Do Language Models Follow Occam's Razor? An Evaluation of Parsimony in Inductive and Abductive Reasoning

Anonymous ACL submission

Abstract

Non-deductive reasoning, encompassing inductive and abductive reasoning, is essential in addressing complex real-world questions. One key feature of inductive and abductive reasoning is that there are many valid hypotheses; the simplest ones (those that adhere to Occam's Razor) are often most useful. However, this aspect is ignored in recent work that evaluates the non-deductive reasoning capabilities of large language models (LLMs). This work fills this gap, focusing on understanding whether the inductive and abductive reasoning capabilities of LLMs adhere to Occam's Razor, while also examining the correctness of their reasoning. To accomplish this goal, we introduce a framework to synthetically generate reasoning questions that *(a)* require inductive reasoning and abductive reasoning simultaneously; *(b)* is readily extended to produce any abductive/inductive reasoning question expressible in first-order logic. The task for the intelligent agent is to produce hypotheses to *explain* observations under a given world model. We also propose a new automated metric to assess whether hypotheses quantitatively adhere to Occam's Razor; those hypotheses that are correct and simplest are considered *high-quality*. Our findings on state-of-the-art LLMs suggest that LLMs can perform inductive and abductive reasoning in simple scenarios, but struggle with complex world models and with producing *high-quality* hypotheses, even with popular reasoning-enhancing techniques such as in-context learning and RLVR. [1]

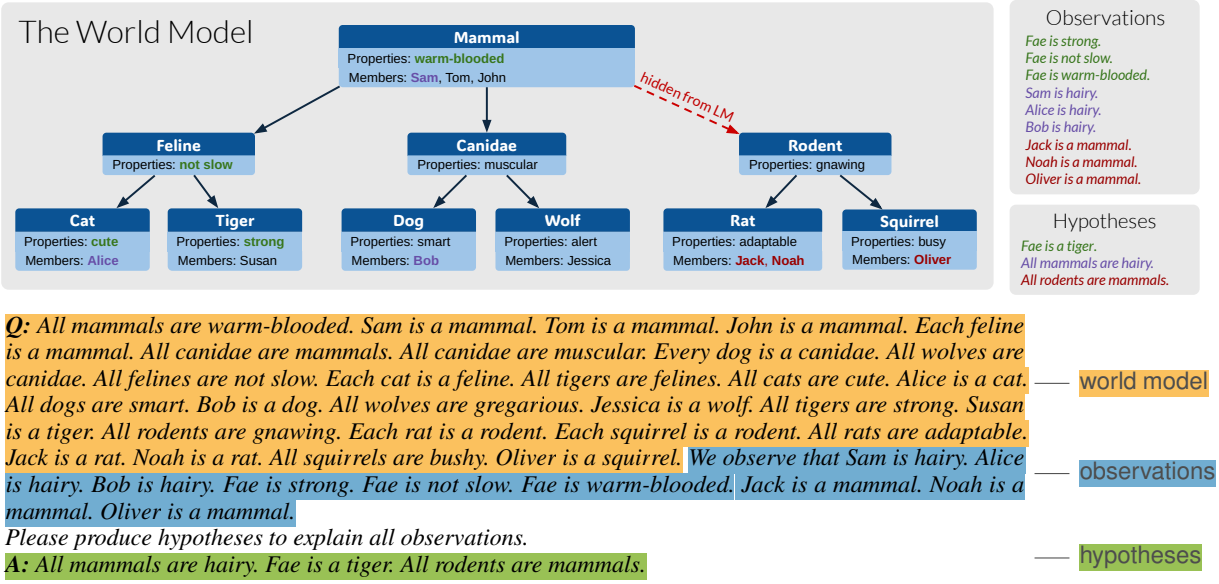
1 Introduction

Reasoning is a hallmark of intelligence. Peirce categorizes reasoning into deductive, inductive, and abductive reasoning (Peirce, 1878; Burch, 2024), all of which are indispensable in real-world scenarios. Deductive reasoning draws a conclusion from a set of premises. Inductive reasoning aims to identify a general principle from multiple individual cases. Abductive reasoning seeks to produce hypotheses to explain observations. See Appendix A for examples of different types of reasoning and for further background.

There is a significant body of work on evaluating the deductive reasoning capabilities of LLMs (Saparov and He, 2023; Saparov et al., 2023; Liu et al., 2020; Tafjord et al., 2021; Han et al., 2024), as well as techniques to improve deductive reasoning capabilities, such as in-context learning and reinforcement learning with verifiable rewards, *RLVR* (Lambert et al., 2024; Su et al., 2025). However, there is relatively little exploration into the inductive and abductive reasoning capabilities of large language models (LLMs). A fundamental difference between deductive and inductive or abductive reasoning is that deductive reasoning is deterministic. However, inductive and abductive reasoning are inherently probabilistic. There may exist many hypotheses, some *more probable* and some *less probable*. For example, if we hear a bird sound, a *more probable* hypothesis is that there is a bird outside; while a *less probable* hypothesis is that someone outside is playing the bird sound on a speaker. Among these possible hypotheses, the simplest ones are often more useful, and we view them as more probable. This principle is known as *Occam's Razor*. We call hypotheses that adhere to Occam's Razor *high-quality* hypotheses. Knowing and applying Occam's Razor is essential in areas such as scientific discovery, which require inductive and abductive reasoning, yet this is often overlooked. Physicists, for example, strive for the simplest explanations to account for as many observations as possible.

The high-level research question we aim to answer is: *Do LLMs reason inductively and abductively? If so, how well do they adhere to Occam's Razor?* This work systematically evaluates whether

[1] All data and generation code are freely available at https://anonymous.4open.science/r/D7F8/



Q: *All mammals are warm-blooded. Sam is a mammal. Tom is a mammal. John is a mammal. Each feline is a mammal. All canidae are mammals. All canidae are muscular. Every dog is a canidae. All wolves are canidae. All felines are not slow. Each cat is a feline. All tigers are felines. All cats are cute. Alice is a cat. All dogs are smart. Bob is a dog. All wolves are gregarious. Jessica is a wolf. All tigers are strong. Susan is a tiger. All rodents are gnawing. Each rat is a rodent. Each squirrel is a rodent. All rats are adaptable. Jack is a rat. Noah is a rat. All squirrels are bushy. Oliver is a squirrel. We observe that Sam is hairy. Alice is hairy. Bob is hairy. Fae is strong. Fae is not slow. Fae is warm-blooded. Jack is a mammal. Noah is a mammal. Oliver is a mammal.* — world model — observations

Please produce hypotheses to explain all observations.

A: *All mammals are hairy. Fae is a tiger. All rodents are mammals.* — hypotheses

Figure 1: An example of a question requiring inductive and abductive reasoning. Note that the dashed arrow means the subtype relationship `All rats are mammals` is hidden and needs to be inferred (i.e., a hypothesis to propose). This example uses a real-world model for illustrative purposes. In our dataset, we use fictional world models.

and to what extent LLMs can reason inductively and abductively and adhere to Occam's Razor. We also assess whether popular reasoning-enhancing techniques, such as in-context learning (Zhou et al., 2022; Dong et al., 2024) and RLVR (Lambert et al., 2024; Su et al., 2025), which primarily aim to improve deductive reasoning capabilities, can also improve inductive and abductive reasoning.

To answer this research question, we introduce a framework to synthetically generate reasoning questions that *(a)* contain inductive reasoning and abductive reasoning simultaneously; *(b)* is readily extended to produce any inductive/abductive reasoning question expressible in first-order logic. Each reasoning example has a fictional *incomplete* world model in the form of an ontology tree and a set of observations, represented in first-order logic. To answer each reasoning example, LLMs must produce hypotheses to explain observations under the incomplete world model. We generate synthetic examples of fictional world models to mitigate the contamination of training data. The ontology tree's height characterizes the difficulty of the world model and the corresponding reasoning example.[2] To assess whether LLMs adhere to Occam's Razor quantitatively, we also introduce an automated metric to quantify the quality of each hypothesis (i.e., *how well each hypothesis adheres to Occam's Razor*). We also introduce a new benchmark dataset, INABHYD(**In**ductive and **Ab**ductive **Hy**pothesis **D**iscovery), generated under this framework. This dataset contains over 2K reasoning questions with different difficulty parameters. The dataset also includes the generation code, which can, in principle, be used to generate an unbounded number of reasoning questions.

We evaluate our dataset on seven widely-used LLMs: GPT-5.4 (Singh et al., 2025), o3 (OpenAI, 2025), LLAMA3-70B (Dubey et al., 2024) , GEMMA3-27B (Kamath et al., 2025), DEEPSEEK-V3 (DeepSeek-AI et al., 2024), GPT-4o (OpenAI, 2023) , and DEEPSEEK-R1-DISTILL-LLAMA-70B (DeepSeek-AI et al., 2025). For each LLM, we test whether it can produce *valid* and *high-quality* hypotheses. We also test whether popular reasoning-enhancing techniques, such as in-context learning and RLVR, which primarily were meant to improve deductive reasoning, can also improve inductive and abductive reasoning. Our findings can be summarized as follows:

1. LLMs can perform inductive and abductive reasoning at high accuracy for shallow ontology trees with a single hypothesis.
2. The accuracy drops significantly when the ontology tree structure is complex and contains multiple hypotheses.
3. In most cases, LLMs may produce *valid* hypotheses (which can explain all observations); but fail to produce *high-quality* hypotheses (which adhere to Occam's Razor).
4. In general, we observe that reasoning-enhancing techniques *moderately* improves inductive and

[2]Refer to Figure 1 and Appendix C for examples.

abductive reasoning.

