina}, \text{Jonathan})$ from F56. Therefore, the statement $r_{14}(\text{Amelie}, \text{Jonathan})$ is true.

658 Statement: $r_{31}(\text{Philipp}, \text{Nina})$
659 Answer: Let's use L28: $\forall A, B, C, D : r_3(B, A) \wedge r_3(C, B) \wedge r_3(C, D) \wedge r_1(A) \rightarrow r_{31}(A, D)$ to
660 see if we can prove the statement $r_{31}(\text{Philipp}, \text{Nina})$. From the facts, we know that $r_3(\text{Luisa}, \text{Nina})$
661 (F52) and $r_3(\text{Emilia}, \text{Philipp})$ (F62). We can use L28 with $A = \text{Philipp}$, $B = \text{Emilia}$, $C = \text{Luisa}$,
662 and $D = \text{Nina}$ to get: $r_3(\text{Emilia}, \text{Philipp}) \wedge r_3(\text{Luisa}, \text{Emilia}) \wedge r_3(\text{Luisa}, \text{Nina}) \wedge r_1(\text{Philipp}) \rightarrow$
663 $r_{31}(\text{Philipp}, \text{Nina})$ Since all the conditions are met, we can conclude that $r_{31}(\text{Philipp}, \text{Nina})$ is True.
664 Therefore, the answer is True.

665 Statement: $r_8(\text{Leonie}, \text{Nico})$
666 Answer: We can use L5 to make a deduction about the statement $r_8(\text{Leonie}, \text{Nico})$. L5 states that if
667 there exist three individuals A , B , and C such that A is related to B , B is related to C , and A has a
668 certain property (r_2), then A is related to C in a certain way (r_8). Using this rule, we can see that we
669 have the following facts: $- r_3(\text{Leonie}, \text{Emily})$
670 $- r_3(\text{Emily}, \text{Nico})$
671 Therefore, we can conclude that $r_8(\text{Leonie}, \text{Nico})$ is true.

672 Statement: $r_7(\text{Patrick}, \text{Alina})$
673 Answer: We can use L4: $\forall A, B : r_3(A, B) \wedge r_1(A) \rightarrow r_7(A, B)$ to determine if $r_7(\text{Patrick}, \text{Alina})$
674 is true or false. From the facts, we know that $r_3(\text{Patrick}, \text{Alina})$ is true (F55). We also know that
675 $r_1(\text{Patrick})$ is true (F9). Therefore, we can apply L4 to conclude that $r_7(\text{Patrick}, \text{Alina})$ is true.
676 Answer: True.

677 Statement: $r_{27}(\text{Jonathan}, \text{Leonie})$
678 Answer: Let's use L4: $\forall A, B : r_3(A, B) \wedge r_1(A) \rightarrow r_7(A, B)$ and F56: $r_3(\text{Patrick}, \text{Jonathan})$ to infer
679 that $r_7(\text{Patrick}, \text{Jonathan})$ is true. Then, we can use L24: $\forall A, B, C : r_3(B, A) \wedge r_3(C, B) \wedge r_1(A) \rightarrow$
680 $r_{27}(A, C)$ and F44: $r_3(\text{Leonie}, \text{Emily})$ to infer that $r_{27}(\text{Jonathan}, \text{Leonie})$ is false, since there is no
681 fact that supports $r_3(\text{Jonathan}, \text{Emily})$. Therefore, the answer is False.

682 A.1.5 Zero-Shot of removing rules setting

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

689 A.2 Inductive reasoning

690 A.2.1 Zero-Shot

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

704 A.2.2 Zero-Shot CoT

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

718 A.2.3 Zero-Shot of removing facts setting

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

728 A.3 Abductive reasoning

729 A.3.1 Zero-Shot

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

739 A.3.2 Zero-Shot-CoT

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

750 A.3.3 Few-Shot-CoT prompt

751 **system:** You are a helpful assistant with abductive reasoning abilities. Please select one
 752 single logical rule and a few facts to explain the following statement.
 753 **user:**
 754 I will provide a set of logical rules L1 to L{number of rules} and facts F1 to F{number of
 755 basic facts}. Please select one single logical rule from L1 to L{number of rules} and
 756 a few facts from F1 to F{number of basic facts} to explain the following statement.
 757 Rules: {logical rules}
 758 Facts: {basic facts}
 759 {Demonstrations}
 760 Statement: {statement}
 761 Answer:

762 A.3.4 Examples of Few-Shot-CoT Demonstrations

763 Statement: $r4(\text{Laura}, \text{Fabian})$
 764 Answer: To explain the statement $r4(\text{Laura}, \text{Fabian})$, we can use the rule L3 and fact F1 and F37.
 765 L3: $\forall A, B : r1(A, B) \wedge r44(A) \rightarrow r4(A, B)$
 766 F1: $r1(\text{Laura}, \text{Fabian})$
 767 F37: $r44(\text{Laura})$
 768 According to L3, if there is a relationship $r1$ between A and B, and A has the property $r44$, then the
 769 relationship $r4$ holds between A and B. In this case, we have a relationship $r1$ between Laura and
 770 Fabian (F1) and Laura has the property $r44$ (F37). Therefore, the relationship $r4$ holds between Laura
 771 and Fabian, which is the given statement.
 772 The selected rule and facts are: L3, F1, and F37.
 773 Statement: $r5(\text{Stefan}, \text{Emily})$
 774 Answer: To explain the statement $r5(\text{Stefan}, \text{Emily})$, we can use the rule L4 and fact F17. L4:
 775 $\forall A, B : r1(A, B) \wedge r43(A) \rightarrow r5(A, B)$
 776 F17: $r1(\text{Stefan}, \text{Emily})$
 777 According to rule L4, if there is a relation $r1(A, B)$ and A has the property $r43$, then the relation
 778 $r5(A, B)$ holds. In our case, we have the fact F17, which states that there is a relation $r1(\text{Stefan},$
 779 $\text{Emily})$. Additionally, Stefan has the property $r43$ (F61). Therefore, we can conclude that the relation
 780 $r5(\text{Stefan}, \text{Emily})$ holds.
 781 The selected rule and facts are: L4, F17, and F61.
 782 Statement: $r23(\text{Felix}, \text{Elias})$
 783 Answer: To explain the statement $r23(\text{Felix}, \text{Elias})$, we can use the rule L22 and fact F5. L22:
 784 $\forall A, B : r1(B, A) \wedge r43(A) \rightarrow r23(A, B)$
 785 F5: $r1(\text{Elias}, \text{Felix})$
 786 According to rule L22, if there exists a relationship $r1(B, A)$ and A has the property $r43$, then the
 787 relationship $r23(A, B)$ holds. In our case, we have the fact F5, which states that there is a relationship
 788 $r1(\text{Elias}, \text{Felix})$. Additionally, we know that Felix has the property $r43$ (F59). Therefore, applying
 789 rule L22, we can conclude that the relationship $r23(\text{Felix}, \text{Elias})$ holds.
 790 The selected rule and facts are: L22, F5, and F59.
 791 Statement: $r3(\text{Felix}, \text{Fabian})$
 792 Answer: To explain the statement $r3(\text{Felix}, \text{Fabian})$, we can use the rule L2 and facts F2, F1, and F59.
 793 L2: $\forall A, B, C : r1(B, A) \wedge r1(B, C) \wedge r43(A) \rightarrow r3(A, C)$
 794 F1: $r1(\text{Laura}, \text{Fabian})$

795 F2: $r1(\text{Laura}, \text{Felix})$
796 F59: $r43(\text{Felix})$
797 By substituting the values into L2, we get:
798 $r1(\text{Laura}, \text{Felix}) \wedge r1(\text{Laura}, \text{Fabian}) \wedge r43(\text{Felix}) \rightarrow r3(\text{Felix}, \text{Fabian})$
799 Since all the premises are true according to the given facts, the conclusion $r3(\text{Felix}, \text{Fabian})$ is also
800 true.
801 The selected rule and facts are: L2, F1, F2, and F59.

802 Statement: $r7(\text{Fabian}, \text{Helga})$
803 Answer: To explain the statement $r7(\text{Fabian}, \text{Helga})$, we can use the rule L6 and facts F13, F21, and
804 F39.
805 L6: $\forall A, B, C : r1(A, B) \wedge r1(B, C) \wedge r43(A) \rightarrow r7(A, C)$
806 F13: $r1(\text{Fabian}, \text{Thomas})$
807 F21: $r1(\text{Thomas}, \text{Helga})$
808 F39: $r43(\text{Fabian})$
809 According to L6, if there exist A, B, and C such that $r1(A, B)$, $r1(B, C)$, and $r43(A)$, then $r7(A, C)$
810 holds. In our case, A = Fabian, B = Thomas, and C = Helga. Since we have $r1(\text{Fabian}, \text{Thomas})$
811 from F13, $r1(\text{Thomas}, \text{Helga})$ from F21, and $r43(\text{Fabian})$ from F39, we can conclude that $r7(\text{Fabian},$
812 $\text{Helga})$ is true.
813 So, the selected rule and facts are: L6, F13, F21, and F39.