2 Related Work

Evaluating non-deductive reasoning capabilities. Table 1 summarizes the similarities and differences between our work and others on evaluating non-deductive reasoning capabilities of both LLMs and earlier models. Representative works on abductive reasoning evaluation, such as ART (Zhao et al., 2023), AbductionRules (Young et al., 2022), and Wino-Why (Zhang et al., 2020), craft abductive reasoning questions from web sources, and ask LMs to explain everyday events. However, this approach confounds the problem of commonsense knowledge acquisition with the problem of abductive reasoning. In contrast, our framework provides all required facts in the context of each prompt, thereby evaluating LM's abductive reasoning capabilities in a more controlled fashion. Representative works on inductive reasoning evaluation, such as InductiveBench (Hua et al., 2025) and SyGus (Odena et al., 2021) focus exclusively on highly-structured domains such as string translation and program synthesis. However, natural language-based inductive reasoning is also essential and is underexplored.

Evaluating deductive reasoning capabilities. There is no consensus on whether LLMs can perform general deductive reasoning or are stochastic parrots (Bender et al., 2021). Researchers have designed different datasets and evaluation methodologies to answer this question (Saparov et al., 2023; Saparov and He, 2023; Liu et al., 2020; Tafjord et al., 2021; Han et al., 2024). LogiQA (Liu et al., 2020) is a multiple-choice QA dataset to test multiple types of deductive reasoning of language models. However, it focuses solely on label accuracy (*the correctness of the final answer*) without considering the correctness of intermediate proof steps. In contrast, PrOntoQA (Saparov and He, 2023; Saparov et al., 2023) evaluates both label accuracy and step-by-step proof correctness. They can synthetically generate proofs with different heights, widths, and deduction rules. Our work is similar to PrOntoQA in spirit. We both use a first-order ontology tree as the world model. However, PrOntoQA does not evaluate non-deductive reasoning.

Techniques to enhance reasoning. A large body of work has focused on improving the (deductive) reasoning capabilities of LLMs and related downstream tasks such as mathematics and coding (Li, 2025; Wei et al., 2022; Yao et al., 2023; Wang et al., 2023; Su et al., 2025; Lambert et al., 2024; Zhou et al., 2022; Dong et al., 2024). The first line of work aims to elicit the emergent capabilities of LLMs using different prompt formats, such as chain-of-thought (CoT; Wei et al., 2022), tree-of-thought (ToT; Yao et al., 2023), self-consistency (Wang et al., 2023), and in-context learning. The second line of work post-trains LLMs with reinforcement learning with verifiable reward (*RLVR*) Lambert et al. 2024; Su et al. 2025 to improve LLM reasoning. Since non-deductive reasoning is inherently ambiguous, it is difficult to define a verifiable reward in non-deductive settings. The third line of work uses neuro-symbolic approaches to enhance the efficiency and reliability of LLM reasoning (Olausson et al., 2023; Cheng et al., 2025; Pan et al., 2023). However, they are either not directly applicable to unstructured natural language input or designed for deductive reasoning applications. This work investigates whether the above methods benefit inductive and abductive reasoning.

3 Approach

3.1 Overview

In this section, we describe how we generate the dataset and evaluate LLMs. The dataset consists of synthetically-generated reasoning examples in natural language. Figure 2 illustrates the pipeline for generating each question. We evaluate LLM's inductive and abductive reasoning capabilities on metrics such as accuracy and their adherence to Occam's Razor. By composing inductive and abductive reasoning steps shown in Table 2, our evaluation framework can readily be extended to produce any abductive/inductive reasoning question expressible in first-order logic.

Each reasoning example contains a fictional world model and a set of observations. We use real person names (such as `Amy`) and real property names (such as `strong`), but we do not use real concept names (such as `tiger`). Instead, we use fictional concept names (such as `wumpus`). In doing so, LLMs will not be confused by these fictional relations, as we never generate facts that contradict those about real concepts.

LLMs must produce hypotheses to *explain* the observations. Our *hypothesize-explain* framework can cover a wide range of inductive/abductive reasoning scenarios.[3] To generate a single reasoning

[3]Traditionally, producing hypotheses to explain observa-

Dataset	Controls for complexity of world models	Natural language description	Tests for inductive reasoning	Test for abductive reasoning	Evaluation of hypotheses quality	Multiple hypotheses
ART (Zhao et al., 2023)	✗	✓	✗	✓	non-quantitative	✗
InductiveBench (Hua et al., 2025)	N\A	✗	✓	✗	✗	✗
AbductionRules (Young et al., 2022)	✗	✓	✗	✓	✗	✗
Wino-Why (Zhang et al., 2020)	✗	✓	✗	✓	✗	✗
SyGus (Odena et al., 2021)	N\A	✗	✓	✗	✗	✗
UniADILR (Sheng et al., 2025)	✗	✓	✓	✓	✗	✗
LogiEval (Liu et al., 2025)	✗	✓	✓	✓	✗	✗
MME-Reasoning (Yuan et al., 2025)	✗	✓	✓	✓	✗	✗
InAbHyD(**ours**)	✓	✓	✓	✓	✓	✓

Table 1: Comparison of existing datasets to evaluate inductive and abductive reasoning capabilities.

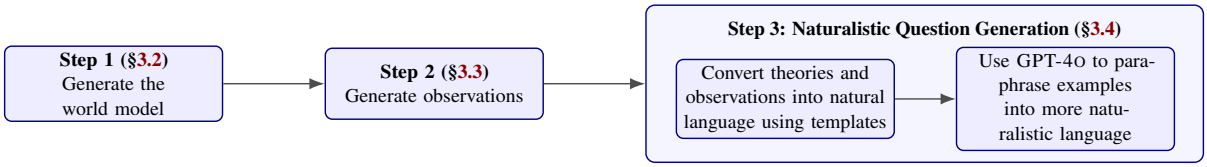


Figure 2: Pipeline for generating each reasoning question in InAbHyD.

example, we go through the following steps:

1. First, we construct a world model as an ontology tree in first-order logic.
2. Second, we traverse the ontology tree and randomly mark some axioms as hidden. Each hidden axiom is the ground truth.
3. Third, for each hidden axiom, we generate observations so that LLMs have enough information to produce ground-truth hypotheses.
4. Fourth, we transform the incomplete world model and observations into natural language.

See Table 2 for an overview of the different subtasks and their details in the simplest case (a height-1 ontology tree with a single hypothesis). We evaluate the output of LLMs (i.e., produced *hypotheses*) on the following metrics:

1. *Strong accuracy*: Hypotheses are strongly correct if they exactly match ground-truth hypotheses (simplest hypothesis).
2. *Weak accuracy*: Hypotheses are weakly correct if they explain all observations.
3. *Hypothesis quality*: A quantitative measurement of the quality of the LLMs' produced hypotheses based on Occam's Razor.

Next, we describe each step in more detail.

3.2 World model generation

In our approach, the world model is an ontology tree in first-order logic. Each node in the ontology tree is a concept (e.g. `cat`) with properties (e.g. `All cats are cute`) and members (e.g. `Amy is a cat`). Each edge in the ontology tree denotes a subtype relationship (e.g., the `ragdoll` node being a child of the `cat` node implies `Each ragdoll is a cat`).

We first generate the topology of the ontology tree with the height as a controllable parameter. Each non-leaf node has a set of child nodes. The height of the ontology tree is a variable that characterizes the complexity of the world model. We then make two passes over the tree using level-order traversal. In the first pass, for each node, we allocate a concept name, properties, and members [4]. After the first pass, we have a *complete* world model with three types of axioms: (1) properties (e.g., `All cats are cute`), (2) membership relations (e.g., `Amy is a cat`), (3) subtype relations (e.g., `Each ragdoll is a cat`). In the second pass, we randomly mark some axioms as hidden. These axioms are the ground-truth hypotheses. After the second pass, we have an *incomplete* world model.

3.3 Observation generation

Next, we generate observations so that LLMs have sufficient information to produce the ground-truth hypotheses. The observations must be complex enough that the reasoning example is non-trivial, yet they should also ensure the problem is feasible.

- To generate an example where the task is to infer a property (e.g., `All cats are cute`), we first

tions is considered abductive reasoning. However, we include inductive reasoning in the same framing. For example, given these axioms as a world model: `Sally is a cat; Polly is a cat; Amy is a cat` and these observations: `Sally is cute; Polly is cute; Amy is cute`, we can produce a hypothesis: `All cats are cute` to explain all observations. This is *inductive reasoning*.

[4] The properties can also be negated, e.g., `All cats are not slow`, which corresponds to the logical form $\forall x(\texttt{cat}(x) \rightarrow \neg\texttt{slow}(x))$ (negation always takes narrow scope). The number of properties, members, and child nodes follows a zipf distribution with $a = 3$.

Subtask	Reasoning type tested	World model	Observations	Ground-truth hypothesis	Proof trees for observations
Infer property	Inductive	o(A) o(B) o(C)	p(A) p(B) p(C)	$\forall x(o(x) \rightarrow p(x))$	$\frac{o(A) \quad \forall x(o(x) \rightarrow p(x))}{p(A)}$
Infer membership relation	Abductive	$\forall x(o(x) \rightarrow p_1(x))$ $\forall x(o(x) \rightarrow p_2(x))$ $\forall x(o(x) \rightarrow p_3(x))$	p1(A) p2(A) p3(A)	o(A)	$\frac{o(A) \quad \forall x(o(x) \rightarrow p_1(x))}{p_1(A)}$
Infer subtype relation	Both	o1(A) o1(B) o1(C)	o2(A) o2(B) o2(C)	$\forall x(o_1(x) \rightarrow o_2(x))$	$\frac{o_1(A) \quad \forall x(o_1(x) \rightarrow o_2(x))}{o_2(A)}$

Table 2: An overview of various subtasks and their details in the simplest case (a height-1 ontology tree with a single hypothesis). Here o denotes an ontology concept (e.g, cat) and o(A) denotes the membership (e.g., cat(Sally) means Sally is a cat). p denotes a property (e.g, cute) and p(A) denotes A has property p (e.g., cute(Sally) means Sally is cute). We only show the proof tree for one observation for simplicity; other observations can be derived similarly. In the proof tree, we show hypotheses in green and axioms in world models in purple.

identify the type in the ontology with this property (e.g., cat), then we sample members of this type and subtypes (e.g., Amy is a cat; Sally is a ragdoll; Polly is a shorthair), and add observations where the members have this property (e.g., Amy is cute; Sally is cute;).

- To generate an example where the task is to infer a membership relation (e.g., Amy is a cat), we first identify the type in the ontology with this member (e.g., cat), then we sample properties of this type and supertypes (e.g., All cats are cute; All felines are not slow; All mammals are hairy), and add observations where this member have these properties (e.g., Amy is cute; Amy is not slow).
- To generate an example where the task is to infer a subtype relation (e.g., All cats are felines), we first identify the subtype in the ontology (e.g., cat), then we sample members of this type and its subtypes (e.g., Amy is a cat; Sally is a ragdoll; Polly is a shorthair), and add observations where these members belong to the supertype (e.g., Amy is a feline; Polly is a feline).

3.4 Naturalistic Question Generation

We first use the same grammar as in PRONTOQA for converting logical forms into natural language using templates. Properties are converted into the form: All cats are cute; subtype relations are randomly converted into one of three forms: Each ragdoll is a cat; Every ragdoll is a cat; All ragdolls are cats; membership relations are converted into the form Amy is a cat. We then use GPT-4o to paraphrase it into more natural language. Details are provided in Appendix D.

3.5 Hypothesis quality

We define a metric to evaluate the quality of hypotheses based on the principles of Occam's Razor (i.e., the priciple that, other things being equal, simpler theories are better; Baker 2022). In this regard, we encourage one hypothesis to explain as many observations as possible and penalize unnecessary hypotheses. Formally, given a world model $\mathcal{T}$, a set of observations $\mathcal{O}$, their associated proof trees $\mathcal{P}$,candidate hypotheses $\mathcal{H}$, and ground-truth hypotheses $\mathcal{H}^*$, for $h \in \mathcal{H}$, $n(h)$ denotes the number of times h appears in the proof trees of $\mathcal{O}$ (i.e., how many times h is used as a premise of a proof step). We first use $|\mathcal{H}|$ to denote the cardinality of $\mathcal{H}$. The quality of $\mathcal{H}$, $q(\mathcal{H})$ is defined by [5]:

$$q(\mathcal{H}) = \begin{cases} 0 & \mathcal{H} \text{ cannot explain } \mathcal{O} \\ \frac{\frac{1}{|\mathcal{H}|}\sum_{h \in \mathcal{H}} n(h)}{\frac{1}{|\mathcal{H}^*|}\sum_{h^* \in \mathcal{H}^*} n(h^*)} & \text{otherwise} \end{cases}$$

According to the definition, the ground-truth hypotheses $\mathcal{H}^*$ have a quality score of 1. Consider the example in Figure 1, here $\mathcal{H}^*$ is {Fae is a tiger, All mammals are hairy, All rodents are mammals} and $|\mathcal{H}^*| = 3$. Each $h^* \in \mathcal{H}^*$ appears three times in the proof trees of $\mathcal{O}$, so $\sum_{h^* \in \mathcal{H}^*} n(h^*) = 9$. Consider candidate hypotheses $\mathcal{H}$ {Fae is a tiger, All mammals are hairy, All rats are mammals, All squirrels are mammals, All cats are hairy}. $|\mathcal{H}| = 5$. Obviously $\mathcal{H}$ can explain $\mathcal{O}$. All rats are mammals appears twice, All squirrels are mammals appears once, and All cats are hairy appear once in the proof trees of

[5] We write the equation as the ratio of two averages: the numerator is the average accuracy of the predicted hypothesis (i.e., the average number of times a hypothesis appears in any valid proof), and the denominator is the average number of times a hypothesis appears in the shortest (i.e., most parsimonious) proof.

$\mathcal{O}$, so $\sum_{h \in \mathcal{H}} n(h) = 3 + 3 + 2 + 1 + 1 = 10$. In this case, $q(\mathcal{H}) = \frac{3 \times 10}{5 \times 9} = 0.67$.

We further motivate the hypothesis metric using Bayesian decision theory in Appendix E.1. We compare the metric to human evaluation of hypotheses in Appendix E.2. Appendix F contains the complete algorithms for generating reasoning examples and evaluating hypotheses.

4 Experiments

4.1 Experimental setup

In each experiment configuration, we generate 100 independent and identically distributed reasoning examples. Each reasoning example is pure text, a user prompt with an incomplete world model and observations. We also specify the desired output format in the system prompt. To answer each reasoning example, LLMs must produce hypotheses to explain all observations. We evaluate hypotheses on three metrics: weak accuracy, strong accuracy, and quality. We compute 95% confidence intervals for both accuracy metrics using the Wilson interval (Wilson, 1927). We use the OpenAI API and together.ai platform.

In our experiments, we vary the following variables: (1) Different types of tasks (i.e., infer property, infer membership relation, or infer subtype relation), (2) Single hypothesis or multiple hypotheses, (3) The height of the ontology tree, (4) Whether the reasoning example contains in-context demonstrations and whether they are in-distribution (5) Different reasoning-augmented prompts such as buffer-of-thought (Yang et al., 2024).

4.2 Zero-shot results

We first investigate the simplest setting with zero-shot prompting.

4.2.1 Single hypothesis

We first investigate results for reasoning examples with one hypothesis for three tasks. This helps us to isolate different types of tasks and understand how LLMs perform in each case. The results are in Figure 3. In general, all models show high accuracy (both weak and strong) greater than 80% for all three tasks under a height-1 ontology tree. When the height of the ontology tree starts to increase, we also observe a general decreasing trend in all three metrics, which is not surprising. One interesting observation is that LLMs perform better in *infer membership relation* tasks than other tasks, especially for GEMMA3-27B. For example, even under a height-4 ontology tree, GEMMA3-27B can achieve near 50% weak accuracy, while the number drops to around 10% for *infer subtype relation* tasks. This may be because the weak correctness for *infer membership relation* tasks only requires searching for the concept that has some property, but do not need to understand the structure of the ontology tree.[6] However, this is not the case for *infer subtype relation* and *infer properties* tasks. We observe a similar trend for *infer subtype relation* and *infer properties* tasks, as we can view the concept's parent as a special property.

4.2.2 Multiple hypotheses

Next, we turn to multi-hypothesis scenarios because they are closer to real-life situations, and we are interested in understanding how LLMs perform in this setting. Each reasoning example can contain multiple hypotheses. To check the difficulty of each reasoning example, we sample the ground-truth hypotheses such that the number of hypotheses grows linearly *w.r.t.* the height of the ontology tree.[7] Invariably, the number of axioms in the world model grows exponentially *w.r.t.* the height of the ontology tree. See Table 3 in Appendix B for statistics under multi-hypothesis scenarios and the **top** row of Figure 4 for results.

Performance in the multi-hypothesis setting differed significantly from that in the single-hypothesis setting. For example, the accuracy of all models drops significantly when the height of the ontology tree is increased from 1 to 2, with accuracy going from above 80% to below 50%, except for GPT-4O. This is surprising because the average number of ground-truth hypotheses increases only from 3.0 to 3.5. We attribute this to the coupling effect of multiple hypotheses to produce and the topology of the ontology tree. Reasoning examples under the ontology tree with heights 3 or 4 achieve lower accuracy and quality scores across all models. We also observe a notable gap between weak accuracy and quality. For example, under the height-4 ontology tree, all models can achieve at least 20% weak accuracy but struggle with extremely low-quality and strong-accuracy scores. We manually inspect some incorrect hypotheses and summarize the most common types of errors in Section 4.5.