814 A.3.5 Few-Shot CoT of removing rules setting

815 **system:** You are a helpful assistant with abductive reasoning abilities. Please select a few
816 facts to explain the following statement.
817 **user:** I will provide a set of facts F1 to F63. Please select a few facts from F1 to F63 to
818 explain the following statement.
819 Facts: {facts}
820
821 Statement: Laura is mother of Felix.
822 Answer: To explain the statement "Laura is mother of Felix", we can use Facts:
823 Fact F2 states: Laura is parent of Felix.
824 Fact F37 states: Laura is female.
825 Using F2 and F37, we can conclude that "Laura is mother of Felix" holds.
826 Therefore, the selected rule and facts are F2, F37.
827
828 Statement: Samuel is brother of Alina.
829 Answer: To infer the statement "Samuel is brother of Alina", we have:
830 F27: Patrick is parent of Samuel.
831 F28: Patrick is parent of Alina.
832 F47: Samuel is male.
833 Based on these facts, we can infer "Samuel is brother of Alina":
834 Therefore, the selected rule and facts are F27, F28, F47.
835
836 Statement: Patrick is grandfather of David.
837 Answer: To explain the statement "Patrick is grandfather of David", we have:
838 F28: Patrick is parent of Alina.
839 F7: Alina is parent of David.
840 F45: Patrick is male.
841 Based on these facts, we can infer "Patrick is grandfather of David":
842 Therefore, the selected rule and facts are F28, F7, F45.
843
844 Statement: Amelie is daughter of Elena.
845 Answer: To explain the statement "Amelie is daughter of Elena", we have:
846 F20: Elena is parent of Amelie.
847 F43: Amelie is female.
848 Based on these facts, we can infer "Amelie is daughter of Elena".
849 Therefore, the selected rule and facts are F20, F43.
850
851 Statement: Claudia is sister of Felix
852 Answer: To prove the statement "Claudia is sister of Felix", we can use facts:
853 F3: Laura is parent of Claudia.
854 F2: Laura is parent of Felix.
855 F40: Claudia is female.
856 Based on these facts, we can infer "Claudia is sister of Felix".
857 Therefore, the selected rule and facts are F3, F2, F40.
858
859 Statement: Laura is mother of Fabian.

860 Answer:

861 A.4 Commonsense reasoning

862 A.4.1 Zero-Shot prompt

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

867 A.4.2 Few-Shot prompt

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

873 A.4.3 Examples of Few-Shot Demonstrations

874 **Standard:**

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

878 **Predict-explain:**

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

887 **Expain-predict:**

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

896 A.5 LLaMA Fine-tuning Prompt

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

903 B Deduction examples of Symbolic Tree datasets

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

907 B.1 Semantics

908 B.1.1 Logic language representations

```
909 Logical rules:
910 L1:  $\forall A,B,C: \text{parentOf}(B, A) \wedge \text{parentOf}(B, C) \wedge \text{female}(A) \rightarrow \text{sisterOf}(A,$ 
911  $C)$ 
912 L2:  $\forall A,B,C: \text{parentOf}(B, A) \wedge \text{parentOf}(B, C) \wedge \text{male}(A) \rightarrow \text{brotherOf}(A,$ 
913  $C)$ 
914 L3:  $\forall A,B: \text{parentOf}(A, B) \wedge \text{female}(A) \rightarrow \text{motherOf}(A,B)$ 
915 L4:  $\forall A,B: \text{parentOf}(A, B) \wedge \text{male}(A) \rightarrow \text{fatherOf}(A,B)$ 
916 L5:  $\forall A,B,C: \text{parentOf}(A, B) \wedge \text{parentOf}(B, C) \wedge \text{female}(A) \rightarrow$ 
917  $\text{grandmotherOf}(A,C)$ 
918 L6:  $\forall A,B,C: \text{parentOf}(A, B) \wedge \text{parentOf}(B, C) \wedge \text{male}(A) \rightarrow$ 
919  $\text{grandfatherOf}(A,C)$ 
920 L7:  $\forall A,B,C,D: \text{parentOf}(A, B) \wedge \text{parentOf}(B, C) \wedge \text{parentOf}(C, D) \wedge \text{female}(A)$ 
921  $\rightarrow \text{greatGrandmotherOf}(A,D)$ 
922 L8:  $\forall A,B,C,D: \text{parentOf}(A, B) \wedge \text{parentOf}(B, C) \wedge \text{parentOf}(C, D) \wedge \text{male}(A) \rightarrow$ 
923  $\text{greatGrandfatherOf}(A,D)$ 
924 L9:  $\forall A,B,C,D: \text{parentOf}(B, A) \wedge \text{parentOf}(B, C) \wedge \text{parentOf}(C, D) \wedge \text{female}(A)$ 
925  $\rightarrow \text{auntOf}(A,D)$ 
926 L10:  $\forall A,B,C,D: \text{parentOf}(B, A) \wedge \text{parentOf}(B, C) \wedge \text{parentOf}(C, D) \wedge \text{male}(A)$ 
927  $\rightarrow \text{uncleOf}(A,D)$ 
928 L11:  $\forall A,B,C,D,E: \text{parentOf}(B, A) \wedge \text{parentOf}(B, C) \wedge \text{parentOf}(C, D) \wedge$ 
929  $\text{parentOf}(D, E) \wedge \text{female}(A) \rightarrow \text{greatAuntOf}(A,E)$ 
930 L12:  $\forall A,B,C,D,E: \text{parentOf}(B, A) \wedge \text{parentOf}(B, C) \wedge \text{parentOf}(C, D) \wedge$ 
931  $\text{parentOf}(D, E) \wedge \text{male}(A) \rightarrow \text{greatUncleOf}(A,E)$ 
932 L13:  $\forall A,B,C,D,E,F: \text{parentOf}(B, A) \wedge \text{parentOf}(C, B) \wedge \text{parentOf}(C, D) \wedge$ 
933  $\text{parentOf}(D, E) \wedge \text{parentOf}(E, F) \wedge \text{female}(A) \rightarrow \text{secondAuntOf}(A,F)$ 
934 L14:  $\forall A,B,C,D,E,F: \text{parentOf}(B, A) \wedge \text{parentOf}(C, B) \wedge \text{parentOf}(C, D) \wedge$ 
935  $\text{parentOf}(D, E) \wedge \text{parentOf}(E, F) \wedge \text{male}(A) \rightarrow \text{secondUncleOf}(A,F)$ 
936 L15:  $\forall A,B,C,D,E: \text{parentOf}(B, A) \wedge \text{parentOf}(C, B) \wedge \text{parentOf}(C, D) \wedge$ 
937  $\text{parentOf}(D, E) \wedge \text{female}(A) \rightarrow \text{girlCousinOf}(A,E)$ 
938 L16:  $\forall A,B,C,D,E: \text{parentOf}(B, A) \wedge \text{parentOf}(C, B) \wedge \text{parentOf}(C, D) \wedge$ 
939  $\text{parentOf}(D, E) \wedge \text{male}(A) \rightarrow \text{boyCousinOf}(A,E)$ 
940 L17:  $\forall A,B,C,D,E,F,G: \text{parentOf}(B, A) \wedge \text{parentOf}(C, B) \wedge \text{parentOf}(D, C) \wedge$ 
941  $\text{parentOf}(D, E) \wedge \text{parentOf}(E, F) \wedge \text{parentOf}(F, G) \wedge \text{female}(A) \rightarrow$ 
942  $\text{girlSecondCousinOf}(A,G)$ 
943 L18:  $\forall A,B,C,D,E,F,G: \text{parentOf}(B, A) \wedge \text{parentOf}(C, B) \wedge \text{parentOf}(D, C) \wedge$ 
944  $\text{parentOf}(D, E) \wedge \text{parentOf}(E, F) \wedge \text{parentOf}(F, G) \wedge \text{male}(A) \rightarrow$ 
945  $\text{boySecondCousinOf}(A,G)$ 
946 L19:  $\forall A,B,C,D,E,F: \text{parentOf}(B, A) \wedge \text{parentOf}(C, B) \wedge \text{parentOf}(D, C) \wedge$ 
947  $\text{parentOf}(D, E) \wedge \text{parentOf}(E, F) \wedge \text{female}(A) \rightarrow$ 
948  $\text{girlFirstCousinOnceRemovedOf}(A,F)$ 
949 L20:  $\forall A,B,C,D,E,F: \text{parentOf}(B, A) \wedge \text{parentOf}(C, B) \wedge \text{parentOf}(D, C) \wedge$ 
950  $\text{parentOf}(D, E) \wedge \text{parentOf}(E, F) \wedge \text{male}(A) \rightarrow \text{boyFirstCousinOnceRemovedOf}$ 
951  $(A,F)$ 
952 L21:  $\forall A,B: \text{parentOf}(B, A) \wedge \text{female}(A) \rightarrow \text{daughterOf}(A,B)$ 
953 L22:  $\forall A,B: \text{parentOf}(B, A) \wedge \text{male}(A) \rightarrow \text{sonOf}(A,B)$ 
954 L23:  $\forall A,B,C: \text{parentOf}(B, A) \wedge \text{parentOf}(C, B) \wedge \text{female}(A) \rightarrow$ 
955  $\text{granddaughterOf}(A,C)$ 
956 L24:  $\forall A,B,C: \text{parentOf}(B, A) \wedge \text{parentOf}(C, B) \wedge \text{male}(A) \rightarrow \text{grandsonOf}($ 
957  $A,C)$ 
958 L25:  $\forall A,B,C,D: \text{parentOf}(B, A) \wedge \text{parentOf}(C, B) \wedge \text{parentOf}(D, C) \wedge \text{female}(A)$ 
959  $\rightarrow \text{greatGranddaughterOf}(A,D)$ 
960 L26:  $\forall A,B,C,D: \text{parentOf}(B, A) \wedge \text{parentOf}(C, B) \wedge \text{parentOf}(D, C) \wedge \text{male}(A)$ 
961  $\rightarrow \text{greatGrandsonOf}(A,D)$ 
962 L27:  $\forall A,B,C,D: \text{parentOf}(B, A) \wedge \text{parentOf}(C, B) \wedge \text{parentOf}(C, D) \wedge \text{female}(A)$ 
963  $\rightarrow \text{nieceOf}(A,D)$ 
964 L28:  $\forall A,B,C,D: \text{parentOf}(B, A) \wedge \text{parentOf}(C, B) \wedge \text{parentOf}(C, D) \wedge \text{male}(A)$ 
965  $\rightarrow \text{nephewOf}(A,D)$ 
966
967 Facts:
968 F1: female(Laura)
969 F2: male(Elias)
970 F3: male(Fabian)
971 F4: female(Claudia)
972 F5: female(Elena)
973 F6: male(Thomas)
974 F7: female(Amelie)
975 F8: female(Luisa)
976 F9: male(Patrick)
```

```

977 F10: female(Emilia)
978 F11: male(Samuel)
979 F12: female(Alina)
980 F13: male(Jonathan)
981 F14: male(Philipp)
982 F15: male(Nico)
983 F16: male(David)
984 F17: female(Emily)
985 F18: male(Konstantin)
986 F19: male(Florian)
987 F20: female(Helga)
988 F21: female(Nina)
989 F22: female(Lea)
990 F23: male(Felix)
991 F24: female(Leonie)
992 F25: male(Stefan)
993 F26: male(Gabriel)
994 F27: male(Tobias)
995 F28: parentOf(Laura, Fabian)
996 F29: parentOf(Laura, Felix)
997 F30: parentOf(Laura, Claudia)
998 F31: parentOf(Elias, Fabian)
999 F32: parentOf(Elias, Felix)
1000 F33: parentOf(Elias, Claudia)
1001 F34: parentOf(Alina, David)
1002 F35: parentOf(Alina, Lea)
1003 F36: parentOf(Nico, David)
1004 F37: parentOf(Nico, Lea)
1005 F38: parentOf(Emily, Nico)
1006 F39: parentOf(Konstantin, Nico)
1007 F40: parentOf(Fabian, Thomas)
1008 F41: parentOf(Fabian, Amelie)
1009 F42: parentOf(Nina, Tobias)
1010 F43: parentOf(Leonie, Emily)
1011 F44: parentOf(Stefan, Emily)
1012 F45: parentOf(Gabriel, Tobias)
1013 F46: parentOf(Elena, Thomas)
1014 F47: parentOf(Elena, Amelie)
1015 F48: parentOf(Thomas, Helga)
1016 F49: parentOf(Thomas, Nina)
1017 F50: parentOf(Thomas, Patrick)
1018 F51: parentOf(Luisa, Helga)
1019 F52: parentOf(Luisa, Nina)
1020 F53: parentOf(Luisa, Patrick)
1021 F54: parentOf(Patrick, Samuel)
1022 F55: parentOf(Patrick, Alina)
1023 F56: parentOf(Patrick, Jonathan)
1024 F57: parentOf(Patrick, Philipp)
1025 F58: parentOf(Patrick, Florian)
1026 F59: parentOf(Emilia, Samuel)
1027 F60: parentOf(Emilia, Alina)
1028 F61: parentOf(Emilia, Jonathan)
1029 F62: parentOf(Emilia, Philipp)
1030 F63: parentOf(Emilia, Florian)
1031
1032 Unknown fact: boyCousinOf(Tobias, David)