[6] e.g., given `Amy is cute`, LLMs only need to search for a concept which has the property `cute` to infer the membership relation of `Amy`.

[7] Each ground-truth hypothesis has ~ 3 observations. So the number of observations also grows linearly *w.r.t.* the height of the ontology tree.

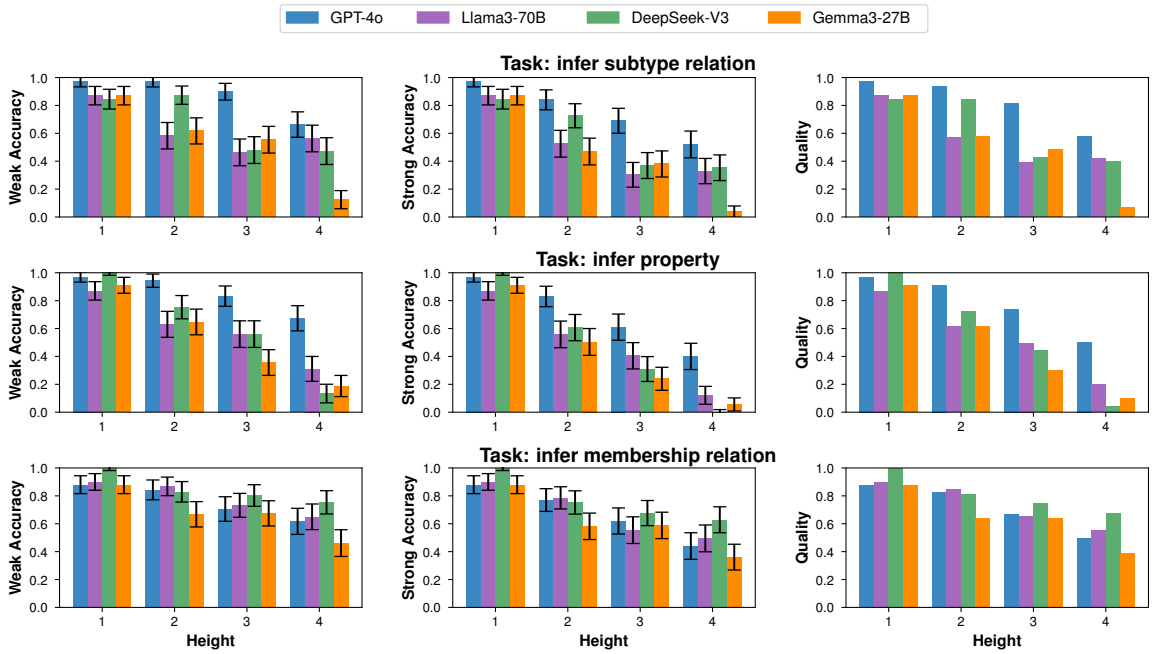


Figure 3: Weak accuracy, strong accuracy, and quality versus the height of the ontology tree with single hypothesis.

4.3 Does in-context learning help?

LLMs have shown extraordinary generalization capabilities with few-shot examples coupled with CoT prompts (Brown et al., 2020). We are also interested in whether in-context learning can help with inductive/abductive reasoning. We add 8 demonstrations to each example, with the demonstrations' ground-truth hypotheses and CoTs. The CoT consists of each observation's step-by-step proof. We have two settings for in-context learning: (1) In-distribution demonstrations, where the demonstration has an ontology tree of the same height as the test question. (2) Out-of-distribution demonstrations, where each demonstration has a single hypothesis with a height-1 ontology tree. In both settings, we ensure demonstrations cover all types of tasks. We present our results in Figure 4.

The in-distribution demonstrations are slightly more helpful than out-of-distribution demonstrations. This result is consistent with the hypothesis that in-context learning mimics Bayesian inference (Xie et al., 2022). We also observe that in-context learning with in-distribution demonstrations improves strong accuracy and quality in ontology trees with heights 3 and 4. We conjecture that LLMs can learn how to navigate the ontology hierarchy and produce high-quality hypotheses through in-context demonstrations. In contrast, we do not observe a significant performance improvement from out-of-distribution demonstrations.

4.4 Do reasoning models and reasoning-augmented prompts help?

Recent developments in large reasoning models, particularly those trained with RLVR, have led to substantial improvements across many evaluations. As shown in Figure 5, stronger models such as GPT5.4 and O3 consistently outperform LLAMA3-70B in both weak and strong accuracy and overall quality across different ontology tree heights. Moreover, reasoning-oriented prompting strategies further enhance performance: buffer-of-thought (BoT; Yang et al., 2024) prompting generally yields higher scores than chain-of-thought prompting (Wei et al., 2022), especially in weak accuracy. However, despite these gains, a persistent gap remains between weak and strong accuracy across all models and settings——for instance, weak accuracy often approaches near-perfect levels at lower heights, while strong accuracy remains notably lower and degrades more rapidly as height increases. This suggests that although recent advances and prompting techniques significantly improve reasoning capabilities, high-quality hypothesis generation remains an open challenge.

4.5 Analysis of common errors

To further understand *how* and *why* LLMs fail, we manually inspect mistaken and low-quality hypotheses. We summarize some of the most common types of errors here for further insights. We provide example responses in each category, quantitative analysis, and the details of the manual inspec-

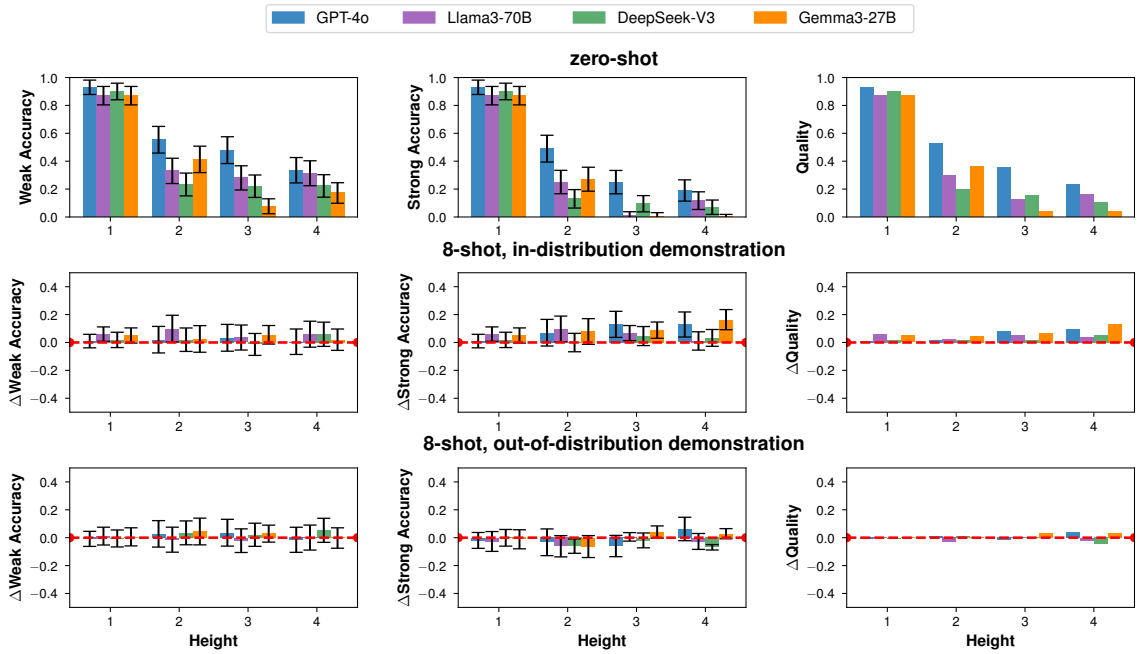


Figure 4: Weak accuracy, strong accuracy, and quality versus the height of the ontology tree with multiple hypotheses under zero-shot (**top**), 8-shot in-distribution (**middle**), and 8-shot out-of-distribution (**bottom**). We list the absolute results of the 8-shot in-distribution and out-of-distribution in Appendix G.