```

1033 B.1.2 Natural language representations

```

1034 Logical rules:
1035 L1: If B is parent of A and B is parent of C and A is female, then A is sister of D.
1036 L2: If B is parent of A and B is parent of C and A is male, then A is brother of D.
1037 L3: If A is parent of B and A is female, then A is mother of C.
1038 L4: If A is parent of B and A is male, then A is father of C.
1039 L5: If A is parent of B and B is parent of C and A is female, then A is grandmother of D.
1040 L6: If A is parent of B and B is parent of C and A is male, then A is grandfather of D.
1041 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
1042     greatGrandmother of E.
1043 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
1044     greatGrandfather of E.

```

1045 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
 1046 aunt of E.
 1047 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
 1048 uncle of E.
 1049 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
 1050 is female, then A is greatAunt of F.
 1051 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
 1052 is male, then A is greatUncle of F.
 1053 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
 1054 is parent of F and A is female, then A is secondAunt of G.
 1055 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
 1056 is parent of F and A is male, then A is secondUncle of G.
 1057 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
 1058 is female, then A is girlCousin of F.
 1059 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
 1060 is male, then A is boyCousin of F.
 1061 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
 1062 is parent of F and F is parent of G and A is female, then A is girlSecondCousin of H.
 1063 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
 1064 is parent of F and F is parent of G and A is male, then A is boySecondCousin of H.
 1065 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
 1066 is parent of F and A is female, then A is girlFirstCousinOnceRemoved of G.
 1067 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
 1068 is parent of F and A is male, then A is boyFirstCousinOnceRemoved of G.
 1069 L21: If B is parent of A and A is female, then A is daughter of C.
 1070 L22: If B is parent of A and A is male, then A is son of C.
 1071 L23: If B is parent of A and C is parent of B and A is female, then A is granddaughter of D.
 1072 L24: If B is parent of A and C is parent of B and A is male, then A is grandson of D.
 1073 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
 1074 greatGranddaughter of E.
 1075 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
 1076 greatGrandson of E.
 1077 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
 1078 niece of E.
 1079 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
 1080 nephew of E.
 1081
 1082 Facts:
 1083 F1: Laura is female.
 1084 F2: Elias is male.
 1085 F3: Fabian is male.
 1086 F4: Claudia is female.
 1087 F5: Elena is female.
 1088 F6: Thomas is male.
 1089 F7: Amelie is female.
 1090 F8: Luisa is female.
 1091 F9: Patrick is male.
 1092 F10: Emilia is female.
 1093 F11: Samuel is male.
 1094 F12: Alina is female.
 1095 F13: Jonathan is male.
 1096 F14: Philipp is male.
 1097 F15: Nico is male.
 1098 F16: David is male.
 1099 F17: Emily is female.
 1100 F18: Konstantin is male.
 1101 F19: Florian is male.
 1102 F20: Helga is female.
 1103 F21: Nina is female.
 1104 F22: Lea is female.
 1105 F23: Felix is male.
 1106 F24: Leonie is female.
 1107 F25: Stefan is male.
 1108 F26: Gabriel is male.
 1109 F27: Tobias is male.
 1110 F28: Laura is parent of Fabian.
 1111 F29: Laura is parent of Felix.
 1112 F30: Laura is parent of Claudia.
 1113 F31: Elias is parent of Fabian.
 1114 F32: Elias is parent of Felix.
 1115 F33: Elias is parent of Claudia.
 1116 F34: Alina is parent of David.

1117 F35: Alina is parent of Lea.
 1118 F36: Nico is parent of David.
 1119 F37: Nico is parent of Lea.
 1120 F38: Emily is parent of Nico.
 1121 F39: Konstantin is parent of Nico.
 1122 F40: Fabian is parent of Thomas.
 1123 F41: Fabian is parent of Amelie.
 1124 F42: Nina is parent of Tobias.
 1125 F43: Leonie is parent of Emily.
 1126 F44: Stefan is parent of Emily.
 1127 F45: Gabriel is parent of Tobias.
 1128 F46: Elena is parent of Thomas.
 1129 F47: Elena is parent of Amelie.
 1130 F48: Thomas is parent of Helga.
 1131 F49: Thomas is parent of Nina.
 1132 F50: Thomas is parent of Patrick.
 1133 F51: Luisa is parent of Helga.
 1134 F52: Luisa is parent of Nina.
 1135 F53: Luisa is parent of Patrick.
 1136 F54: Patrick is parent of Samuel.
 1137 F55: Patrick is parent of Alina.
 1138 F56: Patrick is parent of Jonathan.
 1139 F57: Patrick is parent of Philipp.
 1140 F58: Patrick is parent of Florian.
 1141 F59: Emilia is parent of Samuel.
 1142 F60: Emilia is parent of Alina.
 1143 F61: Emilia is parent of Jonathan.
 1144 F62: Emilia is parent of Philipp.
 1145 F63: Emilia is parent of Florian.
 1146
 1147 Unknown fact: Gabriel is uncle of Lea.

1148 B.2 Symbolization

1149 B.2.1 Logic language representations

1150 Logical rules:
 1151 L1: $\forall A, B, C: r3(B, A) \wedge r3(B, C) \wedge r2(A) \rightarrow r4(A, C)$
 1152 L2: $\forall A, B, C: r3(B, A) \wedge r3(B, C) \wedge r1(A) \rightarrow r5(A, C)$
 1153 L3: $\forall A, B: r3(A, B) \wedge r2(A) \rightarrow r6(A, B)$
 1154 L4: $\forall A, B: r3(A, B) \wedge r1(A) \rightarrow r7(A, B)$
 1155 L5: $\forall A, B, C: r3(A, B) \wedge r3(B, C) \wedge r2(A) \rightarrow r8(A, C)$
 1156 L6: $\forall A, B, C: r3(A, B) \wedge r3(B, C) \wedge r1(A) \rightarrow r9(A, C)$
 1157 L7: $\forall A, B, C, D: r3(A, B) \wedge r3(B, C) \wedge r3(C, D) \wedge r2(A) \rightarrow r10(A, D)$
 1158 \$
 1159 L8: $\forall A, B, C, D: r3(A, B) \wedge r3(B, C) \wedge r3(C, D) \wedge r1(A) \rightarrow r11(A, D)$
 1160 \$
 1161 L9: $\forall A, B, C, D: r3(B, A) \wedge r3(B, C) \wedge r3(C, D) \wedge r2(A) \rightarrow r12(A, D)$
 1162 \$
 1163 L10: $\forall A, B, C, D: r3(B, A) \wedge r3(B, C) \wedge r3(C, D) \wedge r1(A) \rightarrow r13(A, D)$
 1164)\$
 1165 L11: $\forall A, B, C, D, E: r3(B, A) \wedge r3(B, C) \wedge r3(C, D) \wedge r3(D, E) \wedge r2(A) \rightarrow r14(A, E)$
 1166 \$
 1167 L12: $\forall A, B, C, D, E: r3(B, A) \wedge r3(B, C) \wedge r3(C, D) \wedge r3(D, E) \wedge r1(A) \rightarrow r15(A, E)$
 1168 \$
 1169 L13: $\forall A, B, C, D, E, F: r3(B, A) \wedge r3(C, B) \wedge r3(C, D) \wedge r3(D, E) \wedge r3(E, F) \wedge r2(A) \rightarrow r16(A, F)$
 1170 \$
 1171 L14: $\forall A, B, C, D, E, F: r3(B, A) \wedge r3(C, B) \wedge r3(C, D) \wedge r3(D, E) \wedge r3(E, F) \wedge r1(A) \rightarrow r17(A, F)$
 1172 \$
 1173 L15: $\forall A, B, C, D, E: r3(B, A) \wedge r3(C, B) \wedge r3(C, D) \wedge r3(D, E) \wedge r2(A) \rightarrow r18(A, E)$
 1174 \$
 1175 L16: $\forall A, B, C, D, E: r3(B, A) \wedge r3(C, B) \wedge r3(C, D) \wedge r3(D, E) \wedge r1(A) \rightarrow r19(A, E)$
 1176 \$
 1177 L17: $\forall A, B, C, D, E, F, G: r3(B, A) \wedge r3(C, B) \wedge r3(D, C) \wedge r3(D, E) \wedge r3(E, F) \wedge r3(F, G) \wedge r2(A) \rightarrow r20(A, G)$
 1178 \$
 1179 L18: $\forall A, B, C, D, E, F, G: r3(B, A) \wedge r3(C, B) \wedge r3(D, C) \wedge r3(D, E) \wedge r3(E, F) \wedge r3(F, G) \wedge r1(A) \rightarrow r21(A, G)$
 1180 \$
 1181 L19: $\forall A, B, C, D, E, F: r3(B, A) \wedge r3(C, B) \wedge r3(D, C) \wedge r3(D, E) \wedge r3(E, F) \wedge r2(A) \rightarrow r22(A, F)$
 1182 \$
 1183 L20: $\forall A, B, C, D, E, F: r3(B, A) \wedge r3(C, B) \wedge r3(D, C) \wedge r3(D, E) \wedge r3(E, F) \wedge r1(A) \rightarrow r23(A, F)$
 1184 \$

```

1185 L21:  $\forall A, B: r3(B, A) \wedge r2(A) \rightarrow r24(A, B)$ 
1186 L22:  $\forall A, B: r3(B, A) \wedge r1(A) \rightarrow r25(A, B)$ 
1187 L23:  $\forall A, B, C: r3(B, A) \wedge r3(C, B) \wedge r2(A) \rightarrow r26(A, C)$ 
1188 L24:  $\forall A, B, C: r3(B, A) \wedge r3(C, B) \wedge r1(A) \rightarrow r27(A, C)$ 
1189 L25:  $\forall A, B, C, D: r3(B, A) \wedge r3(C, B) \wedge r3(D, C) \wedge r2(A) \rightarrow r28(A, D)$ 
1190  $\wedge r1(A) \rightarrow r29(A, D)$ 
1191 L26:  $\forall A, B, C, D: r3(B, A) \wedge r3(C, B) \wedge r3(D, C) \wedge r1(A) \rightarrow r29(A, D)$ 
1192  $\wedge r1(A) \rightarrow r30(A, D)$ 
1193 L27:  $\forall A, B, C, D: r3(B, A) \wedge r3(C, B) \wedge r3(C, D) \wedge r2(A) \rightarrow r30(A, D)$ 
1194  $\wedge r1(A) \rightarrow r31(A, D)$ 
1195 L28:  $\forall A, B, C, D: r3(B, A) \wedge r3(C, B) \wedge r3(C, D) \wedge r1(A) \rightarrow r31(A, D)$ 
1196  $\wedge r1(A) \rightarrow r31(A, D)$ 
1197
1198 Facts:
1199 F1:  $r2(Laura)$ 
1200 F2:  $r1(Elias)$ 
1201 F3:  $r1(Fabian)$ 
1202 F4:  $r2(Claudia)$ 
1203 F5:  $r2(Elena)$ 
1204 F6:  $r1(Thomas)$ 
1205 F7:  $r2(Amelie)$ 
1206 F8:  $r2(Luisa)$ 
1207 F9:  $r1(Patrick)$ 
1208 F10:  $r2(Emilia)$ 
1209 F11:  $r1(Samuel)$ 
1210 F12:  $r2(Alina)$ 
1211 F13:  $r1(Jonathan)$ 
1212 F14:  $r1(Philipp)$ 
1213 F15:  $r1(Nico)$ 
1214 F16:  $r1(David)$ 
1215 F17:  $r2(Emily)$ 
1216 F18:  $r1(Konstantin)$ 
1217 F19:  $r1(Florian)$ 
1218 F20:  $r2(Helga)$ 
1219 F21:  $r2(Nina)$ 
1220 F22:  $r2(Lea)$ 
1221 F23:  $r1(Felix)$ 
1222 F24:  $r2(Leonie)$ 
1223 F25:  $r1(Stefan)$ 
1224 F26:  $r1(Gabriel)$ 
1225 F27:  $r1(Tobias)$ 
1226 F28:  $r3(Laura, Fabian)$ 
1227 F29:  $r3(Laura, Felix)$ 
1228 F30:  $r3(Laura, Claudia)$ 
1229 F31:  $r3(Elias, Fabian)$ 
1230 F32:  $r3(Elias, Felix)$ 
1231 F33:  $r3(Elias, Claudia)$ 
1232 F34:  $r3(Alina, David)$ 
1233 F35:  $r3(Alina, Lea)$ 
1234 F36:  $r3(Nico, David)$ 
1235 F37:  $r3(Nico, Lea)$ 
1236 F38:  $r3(Emily, Nico)$ 
1237 F39:  $r3(Konstantin, Nico)$ 
1238 F40:  $r3(Fabian, Thomas)$ 
1239 F41:  $r3(Fabian, Amelie)$ 
1240 F42:  $r3(Nina, Tobias)$ 
1241 F43:  $r3(Leonie, Emily)$ 
1242 F44:  $r3(Stefan, Emily)$ 
1243 F45:  $r3(Gabriel, Tobias)$ 
1244 F46:  $r3(Elena, Thomas)$ 
1245 F47:  $r3(Elena, Amelie)$ 
1246 F48:  $r3(Thomas, Helga)$ 
1247 F49:  $r3(Thomas, Nina)$ 
1248 F50:  $r3(Thomas, Patrick)$ 
1249 F51:  $r3(Luisa, Helga)$ 
1250 F52:  $r3(Luisa, Nina)$ 
1251 F53:  $r3(Luisa, Patrick)$ 
1252 F54:  $r3(Patrick, Samuel)$ 
1253 F55:  $r3(Patrick, Alina)$ 
1254 F56:  $r3(Patrick, Jonathan)$ 
1255 F57:  $r3(Patrick, Philipp)$ 
1256 F58:  $r3(Patrick, Florian)$ 

```

1257 F59: \$r3\$(Emilia, Samuel)
 1258 F60: \$r3\$(Emilia, Alina)
 1259 F61: \$r3\$(Emilia, Jonathan)
 1260 F62: \$r3\$(Emilia, Philipp)
 1261 F63: \$r3\$(Emilia, Florian)
 1262
 1263 Unknown fact: \$r9\$(Thomas, Claudia)