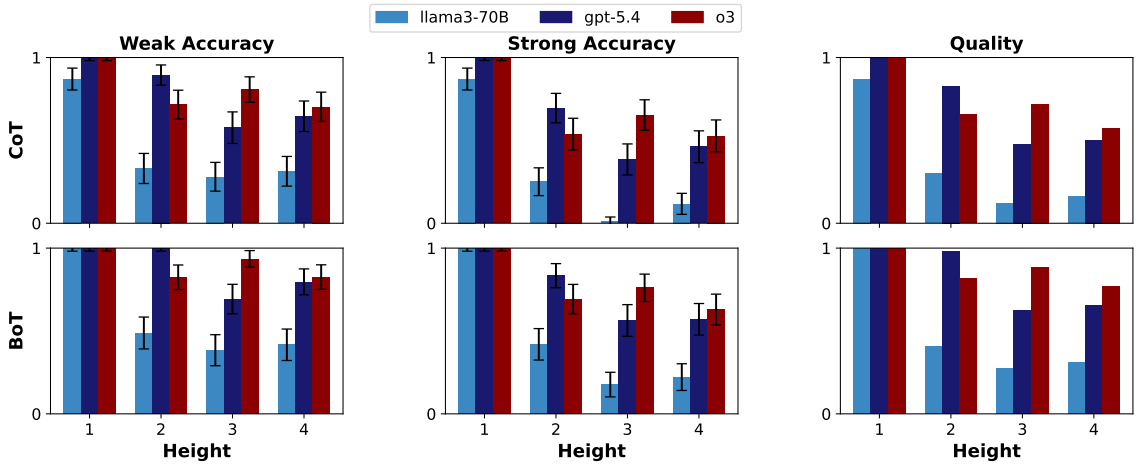


Figure 5: Results for LLAMA3-70B and two large reasoning models GPT-5.4 and O3 with chain-of-thought prompting (**top**) and buffer-of-thought prompting (**bottom**).

tion process in Appendix H. (1) **Wrong ontology direction.** LLMs sometimes produce hypotheses in the wrong direction. For example, the ground-truth hypothesis is `All cats are mammals`, while LLMs will produce the hypothesis `All mammals are cats`. (2) **Ignore the ontology and produce unnecessary hypotheses.** LLMs sometimes can produce unnecessary hypotheses because they ignore the ontology. (3) **Fall back to trivial hypotheses.** Technically, the observation is also a valid hypothesis, but it is considered low-quality because it can only explain itself. LLMs sometimes reuse observations as hypotheses. (4) **Hallucinated entities.** LLMs sometimes produce hypotheses with hallucinated entities (including concepts, properties, members). Usually, this happens when the entity appears in the in-context demonstrations, but LLMs use them in the test question.

5 Conclusion

In this study, we systematically examined the inductive/abductive reasoning capabilities of LLMs, especially whether they follow Occam's Razor. Our results suggest that current LLMs can't generate high-quality hypotheses that match human-level parsimony. Our results can serve to better inform practitioners of LLMs of their limitations, especially in applications that rely heavily on inductive/abductive reasoning. The proposed dataset opens new avenues for future work, such as training models to improve their reasoning capabilities.

Limitations

Our synthetic dataset serves as an out-of-distribution evaluation of the inductive and abductive reasoning abilities of LLMs. We do not claim that success in our benchmark automatically implies success in real-world tasks. Still, our benchmark serves as a checkpoint for developing models that are capable of robust real-world abductive and inductive reasoning.

Another limitation of this work is that first-order logic (FOL) restricts the expressiveness of the world model and the reasoning example. We leave it as future work to utilize higher-order logic (HOL) to generate more complex world models and reasoning examples.

References

Alan Baker. 2022. Simplicity. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Summer 2022 edition. Metaphysics Research Lab, Stanford University.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 610–623. ACM.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Robert Burch. 2024. Charles Sanders Peirce. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Summer 2024 edition. Metaphysics Research Lab, Stanford University.

Fengxiang Cheng, Haoxuan Li, Fenrong Liu, Robert Van Rooij, Kun Zhang, and Zhouchen Lin. 2025. Empowering llms with logical reasoning: A comprehensive survey. *arXiv preprint arXiv:2502.15652*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 81 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 81 others. 2024. Deepseek-v3 technical report. *CoRR*, abs/2412.19437.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 1107–1128. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *The Innovation*.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alexander Wardle-Solano, Hannah Szabó, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, and 16 others. 2024. FOLIO: natural language reasoning with first-order logic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 22017–22031. Association for Computational Linguistics.

Wenyue Hua, Tyler Wong, Sun Fei, Liangming Pan, Adam Jardine, and William Yang Wang. 2025. Inductionbench: Llms fail in the simplest complexity class. *CoRR*, abs/2502.15823.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, and 79 others. 2025. Gemma 3 technical report. *CoRR*, abs/2503.19786.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2024. Tülu 3: Pushing frontiers in open language model post-training. *CoRR*, abs/2411.15124.

Xinzhe Li. 2025. A survey on LLM test-time compute via search: Tasks, LLM profiling, search algorithms, and relevant frameworks. *CoRR*, abs/2501.10069.

Hanmeng Liu, Yiran Ding, Zhizhang Fu, Chaoli Zhang, Xiaozhang Liu, and Yue Zhang. 2025. Evaluating the logical reasoning abilities of large reasoning models. *arXiv preprint arXiv:2505.11854*.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3622–3628. ijcai.org.

Augustus Odena, Kensen Shi, David Bieber, Rishabh Singh, Charles Sutton, and Hanjun Dai. 2021. BUSTLE: bottom-up program synthesis through learning-guided exploration. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. 2023. Linc: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5153–5176.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

OpenAI. 2025. Openai o3 and o4-mini system card. https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. 2023. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3806–3824.

Charles Sanders Peirce. 1878. Deduction, induction, and hypothesis. *Popular Science Monthly*, 13:470–482.

Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Abulhair Saparov, Richard Yuanzhe Pang, Vishakh Padmakumar, Nitish Joshi, Mehran Kazemi, Najoung Kim, and He He. 2023. Testing the general deductive reasoning capacity of large language models using OOD examples. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yu Sheng, Wanting Wen, Linjing Li, and Daniel Zeng. 2025. Evaluating generalization capability of language models across abductive, deductive and inductive logical reasoning. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 4945–4957. Association for Computational Linguistics.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, Zhaopeng Tu, Min Zhang, and Dong Yu. 2025. Crossing the reward bridge: Expanding RL with verifiable rewards across diverse domains. *CoRR*, abs/2503.23829.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. Proofwriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 3621–3634. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Edwin B Wilson. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Ling Yang, Zhaochen Yu, Tianjun Zhang, Shiyi Cao, Minkai Xu, Wentao Zhang, Joseph E Gonzalez, and Bin Cui. 2024. Buffer of thoughts: Thought-augmented reasoning with large language models. *Advances in Neural Information Processing Systems*, 37:113519–113544.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Nathan Young, Qiming Bao, Joshua Bensemann, and Michael Witbrock. 2022. Abductionrules: Training transformers to explain unexpected inputs. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 218–227. Association for Computational Linguistics.

Jiakang Yuan, Tianshuo Peng, Yilei Jiang, Yiting Lu, Renrui Zhang, Kaituo Feng, Chaoyou Fu, Tao Chen, Lei Bai, Bo Zhang, and 1 others. 2025. Mme-reasoning: A comprehensive benchmark for logical reasoning in mllms. *arXiv preprint arXiv:2505.21327*.

Hongming Zhang, Xinran Zhao, and Yangqiu Song. 2020. Winowhy: A deep diagnosis of essential commonsense knowledge for answering winograd schema challenge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5736–5745. Association for Computational Linguistics.

Wenting Zhao, Justin T. Chiu, Claire Cardie, and Alexander M. Rush. 2023. Abductive commonsense reasoning exploiting mutually exclusive explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14883–14896. Association for Computational Linguistics.

Hattie Zhou, Azade Nova, Hugo Larochelle, Aaron C. Courville, Behnam Neyshabur, and Hanie Sedghi. 2022. Teaching algorithmic reasoning via in-context learning. *CoRR*, abs/2211.09066.

A Preliminary background

A.1 Examples of different types of reasoning

For example, given "All humans are moral" and "Socrates is a human", we can conclude that "Socrates is moral". For example, if we see many cute cats, we may inductively reason that all cats are cute. For example, if we hear a bird sound outside, we may abductively reason that there is a bird. The latter two types of reasoning share a common high-level goal: to find explanations for observations. Furthermore, the composition of inductive and abductive reasoning is prevalent in real-world scenarios. For example, given (i) Amy has a skin rash and leg pain and has been diagnosed with a new, unknown virus. (ii) Sam has a skin rash and leg pain and has been diagnosed with the same new unknown virus. (iii) Tom also has a skin rash and leg pain. We can first induce the rule: The new unknown virus causes skin rash and leg pain, and then we can abduce that Tom is infected with this new virus.

A.2 Proof trees

A **proof tree** is a tree-structured representation of a logical derivation where each node corresponds to a statement, and each edge represents the application of an inference rule. The root of the tree is the final conclusion, while the leaves correspond to premises or axioms. Proof trees are commonly used in logic, theorem proving, and formal reasoning to make the structure of derivations explicit.

Formally, let Γ denote a set of premises and φ denote a conclusion. A proof tree demonstrates that $\Gamma \vdash \varphi$ by recursively applying inference rules until all leaf nodes correspond to elements of Γ or axioms.

Example. Consider the following premises:

1. $P \to Q$
2. $Q \to R$
3. P

We can derive R using two applications of *modus ponens*. We first derive Q given the premises P and $P \to Q$. Next, we derive R using the premises Q and $Q \to R$. This proof can be represented as a proof tree, which is shown below.

$$\dfrac{\dfrac{P \to Q \qquad P}{Q} \qquad Q \to R}{R}$$

The tree therefore makes explicit the hierarchical structure of the reasoning process.