1264 B.2.2 Natural language representations:

1265 Logical rules:
 1266 L1: If B is \$r3\$ of A and B is \$r3\$ of C and A is \$r2\$, then A is \$r4\$ of D.
 1267 L2: If B is \$r3\$ of A and B is \$r3\$ of C and A is \$r1\$, then A is \$r5\$ of D.
 1268 L3: If A is \$r3\$ of B and A is \$r2\$, then A is \$r6\$ of C.
 1269 L4: If A is \$r3\$ of B and A is \$r1\$, then A is \$r7\$ of C.
 1270 L5: If A is \$r3\$ of B and B is \$r3\$ of C and A is \$r2\$, then A is \$r8\$ of D.
 1271 L6: If A is \$r3\$ of B and B is \$r3\$ of C and A is \$r1\$, then A is \$r9\$ of D.
 1272 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
 1273 E.
 1274 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
 1275 E.
 1276 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
 1277 E.
 1278 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
 1279 E.
 1280 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\$,
 1281 then A is \$r14\$ of F.
 1282 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\$,
 1283 then A is \$r15\$ of F.
 1284 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\$
 1285 of F and A is \$r2\$, then A is \$r16\$ of G.
 1286 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\$
 1287 of F and A is \$r1\$, then A is \$r17\$ of G.
 1288 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\$,
 1289 then A is \$r18\$ of F.
 1290 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\$,
 1291 then A is \$r19\$ of F.
 1292 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\$
 1293 of F and F is \$r3\$ of G and A is \$r2\$, then A is \$r20\$ of H.
 1294 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\$
 1295 of F and F is \$r3\$ of G and A is \$r1\$, then A is \$r21\$ of H.
 1296 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\$
 1297 of F and A is \$r2\$, then A is \$r22\$ of G.
 1298 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\$
 1299 of F and A is \$r1\$, then A is \$r23\$ of G.
 1300 L21: If B is \$r3\$ of A and A is \$r2\$, then A is \$r24\$ of C.
 1301 L22: If B is \$r3\$ of A and A is \$r1\$, then A is \$r25\$ of C.
 1302 L23: If B is \$r3\$ of A and C is \$r3\$ of B and A is \$r2\$, then A is \$r26\$ of D.
 1303 L24: If B is \$r3\$ of A and C is \$r3\$ of B and A is \$r1\$, then A is \$r27\$ of D.
 1304 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
 1305 E.
 1306 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
 1307 E.
 1308 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
 1309 E.
 1310 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
 1311 E.
 1312
 1313 Facts:
 1314 F1: Laura is \$r2\$.
 1315 F2: Elias is \$r1\$.
 1316 F3: Fabian is \$r1\$.
 1317 F4: Claudia is \$r2\$.
 1318 F5: Elena is \$r2\$.
 1319 F6: Thomas is \$r1\$.
 1320 F7: Amelie is \$r2\$.
 1321 F8: Luisa is \$r2\$.
 1322 F9: Patrick is \$r1\$.
 1323 F10: Emilia is \$r2\$.
 1324 F11: Samuel is \$r1\$.
 1325 F12: Alina is \$r2\$.

1326 F13: Jonathan is \$r1\$.
 1327 F14: Philipp is \$r1\$.
 1328 F15: Nico is \$r1\$.
 1329 F16: David is \$r1\$.
 1330 F17: Emily is \$r2\$.
 1331 F18: Konstantin is \$r1\$.
 1332 F19: Florian is \$r1\$.
 1333 F20: Helga is \$r2\$.
 1334 F21: Nina is \$r2\$.
 1335 F22: Lea is \$r2\$.
 1336 F23: Felix is \$r1\$.
 1337 F24: Leonie is \$r2\$.
 1338 F25: Stefan is \$r1\$.
 1339 F26: Gabriel is \$r1\$.
 1340 F27: Tobias is \$r1\$.
 1341 F28: Laura is \$r3\$ of Fabian.
 1342 F29: Laura is \$r3\$ of Felix.
 1343 F30: Laura is \$r3\$ of Claudia.
 1344 F31: Elias is \$r3\$ of Fabian.
 1345 F32: Elias is \$r3\$ of Felix.
 1346 F33: Elias is \$r3\$ of Claudia.
 1347 F34: Alina is \$r3\$ of David.
 1348 F35: Alina is \$r3\$ of Lea.
 1349 F36: Nico is \$r3\$ of David.
 1350 F37: Nico is \$r3\$ of Lea.
 1351 F38: Emily is \$r3\$ of Nico.
 1352 F39: Konstantin is \$r3\$ of Nico.
 1353 F40: Fabian is \$r3\$ of Thomas.
 1354 F41: Fabian is \$r3\$ of Amelie.
 1355 F42: Nina is \$r3\$ of Tobias.
 1356 F43: Leonie is \$r3\$ of Emily.
 1357 F44: Stefan is \$r3\$ of Emily.
 1358 F45: Gabriel is \$r3\$ of Tobias.
 1359 F46: Elena is \$r3\$ of Thomas.
 1360 F47: Elena is \$r3\$ of Amelie.
 1361 F48: Thomas is \$r3\$ of Helga.
 1362 F49: Thomas is \$r3\$ of Nina.
 1363 F50: Thomas is \$r3\$ of Patrick.
 1364 F51: Luisa is \$r3\$ of Helga.
 1365 F52: Luisa is \$r3\$ of Nina.
 1366 F53: Luisa is \$r3\$ of Patrick.
 1367 F54: Patrick is \$r3\$ of Samuel.
 1368 F55: Patrick is \$r3\$ of Alina.
 1369 F56: Patrick is \$r3\$ of Jonathan.
 1370 F57: Patrick is \$r3\$ of Philipp.
 1371 F58: Patrick is \$r3\$ of Florian.
 1372 F59: Emilia is \$r3\$ of Samuel.
 1373 F60: Emilia is \$r3\$ of Alina.
 1374 F61: Emilia is \$r3\$ of Jonathan.
 1375 F62: Emilia is \$r3\$ of Philipp.
 1376 F63: Emilia is \$r3\$ of Florian.
 1377
 1378 Unknown fact: Nico is \$r27\$ of Stefan.

1379 B.3 Semantics of removing rule setting

1380 I will provide a set of facts. Please predict True/False of the unknown fact based on given
 1381 facts.
 1382 Facts:
 1383 F1: Laura is female.
 1384 F2: Elias is male.
 1385 F3: Fabian is male.
 1386 F4: Claudia is female.
 1387 F5: Elena is female.
 1388 F6: Thomas is male.
 1389 F7: Amelie is female.
 1390 F8: Luisa is female.
 1391 F9: Patrick is male.
 1392 F10: Emilia is female.
 1393 F11: Samuel is male.
 1394 F12: Alina is female.

1395 F13: Jonathan is male.
 1396 F14: Philipp is male.
 1397 F15: Nico is male.
 1398 F16: David is male.
 1399 F17: Emily is female.
 1400 F18: Konstantin is male.
 1401 F19: Florian is male.
 1402 F20: Helga is female.
 1403 F21: Nina is female.
 1404 F22: Lea is female.
 1405 F23: Felix is male.
 1406 F24: Leonie is female.
 1407 F25: Stefan is male.
 1408 F26: Gabriel is male.
 1409 F27: Tobias is male.
 1410 F28: Laura is parent of Fabian.
 1411 F29: Laura is parent of Felix.
 1412 F30: Laura is parent of Claudia.
 1413 F31: Elias is parent of Fabian.
 1414 F32: Elias is parent of Felix.
 1415 F33: Elias is parent of Claudia.
 1416 F34: Alina is parent of David.
 1417 F35: Alina is parent of Lea.
 1418 F36: Nico is parent of David.
 1419 F37: Nico is parent of Lea.
 1420 F38: Emily is parent of Nico.
 1421 F39: Konstantin is parent of Nico.
 1422 F40: Fabian is parent of Thomas.
 1423 F41: Fabian is parent of Amelie.
 1424 F42: Nina is parent of Tobias.
 1425 F43: Leonie is parent of Emily.
 1426 F44: Stefan is parent of Emily.
 1427 F45: Gabriel is parent of Tobias.
 1428 F46: Elena is parent of Thomas.
 1429 F47: Elena is parent of Amelie.
 1430 F48: Thomas is parent of Helga.
 1431 F49: Thomas is parent of Nina.
 1432 F50: Thomas is parent of Patrick.
 1433 F51: Luisa is parent of Helga.
 1434 F52: Luisa is parent of Nina.
 1435 F53: Luisa is parent of Patrick.
 1436 F54: Patrick is parent of Samuel.
 1437 F55: Patrick is parent of Alina.
 1438 F56: Patrick is parent of Jonathan.
 1439 F57: Patrick is parent of Philipp.
 1440 F58: Patrick is parent of Florian.
 1441 F59: Emilia is parent of Samuel.
 1442 F60: Emilia is parent of Alina.
 1443 F61: Emilia is parent of Jonathan.
 1444 F62: Emilia is parent of Philipp.
 1445 F63: Emilia is parent of Florian.
 1446
 1447 Unknown fact: Jonathan is aunt of Thomas.
 1448 The answer (True or False) is:

1449 C Examples of ProofWriter

1450 In this section, we provide examples of deduction experiments conducted on the ProofWriter Depth-1
 1451 dataset. We present examples for both the *Semantics* and *Symbols* settings.

1452 C.1 Semantics

1453 The bear likes the dog.
 1454 The cow is round.
 1455 The cow likes the bear.
 1456 The cow needs the bear.
 1457 The dog needs the squirrel.
 1458 The dog sees the cow.
 1459 The squirrel needs the dog.

```

1460 If someone is round then they like the squirrel.
1461 If the bear is round and the bear likes the squirrel then the squirrel needs the bear.
1462 If the cow needs the dog then the cow is cold.
1463 Does it imply that the statement "The cow likes the squirrel." is True?

```

```

1464 The bear likes the dog.
1465 The cow is round.
1466 The cow likes the bear.
1467 The cow needs the bear.
1468 The dog needs the squirrel.
1469 The dog sees the cow.
1470 The squirrel needs the dog.
1471 If someone is round then they like the squirrel.
1472 If the bear is round and the bear likes the squirrel then the squirrel needs the bear.
1473 If the cow needs the dog then the cow is cold.
1474 Does it imply that the statement "The cow does not like the squirrel." is True?

```

```

1475 Bob is blue.
1476 Erin is quiet.
1477 Fiona is cold.
1478 Harry is cold.
1479 All quiet things are blue.
1480 If Harry is blue then Harry is not young.
1481 Blue things are young.
1482 Blue, round things are cold.
1483 If something is blue and not red then it is round.
1484 If something is young then it is white.
1485 If Erin is red and Erin is not round then Erin is young.
1486 If Erin is red and Erin is not cold then Erin is white.
1487 Does it imply that the statement "Erin is white" is True?
1488 Answer with only True or False. The answer is:

```

```

1489 The bear likes the dog.
1490 The cow is round.
1491 The cow likes the bear.
1492 The cow needs the bear.
1493 The dog needs the squirrel.
1494 The dog sees the cow.
1495 The squirrel needs the dog.
1496 If someone is round then they like the squirrel.
1497 If the bear is round and the bear likes the squirrel then the squirrel needs the bear.
1498 If the cow needs the dog then the cow is cold.
1499 Does it imply that the statement "The cow likes the squirrel." is True?

```

1500 C.2 Symbols

```

1501 The e4 likes the e5.
1502 The e14 is e2.
1503 The e14 likes the e4.
1504 The e14 needs the e4.
1505 The e5 needs the e26.
1506 The e5 sees the e14.
1507 The e26 needs the e5.
1508 If someone is e2 then they like the e26.
1509 If the e4 is e2 and the e4 likes the e26 then the e26 needs the e4.
1510 If the e14 needs the e5 then the e14 is e1.
1511 Does it imply that the statement "The e14 likes the e26." is True?

```

```

1512 The e27 is e7.
1513 The e27 is e15.
1514 The e30 does not chase the e27.
1515 The e30 eats the e27.
1516 The e30 is e1.
1517 The e30 is e15.
1518 The e30 visits the e27.
1519 If something visits the e27 then the e27 does not visit the e30.
1520 If something is e1 and e15 then it visits the e30.
1521 Does it imply that the statement "The e30 visits the e30." is True?

```

1522 The e27 is e7.
 1523 The e27 is e15.
 1524 The e30 does not chase the e27.
 1525 The e30 eats the e27.
 1526 The e30 is e1.
 1527 The e30 is e15.
 1528 The e30 visits the e27.
 1529 If something visits the e27 then the e27 does not visit the e30.
 1530 If something is e1 and e15 then it visits the e30.
 1531 Does it imply that the statement "The e30 visits the e30." is True?

1532 D Different Zero-Shot prompting

1533 We try different Zero-Shot prompts:

1534 (1)

1535 I will provide a set of logical rules L1 to L{number of rules} and facts F1 to F{number of
 1536 basic facts}. Please select one single logical rule from L1 to L{number of rules} and a
 1537 few facts from F1 to F{number of basic facts} to predict True/False of the unknown fact
 1538 using deductive reasoning.
 1539 Logical rules: {rules}
 1540 Facts: {basic facts}
 1541 Unknown fact: {unknown fact}
 1542 The answer (True or False) is:

1543 (2)

1544 I will provide a set of logical rules L1 to L{number of rules} and facts F1 to F{number of
 1545 basic facts}. Please predict True/False of the unknown fact using deductive reasoning.
 1546 Logical rules: {rules}
 1547 Facts: {basic facts}
 1548 Unknown fact: {unknown fact}
 1549 The answer (True or False) is:

1550 (3)

1551 Given a set of rules and facts, you have to reason whether a statement is True or False.
 1552 Here are some rules: {rules}
 1553 Here are some facts: {basic facts}
 1554 Does it imply that the statement "{unknown fact}" is True?
 1555 The answer (YES or NO) is:

1556 The results of the three prompts in the Zero-Shot setting are presented in Table 5. Among the three
 1557 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

1558 E Comparison of memorization abilities of neural-based and symbolic-based methods

1560 We compare fine-tuned language models with the deterministic graph DB Neo4J to explore the
 1561 memorization abilities of neural-based and symbolic-based methods. Language models can implicitly
 1562 store and retrieve facts as "knowledge bases" within their neural parameters. They are trained on a
 1563 snapshot of data and may not have access to the latest or most accurate information. In order to update
 1564 or add facts, specific model parameters need to be modified, or the model needs to be fine-tuned
 1565 with new data. In contrast, symbolic knowledge graphs can directly add or update individual triplets,
 1566 making it easier to incorporate new information. Our comparison affirms the huge advantage of
 1567 using KGs/external DBs to update knowledge rather than finetuning, aligning with the recent trend of
 1568 retrieval-based LLM.

F Introduction of Neo4j

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

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

G Task definitions

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

- *deductive reasoning*: we use *hypothesis classification*, i.e., predict the *correctness* of the *hypothesis* given the *theory* where *theory* consists of basic facts and logical rules, *correctness* can be true or false, and *hypothesis* is a predicted fact, which is one of the inferred facts or negative samples. The accuracy is the proportion of correct predictions.
- *inductive reasoning*: we perform the *rule generation* task. Given multiple facts with similar patterns and a rule template, the goal is to induce a rule that entails these facts. Specifically, for each relation r , we use basic facts and those inferred facts that contain only relation r as provided facts. The induced rule is generated after filling in the rule template. We test the generated rules against the ground truth rules. If the generated rule matches the ground truth rule exactly, we predict the rule to be correct; otherwise, we predict the rule to be incorrect. The precision is the proportion of correct predictions. Note that considering logical rules maybe not all chain rules (e.g., $r_1(y, x) \wedge r_2(y, z) \rightarrow r_3(x, z)$), we add inverse relation for each relation in order to transform them into chain rules and simplify the rule template (e.g., $r_1^{-1}(x, y) \wedge r_2(y, z) \rightarrow r_3(x, z)$). Furthermore, we provide a rule template for each relation. Take *auntOf* as example, its rule template can be $\forall x, y, z : \#(x, y) \wedge \#(y, z) \wedge ++(x) \rightarrow \text{auntOf}(x, z)$ or "If x is $\#$ of y and y is $\#$ of z and x is $++$, then x is aunt of z .", where $\#$ can be *parent* or *inverse_parent*, $++$ can be *female* or *male*.
Besides, a single rule can be equivalent to multiple rules. For example, the rule $\forall x, y, z : \text{parentOf}(x, y) \wedge \text{parentOf}(y, z) \wedge \text{gender}(x, \text{female}) \rightarrow \text{GrandmotherOf}(x, z)$ can be represented as $\forall x, y, z : \text{parentOf}(x, y) \wedge \text{parentOf}(y, z) \rightarrow \text{GrandparentOf}(x, z), \text{GrandparentOf}(x, z) \wedge \text{gender}(x, \text{female}) \rightarrow \text{GrandmotherOf}(x, z)$. We conduct the experiments with both rule representations and find single-longer rules perform better than multiple-short rules. Results are presented in Appendix R. Based on these observations and considering the simplicity of induction evaluation, we rewrite all logical rules by including only the *parentOf* and *gender* relations in the rule body. This also ensures that each inferred relation is implied by a single logical rule, referred to as *grounding truth rule*.
- *abductive reasoning*: We use *explanation generation* to evaluate abductive reasoning abilities. Given a *theory* including basic facts and all logical rules, the task is to select specific facts and a logical rule to explain the *observation*. The *observation* is chosen from inferred facts. We use Proof Accuracy (PA) as an evaluation metric, i.e., the fraction of examples where the generated proof matches exactly any of the gold proofs.
- *memorization*: We use a subset of Symbolic Trees to fine-tune the language model. For the symbolic setting, we use r_1, r_2, r_3 to replace the original relations in the semantic setting. Note that the new dataset does not overlap with the old knowledge base of LLMs, ensuring no disambiguation problem and the influence of pre-existing knowledge. When memorizing, we use *time*, *efficiency* and *forgetting* as metrics: *time* is the cost time of adding/updating facts, *efficiency* is the MRR (mean reciprocal rank [55]) of facts added/updated, and *forgetting* is the MRR of

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.