B Basic statistics of InAbHyD

We list basic statistics of InAbHyD in Table 1.

C More examples of reasoning questions

In this section, we provide additional reasoning examples generated using our proposed framework. For each setting listed below, we provide an example consisting of a world model, list of observations, and ground-truth hypothesis/hypotheses.

Ontology tree height	# of world model axioms	# of observations	# of ground-truth hypotheses	# of examples
1	9.0	10.0	3.0	500
2	14.0	11.8	3.5	500
3	25.5	15.2	4.6	500
4	46.8	20.0	6.6	500

Table 3: Basic statistics of multi-hypothesis examples in INABHYD. We list the average number of world model axioms, observations, and ground-truth hypotheses *w.r.t.* the height of the ontology tree in the same set of reasoning examples we used in evaluation. Note that the number of the ground-truth hypotheses and the number of observations grow linearly with the height of the ontology tree. In contrast, the number of world model axioms grows exponentially.

Single hypothesis example; task: infer property; height-1 ontology tree.

World Model: Amy is a dalpist. Pamela is a dalpist. Jerry is a dalpist.
Observations: Amy is rainy. Jerry is rainy. Pamela is rainy.
Hypotheses: Dalpists are rainy.

Single hypothesis example; task: infer property; height-2 ontology tree.

World Model: Rompuses are dalpists. Sarpers are dalpists. Each gergit is a dalpist. Pamela is a sarper. Jerry is a dalpist. Amy is a rompus.
Observations: Amy is rainy. Pamela is rainy. Jerry is rainy.
Hypotheses: Dalpists are rainy.

Single hypothesis example; task: infer property; height-3 ontology tree.

World Model: Pamela is a porpor. Jerry is a dalpist. Twimpees are sarpers. Each gergit is a dalpist. Each storpist is a gergit. Amy is a storpist. Porpors are rompuses. Sarpers are dalpists. Rompuses are dalpists.
Observations: Jerry is not muffled. Amy is not muffled. Pamela is not muffled.
Hypotheses: Each dalpist is not muffled.

Single hypothesis example; task: infer property; height-4 ontology tree.

World Model: Amy is a felper. Jerry is a dalpist. Pamela is a felper. Each rompus is a dalpist. Each porpor is a rompus. Sarpers are dalpists. Shumples are twimpees. Gergits are dalpists. Each frompor is a porpor. Storpists are gergits. Twimpees are sarpers. Felpers are storpists.
Observations: Amy is translucent. Pamela is translucent. Jerry is translucent.
Hypotheses: Each dalpist is translucent.

Single hypothesis example; task: infer membership relation; height-1 ontology tree.

World Model: Dalpists are liquid. Dalpists are rainy. Each dalpist is brown.
Observations: Jerry is brown. Jerry is rainy. Jerry is liquid.
Hypotheses: Jerry is a dalpist.

Single hypothesis example; task: infer membership relation; height-2 ontology tree.

World Model: Rompuses are dalpists. Jerry is a dalpist. Each sarper is a dalpist. Each dalpist is brown. Edward is a sarper. Gergits are dalpists. Gergits are rainy. Amy is a rompus. Dalpists are liquid.
Observations: Pamela is rainy. Pamela is liquid. Pamela is brown.
Hypotheses: Pamela is a gergit.

Single hypothesis example; task: infer membership relation; height-3 ontology tree.

World Model: Each sarper is a dalpist. Jerry is a dalpist. Barbara is a porpor. Each dalpist is moderate. Rompuses are dalpists. Twimpees are sarpers. Storpists are gergits. Each dalpist is large. Edward is a sarper. Pamela is a gergit. Gergits are dalpists. Debra is a storpist. Porpors are rompuses. Amy is a rompus. Twimpees are not muffled.
Observations: Nicole is moderate. Nicole is large. Nicole is not muffled.
Hypotheses: Nicole is a twimpee.

Single hypothesis example; task: infer membership relation; height-4 ontology tree.

World Model: Amy is a rompus. Each porpor is a rompus. Each sarper is a dalpist. Sharon is a felper. Rompuses are dalpists. Each shumple is a twimpee. Each dalpist is fruity. Barbara is a porpor. Every frompor is a porpor. Every gergit is a dalpist. Edward is a sarper. Raymond is a frompor. Each storpist is a gergit. Jerry is a dalpist. Each dalpist is moderate. Shumples are translucent. Felpers are storpists. Nicole is a twimpee. Pamela is a gergit. Twimpees are sarpers. Debra is a storpist.
Observations: Michael is fruity. Michael is translucent. Michael is moderate.
Hypotheses: Michael is a shumple.

Single hypothesis example; task: infer subtype relation; height-1 ontology tree.

World Model: Amy is a dalpist. Pamela is a dalpist. Jerry is a dalpist.
Observations: Amy is a rompus. Jerry is a rompus. Pamela is a rompus.
Hypotheses: Dalpists are rompuses.

Single hypothesis example; task: infer subtype relation; height-2 ontology tree.

World Model: Gergits are dalpists. Porpors are dalpists. Each sarper is a dalpist. Pamela is a porpor. Jerry is a dalpist. Amy is a gergit.
Observations: Amy is a rompus. Pamela is a rompus. Jerry is a rompus.
Hypotheses: Dalpists are rompuses.

Single hypothesis example; task: infer subtype relation; height-3 ontology tree.

World Model: Amy is a storpist. Jerry is a dalpist. Frompors are porpors. Each sarper is a dalpist. Each twimpee is a sarper. Pamela is a twimpee. Storpists are gergits. Porpors are dalpists. Gergits are dalpists.
Observations: Jerry is a rompus. Amy is a rompus. Pamela is a rompus.
Hypotheses: Each dalpist is a rompus.

Single hypothesis; task: infer subtype relation; height-4 ontology tree.

World Model: Each sarper is a dalpist. Jerry is a dalpist. Amy is a shumple. Pamela is a shumple. Each storpist is a gergit. Dumpuses are frompors. Gergits are dalpists. Felpers are storpists. Each twimpee is a sarper. Frompors are porpors. Shumples are twimpees. Porpors are dalpists.
Observations: Amy is a rompus. Pamela is a rompus. Jerry is a rompus.
Hypotheses: Each dalpist is a rompus.

Multiple hypothesis example; height-1 ontology tree.

World Model: Debra is a dalpist. Amy is a dalpist. Pamela is a dalpist. Edward is a dalpist. Dalpists are liquid. Each dalpist is moderate. Each dalpist is brown. Barbara is a dalpist.
Observations: Barbara is rainy. Jerry is a rompus. Jerry is liquid. Jerry is brown. Pamela is a rompus. Edward is rainy. Debra is rainy. Jerry is moderate. Amy is a rompus.
Hypotheses: Dalpists are rompuses. Dalpists are rainy. Jerry is a dalpist.

Multiple hypothesis example; height-2 ontology tree.

World Model: Nicole is a gergit. Porpors are dalpists. Pamela is a sarper. Sharon is a gergit. Porpors are not cold. Dalpists are rainy. Raymond is a gergit. Each porpor is translucent. Each sarper is a dalpist. Jerry is a dalpist. Edward is a dalpist. Every gergit is a dalpist. Debra is a dalpist. Barbara is a sarper.
Observations: Barbara is liquid. Amy is translucent. Sharon is moderate. Debra is liquid. Amy is rainy. Raymond is moderate. Jerry is a rompus. Amy is not cold. Pamela is a rompus. Nicole is moderate. Edward is liquid. Amy is a rompus.
Hypotheses: Dalpists are rompuses. Dalpists are liquid. Every gergit is moderate. Amy is a porpor.

Multiple hypothesis example; height-3 ontology tree.

World Model: Barbara is a sarper. Sarpers are dalpists. Debra is a shumple. Every felper is a sarper. Raymond is a porpor. Sharon is a storpist. Each storpist is a gergit. Helen is a frompor. Porpors are dalpists. Each dalpist is liquid. Amy is a sarper. Each dalpist is salty. Felpers are rainy. Every shumple is a porpor. Each twimpee is a sarper. Nicole is a gergit. Each frompor is a sarper. Jerry is a dalpist. Dalpists are not muffled. Each gergit is a dalpist. Michael is a twimpee. Sarpers are amenable.
Observations: Debra is large. Amy is a rompus. Pamela is a rompus. Edward is liquid. Pamela is rainy. Edward is not muffled. Barbara is large. Pamela is not muffled. Edward is large. Edward is salty. Jerry is a rompus. Pamela is amenable.
Hypotheses: Each dalpist is a rompus. Dalpists are large. Edward is a dalpist. Pamela is a felper.

Multiple hypothesis example; height-4 ontology tree.

World Model: Every kurpor is a twimpee. Each gergit is pale. Patricia is a dumpus. Timples are frompors. Sharon is a folpee. George is an orgit. Rimpees are felpers. Stephen is a shumple. Mark is an orgit. Joseph is a porpor. Andrew is an orgit. Each tumpus is a felper. Gergits are dalpists. Felpers are not dark. Helen is a timple. Jason is a tumpus. Amy is a pergit. Rimpees are angry. Porpors are dalpists. Raymond is a twimpee. Each rimpee is moderate. Linda is a tumpus. Every shumple is a porpor. Each frompor is a sarper. Every dumpus is a storpist. Gergits are amenable. Barbara is an orgit. Rachel is a frompor. Felpers are sarpers. Jacob is a felper. Janet is a tumpus. Michael is a timple. Steven is a boompist. Edward is a dalpist. Storpists are gergits. Each boompist is a frompor. Debra is a dalpist. Nicole is a sarper. Pergits are felpers. Shirley is a timple. Jerry is a dalpist. Pamela is a kurpor. Every folpee is a frompor. Each sarper is a dalpist. Eric is a storpist. Orgits are shumples. Gergits are not opaque. Twimpees are sarpers.
Observations: Linda is brown. Helen is salty. George is translucent. Jason is brown. Raymond is large. Sharon is large. Jerry is a rompus. William is angry. Barbara is liquid. Edward is liquid. Emily is amenable. Janet is brown. Emily is not opaque. Shirley is salty. William is not dark. Pamela is a rompus. Nicole is large. William is moderate. Andrew is translucent. Michael is salty. Amy is a rompus. Debra is liquid. Emily is pale. Mark is translucent.
Hypotheses: Dalpists are rompuses. Every dalpist is liquid. Emily is a gergit. Sarpers are large. Timples are salty. William is a rimpee. Tumpuses are brown. Orgits are translucent.

D Details of GPT-4o paraphrase

D.1 Prompts

We present our prompts for GPT-4o to paraphrase synthetic reasoning questions into more natural ones in Figure 6.

D.2 Manual inspection

We also manually verify that the generated questions from GPT-4o are sound by randomly sam-

Prompt for Question Rewriting

Instruction

We prompt GPT-4o to rewrite synthetic reasoning questions into more natural, human-readable language while preserving the original semantics and reasoning structure.

Prompt Template

```
You are a professional editor specializing in rewriting synthetic reasoning questions into natural, human-readable language.
Rewrite the following question to be clear, fluent, and natural while preserving all logical structure and meaning. Do not simplify the reasoning steps, remove constraints, or change the problem's difficulty.
Return only the improved question with the exact same semantics, but in more natural language.
Input question: {{QUESTION}}
```

Figure 6: Prompt used to paraphrase synthetic reasoning questions into more natural language.

pling 200 reasoning questions of varying difficulty and manually checking that GPT-4o's output does not alter the problem's semantics. We conduct this study and iteratively improve the prompt until achieving 100% accuracy.

E Details of hypothesis quality metric

E.1 Theoretical justification

Our quality metric can be motivated by an argument from Bayesian decision theory. Let T be a theory which consists of a set of observations $o_1, \ldots, o_n$ and a set of proofs $\pi_1, \ldots, \pi_n$ where each proof π_i explains the observation o_i (i.e., the conclusion of π_i is o_i). Let

$$p(T) = \prod_{i=1}^{n} p(\pi_i \mid \pi_1, \ldots, \pi_{i-1})\, p(o_i \mid \pi_i)$$

be a prior distribution over the theory. Ideally, during inductive/abductive reasoning, given a new observation o_{n+1}, we aim to find the proof π_{n+1} that maximizes the posterior

$$p(\pi_{n+1} \mid \pi_1, \ldots, \pi_n, o) \propto p(\pi_{n+1} \mid \pi_1, \ldots, \pi_n) \cdot p(o_{n+1} \mid \pi_{n+1})$$

Since

$$p(o_i \mid \pi_i) = \mathbb{1}\{o_i \text{ is the conclusion of } \pi_i\},$$

the posterior is equal to

$$p(\pi_{n+1} \mid \pi_1, \ldots, \pi_n)$$

whenever the conclusion of π_{n+1} is o_{n+1}.

Now suppose we make the mild assumption that $p(T)$ is such that $p(T) > p(T')$ whenever T has fewer axioms in its proofs than does T'. Therefore, $p(T)$ is maximized if π_{n+1} minimally introduces new axioms (and equivalently, maximally reuses existing axioms). Note that the quality metric $g(\pi_{n+1})$ is the ratio of two quantities: (1) the average number of times each premise $h \in \pi_{n+1}$ appears in $\pi_1, \ldots, \pi_n$, and (2) the average number of times each premise in the ground-truth proof appears in $\pi_1, \ldots, \pi_n$. We specifically selected each ground-truth proof to be the unique proof with the minimum number of new axioms. If we consider two proofs π_{n+1} and π'_{n+1} such that π_{n+1} has more new axioms than π'_{n+1}, then the sum in the numerator of $g(\pi_{n+1})$ would contain more zero terms than that of $g(\pi'_{n+1})$, and so

$$g(\pi_{n+1}) < g(\pi'_{n+1})$$

necessarily.

Thus, with mild assumptions on the prior over the theory $p(T)$, our proposed quality metric provides an easy-to-compute proxy for the posterior probability of the theory, where more probable theories will have higher quality scores. Furthermore, our quality metric is *agnostic* to the specific choice of the prior $p(T)$.

E.2 Human validation experiments

To evaluate whether the proposed metric aligns with human judgments of hypothesis quality, we conduct a human evaluation study. We recruit three independent annotators with graduate-level STEM degrees in US R1 universities. The evaluation dataset contains 100 reasoning questions, each associated with three candidate hypotheses. For each question, annotators are asked to select the hypothesis that better explains observations and is most

concise. The presentation order of hypotheses is randomized to avoid positional bias. Each reasoning question is evaluated by three annotators, and the final preference label is determined by majority vote.

To quantify alignment with human judgments, we measure agreement accuracy, defined as the percentage when the proposed metric assigns the highest score to the human-preferred hypothesis. The proposed metric achieves an agreement of 79%, substantially outperforming a random baseline (33.3%) and a length-based heuristic baseline (51%). Human annotations exhibit substantial agreement (Fleiss' κ = 0.75), indicating that the evaluation task is reasonably well-defined and that the proposed metric captures meaningful aspects of hypothesis quality.

Please refer to Figure 7 for the instruction to annotators.

F Algorithms

Algorithms 1 and 2 describe how we generate ontology trees and hypotheses for the examples in our framework.

Algorithm 1: Generate a single hypothesis.

```
function generate_hypothesis(ontology tree T)
    if task is "infer property"
        randomly select a member of the root node of T
        randomly select a member of a leaf node of T
        generate the observations where selected
          members have the property
    else if task is "infer membership"
        select a random leaf node of T
        randomly select a property of the leaf node
        randomly select a property of the root node
        randomly select a property of the selected node's
          non-root ancestors
        generate observations where this entity has all
          properties
    else if task is "infer subtype relation"
        randomly select a member of the root node of T
        randomly select a member of a leaf node of T
        randomly select a member of a non-leaf of T
        generate the observations where selected
          members also belong to a supertype
    return all world model axioms, observations, and
      the ground-truth hypothesis
```

G Results with absolute accuracy

Here we list the absolute results of Figure 4 in Figure 8.

Algorithm 2: Generate an abductive/inductive reasoning example.

```
function generate_example()
    /* first, build the ontology tree */
        /* node() initializes an empty node
    object */
            /* add a root node */
    tree_nodes ← [[node()]]
    for i = 1, ..., h − 1 do
        current_nodes ← []
        for parent_node ∈ tree_nodes[−1] do
            u ← random(0,1)
            if u > 1/2
                repeat 2 times
                    current_nodes.append(node())
            else
                repeat 3 times
                    current_nodes.append(node())
        tree_nodes.append(current_nodes)
    for current_nodes ∈ tree_nodes do
        for node ∈ current_nodes do
                /* allocate three members and
            properties */
            repeat 3 times
                node.members .append(
                  random_member())
                node.properties .append(
                  random_property())
    if this is a single-hypothesis example
        randomly select a task type
        generate_hypothesis(tree_nodes)
    else generate 1 to 3 hypotheses per tree layer
```

H Error analysis

H.1 Common error types and example responses

In this section, we detail the most common types of errors that LLMs produce in our experiments. For each of the error types listed below, we provide an example consisting of a world model, list of observations, ground-truth hypotheses, and the mistaken LLM output.

Instruction for annotators

```
- You will be given a set of reasoning questions. Each question includes:
* A description of observations
* Three candidate hypotheses that attempt to explain these observations
- Your task is to select the single best hypothesis according to the following
criteria:
* The hypothesis should account for the given observations coherently and
logically.
* The hypothesis should be as concise as possible.
- Important Guidelines
* Select only one hypothesis per question.
* Do not choose based on writing style, fluency, or formatting alone.
* Focus on the validity and quality of hypotheses, not superficial features.
* The order of hypotheses is randomized; treat each option independently.
* If multiple hypotheses seem similar, choose the one that is most concise.
- Your selection will be used to determine the hypothesis that best explains
the observations.
```

Figure 7: Instruction for annotators.