1627 H Implementation of memorization

1628 We selected 1258 triplets from 4 Symbolic Trees to evaluate the effectiveness of adding knowledge.
 1629 Following the prompting of Taori et al. [58], we use the head entity and relation as instructions and
 1630 provide all candidate tails as input. The model’s training objective is to autoregress toward the true
 1631 tail entities. The detailed prompting is contained in Appendix A.5. In the updating step, we fine-tune
 1632 the model on all 620 triplets from the first two trees whose tail entities are randomly flipped to false
 1633 ones. Besides the effectiveness of updating, we evaluated the forgetting ratio using the remaining
 1634 638 triplets of the least two trees. These triplets have been remembered in the first step and haven’t
 1635 been updated in the second. Noting that, within each tree, the relationships between entities are
 1636 independent, and the entities are distinct. Therefore, we propose that LLM should retain its memory
 1637 of the previously remembered triplets when updating based on the first two trees. We utilized 4 A100
 1638 80G GPUs with batch size 64 for finetuning. The training process involved 100 epochs, employing a
 1639 cosine learning rate schedule with an initial learning rate of $2e-5$. We run these experiments three
 1640 times and recorded their mean and standard MRR.

1641 I Consistency of knowledge base

1642 In the context of updating a knowledge base, it is important to ensure the consistency of the knowledge
 1643 base. When revising a fact, it is necessary to update other related facts accordingly to maintain a
 1644 coherent and accurate knowledge base. For example, if we have the facts: Alice is Bob’s mother,
 1645 Amy is Alice’s sister, Momo is Lily’s sister, and Amy is Bob’s aunt, and we update Bob’s mother as
 1646 Lily, it follows that Bob’s aunt should be updated to Momo instead of Amy to maintain consistency.
 1647 However, in our experiments, we did not consider the situation of updating related facts in order to
 1648 simplify the evaluation process, although it may lead to inconsistencies within the knowledge base.
 1649 In real-world scenarios, ensuring consistency is an essential aspect of maintaining the accuracy and
 1650 reliability of a knowledge base when updating or revising facts.

1651 J Different Few-Shot-CoT examples

1652 We use three different kinds of demonstrations generation:

- 1653 (1) leverage LLMs with the “Let’s think step by step” prompt to generate reasoning chains for
 1654 demonstrations.
- 1655 (2) manually designed each demonstration with a question and a reasoning chain. A reasoning chain
 1656 is composed of a rationale (a series of intermediate reasoning steps) and an expected answer.
- 1657 (3) leverage LLMs with the “Let’s think step by step” prompt to generate reasoning chains and filter
 1658 with those generated chains coming with mistakes for demonstrations.

1659 The results of these approaches are provided in Table 6. From the results, we observe that the
 1660 *Auto-Randomly* approach outperforms both the *Auto-Filter* and *Manually-Designed* approaches. This
 1661 suggests that randomly generating reasoning chains is more suitable for LLMs and can generate a
 1662 diverse range of solutions, resulting in better reasoning performance.

1663 J.1 Few-Shot-CoT examples with Auto-Filter

1664 Statement: $r8(\text{Elena}, \text{Nina})$
 1665 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
 1666 deduce whether the statement $r8(\text{Elena}, \text{Nina})$ is true or false. From the facts, we know that $r3(\text{Elena},$
 1667 $\text{Thomas})$ and $r3(\text{Thomas}, \text{Nina})$. Therefore, we can use L5 with $A = \text{Elena}$, $B = \text{Thomas}$, and
 1668 $C = \text{Nina}$ to see if $r8(\text{Elena}, \text{Nina})$ is true or false. L5 states that if $r3(A, B)$ and $r3(B, C)$ and $r2(A)$

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

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

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 $r3(\text{Elena}, \text{Thomas})$ (F46) and $r3(\text{Thomas}, \text{Amelie})$ (F47). Therefore, we can use L26 with $A = \text{Philipp}$, $B = \text{Thomas}$, $C = \text{Amelie}$, and $D = \text{Elena}$ to get:

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

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

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

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

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

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

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 $r6(\text{Luisa}, \text{Helga})$.

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

Thus, the answer is True.

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

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 $r14(\text{Amelie}, \text{Alina})$ if we can find facts that satisfy the antecedent of the rule.

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

Therefore, the answer is True.

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

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 $r26(\text{Alina}, \text{Luisa})$.

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

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

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

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

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

Therefore, the answer is True.

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

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

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 $A = \text{Amelie}$ and $E = \text{Jonathan}$ to get: $r3(B, \text{Amelie}) \wedge r3(B, C) \wedge r3(C, D) \wedge r3(D, \text{Jonathan}) \wedge r2(\text{Amelie}) \rightarrow r14(\text{Amelie}, \text{Jonathan})$. Now, we need to find all facts that match the conditions $r3(B, \text{Amelie})$, $r3(B, C)$, $r3(C, D)$, $r3(D, \text{Jonathan})$ and $r2(\text{Amelie})$. We have:

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

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

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

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

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

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

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

	Auto-Randomly	Auto-Filter	Manually-Designed
KG ₁	55.6	53.6	40.7

1724 values to B, C, and D:
1725 (1) B = Fabian, D = Patrick.
1726 (2) B = Elena, D = Emilia.
1727 (3) B = Fabian, D = Emilia.
1728 (4) B = Elena, D = Patrick.
1729 Using the values of B and D from the possible assignments above, we can find the value of C that
1730 satisfies the conditions $r3(B, C)$ and $r3(C, D)$. For assignment (1), we have:
1731 F40: $r3(\text{Fabian}, \text{Thomas})$
1732 F50: $r3(\text{Thomas}, \text{Patrick})$
1733 Since all the conditions are satisfied, we can conclude that the statement $r14(\text{Amelie}, \text{Jonathan})$ is
1734 True if we assign B = Fabian and C = Thomas and D = Patrick. Thus, the output is True.

1735 Statement: $r31(\text{Philipp}, \text{Nina})$
1736 Answer: We can use logical rule L28: $\forall A, B, C, D : r3(B, A) \wedge r3(C, B) \wedge r3(C, D) \wedge r1(A) \rightarrow$
1737 $r31(A, D)$. We can substitute A = Philipp and D = Nina to get: $r3(B, \text{Philipp}) \wedge r3(C, B) \wedge$
1738 $r3(C, \text{Nina}) \wedge r1(\text{Philipp}) \rightarrow r31(\text{Philipp}, \text{Nina})$. Now, we need to find all facts that match the
1739 conditions $r3(B, \text{Philipp})$, $r3(C, B)$, $r3(C, \text{Nina})$, and $r1(\text{Philipp})$. We have:
1740 F57: $r3(\text{Patrick}, \text{Philipp})$
1741 F62: $r3(\text{Emilia}, \text{Philipp})$
1742 F49: $r3(\text{Thomas}, \text{Nina})$
1743 F52: $r3(\text{Luisa}, \text{Nina})$
1744 We can see that F57 and F62 match the condition $r3(B, \text{Philipp})$ with B being Patrick and Emilia,
1745 respectively. F49 and F52 match the condition $r3(C, \text{Nina})$ with C being Thomas and Luisa, respec-
1746 tively. Therefore, there is only one single way of assigning values to B and C.
1747 B = Patrick and C = Thomas. We can substitute B and C to get: $r3(\text{Patrick}, \text{Philipp}) \wedge$
1748 $r3(\text{Thomas}, \text{Patrick}) \wedge r3(\text{Thomas}, \text{Nina}) \wedge r1(\text{Philipp}) \rightarrow r31(\text{Philipp}, \text{Nina})$. Since all
1749 the conditions are satisfied, we can conclude that the statement $r31(\text{Philipp}, \text{Nina})$ is True if we
1750 assign B = Patrick and C = Thomas. Thus, the output is True.

1751 Statement: $r8(\text{Leonie}, \text{Nico})$
1752 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
1753 can substitute A = Leonie and C = Nico to get: $r3(\text{Leonie}, B) \wedge r3(B, \text{Nico}) \wedge r2(\text{Leonie}) \rightarrow$
1754 $r8(\text{Leonie}, \text{Nico})$. Now, we need to find all facts that match the conditions $r3(\text{Leonie}, B)$, $r3(B,$
1755 $\text{Nico})$, and $r2(\text{Leonie})$. We have:
1756 F43: $r3(\text{Leonie}, \text{Emily})$
1757 F38: $r3(\text{Emily}, \text{Nico})$
1758 F39: $r3(\text{Konstantin}, \text{Nico})$
1759 F24: $r2(\text{Leonie})$
1760 We can see that F43 matches the condition $r3(\text{Leonie}, B)$ with B being Emily. and F38 matches the
1761 condition $r3(B, \text{Nico})$ with B being Emily. Therefore, we can assign B = Emily. Also, F24 matches
1762 the condition $r2(\text{Leonie})$. Therefore, we can conclude that the statement $r8(\text{Leonie}, \text{Nico})$ is True if
1763 we assign B = Emily. Therefore, the output is True.