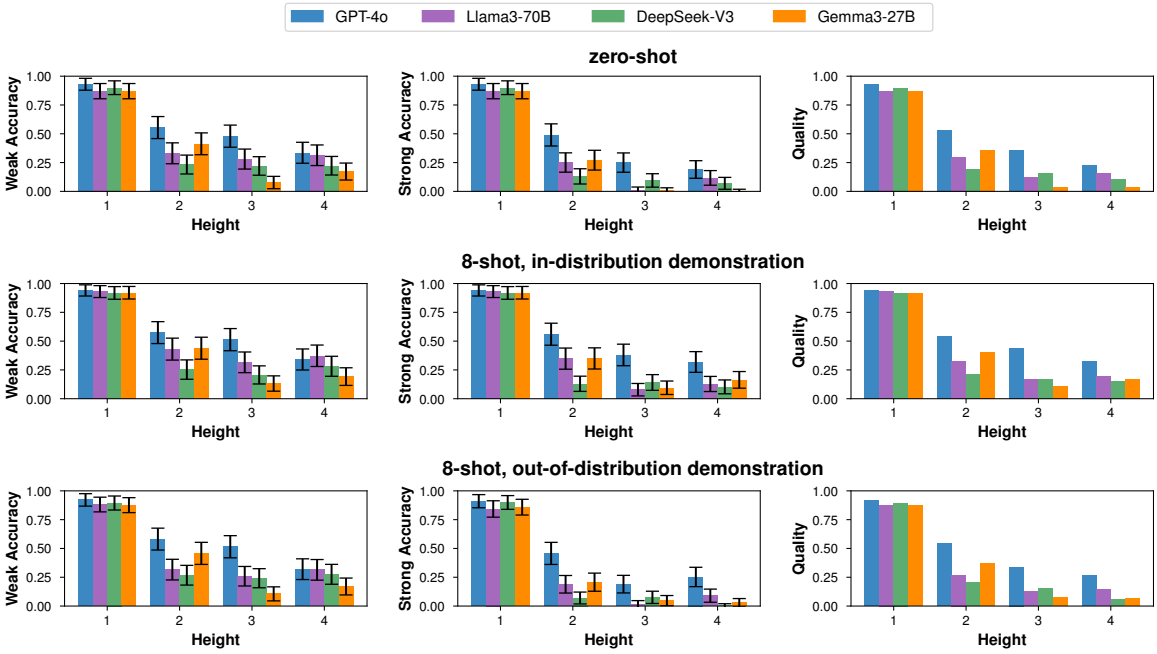


Figure 8: Absolute results of weak accuracy, strong accuracy, and quality versus the height of the ontology tree with multiple hypotheses under zero-shot (**top**), 8-shot in-distribution (**middle**), and 8-shot out-of-distribution (**bottom**).

Error Type 1: *Wrong ontology direction*

World Model: Porpors are dalpists. Nicole is a gergit. Sarpers are dalpists. Every gergit is a dalpist. Debra is a dalpist. Each dalpist is rainy. Every porpor is not cold. Raymond is a gergit. Each porpor is translucent. Amy is a porpor. Jerry is a dalpist. Pamela is a dalpist. Sharon is a gergit. Edward is a dalpist. Michael is a sarper.
Observations: Sharon is moderate. Barbara is not cold. Debra is liquid. Barbara is rainy. Barbara is liquid. Barbara is translucent. Amy is a rompus. Raymond is moderate. Jerry is a rompus. Edward is liquid. Pamela is a rompus. Nicole is moderate.
Ground-truth hypotheses: Dalpists are rompuses. Every dalpist is liquid. Each gergit is moderate. Barbara is a porpor.

Mistaken LLM Output:
... Rompuses are dalpists. ...

Error Type 2: *Ignore the ontology and produce unnecessary hypotheses*

World Model: David is a sorple. Each sorple is a dropant. Each hilper is a thorpin. Rimpees are scrompists. Brian is a sorple. Dropants are gompers. Angela is a gomper. Each thorpin is a gomper. Every scrompist is a gomper.
Observations: David is rainy. Brian is rainy. Angela is rainy.
Ground-truth hypotheses: Every gomper is rainy.

Mistaken LLM Output: ... Every gomper is rainy. Every dropant is rainy. ...

Error Type 3: *Fall back to rival hypotheses*

World Model: Rorpants are dolpers. Every worple is a wumpus. Each drompee is moderate. Ashley is a rorpant. Fimples are dolpers. Ronald is a wumpus. Wumpuses are gwompants. Yompins are dolpers. Brian is a dolper. Michelle is a worple. Each dolper is a gwompant. Shirley is a rorpant. Sandra is a bempin. Linda is a fimple. Drompees are pale. Jack is a yompin. Each drompee is a bempin. Bempins are cold. Bempins are gwompants. Emma is a gwompant. Michael is a gwompant.
Observations: Shirley is a borpin. Christopher is pale. Emma is a borpin. Brian is salty. Michael is salty. Christopher is a borpin. Christopher is cold. Ashley is salty. Christopher is moderate.
Ground-truth hypotheses: Each gwompant is a borpin. Each gwompant is salty. Christopher is a drompee.

Mistaken LLM Output:
... Shirley is a borpin. ...

Error Type 4: *Hallucinated entities*

World Model: Jonathan is a rifpist. Yimples are dull. Samantha is a rifpist. Karen is a gwompant. Scott is a jelgit. Olivia is a frompor. Maria is a lirpin. Twimpees are lirpins. Each rifpist is aggressive. Frompors are jelgits. Jessica is a jelgit. Amanda is a twimpee. Rifpists are jelgits. Each lirpin is a jelgit. Yimples are small. Each gwompant is a frompor. Each yimple is a rifpist.
Observations: Mary is not overcast. Scott is not overcast. Jonathan is not overcast. Samantha is a kurpor. Mary is aggressive. Jessica is a kurpor. Olivia is a kurpor. Mary is small. Mary is dull.
Ground-truth hypotheses: Every jelgit is a kurpor. Jelgits are not overcast. Mary is a yimple.

Mistaken LLM Output:
... Mary is a dolphant. ...

H.2 Details of error analysis

We first randomly sampled 200 LLM example responses to reasoning questions of varying difficulty. Then, we manually inspected them to identify re-

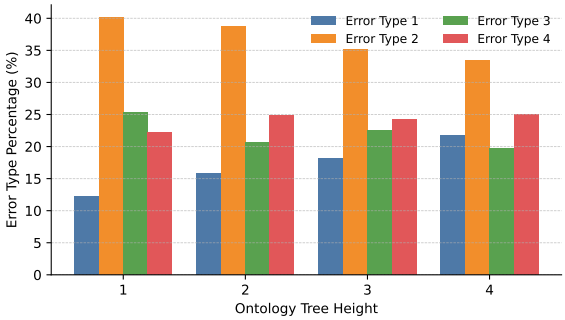


Figure 9: Distribution of error types across ontology tree heights.

curring failure patterns as summarized in the paper. To further quantitatively study the error patterns of all responses, we use an LLM-as-a-judge (Gu et al., 2024) and prompt GPT-4o with in-context examples. We use the following prompt in Figure 10 for the LLM-as-a-judge. We categorize the error pattern distribution by ontology tree height and present the results in Figure 9. As shown in the figure, Error Type 2 (*Ignore the ontology and produce unnecessary hypotheses*) is the most significant error across all ontology tree heights, indicating that LLMs can generate weakly correct hypotheses but fail to consider the given ontology tree and optimize them in the sense of Occam's razor.

Prompt for Error Analysis

Prompt Template

```
You are an expert evaluator of reasoning quality. Your task is to analyze a
model's reasoning process for a question and determine the type of error it
contains.
## Task
Given:
* A question
* The model's reasoning process
* The correct answer
Identify which error type best describes the mistake in the reasoning.
Choose exactly one error type from the following categories.
—
## Error Types
** Error Type 1: Wrong ontology direction** The hypothesis contains the wrong
ontology direction.
** Error Type 2: Fall back to trivial hypotheses** The hypothesis is the
observation itself.
** Error Type 3: Ignore the ontology and produce unnecessary hypotheses** The
hypothesis is technically correct but unnecessary given the ontology.
** Error Type 4: Hallucinated entities** The reasoning relies on non-existent
entities.
—
## Examples
### Example 1 (Wrong ontology direction)
<omitted for space, see §H.1 for examples>
—
### Example 2 (Fall back to trivial hypotheses)
<omitted for space, see §H.1 for examples>
—
### Example 3 (Ignore the ontology and produce unnecessary hypotheses)
<omitted for space, see §H.1 for examples>
—
### Example 4 (Hallucinated entities)
<omitted for space, see §H.1 for examples>
—
## Now evaluate the following case
Question: {{QUESTION}}
Model Reasoning: {{MODEL_REASONING}}
Correct Answer: {{CORRECT_ANSWER}}
—
## Output Format
Error Type: <one of: Wrong ontology direction | Fall back to trivial hypotheses |
Ignore the ontology and produce unnecessary hypotheses | Hallucinated entities>
```

Figure 10: Prompt used to analyze common error patterns.