1764 Statement: $r7(\text{Patrick}, \text{Alina})$
1765 Answer: We can use logical rule L4: $\forall A, B : r3(A, B) \wedge r1(A) \rightarrow r7(A, B)$. We can substitute A
1766 = Patrick and B = Alina to get: $r3(\text{Patrick}, \text{Alina}) \wedge r1(\text{Patrick}) \rightarrow r7(\text{Patrick}, \text{Alina})$. Now,
1767 we need to find all facts that match the conditions $r3(\text{Patrick}, \text{Alina})$ and $r1(\text{Patrick})$. We have:
1768 F55: $r3(\text{Patrick}, \text{Alina})$
1769 F9: $r1(\text{Patrick})$
1770 We can see that F55 matches the condition $r3(\text{Patrick}, \text{Alina})$, and F9 matches the condition
1771 $r1(\text{Patrick})$. Therefore, we can conclude that the statement $r7(\text{Patrick}, \text{Alina})$ is True based on
1772 L4. Thus, the output is True.

1773 K Reasoning results of after human selection

1774 In this section, we perform all experiments after human selection, providing the LLMs with only the
1775 relevant facts related to the predicted fact.

1776 For example, we have facts:

1777 F1: Lea is female.
1778 F2: Philipp is male.
1779 F3: Leonie is female.
1780 F4: Patrick is parent of Alina.
1781 F5: Patrick is parent of Philipp.
1782 F6: Nico is parent of David.
1783 F7: Nico is parent of Lea.
1784 F8: Laura is parent of Fabian.
1785 F9: Laura is parent of Felix.

1786 If we predict the unknown fact "Philipp is the brother of Alina", we filter out the irrelevant facts and
1787 retain only the related facts:

1788 F1: Philipp is male.
1789 F2: Patrick is parent of Alina.
1790 F3: Patrick is parent of Philipp.

1791 K.1 Comparison of Removing rules and Semantics

1792 To explore the ability of LLMs to utilize internal and external knowledge, we conduct an additional
1793 experiment where we provided LLMs with only the relevant facts related to the predicted fact.
1794 We compared the performance of *Removing rules* (leveraging internal knowledge) with *Semantics*
1795 (providing external logical rules). Surprisingly, from Table 7, we find that *Removing rules* performed
1796 better than *Semantics*. This suggests that LLMs possess the necessary internal knowledge to support
1797 answering questions and reasoning tasks, and leveraging internal knowledge is more effective for
1798 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

1799 LLMs’ output examples are provided below:

1800 **Semantics:** According to the given logical rule, for A to be a granddaughter of D, B must be the
1801 parent of A and C must be the parent of B. However, there is no direct connection between Emilia
1802 and Lea in the given facts. Although Emilia is the parent of Alina, who is the parent of Lea, Emilia is
1803 not the parent of Lea’s parent. Therefore, we cannot conclude that Lea is Emilia’s granddaughter
1804 based on the given facts and logical rule.

1805 **Removing rules:** Based on the given facts, Emilia is the parent of Alina, who is the parent of Lea.
1806 Therefore, Emilia is the grandmother of Lea, making Lea Emilia’s granddaughter.

1807 From the example, we can observe that when relying on external logical rules, LLMs need to strictly
1808 adhere to the reasoning process, which can be more challenging for LLMs to predict unknown
1809 answers compared to utilizing the commonsense knowledge already contained within LLMs. This
1810 suggests that leveraging the internal knowledge of LLMs can be more effective for reasoning tasks.

1811 K.2 Reasoning results after human selection

1812 We conduct deductive and inductive reasoning experiments to examine the performance of LLMs
1813 when only provided with the relevant facts related to the predicted fact. The results are presented in
1814 Table 8. They demonstrate that after selecting useful information, LLMs perform reasoning tasks
1815 more effectively. This finding suggests that LLMs face challenges when processing excessively long
1816 in-context information. Selecting relevant facts helps to reduce the memorization load on LLMs
1817 and enables them to focus on the most relevant information for reasoning, leading to improved
1818 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 “ $\forall x, y : \text{parentOf}(x, y) \rightarrow \text{childOf}(y, x)$ ” can be represented as “If x is parent of y, then y is parent of x.”. Through numerous trials, we find that for the *Symbols* or *Counter-CS* setting, LLMs tend to perform better when using logic language representations. Conversely, for the *Semantics* setting, LLMs tend to perform better when using natural language text. The results are presented in Table 10. These observations suggest that natural language representations better stimulate the semantic understanding capabilities of LLMs, while logical language representations are more conducive to symbolic reasoning.

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

		Zero-Shot	Zero-Shot-CoT
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
		Zero-Shot	52.6	50.6	50.5	49.5	55.2	53.1	50.0	53.4	56.6	54.0	52.6
	ChatGPT	Zero-Shot-CoT	56.1	57.0	55.4	57.0	54.5	56.1	55.5	56.9	50.0	58.0	55.7
		Few-Shot-CoT	53.7	56.9	55.2	54.4	55.1	52.0	54.0	55.8	56.8	54.5	54.8
		Zero-Plus-Few-Shot-CoT	53.7	53.6	55.4	51.4	54.0	50.9	54.0	54.2	58.4	54.5	54.0
Semantics	ChatGPT	Zero-Shot	70.0	64.8	70.4	65.8	61.4	63.8	65.8	67.4	63.0	68.9	66.1
		Zero-Shot-CoT	66.7	64.8	64.6	64.1	64.4	67.2	66.5	66.7	64.6	65.4	65.5
		Few-Shot-CoT	71.8	70.4	63.9	69.2	66.7	59.3	68.7	68.3	67.9	64.4	67.1
		Zero-Plus-Few-Shot-CoT	71.3	67.8	66.6	69.5	65.7	60.9	68.4	68.3	66.5	66.8	67.2
		Logic-based	-	100	100	100	100	100	100	100	100	100	100

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

Category	Model	Baseline	S 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	7.14	7.14	10.7	7.14	7.14	7.14	7.14	9.28
		Zero-Shot-CoT	21.4	7.14	17.9	7.14	3.57	7.14	7.14	7.14	7.14	3.57	8.93
Semantics	ChatGPT	Zero-Shot	25.0	32.1	39.3	39.3	42.9	39.3	35.7	32.1	35.7	42.9	36.4
		Zero-Shot-CoT	25.0	28.6	35.7	28.6	35.7	35.7	28.6	35.7	39.3	28.6	32.2
	GPT-4	Zero-Shot	53.6	53.6	50.0	53.6	50.0	53.6	50.0	57.1	53.6	50.0	52.5
		Zero-Shot-CoT	53.6	57.1	53.6	53.6	57.1	53.6	50.0	53.6	57.1	50.0	53.9
Rule-based		-	64.3	60.7	60.7	46.4	67.9	50.0	64.3	57.1	53.6	46.4	57.1

1884 Statement: *icers*(Jonathan, David)
1885 Answer: We can use logical rule L10: $\forall A, B, C, D : inely(B, A) \wedge inely(B, C) \wedge inely(C, D) \wedge$
1886 $atis(A) \rightarrow icers(A, D)$, which states that if A is an artist and B, C, and D are in a line with B in the
1887 middle, then A influences D.
1888 We can use the following facts: - *inely*(Jonathan, Alina) (F55)
1889 - *inely*(Alina, David) (F34)
1890 From these facts, we can conclude that Jonathan, Alina, and David are in a line with Alina in the
1891 middle. Also, we know that Alina is an artist. Therefore, we can apply L10 to conclude that Alina
1892 influences David.
1893 Since Alina influences David, we can use L4: $\forall A, B : inely(A, B) \wedge atis(A) \rightarrow lesai(A, B)$, which
1894 states that if A influences B and A is an artist, then B is a musician. Therefore, we can conclude that
1895 David is a musician.
1896 Therefore, the statement *icers*(Jonathan, David) is False, since David is not an ice skater.
1897 Statement: *indr*(Florian, Nina)
1898 Answer: We can use logical rule L24: $\forall A, B : inely(B, A) \wedge atis(A) \rightarrow indr(A, B)$, which states
1899 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.
1900 From the given facts, we know that *atis*(Florian) and *inely*(Emilia, Nina). Therefore, we can
1901 conclude that Nina is indirectly a fan of Florian, since Emilia is a fan of Florian and Nina is a fan of
1902 Emilia.
1903 Thus, the answer is True.

1904 O More reasoning results of Symbolic Tree

1905 We experiment with 10 sampled trees and report the average results in the main body. In this section,
1906 we provide the reasoning results of each sampled Symbolic Tree, presented in Table Tables 12 to 14.

1907 P Abductive reasoning on smaller datasets

1908 We use smaller Symbolic Tree datasets to conduct the abductive reasoning experiment, which contains
1909 about 12 entities and 100 facts. The results are provided in Table 15. We compare *Symbols* and
1910 *Semantics* and find that the *Semantics* setting still outperforms the *Symbols* setting. This reinforces
1911 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	

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

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

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

Q Replacing entity labels

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

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

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

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

R Multi-short rules

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